THE IMPACT OF OPEN ACCESS MANDATES ON INVENTION

Kevin A. Bryan and Yasin Ozcan*

Abstract—How do barriers to the diffusion of academic research affect innovation? In 2008, the National Institutes of Health (NIH) mandated free online availability of funded research. This policy caused a 50 percentage point increase in free access to funded 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. Nonfunded research, funded research in journals unaffected by the mandate, and academic citations see no change. These estimates are consistent with a model of search for useful knowledge. Inefficiency caused by academic publishing may be substantial.

I. Introduction

UNIVERSITY research is often valuable in industry, particularly in innovative sectors. This research diffuses principally through published research articles (Cohen, Nelson, & Walsh, 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 & 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, Roos, & Lauri, 2009; Khabsa & 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. 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 twelve 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 II, we 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 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 III.

In section IV, 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

Received for publication November 8, 2018. Revision accepted for publication March 23, 2020. Editor: Amitabh Chandra.

*Bryan: University of Toronto; Ozcan: MIT Sloan School of Management.

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.

A supplemental appendix is available online at https://doi.org/10.1162/rest_a_00926.

[1] Similar mandates exist from organizations including the University of California, the Howard Hughes Institute, the Wellcome Trust, and MIT. 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.

[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 quasi-random 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.

The Review of Economics and Statistics, December 2021, 103(5): 954–967

https://doi.org/10.1162/rest_a_00926

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% more often. As the policy only led to a 50 percentage point relative increase in free availability compared with nonfunded 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. We conclude by discussing policy implications.

A. *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%. Murray et al. (2016) 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 (2019), however, find that gene patent grants, instrumented using the variable strictness of patent examiners, do not affect follow-on innovation. They argue that the patent holder optimally allows research that 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 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 (2018) show that entrepreneurship, licensing, and patenting by university researchers fall 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 (Mowery & 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 an invention that itself drew on academic research, and find that at least 60% of all inventions can be traced back to published research. Iaria, Schwarz, and Waldinger (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 saw permanent and severe declines in their research productivity after their access to continuing research from their nations as 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).

II. 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. . . . I've been fortunate to have access to world-class libraries at every stage of my career. As a result,

[3] Earlier research on the direct question of how academic open access affects nonacademic 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 were more likely to begin recommending frontier treatments to the patients. Ware and Monkman (2009) survey private sector researchers in the United Kingdom 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, Swan, and Brown (2011) finds that 68% of Danish firms report access difficulty.

> 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, computational biology, developmental biology, cell biology, immunology and stem cell biology as well as clinical developments.

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 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.

Formally, assume inventors search for knowledge as follows. First, 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. Second, 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 $\sum_j p_{aj} \leq 1$. Third, 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 he or she actually paying for the article. If an inventor is at an institution with a subscription to the journal where j is published, then $c_{ij} = 0$. Finally, let $G \geq 0$ be a multiplier on X, which converts private values of an idea to the social value of that idea. Finally, assume an inventor simultaneously chooses how many articles to purchase and read, given his or her belief about the expected benefit of finding useful academic knowledge a. We discuss these assumptions in more depth and provide proofs in online appendix 4.

Proposition 1. *The value of open access to a given firm i, and to social welfare, 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}.

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.*

When does open access matter in the model? Open access is more likely to improve inventions and, hence, social welfare under four conditions. First, inventors do not have institutional subscriptions. Second, they 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). Third, it is not clear which particular article contains useful knowledge. Fourth, the social value of inventions is much higher than the private value. Further, 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. When individual articles are expensive and the inventor doesn't know precisely which contain useful information, the expected value of reading an additional article is low, even if the information in that article would make the invention much better.

