The Impacts of Open Access on Scientists, Inventors, and the Public[*]

Joseph Staudt[†]

July 2, 2020

Abstract

The main goals of making scientific literature open access are twofold: 1) to speed scientific discovery and 2) to give access to the public who funded the research. In this paper, I use citations from articles, patents, and Wikipedia to examine whether open access achieves these goals by estimating whether innovators (scientists and inventors) or the general public increase their use of articles after those articles become freely available on PubMed Central, the largest repository of free full-text biomedical articles. Estimates suggest that innovators modestly increase their use of the typical article after it becomes freely available, but that the public substantially increases their use. Using the National Institutes of Health's Public Access Policy (PAP) as an instrument for an article becoming freely available suggests that, in contrast to the modest effects for the average article, innovators substantially increase their use of complier articles – those articles that are freely available only because the PAP requires them to be. I unpack the sources of these citation increases by analyzing whether particular subsets of individuals disproportionately increase their citations to an article after it becomes freely available. These subsets include scientists at different types of institutions (firm/university/hospital) or in different countries (upper/middle/lower income) and different types of firms to which patents are assigned (small, young, high-tech, etc.). The latter group will be identified by creating the first-ever link of patent-to-article citation data to confidential firm-level data at the U.S. Census Bureau.

Keywords: economics of science, open access, nih, nih public access policy, policy evaluation.
JEL Classification Numbers: 031, 034, 036, 038

[*] I gratefully acknowledge financial support from NIH Grant P01 AG039347. I would also like to thank Elisabeth Perlman, Shawn Klimek, and participants in the Census Economic Research Brown Bag for helpful comments. All remaining mistakes are my own.

[†] U.S. Census Bureau, Center for Economic Studies (CES), 4600 Silver Hill Rd, Suitland, MD 20746. joseph.staudt@census.gov. Disclaimer: Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. The research in this paper does not use any confidential Census Bureau information.

1 Introduction

Over the past several decades, funders of science have increasingly mandated that supported research be made freely available.[1] One primary rationale for "open access" mandates, closely related to concerns about long-term technological and economic growth, is that freely available research will increase the pace of scientific advancement (Zerhouni, 2008; Tennant et al., 2016). Implicit in this rationale is a vision of innovators – both scientists and inventors – building upon the work of others as they expand the frontiers of science.[2] A second main rationale for making research freely available is that taxpayers, as the ultimate source of funds for this work, have the right to access it (Zerhouni, 2008). Thus, the main justifications for open access rest on the idea that certain stakeholders in the scientific enterprise face meaningful barriers to the scientific literature. In this paper, I study whether making scientific research freely available does, in fact, increase access for these stakeholders – innovators (i.e., scientists and inventors) who use it to advance science and/or the general public who pays for it.[3]

I examine the impact of open access on innovators' ability to make use of scientific literature by measuring whether the rate at which they draw on ideas from an article changes after that article is made freely available. Specifically, I measure whether citations, to a focal article, from other articles (article-to-article citations) and from patents (patent-to-article citations) change after the focal article is made freely available. Intuitively, if innovators have difficulty accessing a gated article, then we should observe citation increases to the article after it becomes open access. Similarly, I examine the impact of open access on the general public's exposure to the scientific literature (which it funded) by measuring whether Wikipedia – one of the most common sources of scientific information directed at lay people (Heilman and West, 2015; Laurent and Vickers, 2009; Mesgari et al., 2015; Spoerri, 2007) – is more likely to reference information contained in an article after the article becomes freely available. Specifically, using information from revision histories, I measure whether citations from Wikipedia entries to a focal article (Wiki-to-article citations) change after the focal article is made freely available. As with article-to-article and patent-to-article

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

[2] Work that treats science as a cumulative process includes Aghion et al. (2008), Aghion and Howitt (1992), Mokyr (2002), Murray et al. (2009), Romer (1990), and Scotchmer (1991).

[3] In testimony to the House Judiciary Committee, then National Institutes of Health (NIH) Director Elias Zerhouni made these justifications explicit, stating: "The policy has two basic premises: 1) the integration and accessibility of biomedical research will speed discoveries, resulting in the prevention of death and disability; and 2) the public has a right to have full access, without charge, to research findings supported by taxpayer dollars, after a reasonable period of embargo." See https://judiciary.house.gov/hearing/hearing-on-h-r-6845-the-fair-copyright-in-research-works-act-0/.

citations, if individuals who edit Wikipedia entries – and thus propagate knowledge to the general public – have trouble accessing a gated article, we should observe increases in Wiki-to-article citations after it becomes freely available.

