The Impact of Open Access Mandates on Invention

Kevin A. Bryan and Yasin Ozcan*

February 20, 2020

Abstract

How do barriers to the diffusion of academic research affect innovation? In 2008, the NIH mandated free online availability of funded biomedical research. This policy caused a 50 percentage point increase in free access to funded relative to unfunded articles. We introduce a novel measure, in-text patent citations, to study how this mandate affected industry use of academic science. After 2008, patents cite NIH-funded research 12 to 27% more often. Non-funded research, funded research in journals unaffected by the mandate, and academic citations see no change. These estimates are consistent with, and interpreted by, a model of search for useful knowledge by industry researchers. Our results imply that inefficiency caused by the academic journal system may be substantial.
JEL Codes: I28, O31, O38.

*Bryan: University of Toronto, Rotman School of Management, Toronto, Ontario, Canada
Ozcan: MIT Sloan School of Management, Cambridge, MA, USA
Correspondence: kevin.bryan@rotman.utoronto.ca. ozcan@alum.mit.edu..
We appreciate comments from seminar audiences at the University of Toronto, the University of Maryland, the NBER, Duke University, Carlo Alberto, the University of British Columbia, and REER. Funding for this project was provided by the Sloan Foundation Contribution Economy program.

1 Introduction

University research is often valuable in industry, particularly in innovative sectors (e.g., Owen-Smith et al. [2002], Stokes [1997]). This research diffuses principally through published research articles (Cohen et al. [2002]). Academic publications generate hypotheses worth exploring, refute unpromising paths, provide tools to speed development, suggest techniques to aid laboratory or statistical work, and create basic pieces of scientific knowledge for recombination. Future researchers more easily build on research that is clearly presented, widely promulgated, and codified in a useful way (Mokyr [2002], Murray and Stern [2007]).

Since a primary vector for industry to learn about frontier research is scientific journals, the academic norms that determine journal access and pricing are particularly important. Unlike the predominant practice in economics, public working papers and freely accessible published journal articles are rare in most fields. In 2006, only 15% of all scientific articles were freely accessible online; by 2013, only 24% were (Björk et al. [2009], Khabsa and Giles [2014]). Why? Promotion and status in academia require publication in elite journals in one's field. Sticky status gives publishers of these journals market power. Private publishers and scientific societies take advantage of this market power, often by charging high per-article fees to ensure institutional libraries maintain subscriptions (Jeon and Rochet [2010]). These costs artificially limit inventor access to academic results.

Do costly journals harm private-sector innovation? We examine this question with a natural experiment. In January 2008, the NIH announced that any funded article accepted for publication after April 7, 2008 must be archived in the open access

PubMed Central ("PMC") database within 12 months of publication.[1] Most non-NIH-funded biomedical and biotech articles were not then, and are not now, free to read.

The NIH mandate proved controversial on two grounds. First, scholarly journals have costs. Mandates shift the costs of these journals from readers, including the private sector, to authors and funders. This is especially problematic for underfunded institutions (Frank [2013]). Second, there is surprisingly little empirical evidence of any positive benefit from open access. The most credible estimates find that open access causes only a small increase in academic citations.[2] The case for open access is limited if its effects are mainly distributional transfers from industry to publishers, with no real change in the rate of innovation: someone must pay the fixed cost of the journal.

In Section 2, we therefore introduce a model of inventor search suggesting how the academic journal pricing structure can generate large welfare harms. In particular, journal market power causes both transfers from industry to publishers and reduces search for useful knowledge by inventors. Even cheap articles - $40 would not be an unusual price - can cause substantial social harm by changing search behavior. Guided by the model, we empirically investigate how the NIH mandate changed the use of research in patented inventions. We use a novel coarse matching approach to search the text of all patent applications for references to any article in 43 top medical journals since 2005. These *in-text citations*, though computationally challenging to extract, have

[1]Similar mandates exist from organizations including the University of California, the Howard Hughes Institute, the Wellcome Trust, and MIT (Suber [2012]). Throughout, we use "open access" and "freely available online" synonymously; of course, there are many definitions of open access, some much more restrictive than ours. See, e.g., Harnad et al. [2004].

[2]Davis et al. [2008] randomize the free journal-website availability of a sample of articles and find no difference in academic citations one year out. Using a large panel of science articles with within-journal open access variation, McCabe and Snyder [2014] find an open access citation advantage of only 8%. Kim [2012] finds a slightly larger effect on social science articles, taking advantage of quasirandom variation in SSRN article acquisition. Gaule and Maystre [2011] control for selection into open access with an instrument based on lab financial resources, and find no effect of open access on citation. Some contrary evidence exists (Gargouri et al. [2010], Evans and Reimer [2009]) although, as Swan [2010] and McCabe and Snyder [2014] point out, studies which find large effects of open access on academic citation tend to have serious identification concerns.

many advantages over the commonly-used "front page" prior art citations. The overlap between in-text citations and front page citations is very low. Further, there are both legal and empirical reasons to believe in-text citations are more correlated with the actual knowledge used by the inventor. To our knowledge, we are the first to extract and use in-text citations in any systematic way. We discuss this data source in detail in Section 3.

In Section 4, we first estimate a difference-in-difference in patent citation propensity for articles published before and after April 2008, with and without NIH funding. Second, we take advantage of a set of journals that make nearly all articles free, no matter what. Because all research is freely available in these journals, the NIH mandate did not change the de facto price of articles. This permits the estimation of a triple difference, looking at how the 2008 mandate affected patent citations to NIH funded articles published in journals affected by the policy versus those that were not. A triple difference ensures that our first estimation strategy does not simply pick up increased NIH funding for more applied projects, among similar concerns. Both estimates give similar results, with open access causing patents to cite articles 12 to 27 percent more often. As the policy only led to a 50 percentage point relative increase in free availability compared with non-funded articles, we argue this is a lower bound on the true effect of open access. With subsample analyses, we rule out that low-quality patents drive our main effect.

In Section 5, we conclude by discussing the implications of our results. Which firms are harmed most by the current academic norms? If industrial innovation is harmed in a substantive way by the academic journal system, what alternative funding mechanism ought we pursue to cover the fixed costs of journals? We argue that the constrained first-best is unlikely to be achieved by industry coordination alone. That said, we discuss a number of options that can increase industry access without negatively

burdening publishers, including scientific societies, who rely on subscription revenue.

1.1 Prior Literature

This paper is, to our knowledge, the first broad empirical investigation of how open access affects industry, with direct implications for the organization of academic publishing. There is a large complementary literature showing how similar openness constraints, in a general sense, limit the use of science.[3] Furman and Stern [2011] show that storing biomaterial in easy-to-access locations increases its use by 50 to 125 percent. Murray et al. [2012] show that transgenic mice with fewer IP restrictions were used more often in studies, especially applied ones. Williams [2013] studies the use of decoded genes from the Human Genome Project and Celera. Genes decoded first by the HGP, which were not bound by any IP, were studied and used in products like diagnostic tests more often. Sampat and Williams [2014], however, find that gene patent grants, instrumented using the variable strictness of patent examiners, do not affect follow-on innovation. They argue that the patentholder optimally allows research which increases the patent's value.

Overall, the existing literature on scientific openness finds harms when the party choosing the extent of openness prefers to limit knowledge diffusion and instead earn rents along an alternate dimension. In our context, publishers earn most of their revenue from institutional subscriptions. Lower per-article prices cause industry to use more science, but also limit pricing power for university subscriptions. Therefore, publishers

[3]Earlier research on the direct question of how academic open access affects non-academic actors is very limited. Hardisty and Haaga [2008] send links to practitioners for new articles in the *Journal of Clinical Psychology*, some of which link to gated articles and some of which link to freely available ones. The practitioners who were sent the freely available article links were much more likely to read the emailed articles, and further were more likely to begin recommending frontier treatments to the patients. Ware and Monkman [2009] survey private sector researchers in the UK and find that over half of the high tech, research-using small businesses surveyed had difficulty accessing academic research useful to their business; a similar survey by Houghton et al. [2011] finds that 68% of Danish firms report access difficulty.

keep per article prices high despite the deadweight loss. Even university researchers who care about the private sector submit research to expensive journals because norms within academia require publication in highly-prestigious rather than highly-accessible venues. Both our theory and our empirical estimates suggest this norm may have serious consequences for industrial use of academic science..

Norms and institutions concerning commercialization of university research have also been widely studied. For instance, Hvide and Jones [2017] show that entrepreneurship, licensing, and patenting by university researchers falls after a Norwegian policy change decreased academic earnings from the commercialization of their research. The Bayh-Dole Act famously encouraged universities to commercialize by changing intellectual property standards (Berman [2008], Mowery and Sampat [2005]).

Although commercialization is a particularly visible venue for the effect of academic research on industry, diffusion of knowledge in scentific documents indirectly affects many more innovative firms. A survey of R&D managers (Cohen et al. [2002]) finds that a third of industry R&D projects use public sector research findings, and over a fifth use public sector instruments and techniques. Their survey respondents claim publications and conferences are much more important than licensing, patents, or the hiring of recent graduates for incorporating research results and tools. Ahmadpoor and Jones [2017] consider the network of citations, where an invention draws on a invention that itself drew on academic research, and find that at least 60% of all inventions can be traced back to published research. Iaria et al. [2018] investigate the collapse in international communication of scientific results during World War I, and find that scientists who were particularly reliant on journal articles from blockaded countries before the War see permanent and severe declines in their research productivity after their access to continuing research from their nations is cut off. Going even further back, the steep decline in the price of books induced by Gutenberg may have caused a welfare increase

more substantial than that of the modern computer (Dittmar [2020]).

2 A Stylized Model of Academic Search

How does the market power possessed by high-prestige journals affect industry researchers? Consider the following letter from a private-sector biopharma consultant, published in the journal *Nature Biotechnology* (Lyman [2011]):

> "The majority of companies have no libraries to speak of and no librarians to help with literature searches. The availability of online journals is insufficient and funds for purchasing access to papers on an individual basis are limited. In one case, a company suffered a six-month setback to a drug development program because a paper was missed in an inaccessible journal. The central question that I raised in my op-ed piece was, at a time when more and more papers are published, when information overload is a given, does a lack of access to the information become an equally large problem? The answer from the community was a vociferous yes.
>
> I've been fortunate to have access to worldclass libraries at every stage of my career. As a result, I learned that being widely read has significant advantages. It enables the formation of new and fruitful collaborations. It facilitates your ability to make connections, to see new relationships and to partake of a bigger view. This larger vision, in turn, can lead to novel insights and spur innovative discoveries. As I noted previously, keeping up with advances in biomedicine has become increasingly difficult in recent years. The overlapping nature of disciplines within the biological sciences means that someone developing a new cancer treatment needs to stay informed about specific areas of biochemistry, genetics, toxicology, computa-

> tional biology, developmental biology, cell biology, immunology and stem cell biology as well as clinical developments. This is in addition to keeping up with general trends in the biotech industry as well as technical advances in experimental reagents, devices, and methodology."

In this mental model of the invention production function, private sector researchers begin with ideas. The reader "needs to stay informed" about developments in many journals to create more valuable inventions. It is ex-ante difficult to know which article will contain a useful piece of knowledge. Therefore, "being widely read" can "lead to novel insights and spur innovative discoveries." Subscriptions are too expensive for small firms since useful information is found in many different journals. Purchasing individual articles is too expensive since many articles must be read to learn which is useful.

Let us expand that qualitative model into a formal model of search. A formal model shows the effect of the journal pricing system on different types of firms and research, and will help interpret some of the empirical parameters we estimate in Section 4. Our model will have three basic properties.

First, journal publishers have market power and hence can price above cost. Jeon and Menicucci [2006] note that journal subscription prices have been rising at more than twice the rate of academic book prices. A reasonable conjecture is that the relative cost of publishing articles versus books has not changed greatly over time. Therefore, the differential inflation is ipso facto evidence of increased markups. It is critical to note that this market power exists solely because university researchers are obligated by the academic incentive structure to send their research to high-prestige journals. Without this market power, academic research is freely available, with the costs of peer review and distribution covered by institutions who either can be funded by non-distortionary taxes or the sale of complements. Market power is not necessarily socially costly, of

course; consider the case of a monopolist that perfectly price discriminates. Whether journals just redistribute from industry to publishers, or cause deadweight loss, will be identifiable in the model.

Second, we permit publishers to both bundle articles into subscriptions and sell access a la carte. The economics of bundling zero marginal cost goods is straightforward (Bakos and Brynjolfsson [1999]). A law of large numbers result implies that bundling is always optimal when the values of individual goods are independent. Further, this bundling lowers consumer surplus. When there are many market segments with correlated demand in each segment, such as large institutions with high willingness-to-pay for all articles and small firms with low willingness-to-pay for most articles, publishers optimally sell the full bundle to the core market, and offer a menu of progressively smaller bundles for the periphery market. In the limit, the smaller bundles become a la carte article pricing. This theory predicts exactly the pricing strategy of academic journals. We therefore do not model the publisher pricing problem directly. Rather, we just assume that firms can either buy a subscription or purchase articles individually. The prices of subscriptions and articles are exogenous to the firm's own demand.

Third, researchers can search academic literature for knowledge that improves the value of their invention. Value increases because the time necessary to invent falls, improvements to the final product are suggested, or dead-end research paths are avoided. The researchers do not know exactly which article might contain that knowledge, if any.

Formally, assume inventors search for knowledge as follows:

Assumption 1. *Let an invention to inventor i in the absence of academic research be worth X_i. Let the value of the invention if academic research a is accessed be $X_{ai} \geq X_i$, where $X_{ai} - X_i$ is a random variable with distribution F.*

Assumption 1 says that useful academic knowledge improves the private prof-

itability of a firm's invention by some random value with known distribution. This distribution will be quite wide if the academic knowledge a is, for instance, the type of idea one might get by browsing the new issue of a journal. It may be quite tight if the researcher is trying to figure out a particular statistical technique, or method of generating a chemical compound, but simply does not know where to look.

Assumption 2. *Let there be a set of journal articles J such that the probability article $j \in J$ includes useful information a is p_{aj}, disjoint across all j, such that $\sum_j p_{aj} \leq 1$.*

Assumption 2 says that the location of this valuable academic knowledge can only be found by searching the academic corpus. If $\sum_j p_{aj}$ is strictly less than 1, then there is some chance that no article contains the useful knowledge.

Assumption 3. *Let $(1 - s_{ij})c_{ij}$ be the de facto cost of accessing information article j, where c_{ij} is the stated cost of j to inventor i and s_{ij} is the probability that the information in j spills over to inventor i without them actually paying for the article. If an inventor is at an institution with a subscription to the journal where j is published, then $c_{ij} = 0$.*

Assumption 3 gives the cost of searching a particular article, which is free if the information spills over locally or the institution has a subscription, and positive otherwise. Of course, researchers may also email authors for an article, or travel to a university library. The model only requires that those with institutional subscriptions access the article at lower cost.

Assumption 4. *Let $G \geq 0$ be a multiplier on X which converts private values of an idea to the social value of that idea.*

Assumption 4 says that the private and social surplus of invention are misaligned. If invention generates spillovers and consumer surplus, then $G \geq 1$. If the invention would have been created by some other firm in the near future anyway, and thus invention is just business stealing, then $G \leq 1$.

As far as timing is concerned, an inventor will simultaneously choose how many articles to purchase and read, given their belief about the expected benefit of finding useful academic knowledge a.[4] That is, inventors solve:

$$\max_{\mathbb{I}_j} \sum_{j \in J} \mathbb{I}_j [p_{aj}((E[X_{ai} - X_i]) - (1 - s_{ij})c_{ij})]$$

where $\mathbb{I}_j$ is the indicator function equal to 1 if inventors read article j. Since p_{aj} are disjoint across j, the maximand involves buying all articles such that

$$p_{aj}(E[X_{ai} - X_i]) \geq (1 - s_{ij})c_{ij}$$

Under an open access regime, all articles have $c_{ij} = 0, \forall i, j$, hence inventors buy all articles such that

$$p_{aj}(E[X_{ai} - X_i]) \geq 0$$

That is, they read everything even potentially useful. Note that c is a transfer from the inventor to the journal and hence does not affect social welfare.

The previous two inequalities imply that the difference in social welfare generated by firm i under an open access regime (the "value of open access") is

$$G \times \int (X_{ai} - X_i) dF \times \sum_{0 \leq p_{aj}(E[X_{ai} - X_i]) \leq (1 - s_{ij})c_{ij}} p_{aj}$$

which is simply the expected private value gain if a is known times the probability a is learned only under open access times the social value multiplier G.

[4] Bryan [2020] solve for the probability a prize is found with sequential, rather than simultaneous, search on a partition. Costly sequential search finds weakly more articles than simultaneous search, hence the benefit of open access is weakly lower. Nonetheless, the comparative statics in Proposition 1 remain identical. In particular, for any reward and search cost, there remains a calculable cutoff article which is not purchased, the number of articles increases in the payoff, decreases in the cost of search, and increases in the coarseness of the partition.

Let $\bar{I}$ be the set of inventors with institutional access to research. For these researchers, the mean value of knowledge transfer from academia to their inventions is

$$G \times E_{i \in \bar{I}}(E[X_{ai} - X_i]) \times \sum p_{aj}$$

which is the expectation over all firms that have institutional subscriptions of the expected increase in idea value due to academic knowledge times the probability the relevant knowledge is contained in some journal times the social value multiplier.

Let us first show which firms benefit most from open access:

Proposition 1. *The value of open access to a given firm i is*

1) increasing and then decreasing in a step function in $X_{ai} - X_i$

2) increasing in the coarsening of p_a

3) increasing in the social value multiplier G

4) increasing in c_{ij}

5) decreasing in s_{ij}

Proof. 1) If

$$E[X_{ai} - X_i] < \min_j \frac{(1 - s_{ij})c_{ij}}{p_{aj}}$$

then increasing $E[X_{ai} - X_i]$ by ϵ does not change which articles are bought, but does increase $G \times E[X_{ai} - X_i]$ and hence the total value of academic knowledge. On the other hand, if

$$E[X_{ai} - X_i] = \min_j \frac{(1 - s_{ij})c_{ij}}{p_{aj}}$$

then increasing $E[X_{ai} - X_i]$ by ϵ means that the least valuable academic article is worth enough that it would have been bought by the inventor even without open access, hence open access has less total value. The step-like function of the value of open access in $E[X_{ai} - X_i]$ can be proven inductively in an analogous manner.

2) Holding $\sum p_{aj}$ constant, but letting p_a be a more coarse partition weakly increases

$$\sum_{0 \leq p_{aj} E[X_{ai} - X_i] \leq (1 - s_{ij}) c_{ij}} p_{aj}$$

and hence weakly increases the value of open access.

3) Trivial.

4) Higher costs per article increase $\sum_{0 \leq p_{aj} E(X_{ai} - X_i) \leq (1 - s_{ij}) c_{ij}} p_{aj}$ and hence the value of open access.