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

III. Data

Our data consist 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.[4] For each article, we extract the country of the first author's affiliation, the affiliated state if the author is in the United States, 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.[5] The FFT category is nearly identical to the set of articles one

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

[5] For 3,002 articles, we are unable to extract author location, and for 2,253, we were unable to extract the number of academic citations. In general,

could find freely available anywhere online, and would include, for example, an article freely available on a publisher's website that was not deposited in PubMed Central.[6] PubMed and PMC are by far the most commonly accessed medical research databases in the world, with PMC searches alone resulting in over 1 million article views per day (Blumenthal & 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 that are public as of March 19, 2015.[7] 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 eighteen 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.[8]

To link the two data sets, we develop a custom coarse matching algorithm that 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* is 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 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 trade-off between type 1 and type 2 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.[9] Seventeen percent of our sample (22,611 academic papers) receive at least one citation; for our oldest cohort of papers, from 2005, more than 28% are cited at least once. The matches are almost entirely medical-related, as would be expected: over 91% of the patents come from just six primary patent classes.[10] No more than 2% of the matches, and by our best estimate much less than that, are "self-cites" where the article author is also a patentee.[11]

A. *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 (Fleming & Sorenson, 2004). 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, examine 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

these missing data refer to editorials and other types of articles that were miscoded as being research oriented.

[6]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 six to eighteen 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.

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

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

[9]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.

[10]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), and 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.

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

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 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, Lemley, & Sampat, 2013; Alcacer, Gittelman, & Sampat, 2009). More important, these 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 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.[12]

This distinction is not unusual. In our sample, restricting to applications that have been granted, 73% of the in-text citations do not appear on the front page of the granted patent. Going the other direction, 82% 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 overlaps almost perfectly, and the exact same matching algorithm is used on both data sets.[13]

Front page citations, of course, have a long and well-validated history among innovation scholars (Jaffe, Trajtenberg, & Henderson, 1993). 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 & Cohen, 2013; Meyer, 2000). In-text citations, purely on legal grounds, ought to measure real knowledge flows better. In a companion paper (Bryan, Ozcan, & Sampat, 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 thirty years and describes more fully the legal interpretation of each type of citation.

Although we contend that in-text citations better measure actual knowledge transfer 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 provide the caveat that our results directly measure only increased citation, not increased innovation. We investigate further in section IVB why the former proxies for the latter.

B. Summary Statistics and Estimation Technique

Summary statistics. Table 1 gives summary statistics for articles and the patent applications that cite them. Articles

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

[13] 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.

TABLE 1.—SUMMARY STATISTICS

	All Articles
Observations	132,872
Mean # of Patent Citations	.475
Mean # of Patent Citations to Year 2005 Papers	1.052
Maximum Number of Citations	248
Pr($\geq$ 1 patent citation)	.170
Available via Free Full Text	.543
Funded by NIH	.367
Mean # of Academic Cites	55.8
	All 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 university)	.284
Pr(first inventor in United States)	.648
Pr(inventors in multiple countries)	.150
Pr(patent granted by August 7, 2017)	.487
Pr(first inventor in same country as first author of cited article)	.491

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

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% of the articles are funded by the NIH, a number that is roughly constant from 2005 to 2012 (online appendix figure A2). Fifty-four 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% 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 time frame of our data set: 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 data set 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.[14] Locations with large government or institutional medical research centers like DC, Maryland, Minnesota, New York, the United Kingdom, 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 that are traditional biotech and pharma hotbeds.

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*. Nonfunded 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."[15]