5) Analogous to 4. □

That is, open access is more valuable if inventors without institutional subscriptions are using knowledge that is neither too unimportant (in which case open access is of little consequence) nor too valuable (in which case the private sector is already buying everything); if it is not clear which particular article contains useful knowledge; if the social value of inventions is much higher than the private value; if articles are costly; and if spillovers are inconsequential. Since social value is simply a multiple of private value, the societal value of open access has the same five comparative statics.

Consider now the expected value of additional knowledge found under open access. Those with institutional access search everything, and always find a if it exists. Therefore, integrating over all institutional researchers $\bar{I}$, the mean expected value of knowledge firms learn from academia is

$$E_{i \in \bar{I}} E[X_{ai} - X_i]$$

Those without institutional access only search if the idea they are looking for is sufficiently valuable to make search worthwhile. The mean expected value of knowledge

learned by these firms when there is no open access is

$$E_{i \notin I | E[X_{ai} - X_i] \geq (1 - s_{ij}) c_{ij}} E[X_{ai} - X_i]$$

Finally, the expected value of knowledge learned only under an open access mandate for researchers without institutional access is

$$E_{i \notin I | E[X_{ai} - X_i] < (1 - s_{ij}) c_{ij}} E[X_{ai} - X_i]$$

Proposition 2. *The expected value of additional knowledge learned only following an open access mandate is*

1) lower than the expected value of knowledge learned by the same firm when access is costly

2) potentially higher than the mean value of knowledge learned by all firms when access is costly

Proof. 1) Immediate; high value knowledge will induce search even when researchers have to pay for access.

2) Without open access, let p_1 be the proportion of all firms with institutional access, and p_2 be the proportion of all firms such that $E[X_{ai} - X_i] \geq (1 - s_{ij})c_{ij}$. The mean value of knowledge found without open access is

$$\frac{p_1}{p_1 + p_2} E_{i \in \bar{I}} E[X_{ai} - X_i] + \frac{p_2}{p_1 + p_2} E_{i \notin I | E[X_{ai} - X_i] \geq (1 - s_{ij}) c_{ij}} E[X_{ai} - X_i]$$

Additional knowledge found following open access has expected value

$$E_{i \notin I | E[X_{ai} - X_i] < (1 - s_{ij}) c_{ij}} E[X_{ai} - X_i]$$

The latter equation can be greater than the former if three necessary conditions hold. First, many inventions come from inventors with institutional access (p_1 is high). Second, inventors without institutional subscriptions are using academic knowledge in at least as valuable ways as those with institutional subscriptions ($E_{i \notin I} E[X_{ai} - X_i] > E_{i \in I} E[X_{ai} - X_i]$). Third, either the spillover-adjusted de facto cost of articles is high or the potential location of useful information is dispersed. □

The second statement in Proposition 2 may be surprising. It says that the additional knowledge found only under open access may be, on average, *more* valuable than the average piece of knowledge found when academic journals are costly.

The intuition behind that result is straightforward. A given firm only searches if the expected value of what they learn exceeds the search cost to learn it. Therefore, if a given firm has to pay to search, they no longer search for and find less valuable knowledge. The additional knowledge learned because of open access will have lower expected value for any given firm than the knowledge they learn when articles are costly. However, open access does not induce extra learning by all firms, but only by firms who found it too expensive to search when articles were costly. If these firms use knowledge in valuable ways on average compared to the mix of firms with institutional subscriptions and firms who perform costly search without open access, then the average knowledge learned due to open access can be more valuable than the average knowledge learned by all firms when search was costly.

This counterintuitive outcome is most likely to occur when firms with journal subscriptions have many low-value uses of knowledge, firms without journal subscriptions have many uses of knowledge that are valuable but not too valuable, the cost of buying articles is high, and the set of journals where useful information may be found is large.

With this theory as a guide, let us examine the case of the NIH open access mandate empirically. We will clarify in the data section how our empirical objects

relate to the theoretical variables above.

3 Data

Our data consists of a sample of academic research articles, dummies denoting article availability in open access repositories, and a sample of patent applications.

We examine 132,872 research articles appearing in 43 prominent medical and biotechnology journals published between 2005 and 2012.[5] For each article, we extract the country of the first author's affiliation, the affiliated state if the author is in the U.S., a dummy indicating whether the author reports funding from the NIH, the journal name, the number of academic citations (cites given in the bibliography of another academic article) as of July 2014, a dummy denoting open access availability via PubMed Central (PMC), in which case we can see the exact date the article was made free-to-read, and a dummy denoting availability via Pubmed's broader "Free Full Text" (FFT) category as of June 2013.[6] The FFT category is nearly identical to the set of articles one could find freely available anywhere online, and would include, e.g., an article freely available on a publisher's website which was not deposited in PubMed Central.[7] PubMed and PMC are by far the most commonly accessed medical research databases in the world, with PMC searches alone resulting in over one million article

[5]The journals consist of prominent general interest publications (e.g., The New England Journal of Medicine, Lancet), top field journals (Hematology, Immunity) and 10 highly-cited biotechnology journals (J. Biotechnology, Tissue Engineering). Exact details of our sample are available in the online appendix.

[6]For 3002 articles, we are unable to extract author location, and for 2253 we were unable to extract the number of academic citations. In general, this missing data refers to editorials and other types of articles which were miscoded as being research-oriented.

[7]Optimally, we would know the exact date every article was available anywhere online, rather than just the fact that it was available freely as of 2013. However, almost all of the NIH-funded articles are deposited directly into PubMed Central, and we can observe that the deposit date is nearly always within 6 to 18 months following publication. For non-NIH-funded articles, anecdotally many of these were made freely available only in 2011 or 2012, meaning that our estimate of the differential open access effect generated by the NIH policy may be too conservative. Cutting off citations as of 2015 means our study is not affected by Sci-Hub and other quasi-legal websites offering free scientific articles.

views per day (Blumenthal and Freiburger [2012]), a number that has been growing rapidly since 2008 (Online Appendix Figure A1).

Our patent application sample consists of the raw text of all U.S. patent applications since 2005 which are public as of March 19, 2015.[8] This sample includes 2,989,005 applications in over 200 gigabytes of weekly XML compilations produced by the USPTO. From this sample, we extract the names and locations of all authors, the name and location of all assignees, and the patent classes and subclasses. We further extract, in May 2015 and August 2017, whether the patent has been granted, and how many related applications have been filed with foreign patent offices. Note that patent applications are generally not made public until 18 months after the application is submitted. Further, many applicants request secrecy for an even longer period. For this reason, as we reach the end of our sample, we are observing fewer and fewer applications. For every assigned patent, we algorithmically construct dummies indicating whether the assignee is a corporation, a major biotech or pharmaceutical corporation, a university, a government agency, or an individual. For 98.5% of the assigned citing patents, we are able to code them into at least one of those categories.[9]

To link the two datasets, we develop a custom coarse matching algorithm which operates on the raw specification text of the patent applications. Citations in the text of a patent are not coded in a standardized way. Instead, references are strewn throughout the specification text in a wide variety of formats, sometimes including article titles and full bibliographic information, and sometimes in a much more informal format. Even journal names are not referred to in a standard way; the New England Journal of Medicine will be referred to as NEJM in one patent, New Eng. J. Med. in another, and with its unabbreviated title in a third. Full details of our matching algorithm are

[8]For readability, throughout we will use "patent" and "patent application" simultaneously, though all of our data refers to patent applications unless noted otherwise.

[9]Patents can have multiple assignees; just over 500 of our patents are assigned both to a corporation and to a university. We discuss the details of the dummy construction in the online Appendix.

left to the online appendix, but the basic idea is to search chunks of patent text for nearly-adjacent appearances of the article year, one of a large number of abbreviations or acronyms for the publishing journal, the first author's last name, and/or the first few words of the article title, tightening the requirements for articles where the first author has a particularly common last name. This method naturally involves a tradeoff between Type I and Type II errors, and we have chosen to be conservative in identifying matches. Manual investigation suggests that over 99% of our claimed patent-paper matches were in fact correctly matched.

Minimizing false positives means that we miss some matches; for instance, "In 1989 Stephan J. Weiss in the New England Journal of Medicine conducted bacterial sensitivity studies on E. Coli and toxicity on tissue in guinea-pigs" in patent application 12/101,775 is too vague, lacking both an article title and a journal issue number, for our algorithm to match with a specific article. However, manually investigating a large sample of patent texts, we found only a small number of matches that would be missed by our algorithm; these Type 2 errors are generally caused by misspellings or special characters in the author name or article title.

The algorithm identifies 28,136 patents citing at least one article in our sample, with 63,106 total citations of academic papers.[10] 22,611 academic papers, or 17 percent of our sample, receive at least one citation; for our oldest cohort of papers, from 2005, more than 28 percent are cited at least once. The matches are almost entirely medical-related, as would be expected: over 91 percent of the patents come from just six primary patent classes.[11] No more than 2 percent of the matches, and by our best estimate much

[10]Naturally, if a single patent application cites the same academic paper multiple times, this counts as only one citation. Further, we drop all applications that are continuations of applications already in our sample.

[11]424 (Drug, bio-affecting and body treating compositions), 435 (Chemistry: molecular biology and microbiology), 506 (Combinatorial chemistry technology: method, library, apparatus), 514 (Drug, bio-affecting and body treating compositions), 600 (Surgery), 800 (Multicellular living organisms and unmodified parts thereof and related processes). 424 and the related class 514 alone make up 63% of the citing patents.

less than that, are "self-cites" where the article author is also a patentee.[12]

3.1 Why In-Text Rather Than Front Page Citations

The most common proxy for the scientific base on which an invention is built are the "front page" prior art citations, particularly citations to academic research (e.g., Fleming and Sorenson [2004], Azoulay et al. [2015]). Front page citations are derived from documents listed by patent applicants on their Invention Disclosure Statement, or are added by patent examiners. We use in-text citations, extracted from the specification text of the patent, rather than front page citations for both practical and substantive reasons.

The practical reason is the long lag between application and patent grant. Many studies, including ours, study very recent policy changes for which the application-to-grant delay binds. Patent applications do not contain front page references. In-text citations allow us to investigate the "paper trail of knowledge" even when all we have are patent applications. The substantive reason concerns the meaning of a patent citation. The closest object to the learned knowledge "a" in our theoretical model is any knowledge learned from academia, by the inventor, which increases the value of the patent in some way.

Consider first front page citations. Examiner-added citations, of course, make up a portion of front page prior art, and they are by definition not known by the inventor (e.g., Cotropia et al. [2013], Sampat [2010], Alcacer et al. [2009]). More importantly, front page citations are legally consequential and hence are often added by patent drafters and patent attorneys well after the actual invention in question has been created. The reason is that U.S. patent applicants face a "duty of disclosure." This duty requires disclosure of any known invention or publication relevant to the patentability

[12]The online appendix contains further details on self-citations.

of the patent's claims. To put it in academic terms, front page prior art resembles a list of papers similar to one's own, as determined by the authors, their conference attendees, and the journal editor they send the paper to.

The situation with in-text citations is very different. The specification is legally required to include the background of the invention, show how the invention solves a useful problem, and show how a person skilled in the art can make and use the invention without excessive experimentation. Though the applicant can describe the invention's background and method of construction using text and graphics, it is often easier to "incorporate by reference" (U.S. 37 CFR 1.57). That is, an applicant can simply refer to an earlier patent or an academic article when pointing to details necessary to understand or construct their invention. As these references are both technical and not as legally consequential as front page references, they are less likely to be added by patent attorneys. To again put things in academic terms, in-text citations play a role much closer to how citations are used in academic papers: a list of motivations, tools, similar work, and so on.

The difference between front page and in-text citations is not merely theoretical. Consider as an example patent application 11/407,702:

> "The requirement of positive GLI function for RAS action in human melanomas raised the possibility that tumor induced by direct oncogenic activation of RAS signaling could require SHH-GLI pathway function. To test this idea primary and metastatic melanomas were collected from mice expressing oncogenic NRASQ61K from the tyrosinase promoter (Ackermann, J. et al. Metastasizing melanoma formation caused by expression of activated N-RasQ61K on an INK4a-deficient background. Cancer Res. 65, 4005-4011 (2005))."

This 2005 article by Ackermann et al, on a technique used to generate oncogenic mice,

is cited *seven times* at various parts of the patent application specification, and the specification of the granted patent retains all of these. Nonetheless, the prior art for this patent does not include the Ackermann article.[13]

This distinction is not unusual. In our sample, restricting to applications that have been granted, 73 percent of the in-text citations do not appear on the front page of the granted patent. Going the other direction, 82 percent of the front page citations do not appear in the patent specification. These discrepancies exist even though the matched list of papers in the application specification and grant specification overlap almost perfectly, and the exact same matching algorithm is used on both datasets.[14]

Front page citations, of course, have a long and well-validated history among innovation scholars (e.g., Jaffe et al. [1993], Narin [1994]). They also have a number of skeptics, who have shown empirically that, for the reasons mentioned above, front page citations do not measure knowledge flow in the same manner as academic citations (Roach and Cohen [2013], Tijssen [2002], Meyer [2000]). In-text citations, purely on legal grounds, ought better measure real knowledge flows. In a companion paper (Bryan et al. [2020]), we empirically show that in-text citations are more closely linked to the knowledge of inventors and the firm's reliance on academia as a source of spillovers, while front page citations are more closely linked to patent value. This paper also documents the empirics of in-text citations across a variety of academic fields going back more than 30 years, and describes more fully the legal interpretation of each type of citation.

Although we contend that in-text citations better measure actual knowledge trans-

[13]The initial list of references forming the base of the non-patent prior art list was not even submitted to the USPTO until more than three years after the original patent application. The USPTO Public PAIR dataset includes the Image File Wrappers with these dates.

[14]Over 100 randomly selected patents were also investigated by hand, to ensure that these figures do not simply reflect error in the matching algorithm; from that sample, we found zero discrepancies relevant to the two comparisons described above. In-text and front page references do share some properties in common, such as their skewness: see Appendix Figure A6.

fer to inventors than front page citations, the broader question of how knowledge transfer relates to the level or direction of innovation is one that has long bedeviled the innovation literature. We do not claim to have solved the problem of identifying how much given knowledge inputs contribute to a given invention. Though revealed preference as in our model suggests that cited academic knowledge must have some value - otherwise why would inventors spend time and money acquiring it? - the nonexistent "paper trail of knowledge" is challenging to track. For this reason, it is important to caveat that our results directly only measure increased citation not increased innovation. We investigate further in Section 4.2 why the former reasonably proxies for the latter.

3.2 Summary Statistics and Estimation Technique

3.2.1 Summary Statistics

Tables 1 and 2 give summary statistics for articles and the patent applications which cite them. Articles in our sample receive a mean of .48 patent citations. For articles written in 2005, which have had the most time to collect citations, the mean number of citations is just over 1. Nearly 37 percent of the articles are funded by the NIH, a number which is roughly constant from 2005 to 2012 (Online Appendix Figure A2). 54 percent of the articles are eventually freely available on the internet, though this figure masks substantial heterogeneity across journals; for instance, the New England Journal of Medicine has made its articles freely available six months after publication throughout our sample period, while the Journal of Neurochemistry generally makes archives freely available only when required by a funder.

Among patent applications, 62.3% are assigned in the initial application. Of those, corporations and universities make up over 96 percent of all assignees. The first inventor

is in the United States on 64.8% of the citing patents. Most knowledge transfer from academic articles to patents takes place at a distance; on only 49% of the citing patents are the first inventor and the article first author in the same country, and only 18% in the same state (if American) or same country (otherwise). Most of the applications are not granted within the timeframe of our dataset: 31.2% are granted by March 2015, and 48.7% by August 2017.

To ensure that our patent-paper matches are not simply reflecting low-value or unusual patent applications, we can investigate geographic and other characteristics of the matched sample. Online Appendix Table A11 shows which countries and states do the most medical research in top journals, and which produce the most patents in our dataset citing that frontier research. The following facts are of note. First, Massachusetts, especially when it comes to patented science, stands out. If Massachusetts were a country, it would produce five times more research-citing patents per capita than any other country. Second, though there is a correlation between research output and patenting activity, it is not one-to-one. New Jersey, New Hampshire, California, Israel, Singapore and Belgium all produce many more research-citing patents than would be expected given their academic research output.[15] Locations with large government or institutional medical research centers like D.C., Maryland, Minnesota, New York, the UK and the Netherlands all produce less than would be expected. These geographies clarify that our patent-paper matches are generally capturing medical patents written in regions which are traditional biotech and pharma hotbeds.

[15]Note also that the differences in locations that do lots of academic biomedical research and lots of invention using that research further motivates focusing on article-to-patent transfers of knowledge. It is not the case that locations which are good in one are necessarily strong in the other.

3.2.2 The NIH Mandate

The NIH mandate requires funded research to be placed in an open access repository within one year of publication. It binds on all research first published after April 7, 2008. Of the 43 academic journals in our sample, 13 make more than 80% of their articles across the sample freely available as of 2013. In most cases, they are making nearly 100% freely available, so the NIH mandate caused no de facto change in accessibility.

The other 30 journals in our sample "gate" their archives in the absence of an open access mandate. Among articles published in those 30 journals, Figure 1 and Online Appendix Figure A3 show that the NIH-funded articles became 55 percentage points more likely to be freely available following the mandate, depending on the precise definition of "open access". Non-funded articles in those journals, on the other hand, became only 5 percentage points more likely to be freely available. For this reason, we refer to these 30 journals as being "affected" by the NIH mandate, and the other 13 journals as being "unaffected".[16]

Mandate compliance is less than perfect on both sides of the April 2008 boundary. In general, and especially at journals that do not make articles freely available unless required by an institutional mandate, authors are themselves responsible for uploading their research to PubMed Central. Less than perfect compliance after 2008, when only about 80 percent of NIH-funded research in affected journals is freely available, is driven by authors being unaware of the mandate, believing the mandate does not apply to them, simple forgetfulness, or attempts to avoid open access due to the fact that some journals charge fees on the order of $2000 to $5000 per article to permit free availability for readers.[17] Beginning in early 2013, the NIH began toughening

[16]Recall that funders other than the NIH also implemented open access policies during this period, so some small increase is to be expected.

[17]In general, funders permit grants to be used to pay these fees, but nonetheless the fees require diverting funds that could be used for other lab expenses.

enforcement, threatening delays on future grants for authors who don't make their previously-funded articles available. This policy caused a jump in free availability for articles published after 2013, but the policy was predicated on the stagnant and less-than-perfect mandate compliance for articles published between 2008 to 2012. That is, the nature by which the NIH enforced its mandate between 2008 and 2012, our "post-treatment period", was roughly constant (see, e.g., van Noorden [2013]).[18]

In the year before the mandate began, Figures 1 and A3 show that there was already a slow increase in the probability an NIH-funded article was freely available online. This reflects both that there was a voluntary, relatively unsuccessful, attempt to encourage NIH authors to make work freely available before April 2008, and that some authors may have assumed that the NIH mandate, stating that work published *after* that date must be made freely available *within one year*, referred to all research that had been published within a year of the mandate start date. This fuzzy compliance will be immaterial given our empirical strategies, which will require only that the mandate made a certain set of publications *more likely* to be freely available online, as Figures 1 and A3 make clear was the case. We will never use actual article-level availability or non-availability in these estimates.

3.3 Estimation Technique and Statistical Inference

Online Appendix Table A1 and Figure A4 show that open access articles are much more likely to be cited both by patents and other academic articles even after controlling for the journal, publication date, funder, and author country. This effect should not be interpreted causally, however. The causal effect may be overstated if articles subject to an open access mandate, such as those written at prominent institutions which support

[18] Also note that websites like Sci-Hub, which permit non-subscribers to access gated research illicitly, did not exist during the time period of our study.

OA, are inherently more likely to be cited, if journals made their archives open access under editorial leadership that was more generally concerned with applied science, or if journals selectively made high-profile results open access. Broadly speaking, it is difficult to assign causality without knowing why some articles were freely available and others were not.

A perfectly designed open access experiment goes beyond simply randomizing the free availability of articles. Open access will naturally only affect behavior if inventors we intend to treat actually know of and can find the article. Since potential users always have the option of buying access to an article, either individually or via a subscription, mandated open access is equivalent to a reduction in search cost, and the reduction in search cost is consequential only if there are many free-to-read articles in a centralized and easy to search location.[19] Therefore, an optimal experiment would construct a large database of scientific research, some of which is free to read and some available only at a cost, with random assignment to the two groups.

The NIH mandate, which affected 37 percent of articles published in top journals and led to deposit of these articles in the widely known Pubmed database, did not lead to assignment at random. Controlling for journal and time of publication, NIH funded articles before the mandate even began are 26 to 28 percent more likely to be cited by a patent, reflecting both the more US-heavy authorship and potentially the higher quality of the research (Online Appendix Table A2). That the NIH mandate affects only research published after April 2008, however, allows that time cutoff to help causally identify the effect of open access. As noted, compliance with the mandate was imperfect: in the 30 journals which gate nearly all of their articles in the absence of a mandate, NIH funding increases the probability a given article is freely available after April 2008 by around 50 percent, as was seen in Figures 1 and A3. Therefore, if the

[19] That results not only see their monetary cost fall, but can be found at that lower cost, was implicit in our search model in Section 2.

effect of treatment is linear in the probability of being made free to read, the true effect of the NIH policy may be as much as twice the treatment effects we report. We discuss noncompliance and its effect on our results further in Section 4.

We will estimate

$$y_i = f(PostApril08 \times NIH, PostApril08, NIH, X_i) \tag{1}$$

where y_i is a measure of article-level citations such as academic citations, total patent citations, the probability of at least one patent citation, or citations within a given time period following article publication, and X_i are article-level covariates such as publication time, journal, and first author location. Identification with the NIH mandate ensures that benefits ascribed to open access do not reflect selection into open access on the basis of journal policies (a journal that switches to open access may have a better editorial board, or a more applied focus) or home institution rules (elite universities may be more likely to require open access from their faculty).

This identification strategy requires that the use of NIH-funded research by industry did not differentially change in 2008 for reasons unrelated to open access. For instance, if the NIH itself was becoming relatively more likely to fund applied research around the same time as they began their open access mandate, we would be wrongfully conflating open access with this general applied reorientation.[20] We use two methods to account for this.

First, we estimate a placebo of Equation 1 using only the 13 journals in our sample which make nearly all articles free to read, whether NIH funded or not. If there is a general increase in the relative use of NIH-cited medical research compared to other research, then even NIH-funded articles in these 13 placebo journals should see

[20] We do not know of any NIH policy along these lines in 2008, but there was a general push toward applied impact within the NIH in the mid-2000s. See http://ncats.nih.gov.

a citation bump after April 2008 compared to unfunded articles. The placebo is also useful for investigating substitution. If the NIH mandate causes industry researchers to simply substitute easily found references to, say, a basic scientific fact, then the value of the increased citations caused by open access would be small. If, however, articles under open access see more citations while those with no change in access see no decrease in citations, then additional citations are more likely to represent real knowledge flows than citations of convenience.

Second, we formally estimate the triple difference

$$y_i = f(PostApril08 \times NIH \times Affected, X_i) \tag{2}$$

where X_i includes the covariates from Equation 1 as well as full saturation of the elements of the triple difference. That is, we investigate the relative change in i) citations to NIH-funded articles published after the mandate in journals which do not make everything free-to-read, compared to ii) citations for funded articles published after the mandate in unaffected journals.

A brief statistical caveat: in both estimates we are interested in the *percentage increase* in citation propensity (or total citations) conditional on open access status. In terms of statistical inference, then, we are investigating *multiplicative treatment effects*.[21] The reason for this is the parallel trends assumption underlying identification with a difference-in-difference approach. Our prior is that, if there were no open access mandate, NIH-funded articles would be more likely to be cited by a multiplicative rather than an additive factor compared to non-funded articles. That is, if 10% of unfunded

[21] Of course, the assumption must either be that open access generates a multiplicative increase in total cites *or* propensity to cite at least once. Truncation of cites at 1 and the fact that total cites is higher in the pre- than the post-period implies that if the multiplicative treatment assumption is true for total cites - for instance, if cites arrive according a possibly zero-inflated Poisson process at rate C for non-NIH and λC for NIH articles - then an estimated treatment effect of the NIH policy using truncated cites will *underestimate* the true effect. We return to this point in the conclusion.

articles and 20% of funded articles published in 2005 are cited by a patent, then we would not expect relative citation for articles published in a counterfactual 2012 without a mandate to be 2% and 12%. Rather, we would expect that if 2% of unfunded 2012 articles have been cited, then something like 4% of funded articles should have been cited.

If the outcome of interest is always positive, many researchers just log variables to convert multiplicative parallel trends to additive parallel trends, then use standard diff-in-diff techniques. In the cases like ours where the outcome variable is equal to zero for the majority of entries, log linearization is not possible. The problems with log linearization and the solution even in the case with many zeroes is well-studied in the international trade literature (e.g., Santos-Silva and Tenreyro [2006], Ciani and Fisher [2014]). Generically, with non-smooth dependent variables like a "was there a citation or not?" binary, point identification of treatment effects with nonlinear versions of the parallel trends assumption is impossible (Athey and Imbens [2006]). However, imposing somewhat stronger assumptions on the nature of the link function, coefficients of the nonlinear model can be estimated using poisson pseudo-maximum likelihood (ppml). Standard errors are asymptotically correct even with overdispersion (e.g., Santos-Silva and Tenreyro [2010], Santos-Silva and Tenreyro [2011], Hilbe [2007]).[22] We will use this model even when the dependent variable is a binary for comparability of results, and because the coefficients of logistic models are widely-misunderstood odds ratios rather than percentage increases (e.g., Zou [2004]). In Online Appendix 2, we show that alternative forms of estimating a multiplicative treatment effect are misleading. In particular, we show that the commonly-used $\ln(n+1)$ transformation, when used on binary or zero-inflated data, not only does not measure a multiplicative treatment effect, but rather estimates $\ln(2)$ times the OLS diff-in-diff treatment effect under the

[22]This estimation technique is much more common in the trade literature than in management, although it is not entirely unknown in the latter field; see, e.g., Agrawal et al. [2014].

assumption of additive parallel trends.

4 Results

Figure 2 displays the *ratio* of citations received by NIH funded compared to non-funded articles in the thirty journals affected by the 2008 NIH policy. This ratio, whether measured using total citations or the less skewed probability of at least one citation, is roughly constant before the NIH policy was implemented, albeit with nontrivial month-to-month variation. Following the mandate, the ratio slowly and continuously rises.[23]

Table 3 presents our primary estimates. Controlling for journal and publication month, moving from zero to complete open access would increase patent citations of academic research by 25.3%, increase the probability of at least one patent citation by 21.3%, and increase the the probability of at least one patent citation within 3 years of publication by 12.3%. Online Appendix Tables A3 and A4 show robustness of these estimates to alternative methods of controlling for the decay in citations over time, to additional covariates like the home country or state of the article author, and to restricting the diff-in-diff kernel to articles published within 24 months of the NIH mandate implementation. Online Appendix Figure A5 shows that our result is not being driven by articles in a single, or a small number, of journals.

Confirming prior research like McCabe and Snyder [2014], we find a precisely estimated zero increase in *academic* citations due to the NIH open access policy; this is not surprising given that biomedical academics tend to have both institutional access to journals and competent research assistants to help search the literature. The bottom panel of Figure 3 shows the null result within academia graphically.

[23]The increasing variance, rather than increasing trend, over time in this ratio is a result of lower propensity to be cited by patents for both funded and unfunded articles later in the sample. Recall again that patent applications are kept secret for a period, usually 18 months but often longer, hence the number of cites we observe as we become closer to the present is falling.

As discussed in the previous section, a general reorientation of NIH funding toward more applied projects around 2008, among similar concerns, may have generated our primary results even if open access actually did not affect patent citations. In order to rule this out, Table 4 and the top panel of Figure 3 investigate the change in citations to NIH-funded articles relative to non-funded articles within the 13 journals that make the vast majority of their back catalog freely available. For instance, the New England Journal of Medicine has made all research articles free-to-read online six months after publication since 2001 (Campion et al. [2001]). If the NIH was funding more applied projects after 2008, then a positive treatment effect of "open access" should be evident even in journals like the New England Journal of Medicine.

The top panel of Figure 3 shows that, in fact, there was no such increase in the citation advantage for NIH-funded work after 2008 in the journals unaffected by the mandate. The formal ppml estimates in Table 4 show precisely estimated null effects of open access in these placebo journals. Table 5 estimates a multiplicative triple difference of the relative increase in citations for NIH-funded articles published after April 2008 in journals that are expected to be affected by the mandate compared to NIH-funded articles published after April 2008 in unaffected journals. The triple-diff estimates accord nearly exactly with the estimates in our primary regression, finding a 26.5 percent increase in total patent citations, and a 14 to 20 percent in the probability of at least one citation. Again, citations within academia are relatively unaffected by the mandate.

Figure 4 summarizes our main results graphically.[24] Each panel shows the relative citation advantage for NIH-funded articles published in a given half year period, normalized to the citation advantage of NIH-funded articles in 2005. The top left panel shows that the patent citation advantage of NIH-funded articles is constant until 2008,

[24] A table with the estimates used in Figure 4 can be found in Online Appendix Table A5.

and that the advantage is positive in every half-year period after the first half of 2009.[25] On the other hand, the bottom left panel and two right panels show that there is neither an abrupt change nor a trend in the relative academic citation advantage or in the patent citation advantage for articles published in unaffected journals.

Table 6 and Online Appendix Figure A7 estimate our main results using only granted patents, in order to compare the treatment effect on front page citations (which only appear in grants and not applications) to in-text citations. While the effect of the NIH policy on in-text citations to granted patents is similar to our main results, the effect on front page citations is statistically indistinguishable from zero. The point estimate is that the NIH policy led to 4% higher probability of an article being cited on the front page, and 9% fewer total cites, though the latter measure is particularly noisy. This result is consistent with our discussion of the origin of in-text versus front page cites. Front page citations have a legal rationale, and only must be disclosed when the applicant is aware of the potential for the reference to relate to their patent claims. A lawyer would not have the incentive to actively search literature for potential references of this type. We return to this distinction when discussing limitations of our results in Section 4.

Table 7 and Online Appendix Tables A7 and A8 investigate the effect of open access within various subgroups. Table 7 shows that the main treatment effect is not being driven by low-value patents. The effect of open access is qualitatively similar to our primary estimates even if we restrict to patents assigned upon application (Table 7, Column 1 and 2) and patents with at least one related application filed to a foreign patent office (Column 5). All three measures proxy for high-value patents.[26] Patent

[25] Again, since Online Appendix Figure A1 shows that PubMed Central became more visible and more frequently used between 2008 and 2012, we should expect the citation advantage of open access articles to be growing over time, not constant throughout the post-mandate period.

[26] Patents assigned on application are correlated with patents assigned upon being granted in our data.

applicants in the same geographic region as the research they cite see the same effect of open access as those from more distant regions; this is perhaps not surprising given that spillovers are often highly localized, while our "regions" are at the level of a state or country (Columns 3 and 4). Online Appendix Table A7 attempts to identify the type of firm, rather than the quality of patent, that is associated with increased patenting, without consistent differences by assignee type. Online Appendix Table A8 suggests that open access affects patents with few inventors more than those with many inventors, although the differences are not themselves statistically significant. That said, even restricting to citations from patents with five or more inventors, there remains a large, positive impact of open access on patent citations. This evidence, though limited, is again consistent with the idea that the additional cites from open access are not merely coming from low-value patents.

Finally, Online Appendix Table A9 examines the effect of the NIH policy on patent citations when we weight the patents by the number of forward citations they themselves receive from further patents. Patents with forward citations are well-established as being more valuable inventions. Just under 30 percent of all articles which are cited are cited by a patent with a forward citation. These forward citations are highly skewed. The combination of these facts means weighed patent citations will be relatively noisy compared to our primary estimates. Nonetheless, the point estimates of the effect of the NIH policy - 20.8% more weighted patent citations and a 14.0% increase in the probability of being cited by at least one patent which is itself cited by future patents - are quite similar to our primary estimates. However, restricting to articles with at least one citation, the average weighted quality of citing patents conditional on total citing patents is statistically no different for treated articles. That is, the marginal knowledge in patents caused by open access mandates does not appear to shift the quality of the citing inventions. We note that this statement should be heavily caveated by the

noisiness of these estimates.

4.1 Threats to Identification and Interpretation

We have identified the effect of open access mandates on the use of academic knowledge in patents using two techniques, taking advantage of the large exogenous jump in the propensity an NIH funded article is open access after mid-2008, and the fact that some journals ought not be affected by this policy since they make their archives freely available no matter who funds the published research. The primary threats to identification and interpretation are threefold. First, the NIH may have changed other policies in the late 2000s which affect the citation of research in patents, and which our triple difference does not suitably control for. Second, the increase in patent citations may simply reflect low-value substitution, whereby a patent attorney or low-level employee of a lab is tasked with finding relevant scientific background for a patent and simply cites what is easiest to find. Third, since inventors always had the option to purchase journal subscriptions, or to purchase individual articles, the marginal value of induced extra citations may be low compared to the average knowledge flow overall in a patent. We handle these concerns in turn.

The first threat, that of NIH programs other than open access occurring at the same time, could most aptly be handled by taking advantage of the panel data nature of citations. A natural way to investigate the impact of open access policies is to look at articles which spent, for excludable reasons, more or less time as part of the PubMed database, or to look at within-article differences in citation probability before and after the article is added to the database. For example, in prior studies of open science more generally, Furman and Stern [2011] have taken advantage of the random accession of biomaterial into a centralized database, where biomaterial from some older studies and some new studies was added simultaneously, and Williams [2013] used quasirandom

variation in the amount of time individual parts of the human genome were restricted by Celera's license.

Since the NIH mandate relied on individual authors or their publishing journal to actually upload articles bound by the mandate, there is some minor variation in the exact delay between publication and free online availability. For instance, some articles were added after only 10 months, while others were not free online until 14 months after publication. In principle, then, we could investigate the month-by-month hazard rate of patent citation for articles that either are or are not yet open access, or could investigate whether longer delays attenuate our estimate of the effects of open access. The problem is both that this variation is so minor, particularly given the fact that very few citations come within a year of article publication, and that the underlying source of variation is likely to be connected to an article's propensity to be cited for other reasons. For instance, large labs, or authors who are very proud of a particular piece, may be less likely to absentmindedly submit their article to PMC later than required by the mandate.

Since a panel setup is infeasible, one might be concerned that our estimates, particularly our diff-in-diff, may simply be picking up other policies that affect NIH-funded research in the late 2000s. Although our placebo and triple difference should help mitigate this concern - recall that NIH funded research in journals whose open access status is unaffected by the mandate do not appear to gain any patent citation advantage - it would potentially be useful to take advantage of mandates other than the NIH rule which occur at times other than 2008. There are two reasons we do not try to take advantage of these mandates. First, all PubMed accessions of institutional or funded research we are aware of, other than articles affected by the NIH policy, are either very small in size or are very challenging to link to individual articles. The small potential size of alternative mandates can be seen in Figures 1 and A2, where only 6%

of non-NIH funded research even by 2012 in the thirty journal subset is freely available online, with close to zero availability prior to 2008. This 6% represents the maximal total number of articles bound by some mandate other than the NIH mandate. Second, we want to estimate the effect of open access *relative* to the article's citation pattern if it were gated. Therefore, we need a base rate of articles unlikely to be treated by any mandate. Hence, even if we had a large sample of articles treated by non-NIH mandates, we would only be able to estimate the differential effect of that mandate relative to what is, following the 2008 NIH mandate, an ever-smaller sample of untreated articles.

4.2 Interpretation of Treatment Effects

To interpret our empirical results, let us return to the model in Section 2. In particular, we want to understand how the relatively minor impediment of paying to read research could possibly generate meaningful economic distortions. As of March 2016, articles in the Journal of Biotechnology cost $37.95 for nonsubscribers. If these articles were free, would they be cited more by inventors? The empirical evidence suggests that they indeed would be, and not just in low-value inventions. But why? Are these references simply throwaway citations of no importance? Do these citations simply substitute for other references, leading to no net increase in the use of academic work?

The model suggests that in the absence of open access, authors will only read articles where the probability the article contains useful knowledge times the expected value of the increased private profit generated by the invention due to that knowledge exceeds the cost of the article. Consider a particular piece of knowledge that would increase the expected profitability of the invention by $10,000. If there are 300 articles that potentially contain that knowledge, and they cost $37.95 each, the inventor will not bother to search the literature. This remains true even if the *social value* of the invention, inclusive of consumer surplus and spillovers, is a multiple of that $10,000.

That is, the model suggests that wholly rational inventors will skip reading scientific literature even when the gains from doing so are quite large. A corollary is that the knowledge incorporated as a result of open access can be valuable. Indeed, theory suggests that these potential $10,000-or-more citations induced by open access can be more valuable than the average contribution of knowledge cited in by patents in the absence of open access.

Are these numbers reasonable? Placing a precise dollar figure which translates the treatment effects into a social loss demands far too heroic an interpretation of the model. That said, five features are important for bringing the model to data qualitatively. First, we must have an empirical analogue for the "piece of knowledge" our theoretical researcher was trying to find. Second, we need to know the value an additional piece of knowledge has in expectation for researchers with institutional access and those without. Third, we must estimate the difficulty of locating useful knowledge; that is to say, how many journals will you need to read before finding something worthwhile. Fourth, we need the effective cost of accessing an article if you don't have an institutional subscription. Fifth, we need the difference between the private value of an invention and its social value.