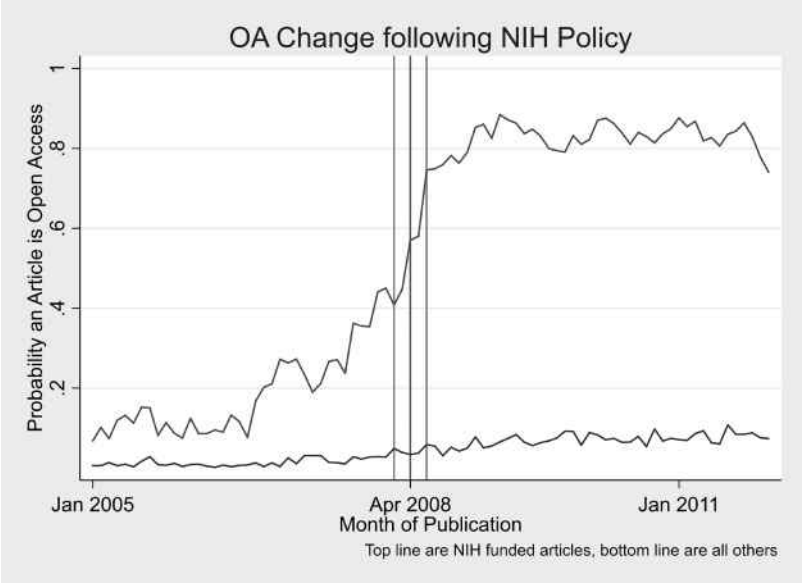
Mandate compliance is less than perfect on both sides of the April 2008 boundary. In general, and especially at journals that do not make articles freely available unless required by an institutional mandate, authors are themselves responsible for uploading their research to PubMed Central. Less-than-perfect compliance after 2008, when only about 80% of NIH-funded research in affected journals is freely available, is driven by authors 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 $2,000 to $5,000 per article to permit free availability for readers.[16] Beginning in early 2013, the NIH began toughening 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

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

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

[16] 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.

FIGURE 1.—SAMPLE CONSISTS OF ALL MEDICAL RESEARCH ARTICLES IN THE SUBSET OF THIRTY 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 vertical lines represent two months on either side of the April 2008 start date of the start of the NIH mandate.

policy was predicated on the stagnant and less-than-perfect mandate compliance for articles published between 2008 and 2012. That is, the nature by which the NIH enforced its mandate between 2008 and 2012, our "posttreatment period," was roughly constant (see, e.g., van Noorden, 2013).[17]

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 nonavailability in these estimates.

C. *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 that support 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 affect behavior only 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 by 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.[18] 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% of articles published in top journals and led to the 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 25% to 27% more likely to be cited by a patent, reflecting both the more U.S.-heavy authorship and potentially the higher quality of the research (online appendix table A2). That the NIH mandate affects only research published

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

[18] 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 II.

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 thirty journals that 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%, as was seen in figures 1 and A3. Therefore, if the 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 IV.

We will estimate

$$y_i = f(PostApril08 \times NIH, PostApril08, NIH, X_i), \quad (1)$$

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.[19] We use two methods to account for this.

First, we estimate a placebo of equation (1) using only the thirteen journals in our sample that 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 thirteen placebo journals should see 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

$$y_i = f(PostApril08 \times NIH \times Affected, X_i), \quad (2)$$

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 citations to NIH-funded articles published after the mandate in journals that do not make everything free-to-read, compared to 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*.[20] 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 nonfunded articles. That is, if 10% of unfunded 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.

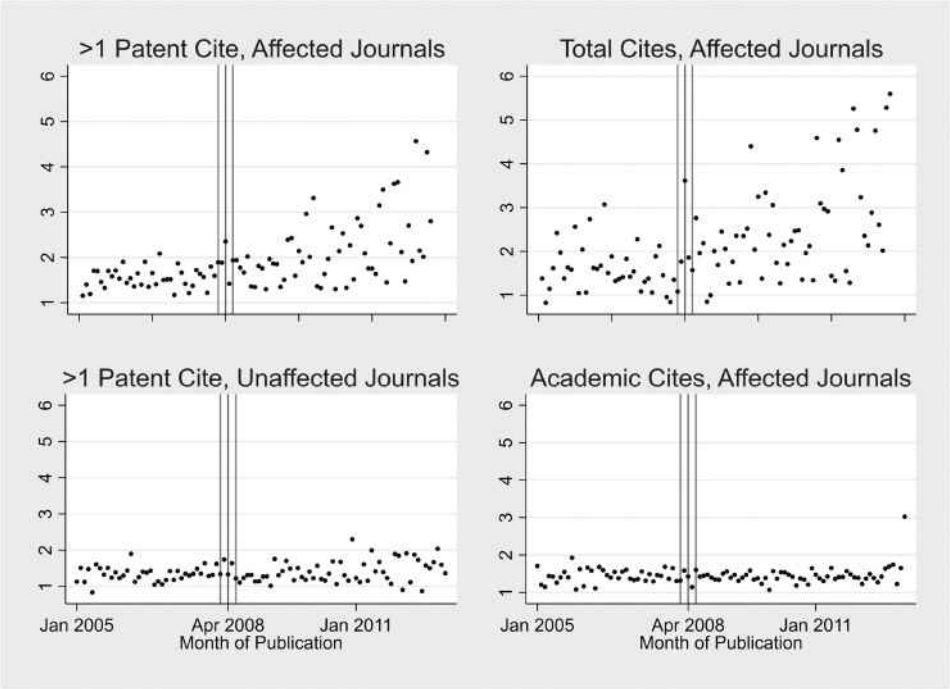
If the outcome of interest is always positive, many researchers just log variables to convert multiplicative parallel trends to additive parallel trends, then use standard diff-in-diff techniques. In the cases like ours where the outcome variable is equal to 0 for the majority of entries, log linearization 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 (Santos-Silva & Tenreyro, 2006; Ciani & Fisher, 2019). Generically, with nonsmooth 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 & 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 (Santos-Silva & Tenreyro, 2010, 2011; Hilbe, 2007).[21] 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