On the first measure, we argue that in-text citations fit the model quite well. As we have noted, the nature of in-text citations means that they will generally be added by the inventor themselves. They can incorporate a broad range of valuable knowledge inputs, including background facts, tools, techniques, motivations, and so on. Examining which journals are cited most frequently by patents, the highest per-article citation average is for articles in Nature Immunology and Cell Stem Cell. Articles in both of these journals are cited much more heavily than articles in "prominent" journals with high impact factors like JAMA or the New England Journal of Medicine. The fact that journals with a more applied orientation are cited more heavily is empirical

evidence, in addition the legal theory already discussed, supporting the validity of in-text citations as a real knowledge flow. Table 4 also shows that as open access increased citation to affected journals, it did not change citation in unaffected journals. This is consistent both with the search model and with the notion that in-text citations do not just represent ceremonial references.

On the relative value of knowledge flow for inventors without institutional access versus those with access, it will naturally depend on what industry is being examined. However, in biomedical research, small firms perform a great deal of early stage work where intellectual rather than regulatory or manpower bottlenecks are most severe. Nonetheless, small biomedical firms rarely have their own institutional subscription, which suggests that the value of academic knowledge they might obtain is not so high as to make the subscription model worthwhile. Proposition 1 shows that it is precisely these inventors - too small to make subscriptions worthwhile, yet still requiring knowledge neither too important nor trivial - who benefit the most from open access.

The extent of search required to find useful knowledge and the cost of accessing research without a subscription again will depend on the industry. On these points, we return to Lyman [2011], the correspondent to Nature Biotechnology we met earlier:

> The number of published biological science journals has been expanding for decades, driven by both scientific societies and for-profit publishers like Nature Publishing Group (NPG). Some of these journals have grown and divided like the bacteria that they often report on. NPG, for example, publishes not just Nature but also Nature Biotechnology, Nature Cell Biology, Nature Chemical Biology, Nature Genetics, Nature Immunology, Nature Medicine and Nature Neuroscience, to name a few, and a wide spectrum of Nature Review journals.

That is, the number of good journals, especially in biology, has expanded rapidly,

and the number of fields that must be covered by a biomedical researcher searching for useful knowledge has grown as well. The increasing burden of knowledge to reach the frontier means that surface-level investigations of neighboring fields have become tougher. On the size of spillovers, the fact that there is any increase in citation behavior at all due to open access means that, taking the model seriously, word-of-mouth is an insufficient substitute for scientific journals.

Two final caveats should be kept in mind. First, our sample is medical and biotech invention. Inventors in this class are particularly likely to have technical backgrounds, and to be familiar with reading academic research. It is not clear that the magnitudes we find here would translate to industries where inventors are less connected to academia. Second, we do not have direct evidence that the open access policy led to more or better invention. It is a longstanding problem in the economics of innovation to measure true knowledge flows, and an even harder problem to measure the relative contribution of particular pieces of knowledge in an invention to its social value.

5 Discussion and Conclusion

Institutional open access mandates have become increasingly common even though they appear to have only minor effects within academia. Academics, especially at top universities, have institutional access to published research. In the past few years, the US, UK and EU have all considered legislation which would either greatly expand mandated open access requirements, or greatly roll back existing mandates.

We show that open access causes patents to cite academic knowledge much more frequently. We measure citations with the novel tool of extracted in-text citations, which ought be more closely linked to the knowledge of the inventor themselves than the commonly-used front page patent citation. A theoretical model of search by inventors

suggests that these citations can represent real, valuable knowledge flows even when the cost of a journal article is relatively low. Inventors do not consume enough research because it is artificially costly. The proximate source of this cost is academic norms around publishing in high-prestige journals. Given the importance of access to research, what can be done? We can consider this question at four different constituent levels: managers of funders, universities, firms, and journals.

For firms, the main takeaway is that limited access to research is consequential for the quality of the firm's innovations. Referring the inventors to colleagues or friends with subscriptions does not constitute an effective solution, especially when the inventor needs to keep up to date with an ever-growing literature base. As a more comprehensive solution, the natural responses for any firm whose input supplier is generating inefficiency in the value chain by pricing above marginal cost is to internalize the externality. The difficulty here is that the market power of scientific journals derives not from some economic consideration, but from a non-market norm within academia about the types of venues where serious research needs to be published. Even if a consortium of small biotech firms started an academic journal which was free to readers, what academic would submit there instead of Lancet or JAMA or Cell? That is, academic norms create barriers to internalization.

An alternative for the firm managers is to pressure academics to shift toward more open publishing. This is unlikely with only pressure from industry. Martin Frank, the executive direction of the American Physiological Association, considered this in an essay in the New England Journal of Medicine. His verdict? "At a time of limited resources, should we be diverting funds from research in order to fund open-access publishing? Personally, I think not" (Frank [2013]). That is, though there are clear efficiency harms from the current structure of academic publishing, the distribution of winners and losers favors precisely the academic societies who would need to be pres-

sured to change norms. Therefore, a coordinated effort between various distributional losers, including industry, leading academics, and funders, may prove more fruitful than pressuring publishers and scholarly societies alone. A number of recent cases have seen editorial boards at profit-maximizing journals defect to open access journals, taking their personal prestige to the less distortionary new title.[27]

A third option for firms, directly suggested by the model, is to work with complementors who lower the cost of figuring out which journal articles contain useful information before purchasing the article. Automated or assisted literature search companies are now widespread. Proposition 1 showed that the harm of open access was strictly increasing in the coarseness of the partition of potential articles that might contain the information a firm needs. The higher the cost of articles and the coarser the information spread, the more valuable automated curation and literature search tools become.

A final option, and one that has become much more common since the timeframe of the data in this paper, is theft. Pirate websites like Sci-Hub and LibGen, with illicit pdfs serving as scientific samizdat, have become mainstream very quickly. Essentially any article in any journal can be read simply by copying the article URL into scientific piracy sites. The existence of scientific piracy may be welfare-improving, since in addition to pure transfers of surplus from publishers to readers, it reduces deadweight loss.[28] The deadweight loss in question is the economic benefit from innovation that is improved with knowledge found while searching the academic literature. The empirical effects of the 2008 NIH mandate suggest these deadweight losses are not trivial. For this reason, even if the academic norms that give publishers market power continue, we may see that market power decline, and hence legally accessible research become easier to obtain, because of competitive pressure from piracy.

[27] Lingua in 2015 and the Journal of Algebraic Combinatorics in 2017 are prominent examples.

[28] That is, these services will play an analogous role to file sharing in the music industry. See Waldfogel [2012].

Piracy may make it more difficult to sustain the status quo system of journal financing. As was seen in the music industry, using price discrimination to protect legacy income streams can increase the demand for piracy, and eventually harm core revenue. Hence, out of concern for the dynamic interaction of journal pricing and piracy demand, publishers may wish to shift pricing toward one that is more accessible for non-academics like research-intensive firms. It is also possible to accomplish this through a coordination effort hinted at above. One example is the Sponsoring Consortium for Open Access Publishing in Particle Physics (SCOAP3), which is a partnership of over three thousand libraries, funding agencies and research centers in 44 countries and 3 intergovernmental organizations. Member countries contribute funding commensurate with their publications, and SCOAP3 distributes this funding to publishers for costs involved in providing Open Access. Publishers, then, convert key journals in the field of High-Energy Physics to Open Access at no cost for authors, and reduce subscription fees for all customers, which enables contributions to SCOAP3. Such an innovative pricing mechanism would not have been possible without the buy-in of the publishers.

At the policymaker, or funder, level, decisions about open access need to account for its effects outside academia in addition to within. The high price of individual academic articles required to maintain incentives for institutions to purchase subscriptions is disproportionately damaging to inventors who would otherwise build sequentially on the existing base of scientific results. Therefore, if the objective of the funder is creating a public good and ensuring its seamless dissemination, then taking steps to limit externalities created by the market power of journals is paramount. Mandating open availability of publications resulting from funded research, as in the NIH rule, is one method. Creating or supporting alternative dissemination mechanisms that are in line with the incentives of academics, such as creating a new journal with the coordination of leading faculty is yet another method.

References

Ajay Agrawal, Carlos Rosell, and Timothy S. Simcoe. Do tax credits affect r&d expenditures by small firms? evidence from canada. *NBER Working Paper 20615*, 2014.

Mohammad Ahmadpoor and Benjamin F. Jones. The dual frontier: Patented invention and prior scientific advance. *Science*, 2017.

Juan Alcacer, Michelle Gittelman, and Bhaven Sampat. Applicant and examiner citations in u.s. patents: An overview and analysis. *Research Policy*, 38(2), March 2009.

Susan Athey and Guido W. Imbens. Identification and indifference in nonlinear difference-in-difference models. *Econometrica*, 74(2), 2006.

Pierre Azoulay, Joshua Graff Zivin, Danielle Li, and Bhavan Sampet. Public r&d investment and private sector patenting: Evidence from nih funding rules. *NBER Working Paper 20889*, 2015.

Yannis Bakos and Erik Brynjolfsson. Bundling information goods: Pricing, profits, and efficiency. *Management Science*, 1999.

Elizabeth Popp Berman. Why did universities start patenting?: Institution-building and the road to the bayh-dole act. *Social Studies of Science*, 2008.

Bo-Christer Björk, Annikki Roos, and Mari Lauri. Scientific journal publishing: yearly volume and open access availability. *Information Research*, 2009.

J. Blumenthal and G. Freiburger. MLA/AAHSL Sequestration Letter. *Unpublished*, 2012.

Kevin A. Bryan. Sequential search on a partition. *Working Paper*, 2020.

Kevin A. Bryan, Yasin Ozcan, and Bhaven Sampat. A user's guide to in-text citations. *Research Policy*, 2020.

Edward W. Campion, Kent R. Anderson, and Jeffrey M. Drazen. A new web site and a new policy. *New England Journal of Medicine*, 344:1710–1711, 2001.

Emmanuel Ciani and Paul Fisher. Dif-in-Dif Estimates of Multiplicative Treatment Effects. *ISER Working Paper*, 2014.

Wesley M. Cohen, Richard R. Nelson, and John P. Walsh. Links and impacts: The influence of public research on industrial r&d. *Management Science*, 2002.

C.A. Cotropia, M. Lemley, and B. Sampat. Do Applicant Patent Citations Matter? *Working Paper*, 2013.

P. Davis, B. Lewenstein, D. Simon, J. Booth, and M. Connolly. Open access publishing, article downloads, and citations: randomised controlled trial. *British Medical Journal*, 2008.

Jeremiah Dittmar. The welfare impact of a new good: The printed book. *Working Paper*, 2020.

J. Evans and J. Reimer. Open Access and Global Participation in Science. *Science*, 2009.

Lee Fleming and Olav Sorenson. Science as a map in technological search. *Strategic Management Journal*, 2004.

M. Frank. Open But Not Free - Publishing in the 21st Century. *The New England Journal of Medicine*, 2013.

J. Furman and S. Stern. Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research. *AER*, 2011.

Y Gargouri, C. Hajjem, V. Lariviere, L. Carr, Y. Gingras, T. Brody, and S. Harnad. Self-selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. *PLoS ONE*, 2010.

Patrick Gaule and Nicolas Maystre. Getting cited: Does open access help? *Research Policy*, 2011.

David J. Hardisty and David A. F. Haaga. Diffusion of Treatment Research: Does Open Access Matter? *Journal of Clinical Psychology*, 2008.

S. Harnad, T. Brody, F. Vallieres, L. Carr, S. Hitchcock, Y. Gingras, C. Oppenheim, H. Stamerjohanns, and E. R. Hilf. Green and Gold Roads to Open Access. *Nature*, 2004.

Joseph M. Hilbe. *Negaitve Binomial Regression*. Cambridge University Press, 2007.

John Houghton, Alma Swan, and Sheridan Brown. Access to Research and Technical Information in Denmark. *Working Paper*, 2011. URL `http://www.fi.dk/publikationer/2011/adgang-til-forskningsresultaterog-teknisk-inform`

Hans K. Hvide and Benjamin F. Jones. University innovation and the professor's privilege. *Working Paper*, 2017.

Alessandro Iaria, Carlo Schwarz, and Fabian Waldinger. Frontier knowledge and scientific production: Evidence from the collapse of international science. *Quarterly Journal of Economics*, 2018.

Adam B. Jaffe, Manuel Trajtenberg, and Rebecca Henderson. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics*, 108(3), 1993.

Doh-Shin Jeon and Domenico Menicucci. Bundling electronic journals and competition among publishers. *Journal of the European Economic Association*, 2006.

Doh-Shin Jeon and Jean-Charles Rochet. The Pricing of Academic Journals: A Two-Sided Market Perspective. *American Economic Journal: Microeconomics*, 2010.

Madian Khabsa and C. Lee Giles. The number of scholarly documents on the public web. *PLOS One*, 2014.

Heekyung Hellen Kim. The effect of free access on the diffusion of scholarly ideas. *Working Paper*, 2012.

Stewart Lyman. Industry access to the literature. *Nature Biotechnology*, 2011.

Mark J. McCabe and Christopher M. Snyder. Identifying the Effect of Open Access on Citations Using a Panel of Science Journals. *Economic Inquiry*, 2014.

Martin Meyer. What is Special About Patent Citations? Differences Between Scientific and Patent Citaitons. *Scientometrics*, 2000.

Joel Mokyr. *The Gifts of Athena*. Princeton University Press, 2002.

David Mowery and Bhaven Sampat. *Essays in Honor of Edwin Mansfield*, chapter The bayh-dole act of 1980 and university-industry technology transfer: A model for other OECD governments?, pages 233–245. 2005.

F. Murray, P. Aghion, M. Dewatripont, J. Kolev, and S. Stern. Of Mice and Academics: The Role of Openness in Science. *NBER Working Paper 14819*, 2012.

Fiona Murray and Scott Stern. Do formal intellectual property rights hinder the free flow of scientific knowledge? an empirical test of the anti-commons hypothesis. *Journal of Economic Behavior & Organization*, 2007.

Francis Narin. Patent bibliometrics. *Scientometrics*, 1994.

Jason Owen-Smith, Massimo Riccaboni, Fabio Mammolli, and Walter W. Powell. A comparison of u.s. and european university-industry relations in the life sciences. *Management Science*, 2002.

Patrick A. Puhani. The treatment effect, the cross difference, and the interaction term in nonlinear 'difference-in-differences' models. *Economics Letters*, 115(1), 2008.

Michael Roach and Wesley M. Cohen. Lens or prism? patent citations as a measure of knowledge flows from public research. *Management Science*, 59(2), February 2013.

Bhavan Sampat and Heidi Williams. How Do Patents Affect Follow-on Innovation? Evidence from the Human Genome. *Working Paper*, 2014.

Bhaven Sampat. When do applicants search for prior art? *Journal of Law and Economics*, 53(2), May 2010.

J.M.C. Santos-Silva and Silvana Tenreyro. The log of gravity. *Review of Economics and Statistics*, 88(4), November 2006.

J.M.C. Santos-Silva and Silvana Tenreyro. Further simulation evidence on the performance of the poisson pseudomaximum likelihood estimator. *Economics Letters*, 112 (2), 2011.

Joao M. C. Santos-Silva and Silvana Tenreyro. Currency Unions in Prospect and Retrospect. *Annual Review of Economics*, 2010.

Donald E. Stokes. *Pasteur's Quadrant*. Brookings Institution Press, 1997.

P. Suber. Ensuring Open Access for Publicly Funded Research. *British Medical Journal*, 2012.

A. Swan. The Open Access Citation Advantage: Studies and Results to Date. *Technical Report, University of Southampton*, 2010.

Robert J. W. Tijssen. Science dependence of technologies: evidence from inventions and their inventors. *Research Policy*, 2002.

Richard van Noorden. Nih sees surge in open-access manuscripts. *Nature News Blog*, 2013. http://blogs.nature.com/news/2013/07/nih-sees-surge-in-open-access-manuscripts.html.

Joel Waldfogel. *Innovation Policy and the Economy, Volume 12*, chapter Music Piracy and Its Effects on Demand, Supply, and Welfare. University of Chicago Press, 2012.

M. Ware and M. Monkman. Access by UK Small and Medium-sized Enterprises to Professional and Academic Information. *Publishers Research Consortium Report*, 2009. URL `http://www.publishingresearch.net/documents/PeerReviewFullPRCReport-final.pdf`.

H. Williams. Intellectual Property Rights and Innovation: Evidence from the Human Genome. *The Journal of Political Economy*, 2013.

Guangyong Zou. A modified poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology*, 159(7), 2004.

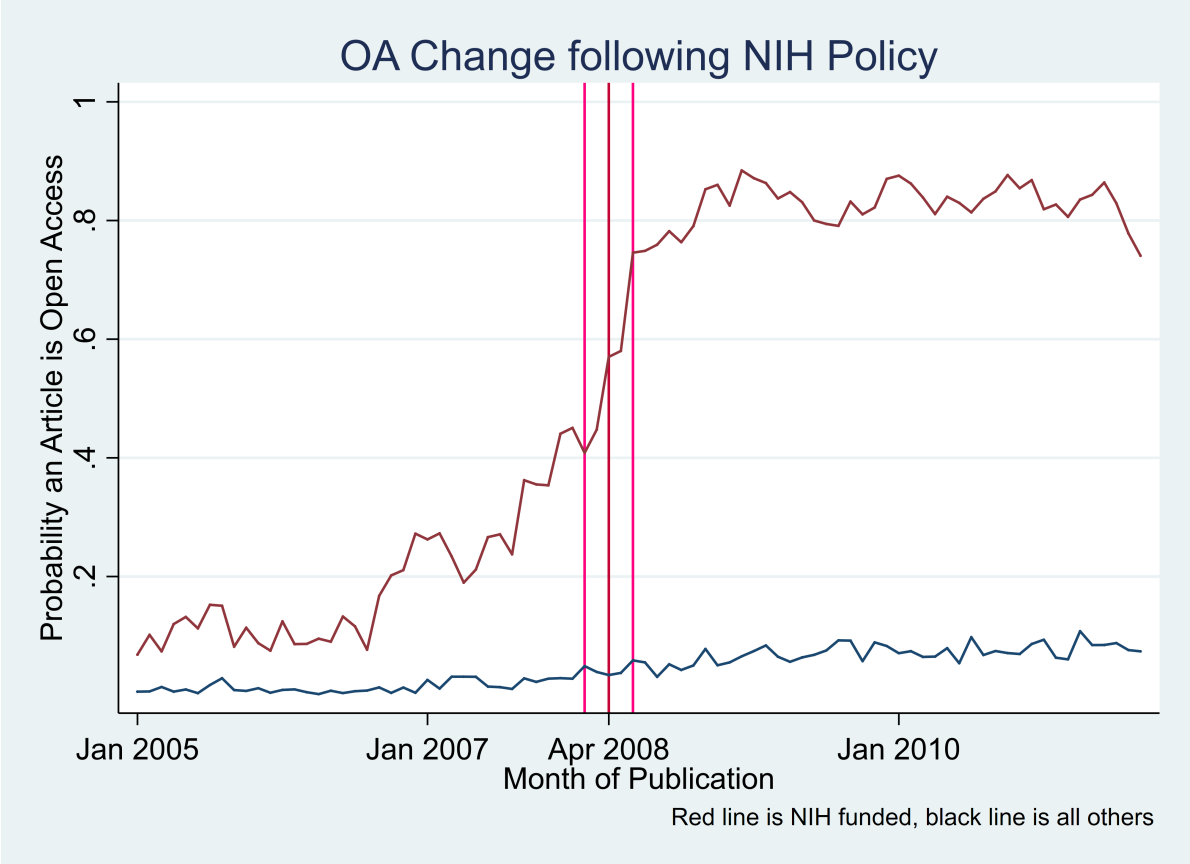

Figure 1: Sample consists of all medical research articles in the subset of 30 journals that generally do not make research freely available unless forced to. "Open access" refers to the article being freely available anywhere on the internet (the "Free full text" category on PubMed) as of July 2013. The red (center) line represents the April 2008 NIH policy, and the pink (left and right) lines represent two months before and after the official beginning of the policy.