[19] 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.

[20] 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 are higher in the pre- than the postperiod implies that if the multiplicative treatment assumption is true for total cites (e.g., if cites arrive according a possibly 0-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.

[21] 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 (Zou, 2004).

FIGURE 2.—RATIOS OF CITATIONS FOR NIH-FUNDED ARTICLES VERSUS NONFUNDED ARTICLES, BY ARTICLE PUBLICATION MONTH



The top panels give the ratios for propensity of at least one cite, and for total cites, of funded versus unfunded articles. Articles restricted to the thirty journals that generally do make articles freely available unless required by a mandate. The bottom charts are a placebo estimate of the top figure, restricting to the thirteen journals that make nearly all research freely available and hence are unaffected by the mandate, and the ratio of academic citations before and after the mandate. The vertical lines represent two months on either side of the April 2008 start date of the start of the NIH mandate.

multiplicative treatment effect but rather estimates ln(2) times the OLS diff-in-diff treatment effect under the assumption of additive parallel trends.

IV. Results

The top panels of figure 2 display the ratio of citations received by NIH-funded compared to nonfunded 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.[22]

Table 2 presents our primary estimates. Controlling for journal and publication month, moving from 0 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 probability of at least one patent citation within three years of publication by 12.3%. Online appendix table A3 shows robustness of these estimates 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 journal or a small number of them.

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-right panel of figure 2 shows the null result within academia graphically.

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, the bottom panel of table 2 and the bottom-left panel of figure 2 investigate the change in citations to NIH-funded articles relative to nonfunded articles within the thirteen 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, Anderson, & Drazen, 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*.

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 the bottom panel of table 2 show precisely estimated null effects of open access in these placebo journals. Table 3 estimates a multiplicative

[22] 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 eighteen months but often longer; hence, the number of cites we observe as we become closer to the present is falling.

TABLE 2.—DIFFERENCE-IN-DIFFERENCE ESTIMATES FOR TREATED JOURNALS

	Pat. Cites	Pr(≥ 1 Pat. Cite)	Pr(≥ 1 Pat. Cite in 3 yr)	Acad. Cites
NIH × 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	71,337	71,337	71,337	70,184
Placebo				
NIH × 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	61,408	61,408	61,408	60,310

The unit of observation is the academic article. Top panel estimates restrict to the thirty journals that rarely make research free-to-read in the absence of a mandate; bottom panel estimates restrict to the thirteen journals that make almost all research free-to-read, and hence 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}. Statistically significant at *.1, **.05, ***.01.

TABLE 3.—TRIPLE DIFFERENCE ESTIMATES

	Pat. Cites	Pr(≥ 1 Pat. Cite)	Pr(≥ 1 Pat. Cite in 3 yr)	Acad. Cites
NIH×post 04/2008×Affected	.2354**	.1780***	.1323**	.0443
(in % terms)	26.5	19.5	14.1	4.5
NIH dummy	.1981***	.2229***	.2467***	.1290***
	(.0395)	(.0183)	(.0272)	(.0197)
Observations	132,745	132,745	132,745	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 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}. Statistically significant at *.1, **.05, ***.01.

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% increase in total patent citations and a 14% to 20% in the probability of at least one citation. Again, citations within academia are relatively unaffected by the mandate.

Figure 3 summarizes our main results graphically.[23] 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 and that the advantage is positive in every half-year period after the first half of 2009.[24] 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.

Online appendix table A12 and 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 0. 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 must be disclosed only when the applicant is aware of the potential for the reference to relate to his or her 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 IVA.

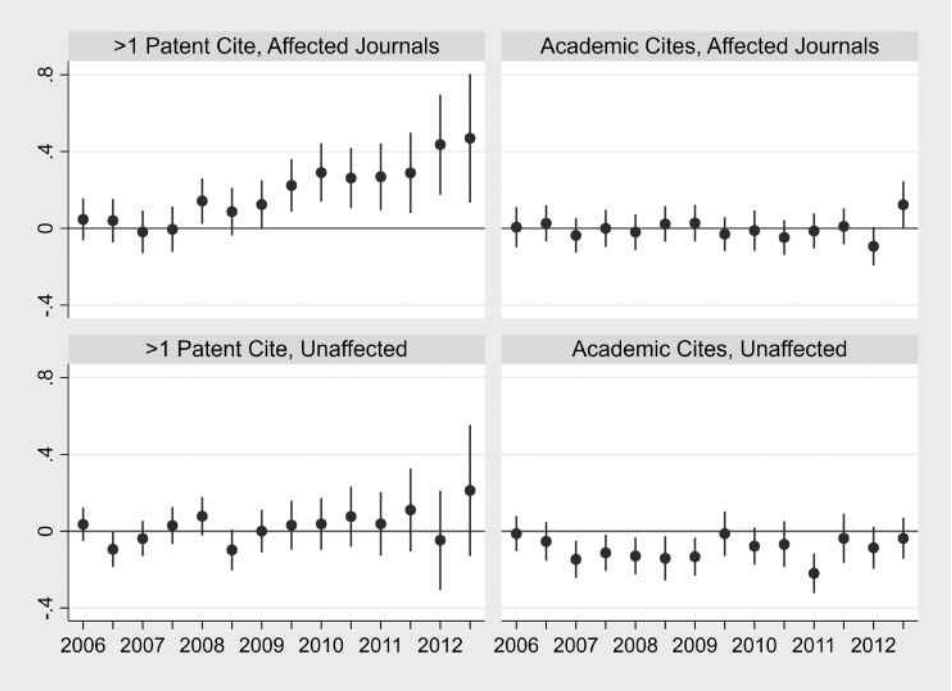
Online appendix tables A6, A7, and A8 investigate the effect of open access within various subgroups. Table A6 shows that the main treatment effect is not being driven by low-value patents. The effect of open access is qualitatively similar to our primary estimates even if we restrict to patents assigned upon application (table A6, columns 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.[25] Patent 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

[23] A table with the estimates used in figure 3 is in online appendix table A4.

[24] 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 postmandate period.

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

FIGURE 3.—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



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 III, reflect the estimated effect of the NIH mandate rather than the effect of going from 0 to complete open access. The top-left panel is essentially the difference-in-difference of table 2 in event study form, the bottom-left panel the placebo using the "unaffected" thirteen journals that 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.

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% of all articles that 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 that 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.

A. 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 that affect the citation of research in patents and that 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 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 that 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 quasi-random 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 ten months, while others were not free online until fourteen 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 that 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 0 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.

B. Interpretation of Treatment Effects

To interpret our empirical results, let us return to the model in section II. 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 that 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 you will 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. 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 to the legal theory already discussed, supporting the validity of in-text citations as a real knowledge flow. Table 2 also shows that as open access increased citation to affected journals, it did not change citation in unaffected journals. This is consistent with both the search model and 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 long-standing 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.