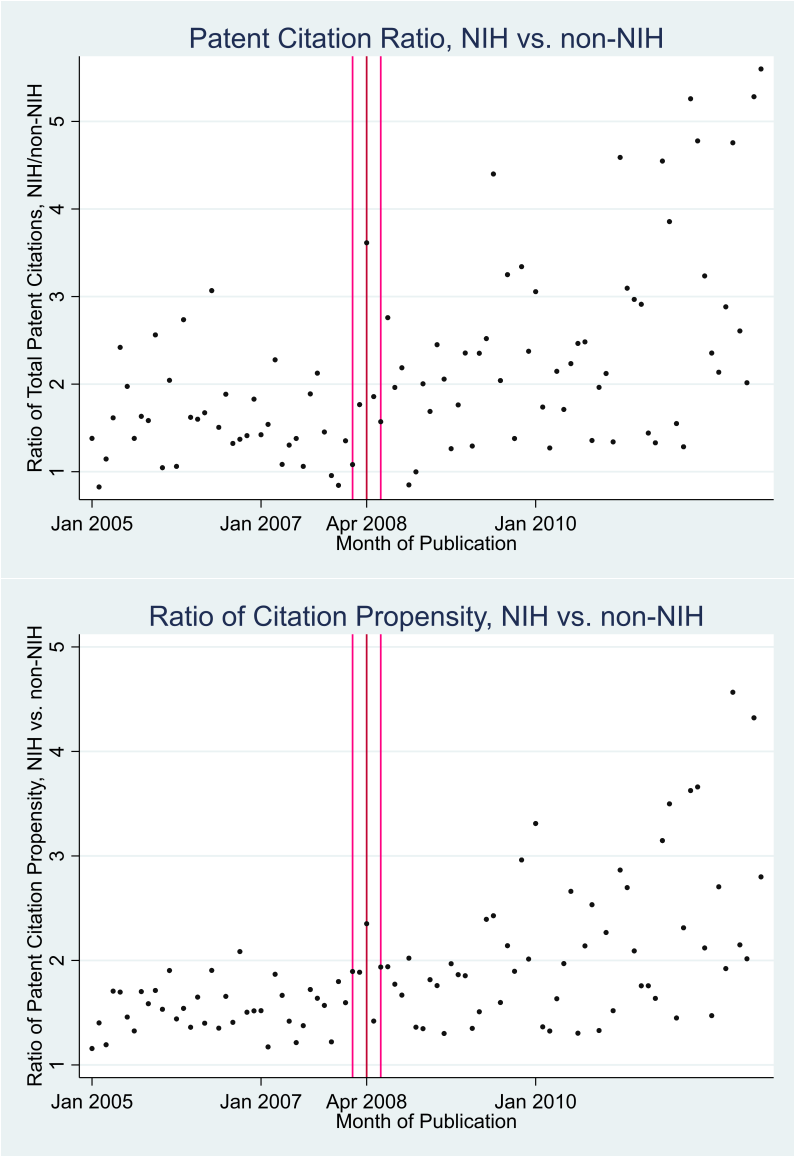

Figure 2: Ratios of patent application citations for NIH funded articles versus non-funded articles, by article publication month. The top panel gives the ratio of total patent application citations. The bottom panel gives the ratio of propensities to have at least one patent application citation. Articles restricted to the thirty journals which generally do make articles freely available unless required by a mandate. The red (center) line represents April 2008, and the pink (left and right) lines represent two months before and after the official beginning of the policy.

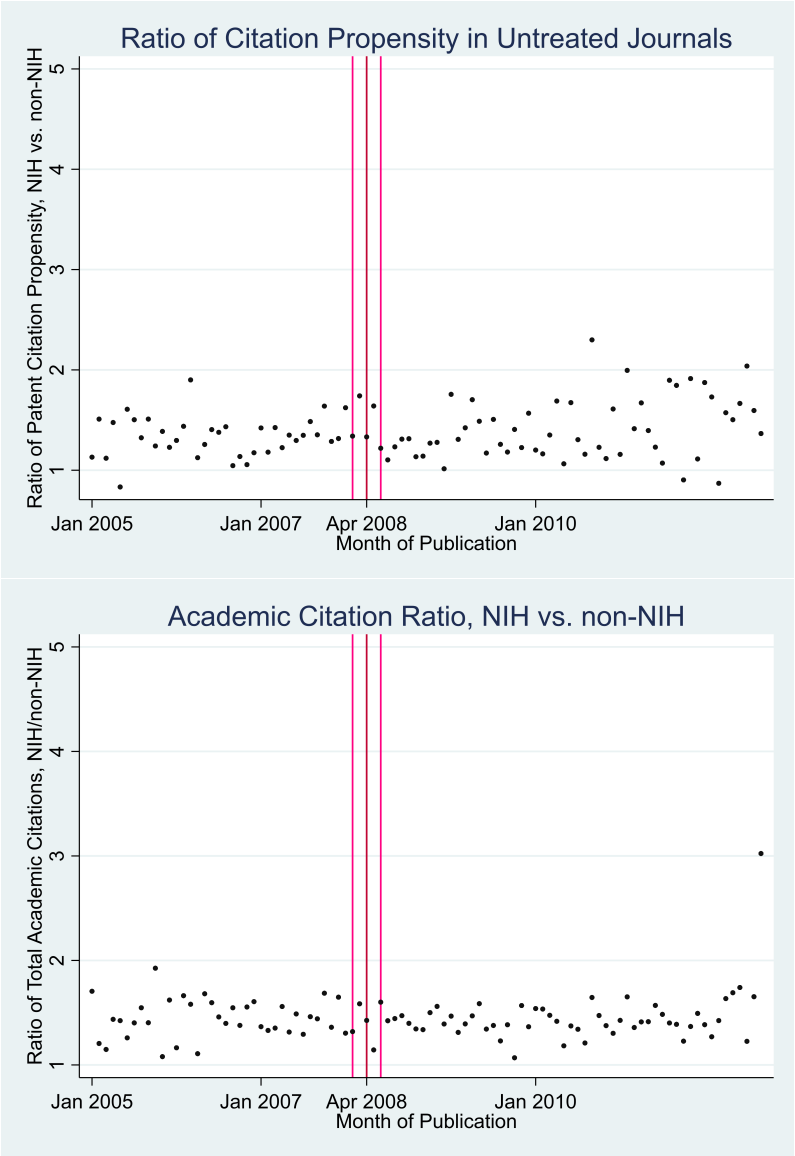

Figure 3: Ratios of the propensity to be cited for NIH funded articles versus non-funded articles, by article publication month. The top chart is a placebo estimate of the previous figure, restricting to the thirteen journals which make nearly all research freely available and hence are unaffected by the mandate. The bottom figure considers academic citations before and after the mandate. The red (center) line represents April 2008, and the pink (left and right) lines represent two months before and after the official beginning of the policy.

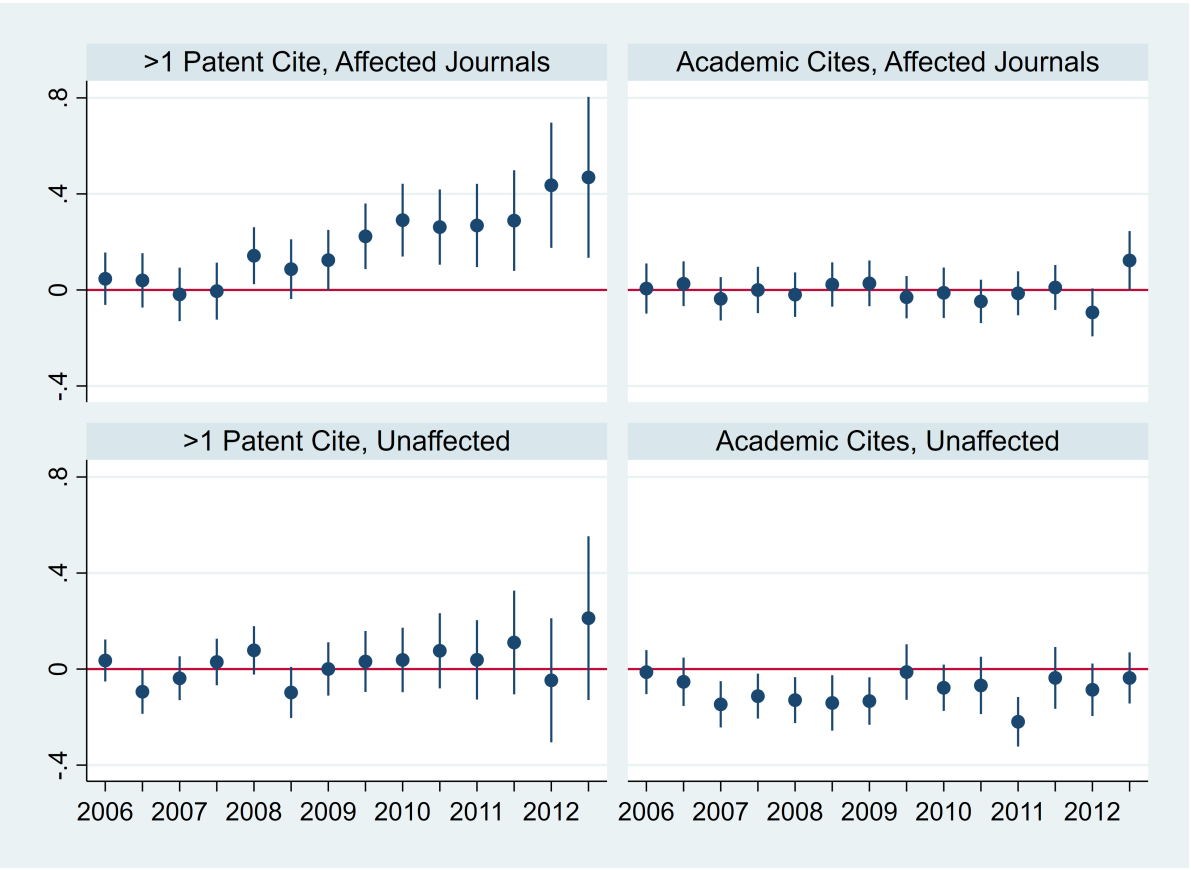

Figure 4: By half year, the estimated percentage difference in the ratio of the independent variable for NIH versus non-NIH funded research, relative to the ratio in 2005, where estimates are ppml controlling for journal and polynomial of publication month. These percentages are not scaled by 2, and hence following the discussion in Section 3, reflect the estimated effect of the NIH mandate rather than the effect of going from zero to complete open access. The top left panel is essentially the difference-in-difference of Table 3 in event study form, the bottom left panel the placebo using the "unaffected" thirteen journals which generally make all research freely available and hence are unaffected by the mandate, and the right hand side panels show that academic citations are generally unaffected by the open access mandate.

Table 1: Summary Statistics for Articles

	All Articles
Observations	132,872
Mean # of Patent Citations	.475
Mean # of Patent Citations to Year 2005 Papers	1.052
Minimum Number of Citations	0
Maximum Number of Citations	248
Pr($\geq$ 1 patent citation)	.170
Available via PubMed Central	.265
Available via Free Full Text	.543
Funded by NIH	.367
Mean # of Academic Cites	55.8
Pr(First author in United States)	.474

Includes all research articles published between January 2005 and December 2012, matched to the universe of public US patent applications from January 2005 to March 2015.

Table 2: Summary Statistics for Patent Applications

Total patents in sample	2,898,005
Unique citing patents	28,136
Total Cites	63,106
Mean # of Patent Authors	3.65
Pr(patent is assigned)	.623
Pr(assigned to a corporation)	.333
Pr(assigned to a major biotech or pharma firm)	.059
Pr(assigned to a university)	.284
Pr(assigned to an individual)	.003
Pr(assigned to a government, excl. universities)	.014
Pr(first inventor in United States)	.648
Pr(inventors in multiple countries)	.150
Pr(application submitted in >1 country)	.825
Pr(patent granted by March 13, 2015)	.314
Pr(patent granted by August 7, 2017)	.487
Pr(first inventor in same country as first author of cited article)	.491
Pr(first inventor in same region as first author of cited article)	.180

"Major biotech or pharma firm" includes 27 high-revenue firms listed in the appendix. Region means "same country if outside the US" or "same state if both inside the US". Probabilities all refer to the sample of patent applications which cite at least one medical research article.

Table 3: Difference in Difference estimates

	Pat. Cites	Pr($\geq$ 1 Pat. Cite)	Pr($\geq$ 1 Pat. Cite in 3 yr)	Acad. Cites
NIH $\times$ post 04/08	.2253**	.1930***	.1160**	-.0046
	(.0845)	*(.0358)*	*(.0454)*	*(.0249)*
(in % terms)	25.3	21.3	12.3	-0.4
NIH dummy	.3075***	.2832***	.3557***	.2055***
	(.0617)	*(.0236)*	*(.0337)*	*(.0197)*
Observations	71337	71337	71337	70184

The unit of observation is the academic article, restricting to the thirty journals which rarely make research free-to-read in the absence of a mandate. All estimates are Poisson pseudo-maximum likelihood (errors are robust by construction), and all include journal and publication month dummies. "In % terms" is equal to e^{β}.

Statistical significance indicators: *: .1, **: .05, ***: .01

Table 4: Placebo Difference in Difference estimates

	Pat. Cites	Pr($\geq$ 1 Pat. Cite)	Pr($\geq$ 1 Pat. Cite in 3 yr)	Acad. Cites
NIH $\times$ post 04/08	-.0217	.0136	-.0186	-.0509*
	(.0575)	*(.0318)*	*(.0409)*	*(.0270)*
(in % terms)	-2.1	1.4	-1.8	-5.0
NIH dummy	.2026***	.2242***	.2480***	.1295***
	(.0394)	*(.0183)*	*(.0273)*	*(.0198)*
Observations	61408	61408	61408	60310

The unit of observation is the academic article, restricting to the thirteen journals which make almost all research free-to-read, and hence which ought be unaffected by the 2008 NIH mandate. All estimates are Poisson pseudo-maximum likelihood (errors are robust by construction), and all include journal and publication month dummies. "In % terms" is equal to e^{β}.

Statistical significance indicators: *: .1, **: .05, ***: .01

Table 5: Triple Difference Estimates

	Pat. Cites	Pr($\geq$ 1 Pat. Cite)	Pr($\geq$ 1 Pat. Cite in 3 yr)	Acad. Cites
NIH×post 04/2008×Affected	.2354**	.1780***	.1323**	.0443
	(.1016)	*(.0479)*	*(.0611)*	*(.0367)*
(in % terms)	26.5	19.5	14.1	4.5
NIH dummy	.1981***	.2229***	.2467***	.1290***
	(.0395)	*(.0183)*	*(.0272)*	*(.0197)*
Observations	132745	132745	132745	130494

The unit of observation is the academic article. All estimates are Poisson pseudo-maximum likelihood (errors are robust by construction), and all include journal and publication month dummies, and full saturation of post-April 2008 dummies, NIH funding status, and a dummy indicating whether a journal is expected to be affected by the open access mandate or whether it generally makes all or almost all archived articles free-to-read. "In % terms" is equal to e^{β}.

Statistical significance indicators: *: .1, **: .05, ***: .01

Table 6: Effect is Much Stronger for In-Text Citations in Granted Patents

	Front Page Cites	≥ 1 Front Page Citation	In-Text Cites	≥ 1 In-Text Citess
NIH $\times$ post 04/08	-.0970	.0349	.1430***	.2091**
	(.0983)	*(.0422)*	*(.0512)*	*(.0987)*
(in % terms)	9.2	3.6	15.4	23.3
NIH dummy	.4070***	.3058***	.3189***	.3603***
	(.0678)	*(.0289)*	*(.0328)*	*(.0707)*
Observations	71337	71337	71337	71337

The unit of observation is the academic article. Dependent variable is number of citations or probability of at least one citation *in a granted patent*, using either front page or in-text citations. All estimates are Poisson pseudo-maximum likelihood (errors are robust by construction), and all include journal and publication month dummies, and full saturation of post-April 2008 dummies and NIH funding status.

Statistical significance indicators: *: .1, **: .05, ***: .01

Table 7: Difference in Difference Subgroup estimates

	Assigned	Unassigned	Same Region	Diff. Region	Big Family
NIH×Post-04/08	.1952***	.1837***	.2213***	.1730***	.1697***
	(.0429)	*(.0527)*	*(.0770)*	*(.0395)*	*(.0404)*
(in % terms)	21.6	20.2	24.8	18.9	18.5
NIH dummy	.2966***	.3465***	.6200***	.2459***	.2682
	(.0282)	*(.0334)*	*(.0514)*	*(.0258)*	*(.0260)*
Observations	71337	71337	71337	71337	71337

The dependent variable in all estimates is the probability of at least one cite of the listed type. All estimates are Poisson pseudo-maximum likelihood with Huber-White robust standard errors, and all include journal and publication month dummies. "In % terms" is equal to e^{β}.

Statistical significance indicators: *: .1, **: .05, ***: .01

Online Appendix 1: Data Construction

Data Sources

Our data consists of a sample of 132,872 academic research articles published in 43 prominent medical and biotechnology journals between 2005 and 2012, and the universe of public patent applications from January 1, 2005 to March 13, 2015, a sample of 2,989,005 applications. We select all research articles, omitting notes, summaries and editorials, from the following medical journals: The New England Journal of Medicine, Lancet, The Journal of the American Medical Association, The Journal of Experimental Medicine, The Journal of Clinical Investigation, Neuron, Nature Medicine, Circulation, The Journal of Clinical Oncology, Nature Immunology, Immunity, Blood, Gastroenterology, The Journal of the American College of Cardiology, The Journal of Neuroscience, Nature Neuroscience, Neuroimage, Cancer Cell, Oncogene, Hepatology, Genome Research, Biological Psychiatry, Cancer Research, Journal of Neurochemistry, Arthritis and Rheumatism, Lancet Neurology, Clinical Cancer Research, Clinical Infectious Diseases, Brain, Journal of Allergy and Clinical Immunology, Neurology, Cell Stem Cell and Lancet Oncology. We also select all research articles from the following biotechnology journals: Nature Biotechnology, Trends in Biotechnology, Applied Microbiology and Biotechnology, Biotechnology and Bioengineering, Tissue Engineering, Journal of Biotechnology, Journal of Neural Engineering, Biotechnology Progress, Biotechniques, and Transgenic Research. These journals were selected by searching for high impact factor general interest biomedical, specialty biomedical, and biotechnology journals.