V. 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 United States, United Kingdom, and EU have all considered legislation that 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 to be more closely linked to the knowledge of the inventor than to 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?

Decisions about open access need to account for its effects both within and outside academia. 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 another. The distributional consequences of the academic journal system have traditionally been seen as pure

transfers from private industry to publishers and academic societies. Our results suggest that the deadweight loss created by the price discrimination that permits these transfers is substantial.

REFERENCES

Ackermann, J., M. Frutsch, K. Kaloulis, T. McKee, A. Trumpp, and F. Beermann, "Metastasizing Melanoma Formation Caused by Expression of Activated N-RasQ61K on an INK4a-Deficient Background," *Cancer Research* 65 (2005), 4005–4011. 10.1158/0008-5472.CAN-04-2970, PubMed: 15899789

Ahmadpoor, Mohammad, and Benjamin F. Jones, "The Dual Frontier: Patented Invention and Prior Scientific Advance," *Science* 357 (2017), 583–587. 10.1126/science.aam9527, PubMed: 28798128

Alcacer, Juan, Michelle Gittelman, and Bhaven Sampat, "Applicant and Examiner Citations in U.S. Patents: An Overview and Analysis," *Research Policy* 38:2 (2009). 10.1016/j.respol.2008.12.001

Athey, Susan, and Guido W. Imbens, "Identification and Indifference in Nonlinear Difference-in-Difference Models," *Econometrica* 74 (2006).

Björk, Bo-Christer, Annikki Roos, and Mari Lauri, "Scientific Journal Publishing: Yearly Volume and Open Access Availability," *Information Research* 14:1 (2009).

Blumenthal, Jane, and Gary Freiburger, "MLA/AAHSL Sequestration Letter," unpublished manuscript (2012).

Bryan, Kevin A., Yasin Ozcan, and Bhaven Sampat, "A User's Guide to In-Text Citations," *Research Policy* (2020).

Campion, Edward W., Kent R. Anderson, and Jeffrey M. Drazen, "A New Web Site and a New Policy," *New England Journal of Medicine* 344 (2001), 1710–1711.

Ciani, Emmanuel, and Paul Fisher, "Dif-in-Dif Estimates of Multiplicative Treatment Effects," *Econometric Methods* 8:1 (2019), 1–10. 10.1515/jem-2016-0011

Cohen, Wesley M., Richard R. Nelson, and John P. Walsh, "Links and Impacts: The Influence of Public Research on Industrial R&D," *Management Science* 48 (2002), 1–23. 10.1287/mnsc.48.1.1.14273

Cotropia, Chris A., Mark Lemley, and Bhaven Sampat, "Do Applicant Patent Citations Matter?," *Research Policy* 42 (2013), 844–854. 10.1016/j.respol.2013.01.003

Davis, P., B. Lewenstein, D. Simon, J. Booth, and M. Connolly, "Open Access Publishing, Article Downloads, and Citations: Randomised Controlled Trial," *British Medical Journal* 337 (2008). 10.1136/bmj.a568

Dittmar, Jeremiah, "The Welfare Impact of a New Good: The Printed Book," working paper (2020).

Fleming, Lee, and Olav Sorenson, "Science as a Map in Technological Search," *Strategic Management Journal* 25 (2004), 909–928. 10.1002/smj.384

Frank, Martin, "Open but Not Free: Publishing in the 21st Century," *New England Journal of Medicine* 368 (2013), 787–789. 10.1056/NEJMp1211259, PubMed: 23445089

Furman, Jason, and Scott Stern, "Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research," *American Economic Review* 101 (2011), 1933–1963. 10.1257/aer.101.5.1933

Gaule, Patrick, and Nicolas Maystre, "Getting Cited: Does Open Access Help?," *Research Policy* 40 (2011), 1332–1338. 10.1016/j.respol.2011.05.025

Hardisty, David J., and David A. F. Haaga, "Diffusion of Treatment Research: Does Open Access Matter?," *Journal of Clinical Psychology* (2008). 10.1002/jclp.20492, PubMed: 18425790

Hilbe, Joseph M., *Negative Binomial Regression* (Cambridge: Cambridge University Press, 2007).