The bibliographic data on academic research articles come from the PubMed database, and include the full citation of each article (author(s), title, journal, page numbers, country of first author address, year of publication). We observe the first author country or U.S. state of origin in all but 2.26% of the articles. The data also includes the date at which the full text of the article became public on the PubMed database. For articles not on PubMed Central, we extract whether the article was available via the "Free Full Text" link on June 30, 2013, where FFT denotes articles freely available on the internet though not on the PubMed Central server. In 13 of these journals, over 80% of their archives were freely available online as of mid-2013; in the others, the vast majority of the back catalog was not free to read as of the same date.[29] We also extract the total number of academic citations as of June 2014 from Thomson Reuters' Web of Science.

The data on patent applications come from the publicly available USPTO Patent Application Publication Full Text files. These files include the full text of the patent applications, as well as bibliographic information, including the application and publication dates, and the inventor names and locations.

[29] The exact 80% cutoff is unimportant.

Matching Papers and Patents

The main challenge for our study is to identify the citations patent applications make to research articles. Unlike granted patent data, the patent application files do not have a separate section listing patent and non-patent citations in a standard format as prior art; instead, there are references within the application specification, some of which will later be considered prior art, and others of which will remain in the granted patent but will never appear in the prior art list. As we discuss in the body of this article, we believe the in-specification references may more accurately track what management researchers have in mind when they think of the "paper trail" of knowledge flows, but extracting these non-standardized references is a difficult problem.

To identify whether an article is cited by a patent, one needs to search for parts of the article information within the full text of the patent application. Given the 132,872 articles in our sample, just searching for the article first author name and article year would result in more than a quarter of a million queries to 533 weekly xml files, with each file having an average size of 0.5 GBs. Therefore, we have developed the following algorithm to identify the matches within a reasonable timeframe.

1 - The patent applications are provided as xml files. A single line in this xml file may contain an entire paragraph in the patent application text, hence a line may contain thousands of characters. Investigating a subset of the files, we have identified that there are very few lines longer than 7000 characters in length; therefore, we have kept only the first 7000 characters of each line in the xml file.

2 - Each line in the xml file starts with an xml tag identifying the information in that line. Through investigation of a subset of the files, we have found that references are nearly always included in the lines with tags "p" and "li", which contain the body paragraphs and list elements, respectively. Therefore we dropped the remainder of the files, and kept only lines with these tags. The remaining portions of the files also contain a minimal amount of citations, but investigation by hand suggests that these are mostly repetitions of citations also made elsewhere within the same patent application. In any case, we have no reason to believe that the citations from these lines have any relationship to open access status of the paper, hence should not contaminate our results.

3 - We then identify lines that contain any of the journal names in our list of 43 journals. For the purposes of this search, the journal list is augmented by various common abbreviations of the same journal name. For example, to capture New England Journal of Medicine, eight different abbreviations were searched for, including the following: "NEJM", "N.E.J.M", "N. Engl. J. Med", and "New England J. Medicine". In total, to identify the 43 journals in our sample, we have searched for 186 different abbreviations of these journal names.

4 - We eliminate any lines not containing the four digits of at least one year from 2005 to 2012 within 200 characters of a journal name identified in the previous step.

5 - We eliminate lines that do not contain the first author's last name within 150 characters of the journal title. In this step, we are only identifying the first citation

to a single journal within a single line. In other words, if two different articles from the same journal are cited within a single line of the patent application, then we may or may not capture the second one depending on how far apart it is located from the first citation. We have no reason to believe that missing such citations would bias our results.

6 - Among the matches identified so far, we eliminate matches which include neither the article page numbers nor the first four words of the article title within 150 characters of the journal title.

7 - Finally, we manually investigated a sample of citations to papers where author last names appeared frequently in our dataset: Brown, Chan, Chang, Chen, Cheng, Choi, Guo, Hu, Huang, Jiang, Johnson, Jones, Kim, Lee, Li, Lin, Liu, Lu, Ma, Park, Singh, Smith, Song, Southgate, Sun, Tang, Wang, Williams, Wong, Wu, Xu, Yang, Yu, Zhang, Zhao, Zheng, Zhou and Zhu.

This algorithm identified 63,106 citations made from patent applications to academic articles in our sample, coming from 28,136 unique patents.

Any algorithm of this type needs to balance between Type I and Type II errors. In this context, a Type I error is erroneously claiming the existence of a citation. Investigation by hand suggests that the matches identified by the algorithm contain less than one percent Type I errors. A Type II error happens if the algorithm fails to identify an existing citation. For instance, "In 1989 Stephan J. Weiss in the New England Journal of Medicine conducted bacterial sensitivity studies on E. Coli and toxicity on tissue in guinea-pigs" in patent application 12/101,775 is too vague, lacking both an article title and a journal issue number, for our algorithm to match it with a specific article. The extent of Type II errors of this kind is difficult to quantify, but we have no reason to believe that missing matches are correlated with the open access status of articles, and hence they ought not bias our results. We investigated a number of less restrictive algorithms, but generally they resulted in many more Type I errors with very few additional legitimate matches.

We also check for self-citations, where patent applicant authors cite their own academic article. Such cites are at most 1.5% of our sample and likely far less, where 1.5% represents cites from patentees with the same last name as a paper author in the same country or state citing within 12 months of the paper publication date. Many of these potential self-cites represent continued research by the same scientist, or coincidences with common names. Such a low number of self-cites is to be expected since we are investigating patent applications made *after* the paper publication date.

Note that we only observe publicly available patent applications. The modal patent is kept secret for 18 months after its application is made, though a combination of patent applicant requests, foreign patent office rules, and rapid grant dates means that there is a lot of heterogeneity in this delay. Since our patent data is through March 2015, this means that we only observe the modal patent applied for in months before October 2013, and hence for mechnical reasons the closer an article date gets to the present, the fewer patent citations we will observe. We have examined all of estimates restricting to citations within three years of the article publication date, and aside from

adding noise the estimates are nearly identical to our preferred estimates.

Identifying Assignee Type

It is not obvious how to assign patent applications as corporate, university, or otherwise. Our technique was to manually examine our patent matches to generate a list of case- and spacing-dependent strings common for university assignees (the word "university" in many languages, the names of large research centers, etc.) and corporate assignees (the assignee name used by common patentees, the words for "Inc." or "LLC" in many languages, etc.). This technique allowed us to sort over 98% of the assigned organizations (a single patent may have multiple assignees, and we attempt to sort each assignee on each patent) into either a University/Research Center, Government, Corporation, Other Hospital or Individual. The remaining 2% or so could be assigned by hand, but we prefer for replicability reasons to use only automated assignation. Note that the particular strings below are uniquely chosen for medical-related patents from 2005 to 2015, so this technique is not a broadly applicable automatic categorization process.

"University" was a designation given to patents with any of the following in one of their patent assignee strings: "university", "alumni", " univ", "national cancer", "brigham", "jackson lab", "research center", "akademie", "vib ", "RIKEN", "Eye & Ear", "medical school", "national jewish health", "eth zurich", "Center for", "univeristy", "higher education", "cold spring harbor", "akadamie", "centre for", "fundacio", "Université", "centre", "planck", "universuty", "Universitât", "fundacion", "UNIVERSITÀ", "agence nationale", "insitute", "UNIVERSITÉ", "eye and ear infirmary", "Society for", "Unversity", "cancer centre", "universite", "institue", "istituto", "cancer center", "fondation", "universiteit", "universitet", "universitaet", "city of hope", "educational fund", "zentrum", "consejo", "ecole", "universtiy", "centro", "kettering", "mayo", "schule", "institucio", "centrum", "hospital for sick", "children's hospital", "academisch", "universita", "universit´at", "unviersity", "georgia tech", "school of", "consiglio nazionale", "intellectual properties", "fondazione", "national centre", "centro nacional", "centre national", "foundation", "regents", "council", "fred hutchinson", "general hospital corporation", "universidade", "research hospital", "medical center", "foundation", "universitat", "universidad", "colegio", "univerisite", "institut", "institute", "instituto", "trustees", "academia", "academy", or "college". These strings were picked following manual investigation in order to limit type I and type II errors, and attempt to capture academic research hospitals as well as universities themselves.

"Government" was a designation given to patents with any of the following in one of their patent assignee strings, if that patent assignee string was not previously denoted "University": "her majesty", "as represented by", "agency", "department of", "dept. of", "dept of", " NIH", "NHS ", " NHS", "prefecture", "global alliance", "commonwealth scientific", "international aids", "Commisariat", or "Commissariat".

"Corporation" was a designation given to patents with any of the following in one of their patent assignee strings, if that patent assignee string was not previously assigned to "University" or "Government": " LLC", " Inc", " Gmbh", " Ltd", "Corporation", " Corp",

"inc.", "l'oreal", "biomerieux", "s.p.a", "pharnext", "nektar", "janssen", "gingko bioworks", "ooo ", " SL.", "Galderma", "Moderna ", "bio-rad", " Co", "B. V", " LLC", "d.o.o.", " aps", "a.r.l.", "n.v.", "GlaxoSmithKline", " Pharma", "L.L.C.", "merck", "law group", "pierre fabre", "gesellschaft", " AB", "B.V.", " AG", "wyeth", "S.L.", "S.A.", "Ltd.", "G.m.b.h.", " SE", " Kaisha", "z o.o.", "s.l.u.", "AstraZeneca", ",LLC", "BV", " holdings", "K.K.", " KK", " SA", " GmhH", ",Inc", " Spa", " NV", " N.V.", "venture capital", " Oy", ",Ltd", " ehf", " s.p.a.", " srl", " s.r.l.", "Sanofi", " AS", " S.A.", "A/S", "Pharmaceuticals", " Limited", "Laboratories", or " plc".

"Hospital" was a designation given to patents with any of the following in one of their patent assignee strings, if that string was not previously assigned to "University", "Government" or "Corporation": "hospital", "hopital", "hopitaux", "hospita", "HÔPITAUX", "Red Cross", "punainen risti", or "health system".

"Individual" was a designation given to patents with an individual assignee name.

We also denoted separately corporate-assigned patents that were assigned to one of the 27 largest pharma or biotech firms, by revenue. Thus, "Major biotech firm" was a designation given to patents assigned to Novo Nordisk, Baxter, Gilead, Biogen Idec, Teva, Celgene, Merck, GlaxoSmithKline, CSL, Alexion, Regeneron, Squibb, Genzyme, Pfizer, Novartis, Sanofi, AstraZeneca, Bayer, Eli Lilly, Wyeth, Hoffmann, La Roche, Boehringer, Takeda, Amgen, Sankyo or Astellas, permitting common abbreviations in patent applications by these firms. Assignation to wholly owned subsidiaries with a different name would not be captured by this measure.

Identifying Granted Patents

Patent applications are linked to granted patents in two ways. First, the grant bulk data was individually parsed. Second, Google Patents was scraped for the "also published as" field on each patent application in our sample, then scraped to determine whether that corresponds to a granted patent with the application number we started with. In over 99.9% of our sample, these two methods give identical results; it appears clerical errors explain the handful of discrepancies.

Online Appendix 2: Alternative Identification Models

In this appendix, we show our main effects of interest using a simple difference in means, and note issues using a log-linear OLS model or probit/logit difference-in-difference estimations. Since most articles are never cited by a patent, our data contains many zeroes, hence log-linear OLS is estimated using $ln(n+1)$ as the dependent variable. We also discuss precisely what is being estimated in our main effect.

We discuss the following four facts. First, a simple difference in means closely matches our ppml estimates in the main results. This is because the control variables - journal code, publication date, etc. - are roughly symmetric in our data on either side of the NIH mandate. Second, a log-linear OLS model estimates a treatment effect of the NIH policy on citation behavior of the *opposite sign* that we find in our main results. We prove that when estimating difference-in-difference on a binary outcome with a multiplicative true treatment effect, the sign of the log-linear OLS will be *identical* to the sign of a simple linear OLS model assuming an additive treatment effect. We show why linear OLS gets the sign wrong in our context, and hence why log-linear OLS will as well. To our knowledge, this issue with adding 1 to the dependent variable and estimating diff-in-diff using a log-linear OLS has not been explicitly discussed in prior literature.

Third, we show that logit and probit coefficients on the NIH policy are qualitatively aligned with our ppml estimate, although the interpretation of this variable as a treatment effect, and inference properties, are less well-established. Finally, we mention that our primary treatment effect should be interpreted as the treatment effect on the average, not the average treatment effect. Further, we note that the necessary identification assumptions for the extensive margin and intensive margin treatment effect estimates in the main results are not the same, but that the assumptions converge as the extensive margin average becomes close to zero.

Raw Difference in Means

Without controlling for covariates, a naive average multiplicative treatment effect can be computed under the assumption that NIH-funded articles are cited x times more frequently than unfunded articles in the absence of an open access mandate, or alternatively that NIH-funded articles are x times more likely to be cited than unfunded articles in the absence of a mandate. Under that assumption, open access increases patent citations by

$$\begin{aligned} \frac{E(y_{FundedOAt})}{E(y_{FundedNoOAt})} &= \frac{E(y_{FundedOAPost})}{E(y_{UnfundedNoOAPost})} \times \frac{E(y_{UnfundedNoOAPost})}{E(y_{FundedNoOAPost})} \\ &= \frac{E(y_{FundedOAPost})}{E(y_{UnfundedNoOAPost})} \times \frac{E(y_{UnfundedNoOAPre})}{E(y_{FundedNoOAPre})} \end{aligned} \tag{3}$$

where y_{abc} represents citations to (probability of being cited by) articles of NIH-funding status a that are bound by an OA mandate b in time period c, where Pre and Post refer to articles published before or after the NIH mandate. The final equality holds by the multiplicative parallel trends assumption. All four terms are observable in the data, hence under the mulitplicative treatment parallel trends assumption, the average treatment effect can be identified.

In the data, the mean post-mandate citation propensity for funded articles bound by the mandate $E(y_{FundedOAPost})$ is 14.20% and the mean post-mandate citation propensity for unfunded articles not bound by the mandate $E(y_{UnfundedNoOAPost})$ is 7.36%. The mean pre-mandate citation propensity for funded articles $E(y_{FundedNoOAPre})$ is 27.28% and the mean pre-mandate citation propensity for unfunded articles$E(y_{UnfundedNoOAPre})$ is 17.79%. The overall treatment effect of open access on citation propensity is therefore 25.85%, very similar to the ppml treatment effect with covariates of 21.29% found in Table 3 of the main results. Likewise, the four expectations using total patent citations instead of citation propensity are, respectively, .3332, .1587, .9252, and .6124. The estimated treatment effect is therefore 33.02%, again similar to the ppml estimate in Table 3 of the main results of 25.26%. As in the main results, these are lower bounds, since the NIH mandate only increased open access probability by roughly 50 percent.

Linear Models

Note that the baseline number of citations declines over time, so the post-treatment number of citations under any open access rule will be lower than the pre-treatment number under that same status. It is evident from Online Appendix Figure A4, by examining pre-April 2008 averages, that an additive treatment effect assumption is likely to be counterfactual: NIH funded articles published in the first year of our sample recieve .55 more citations than non-funded articles, but by the final year of our sample, the average article receives .065 citations. It would not be plausible - indeed, would be mathematically impossible, for there to be a .55 additive gap in citation rates in that period. Ignoring this issue and estimating a standard diff-in-diff with OLS generates a *negative* treatment effect, as seen in Online Appendix Table A10.

To understand why, look at the conditional means calculated above. The absolute pre-period difference in citation propensity for NIH-funded versus unfunded articles is 9.49 percentage points. Unfunded articles in the post-period have a citation propensity of 7.36%. If we assumed additive parallel trends, we would expect funded articles without open access to have a citation propensity of 7.36+9.49=16.85%. Since the observed probability is 14.20%, OLS without covariates would estimate a treatment effect of -2.65%, very close to the negative full model estimate in column 2 of A10.

Of course, one can in theory take logs and estimate multiplicative treatment effects using OLS. Since the dependent variable using either total cites or the probability of being cited at least once is equal to zero for most observations, this cannot be done directly with our data. Taking logs of the number of cites plus one, or the citation propensity plus one, *does not properly estimate multiplicative treatment effects.* Why?

Consider a binary dependent variable. Taking logs of that variable plus 1, the dependent variable is either ln(2) or ln(1) = 0. Hence $E(\ln(y_{abc}))$ for any funding status, time period, and treatment status is equal to the probability $y_{abc} = 1$ times ln(2). Therefore, the log linear additive treatment effect will be estimated to be

$$[E(y_{FundedOAPost}) - E(y_{UnfundedNoOAPost})] - [E(y_{FundedNoOAPre}) - E(y_{UnfundedNoOAPre})] \times \ln(2)$$

which is exactly ln(2) times the linear OLS treatment effect, precisely what we observe in columns 2 and 4 of Table A10.[30] That is, *with binary dependent variables, we "undo" the log-linearization of the model by adding 1 to the dependent variables.* Since the counterfactual assumption of additive parallel trends in the standard OLS estimate led to the wrong sign on the coefficient, this wrong signed coefficient will be retained in the log linearized model. When the dependent variable is a zero-inflated count variable like total citations in columns 1 and 3 of Table A10, the link between OLS and the log linearized model is not as tightly linked, but the same conceptual issue will occur.

Online Appendix Table A10 also estimates logistic and probit versions of our primary model. While the coefficient on the interaction terms are positive, is it known in the literature both that this interaction term does not represent a treatment effect (Puhani [2008]). Worse, even if that coefficient represented a treatment effect under *some* identifying assumption, the relevant assumption necessary to interpret the output of a logistic or probit regression is not multiplicative parellel trends.

Finally, what precisely is ppml estimating? It is known that with heterogeneous treatment effects, ppml is estimating the treatment effect on the average, not the average treatment effect. More critically, we need to be careful about correctly specifying our identifying assumption. Our two primary dependent variables of interest are total citations and the probability of at least one citation ("cited once"). We may be interested in both if we believe the skewed nature of total citations adds noise. Since cited once truncates total citations, in general it cannot be the case that the multiplicative parallel trend will hold for both outcome variables. However, as the citation rate gets sufficiently small, the two assumptions become equivalent. In particular, assume that citations are generated by a Poisson process, where gated articles receive a mean of p citations and free-to-read articles receive a mean of λp. The probability of exactly one citation is $\exp^{-p} p$ and $\exp^{-\lambda p} \lambda p$. The multiplicative parallel trends assumption is that $\frac{\lambda p}{p} = \lambda$ is constant. In the limit as p goes to zero, $\frac{\exp^{-\lambda p} \lambda p}{\exp^{-p} p} \to \lambda$. If the base probability p is positive, then depending on whether the truncated citations or the total citations have a constant multiplicative ratio for the treated and control groups, either the total citation treatment effect is overstated, or the cited once treatment effect is understated.

[30] Because the OLS model includes covariates for time trends and journal of publication, the covariates need not be in precisely that proportion; in our case, however, those covariates are balanced enough across groups that we, to three decimal places, do in fact have exactly the ln(2) relationship between treatment effects in the linear and log-linear models.

Online Appendix 3: Additional Tables/Robustness

This appendix contains the following auxiliary estimates and robustness checks.

Table A1: Open access articles are cited more often in the raw data

Table A2: NIH funded articles are more cited than unfunded articles in the pre-period

Table A3: Primary estimates are robust to alternate time trends and including additional covariates like location of article authors

Table A4: Primary estimates are robust to restricting to the 24 months before and after April 2008

Table A5: Primary treatment effect can be seen in each half-year period, as in Figure 4

Table A6: Subgroup estimates using placebo journals show consistent null effects

Table A7: Primary estimates splittling citations assigned to universities, corporations, and small corporations.

Table A8: Primary estimates splitting citations from small, medium, and large numbers of co-inventors

Table A9: Primary estimates weighing citations by forward patent citations

Table A10: Primary estimates using OLS, log-linear OLS, logit, and probit

Table A11: The geography of medical researchers who publish in top journals, and of inventors who cite this research, are not totally aligned

Figure A1: Monthly downloads from PubMed Central are increasing over time, as the service becomes more well-known

Figure A2: NIH funding probability is constant over time during our sample period

Figure A3: Alternate definition of "open access" shows an even starker shift in open access availability of NIH funded articles after April 2008, as compared to Figure 1

Figure A4: In the raw data, there is a large patent and academic citation gap between open access and gated articles

Figure A5: Investigating the effect of open access on a journal-by-journal basis, our main estimates are not driven by a small number of journals

Figure A6: In-specification citation properties like skewness look very similar to the properties seen in prior art citations to academic literature

Figure A7: Front page citation propensities show only a small treatment effect from open access, if anything

Table A1: Open Access in the Raw Data

	Patent Cites	Pr($\geq$ 1 Pat. Cite)	Pr($\geq$ 1 Pat. Cite)	Academic Cites
Open Access	.4337***	.2441***	.2140***	.2897***
	(.0623)	*(.0218)*	*(.0218)*	*(.0153)*
(in % terms)	54.3***	27.6***	23.9***	33.6***
NIH dummy	.2071***	.2445***	.1152***	.0947***
	(.0285)	*(.0125)*	*(.0162)*	*(.0107)*
Country Dummies?	N	N	Y	N
Observations	132,745	132,745	129,749	130,494

The unit of observation is the academic article. All estimates are Poisson pseudo-maximum likelihood (errors are robust by construction), and all include journal and article publication month dummies. "Open Access" is a dummy equal to one for articles freely available via the PubMed FFT designation as of June 2013.

Statistical significance indicators: *: .1, **: .05, ***: .01

Table A2: NIH Funded Articles are More Cited in Pre-Period

	Pr($\geq$ 1 Pat. Cite)	Total Pat. Cites
NIH Dummy	.2460***	.2337***
	(.0147)	*(.0354)*
(in % terms)	27.9	26.3
Observations	56,650	56,650

The unit of observation is the academic article, restricted to those published before the NIH mandate begins. All estimates are Poisson pseudo-maximum likelihood with Huber-White robust standard errors, and all include journal and publication month dummies.

Statistical significance indicators: *: .1, **: .05, ***: .01

Table A3: Primary Estimates with Alternate Time Trends and Covariates

	[1]	[2]	[3]	[4]	[5]	[6]
NIH×Post-04/08	.2147***	.2970***	.1932***	.1927***	.2597***	.2081***
	(.0844)	*(.0898)*	*(.0358)*	*(.0380)*	*(.0843)*	*(.0359)*
(in % terms)	23.9	34.6	21.3	21.3	29.7	23.1
NIH dummy	.3173***	.2810***	.2834***	.2801***	-.0247	.1167***
	(.0622)	*(.0640)*	*(.0236)*	*(.0241)*	*(.0772)*	*(.0276)*
Pub Month Quadratic					Y	Y
Pub Month Quartic	Y		Y			
Journal-Spec. Time Trend		Y		Y		
Article Author Location					Y	Y
Observations	71337	71337	71337	71337	69223	69223

[1],[2],[5]: Total patent citations
[3],[4],[6]: Pr(≥1 patent citation)

The unit of observation is the academic article, and the sample restricts to the 30 journal subset as in Table 5 in the main paper. All estimates are Poisson pseudo-maximum likelihood with Huber-White robust standard errors, and all include journal dummies and a publication month quadratic. Location dummies are state and country fixed effects linked to the location of the first author for the article in question. "In % terms" is equal to e^{β}.

Statistical significance indicators: *: .1, **: .05, ***: .01

Table A4: Primary Estimates Restricting to +/-24 Months from April 2008

	Total Cites	>1 Cite	Total Cites	>1 Cite
NIH×Post-04/08	.2098**	.1560***	-.0413	.0129
	(.1003)	*(.0450)*	*(.0715)*	*(.0397)*
(in % terms)	23.3	16.9	-4.0	1.3
NIH dummy	.3151***	.2980***	.1590***	.1970***
	(.0694)	*(.0314)*	*(.0483)*	*(.0254)*
Affected Journals	Y	Y		
Unaffected (Placebo) Journals			Y	Y
Observations	35887	35887	31830	31830

The unit of observation is the academic article. Estimates restricted to sample of articles published between April 2006 and March 2010, or two years before and after the NIH policy was implemented. All estimates are Poisson pseudo-maximum likelihood with Huber-White robust standard errors, and all include journal dummies and a publication month quadratic. "In % terms" is equal to e^{β}.

Statistical significance indicators: *: .1, **: .05, ***: .01

Table A5: Primary Estimates on a Half-year by Half-year Basis

	>1 Cite, Affec.	>1 Cite, Unaffec.	Acad., Affec.	Acad., Unaff.
NIH×2006H1	1.048	1.036	1.006	.987
	(.0695)	*(.0551)*	*(.0638)*	*(.0550)*
NIH×2006H2	1.041	.910*	1.026	.948
	(.0716)	*(.0509)*	*(.0582)*	*(.0582)*
NIH×2007H1	.982	.963	.964	.864**
	(.0665)	*(.0535)*	*(.0530)*	*(.0506)*
NIH×2007H2	.995	1.030	1.000	.894**
	(.0717)	*(.0607)*	*(.0587)*	*(.0509)*
NIH×2008H1	1.153**	1.081	.9807	.879**
	(..0832)	*(.0663)*	*(.0553)*	*(.0511)*
NIH×2008H2	1.091	.907	1.023	.868**
	(.0824)	*(.0587)*	*(.0574)*	*(.0609)*
NIH×2009H1	1.132*	1.000	1.028	.875**
	(.0867)	*(.0676)*	*(.0595)*	*(.0527)*
NIH×2009H2	1.250***	1.032	.970	.988
	(.1038)	*(.0798)*	*(.0520)*	*(.0695)*
NIH×2010H1	1.337***	1.039	.988	.925
	(.1233)	*(.0849)*	*(.0632)*	*(.0542)*
NIH×2010H2	1.299***	1.079	.953	.934
	(.1238)	*(.1027)*	*(.0524)*	*(.0678)*
NIH×2011H1	1.308**	1.039	.986	.803***
	(.1379)	*(.1044)*	*(.0548)*	*(.0504)*
NIH×2011H2	1.335**	1.117	1.010	.964
	(.1701)	*(.1467)*	*(.0576)*	*(.0755)*
NIH×2012H1	1.547***	.9543	.911	.917
	(.2454)	*(.1499)*	*(.0553)*	*(.0610)*
NIH×2012H2	1.599**	1.236	1.131*	.964
	(.3258)	*(.2564)*	*(.0843)*	*(.0624)*
NIH dummy	1.303***	1.253***	1.230***	1.199***
	(.0493)	*(.0357)*	*(.0426)*	*(.0427)*
Observations	71337	61408	70184	60310

IRR of NIH funding on patent cites (in terms of "probability of at least one cite to a given article") and academic cites, by Half-Year, for articles in journals affected and unaffected by the 2008H1 NIH policy, as in Figure 4, where coefficients are relative to Year 2005 articles. The unit of observation is the academic article. The dependent variable in all estimates is the probability of at least one cite of the listed type. All estimates are robust ppml with journal dummies and a publication month quadratic. Translation of treatment effects into % terms is omitted for brevity.

Statistical significance indicators: *: .1, **: .05, ***: .01

Table A6: Difference in Difference Subgroup Estimates in Placebo Journals

	[1]	[2]	[3]	[4]
NIH×Post-04/08	.0430	-.0504	-.0241	.0545
	(.0383)	*(.0483)*	*(.0633)*	*(.0476)*
(in % terms)	4.4	-4.9	-2.4	5.6
NIH dummy	.2230***	.2592***	.0578	.3295***
	(.0220)	*(.0267)*	*(.0361)*	*(.0277)*
Observations	61408	61408	61408	61408

[1]: Assigned Patents
[2]: Unassigned Patents
[3]: Corporate Assignee
[4]: University Assignee

The unit of observation is the academic article. The dependent variable in all estimates is the probability of at least one cite of the listed type. All estimates are Poisson pseudo-maximum likelihood with Huber-White robust standard errors, and all include journal dummies and a publication month quadratic. "In % terms" is equal to e^{β}.

Statistical significance indicators: *: .1, **: .05, ***: .01

Table A7: Difference in Difference Subgroup estimates II

	University	University	Corporate	Corporate	Small Corp.	Small Corp.
NIH×Post-04/08	.2565***	.1548*	.0390	.2732**	.0761	.2290**
	(.0519)	*(.0910)*	*(.0711)*	*(.1274)*	*(.0866)*	*(.1141)*
(in % terms)	29.2	16.7	4.0	31.4	7.9	25.7
NIH dummy	.3992***	.5343***	.1902***	.0099	.2143	.1728
	(.0343)	*(.0634)*	*(.0452)*	*(.0923)*	*(.0550)*	*(.0752)*
Pr($\geq$ 1 patent cite)	Y		Y		Y	
Pr(total patent cites)		Y		Y		Y
Observations	71337	71337	71337	71337	71337	71337

[The unit of observation is the academic article. Small corporations are cites from firms other than the 27 large biotech and pharma firms described in Online Appendix 1. All estimates are Poisson pseudo-maximum likelihood with Huber-White robust standard errors, and all include journal and publication month dummies. "In % terms" is equal to e^{β}.

Statistical significance indicators: *: .1, **: .05, ***: .01

Table A8: Difference in Difference Subgroup estimates III

	≤ 2 Inventors	≤ 2 Inventors	≥ 3 Inventors	≥ 3 Inventors	≥ 5 Inventors	≥ 5 Inventors
NIH×Post-04/08	.2535***	.2472***	.1680**	.2057**	.1463**	.1950*
	(.0508)	*(.0920)*	*(.0452)*	*(.0918)*	*(.0707)*	*(.1191)*
(in % terms)	28.9	28.0	18.3	22.8	15.8	21.5
NIH dummy	.3277***	.3584***	.2884***	.2759***	.2386***	.1111
	(.0329)	*(.0648)*	*(.0299)*	*(.0671)*	*(.0459)*	*(.0839)*
Pr(≥ 1 patent cite)	Y		Y		Y	
Pr(total patent cites)		Y		Y		Y
Observations	71337	71337	71337	71337	71337	71337

The unit of observation is the academic article. All estimates are Poisson pseudo-maximum likelihood with Huber-White robust standard errors, and all include journal dummies and a publication month quadratic. "In % terms" is equal to e^{β}.

Statistical significance indicators: *: .1, **: .05, ***: .01

Table A9: Treatment Effect Weighting Cites by Forward Patent Citations

	[1]	[2]	[3]
NIH×Post-04/08	.1890	.1309*	-.0630
	(.1848)	*(.0727)*	*(.0812)*
(in % terms)	20.8	14.0	-6.1
Total Citations			.0096***
			(.0018)
NIH dummy	.3106**	.3935***	-.0319
	(.1273)	*(.0453)*	*(.0672)*
Observations	71337	71337	9987

[1]: Patent Citations Weighted by Forward Patent Citations
[2]: Binary that article is cited at least once by at least one patent with at least one forward citation
[3] Average quality of citing patents (weighted cites divided by total cites), restricting estimation to articles with at least one patent citation

The unit of observation is the academic article. All estimates are Poisson pseudo-maximum likelihood with Huber-White robust standard errors, and all include journal and publication month dummies. "In % terms" is equal to e^{β}.

Statistical significance indicators: *: .1, **: .05, ***: .01

Table A10: Primary Estimates Using OLS, Logit, Probit

	Total Cites	≥ 1 Cite	Total Cites	≥ 1 Cite	Total Cites	≥ 1 Cite
NIH×Post-04/08	-.1864***	-.0324***	-.0619***	-0225***	.0427*	.1413***
	(.0490)	*(.0061)*	*(.0080)*	*(.0043)*	*(.0264)*	*(.0480)*
(in % terms)					2.5	10.7
NIH dummy	.2794***	.0755***	.1038***	.0523***	.2489***	.4305***
	(.0472)	*(.0054)*	*(.0072)*	*(.0037)*	*(.0648)*	*(.0338)*
Logit						Y
Probit					Y	
OLS	Y	Y				
OLS ln(n+1)			Y	Y		
Observations	71337	71337	71337	71337	71337	71337

The unit of observation is the academic article. All estimates with Huber-White robust standard errors, and all include journal dummies and a publication month quadratic. "In % terms" is equal to the average marginal effect over the post-period NIH-funded average number of cites or propensity to be cited.

Statistical significance indicators: *: .1, **: .05, ***: .01

Table A11: Geography of Medical Research and Frontier-Citing Patents

State	Articles	Patents	Rank	Country	Articles	Patents	Rank
MA	1139.7	382.3	1	Switz.	253.8	58.0	2
DC	1133.3	55.0	13	Nether.	226.0	18.9	9
MD	964.4	164.2	2	US	191.0	57.3	3
CT	361.1	96.0	4	Denmark	159.5	38.2	4
MN	355.7	40.7	18	Sweden	157.1	31.4	5
NY	302.9	66.6	10	Canada	152.0	25.2	7
PA	284.3	83.1	5	UK	144.2	14.2	16
RI	279.0	52.0	14	Finland	137.0	14.8	14
CA	210.1	130.4	3	Israel	123.5	63.5	1
NC	210.1	50.7	15	Australia	107.8	15.1	12
MO	208.5	35.3	20	Germany	105.4	18.0	11
NH	206.9	81.5	7	Austria	103.9	18.4	10
WA	201.7	80.0	8	Belgium	101.9	24.7	8
OH	180.5	27.0	26	Singapore	99.8	30.9	6

Articles and citing patents are reported per 100,000 population. Rank refers to the rank of the state or country in terms of frontier-citing patents per capita. Country list omits those with population below 500,000 (in which case Iceland would rank #1 in patents per capita).

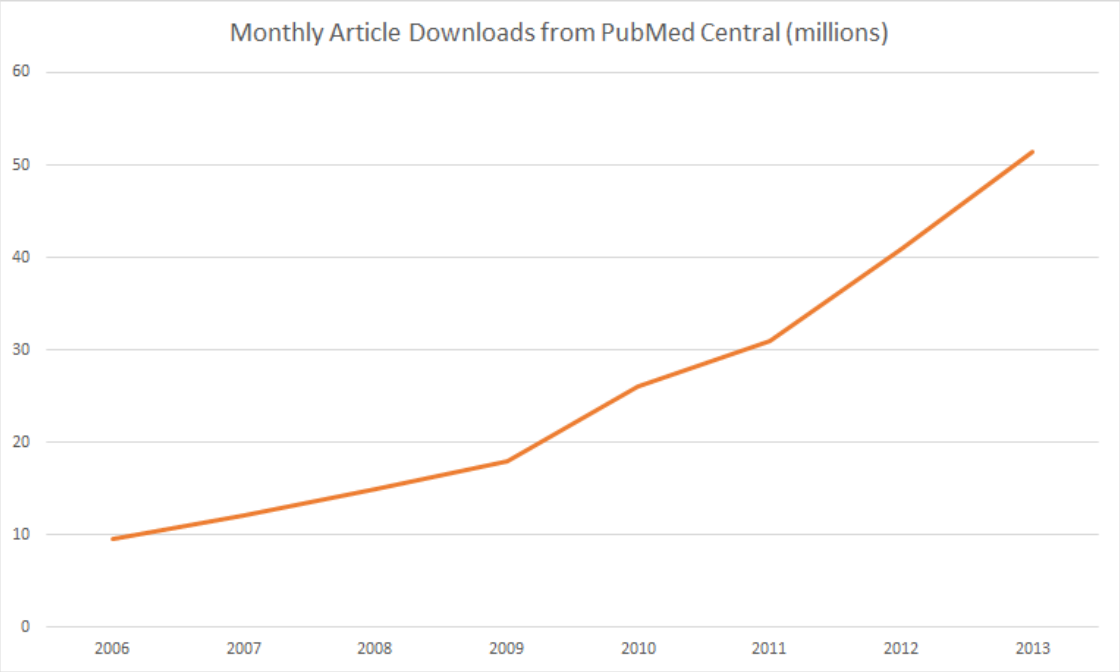

Figure A1: Monthly PubMed Central downloads, sampled each year in October to isolate the trend from seasonal variation. Data courtesy the National Insitutes of Health.