Houghton, John, Alma Swan, and Sheridan Brown, "Access to Research and Technical Information in Denmark," report to the Danish Agency for Science, Technology and Innovation (2011).

Hvide, Hans K., and Benjamin F. Jones, "University Innovation and the Professor's Privilege," *American Economic Review* 108 (2018), 1860–1898. 10.1257/aer.20160284

Iaria, Alessandro, Carlo Schwarz, and Fabian Waldinger, "Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science," *Quarterly Journal of Economics* 133 (2018), 927–991. 10.1093/qje/qjx046

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson, "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," *Quarterly Journal of Economics* 108 (1993). 10.2307/2118401

Khabsa, Madian, and C. Lee Giles, "The Number of Scholarly Documents on the Public Web," *PLoS One* (2014). 10.1371/journal.pone.0093949, PubMed: 24817403

Kim, Heekyung Hellen, "The Effect of Free Access on the Diffusion of Scholarly Ideas," in *Proceedings of the International Conference on Information Systems* (Atlanta, GA: Association for Information Systems, 2012).

Lyman, Stewart, "Industry Access to the Literature," *Nature Biotechnology* 29 (2011), 571–572. 10.1038/nbt.1909

McCabe, Mark J., and Christopher M. Snyder, "Identifying the Effect of Open Access on Citations Using a Panel of Science Journals," *Economic Inquiry* 52 (2014), 1284–1300. 10.1111/ecin.12064

Meyer, Martin, "What Is Special about Patent Citations? Differences between Scientific and Patent Citations," *Scientometrics* 49 (2000), 93–123. 10.1023/A:1005613325648

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

Mowery, David, and Bhaven Sampat, "The Bayh-Dole Act of 1980 and University-Industry Technology Transfer: A Model for Other OECD Governments?" (pp. 233–245), in Arthur N. Link and F. M. Scherer, eds., *Essays in Honor of Edwin Mansfield* (Berlin: Springer, 2005). 10.1007/s10961-004-4361-z

Murray, Fiona, Philippe Aghion, Mathias Dewatripont, Julian Kolev, and Scott Stern, "Of Mice and Academics: The Role of Openness in Science," *American Economic Journal: Economic Policy* 8 (2016), 212–252. 10.1257/pol.20140062

Murray, Fiona, 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 and Organization* 63 (2007), 648–687. 10.1016/j.jebo.2006.05.017

Roach, Michael, and Wesley M. Cohen, "Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research," *Management Science* 59 (2013). 10.1287/mnsc.1120.1644, PubMed: 24470690

Sampat, Bhaven, and Heidi Williams, "How Do Patents Affect Follow-on Innovation? Evidence from the Human Genome," *American Economic Review* 109 (2019), 203–236. 10.1257/aer.20151398

Joao, M. C. Santos-Silva, and Silvana Tenreyro, "The Log of Gravity," this REVIEW 88 (2006). 10.1162/rest.88.4.641

——— "Currency Unions in Prospect and Retrospect. *Annual Review of Economics* 2 (2010), 51–74. 10.1146/annurev.economics.102308.124508

——— "Further Simulation Evidence on the Performance of the Poisson Pseudomaximum Likelihood Estimator," *Economics Letters* 112 (2011). 10.1016/j.econlet.2011.05.008

van Noorden, Richard, "NIH Sees Surge in Open-Access Manuscripts," *Nature News Blog* (2013), http://blogs.nature.com/news/2013/07/nih-sees-surge-in-open-accessmanuscripts.html.

Ware, Mark, and Mike Monkman, "Access by UK Small and Medium-Sized Enterprises to Professional and Academic Information," *Publishers Research Consortium Report* (2009).

Williams, Heidi, "Intellectual Property Rights and Innovation: Evidence from the Human Genome," *Journal of Political Economy* 121 (2013). 10.1086/669706

Zou, Guangyong, "A Modified Poisson Regression Approach to Prospective Studies with Binary Data," *American Journal of Epidemiology* 159 (2004). 10.1093/aje/kwh090, PubMed: 15033648