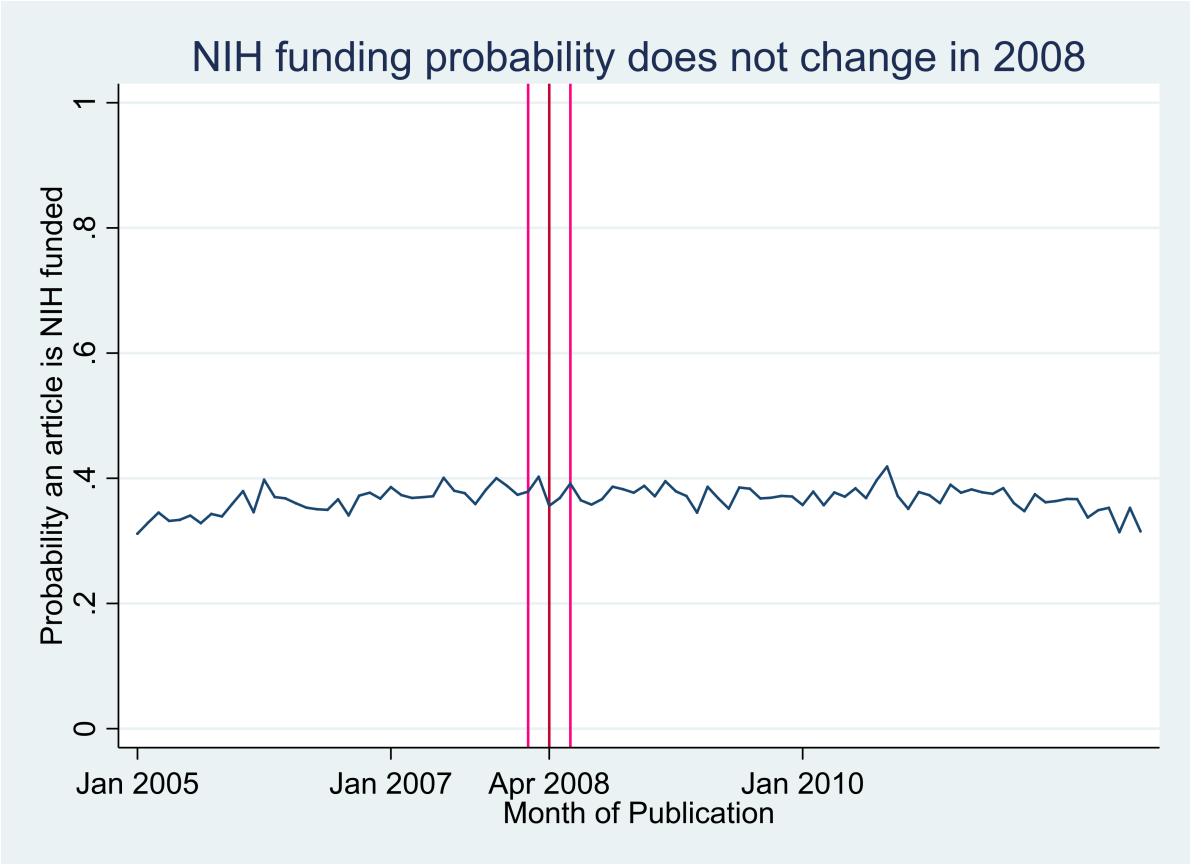

Figure A2: Change in probability of NIH funding over time. The red (center) line represents April 2008, and the pink (left and right) lines represent two months before and after the official beginning of the policy.

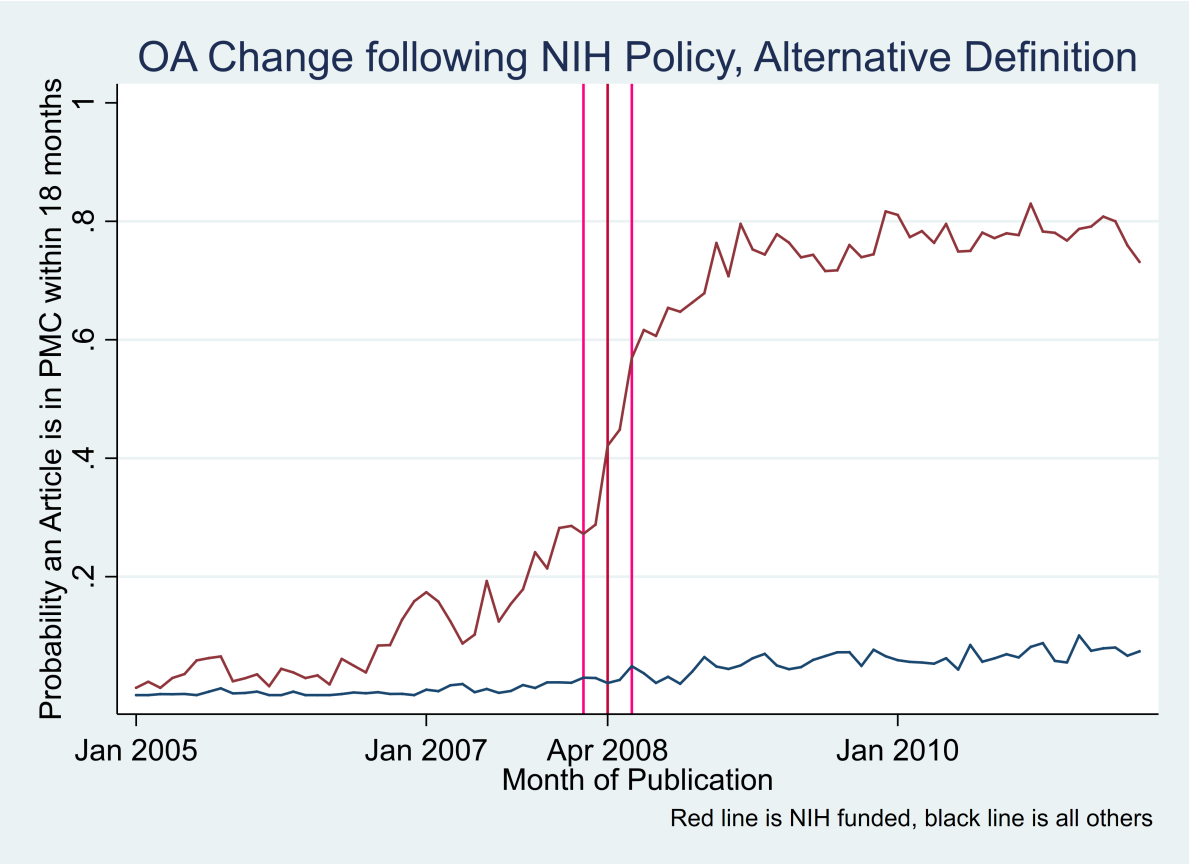

Figure A3: Sample consists of all medical research articles in the subset of 30 journals that generally do not make research freely available unless forced to. "Open access" refers to the article being freely available in the PubMed Central repository within 18 months of publication. As opposed to Figure 1, this restriction better accounts for articles that were not made freely available until years after publication, but does not account for articles freely available via a publisher website or an academic repository only. The red (center) line represents the April 2008 NIH policy, and the pink (left and right) lines represent two months before and after the official beginning of the policy.

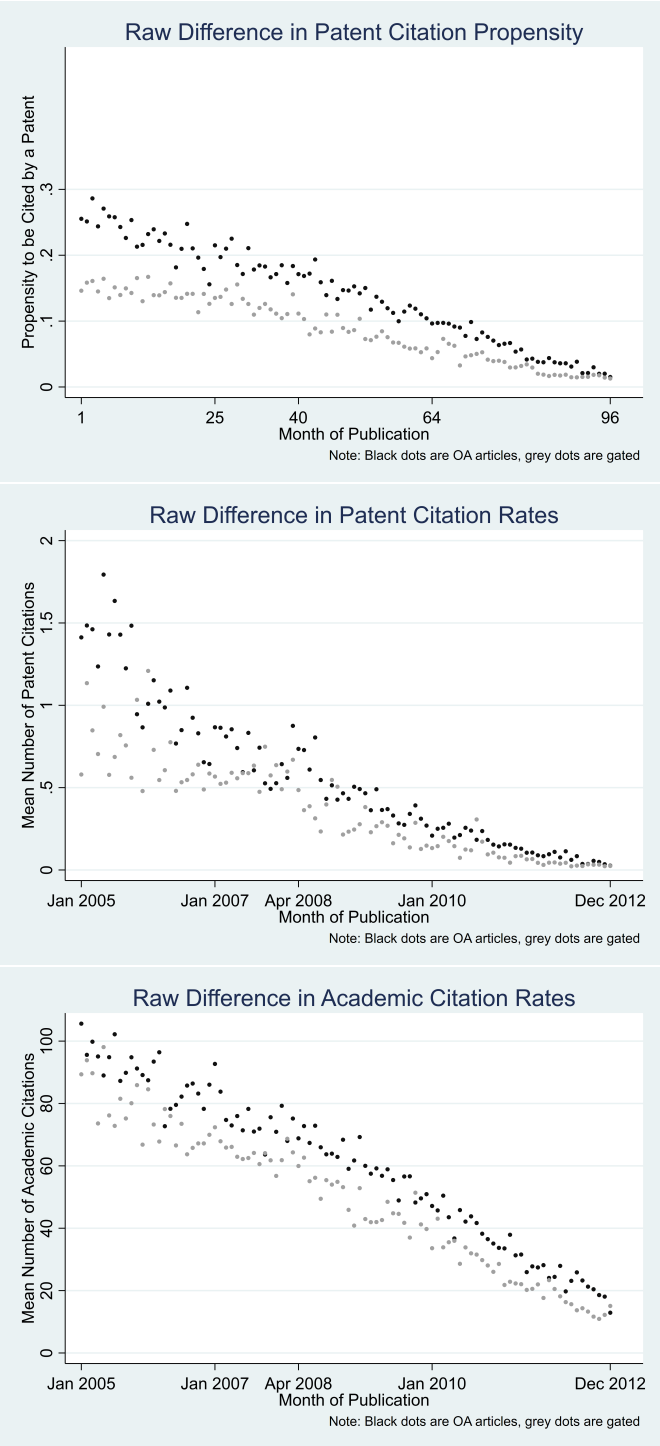

Figure A4: Raw difference in patent and academic citation rates between open access and gated articles, by publication month. The open access advantage in the raw data remains even when controlling for journal, funder, and month fixed effects, as seen in Online Appendix Table A1.

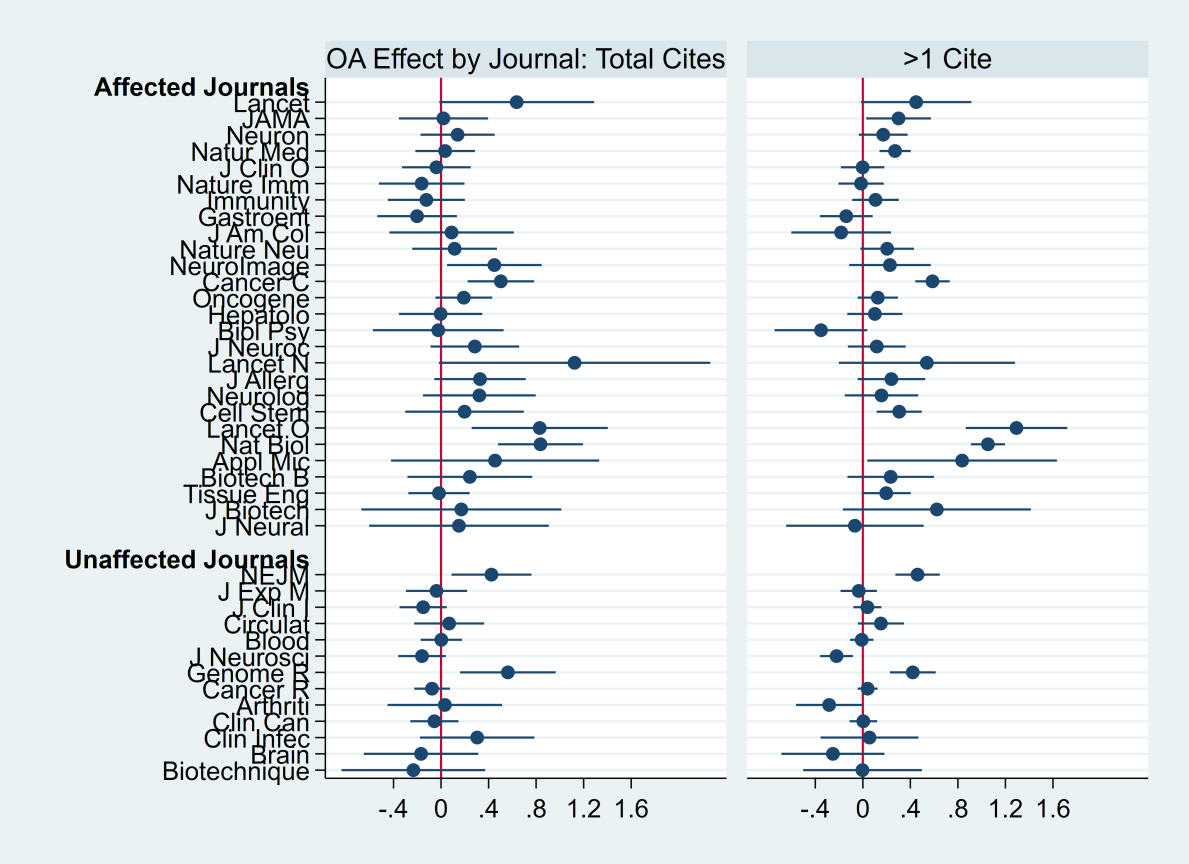

Figure A5: Estimated ppml coefficient on the interaction term After April 2008 × NIH-funded, constructed on a journal-by-journal basis. Publication month, journal and funder fixed effects are constant across journals as in the primary regressions. Three journals for which the small sample size generates very large standard errors were dropped from the above chart.

Note that the positive effect of open access, whether measured in terms of total patent citations or the probability an article has at least one patent citation, can be seen across a wide swath of journals. Among "unaffected" journals, only the New England Journal of Medicine and Genome Research have positive treatment effects. The New England Journal of Medicine began making their archives free-to-read without registration in December 2007 (they had been free after a registration process since 2001) just four months before the NIH policy began, which may explain why the NIH policy appears to positively affect the NEJM in the diff-in-diff.

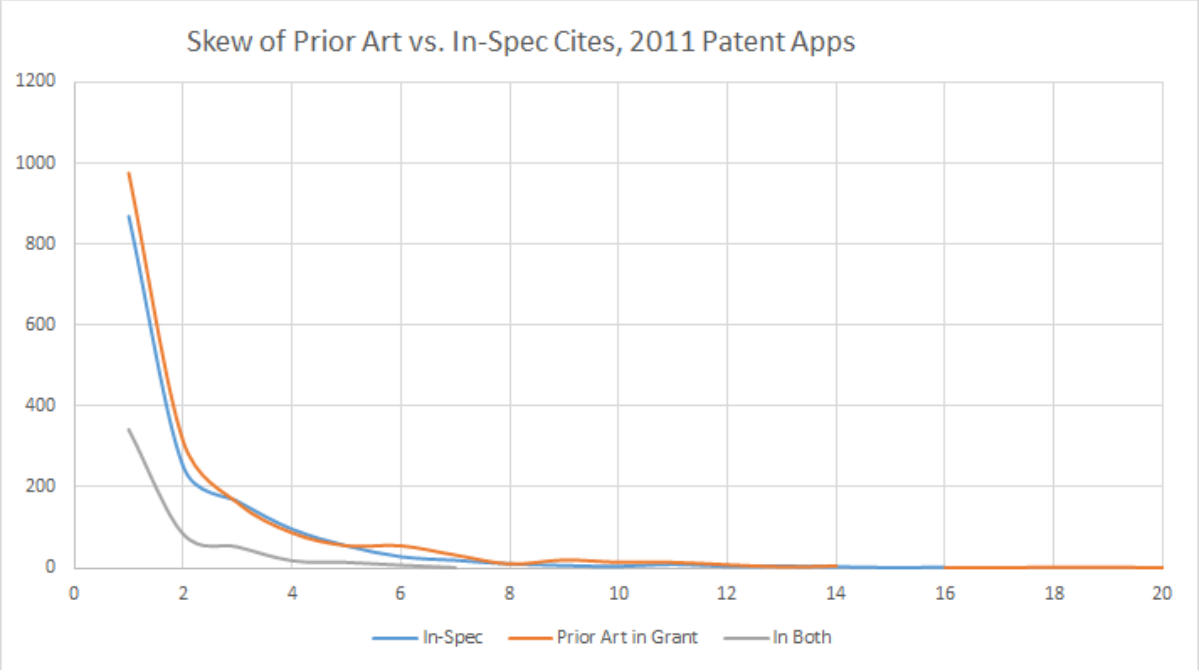

Figure A6: Comparison of the skewness of in-specification citations versus prior art citations. The figure includes in-specification citations made by patent applications in 2011 to academic papers, and the prior-art citations made by grants of the same patent applications from 2011. Note that the skew of these citations is quite similar, and that there is very little overlap between the citation types.

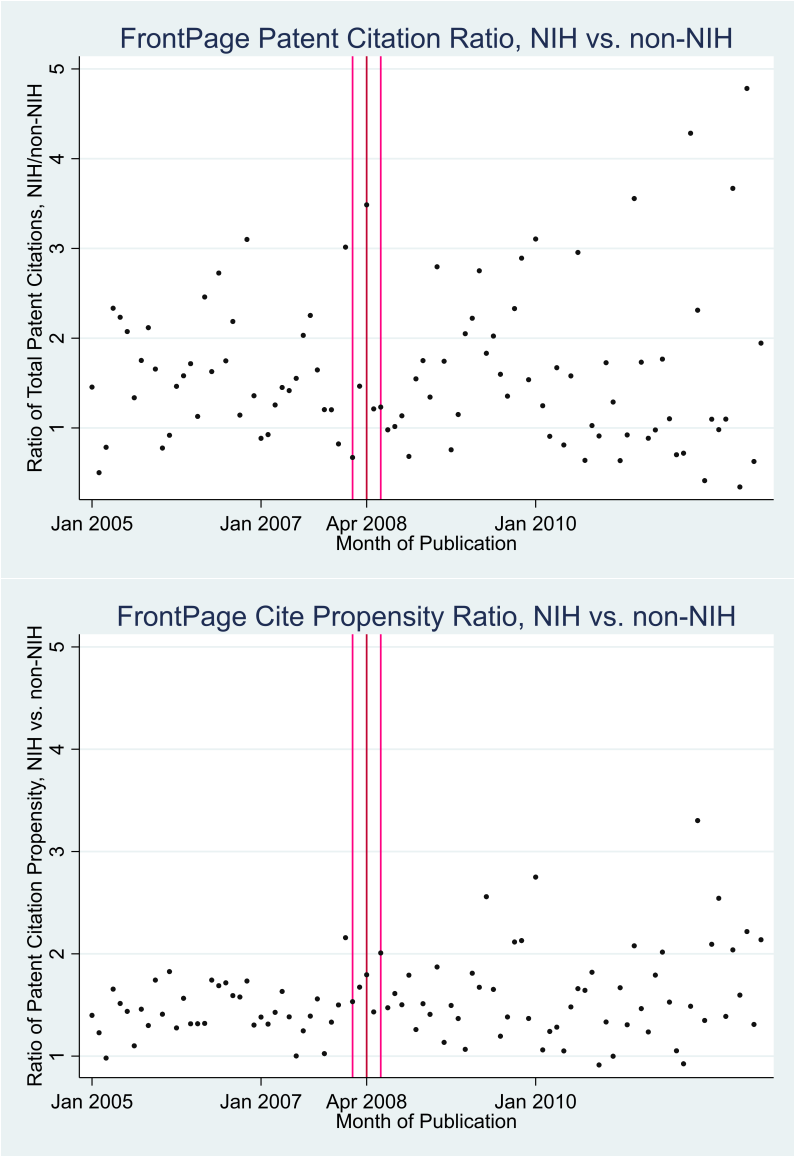

Figure A7: Ratios of front page citations for NIH funded articles versus non-funded articles, by article publication month. The top panel gives the ratio of total patent citations. The bottom panel gives the ratio of propensities to have at least one patent citation. Articles restricted to the thirty journals which generally do make articles freely available unless required by a mandate. The red (center) line represents April 2008, and the pink (left and right) lines represent two months before and after the official beginning of the policy. Note that front page citations can only occur on granted patents.