In order to better understand the sources of changes in total article-to-article citations, I also examine citations from subsets of individuals. These include researchers at different types of institutions – commercial enterprises, universities, or hospitals – as well as individuals located in different countries at different points in the GDP per capita distribution. The idea is that some subsets of individuals may be particularly likely to face access barriers unless an article is freely available (Ware and Mabe, 2015; Houghton et al., 2011). In the same spirit, I examine changes in patent-to-article citations located in the body of the patent, which represent knowledge actually drawn upon when creating the invention, as opposed to citations inserted in the front-matter for legal reasons. In addition, I have linked patents to confidential firm-level data at the U.S. Census Bureau and future versions of this paper will examine how patent-to-article citations from particular types of patent assignees change after an article becomes open access. This linkage allows me to analyze whether there are, for instance, differential changes in patent-to-article citations from young firm assignees, small firm assignees, or high-tech firm assignees. The creation of these unique new data will concretely link, for the first time, changes in the availability of scientific literature to real economic actors (firms) at the micro-level.

To identify article transitions from gated to free availability, I use newly available data from the NIH that indicates whether and on what date focal articles are posted on PubMed Central, the NIH's repository of freely accessible scientific articles. Posting an article to PubMed Central meaningfully increases its availability to individuals without journal subscriptions, both because it is the world's largest and most widely searched repository of biomedical articles and because other search engines are then able to provide an open access version of the article in search results.[4] A freely available copy on PubMed is especially

[4] De Groote and Dorsch (2003) sent surveys to 471 faculty, residents, and students at the University of Chicago Peoria in November of 2000, and find that, of the 188 respondents, 53% use MEDLINE once per week, with all other databases experiencing much lower usage rates. In a follow-up study, De Groote et al. (2014) survey 754 health sciences faculty at the University of Illinois at Chicago (UIC) in November of 2011, and find that, of the 198 respondents, 48% use MEDLINE daily, 79% weekly, and 81% report using it as a starting point to find articles. Only 2% report never using MEDLINE. Google and Google Scholar are the next most commonly used research databases, with all others being used very infrequently. Haines et al. (2010) also find that, in 2010, most basic science researchers at the University of Vermont Medical College start their search for articles either using PubMed/MEDLINE or Google. Tenopir et al. (2004) survey faculty members at the University of Tennessee (UT) during the 2000/2001 academic year and find that 89% of respondents were aware of PubMed, and of those 87% had used it in the past year. Tenopir et al. (2007) survey a random sample of 2,000 members of the American Academy of Pediatrics (AAP) in 2004 and find that 461 of 650 respondents (70.9%) had used PubMed and the typical user accessed it 34 times in the previous year. Islamaj Dogan et al. (2009) examine one month (March 2008) of PubMed logs and identify 23

important for accessing biomedical articles because, unlike economists, it is uncommon for life scientists to circulate working papers prior to publication (Gargouri et al., 2012).

To identify the effect, on citations, of an article going live on PubMed Central, I use articles that are *never* posted on PubMed Central as a set of control articles. In addition, I construct a rich set of covariates for all articles, including NIH funding status, publication month, age at which the article went live on PubMed Central, and pre-treatment outcomes. To my knowledge, this is the first paper to examine the impacts of open access by tracking citations to the same article over time and exploiting within-*article* changes in accessibility (as opposed to within-*journal* changes – e.g., see McCabe and Snyder (2014) and Evans and Reimer (2009)), allowing me to control for the underlying propensity of an article to be cited using pre-treatment outcomes.[5]

Of course, the choice to submit an article to PubMed Central is endogenous, involving complicated interactions between authors, publishers, editors, and funders.[6] For instance, authors may push to submit a high quality article to maximize exposure and publishers may resist submission in order to retain control of a valuable asset. I deal with the endogeneity of submission by constructing an instrument using a policy change by the NIH – the Public Access Policy (PAP) – that requires all NIH-funded articles accepted for publication on or after April 8, 2008 to be made available, in final peer-reviewed form, on PubMed Central within 12 months of publication.

The effects of the policy on an article's probability of being submitted to PubMed Central – the first stage of the IV – can be seen in Figure 1. Before implementation of the PAP, both NIH and non-NIH articles were becoming gradually more likely to be available on PubMed Central (though the probability is always higher for NIH articles). After implementation, the probability of being available on PubMed Central continues to gradually increase for non-

million user sessions, 58 million searches, 67 million abstract views, and 28 million full-text views. Analysis of the referring URLs revealed that "over 80% of retrievals resulted from PubMed searches while the rest were redirected to PubMed from other search engines (e.g. Google) or websites (e.g. Wikipedia.com)." Bryan and Ozcan (2016) report that over 1 million articles are viewed per day on PubMed or PubMed Central and also find (see Figure A1 in their appendix) that monthly article downloads from PubMed Central increased from about 10 million in 2006 to 50 million by 2013.

[5] I have also constructed an article-month panel and estimated two-way fixed effects models. Article fixed effects eliminate constant (over the article's lifecycle) article attributes that affect the decision to submit to PubMed Central, such as unobserved article quality. Age fixed effects eliminate any life-cycle effects that apply to all articles as they age. Unfortunately, outcomes for treated (those submitted to PubMed Central) and control articles are not on parallel paths prior to treatment. Future versions of the paper will include difference-in-differences estimates using matched samples with parallel pre-trends. For discussions of estimates obtained using differences-in-differences versus estimates obtained using regression with pre-treatment outcomes as covariates, see (Imbens and Wooldridge, 2009, p. 70), Lechner (2011), McKenzie (2012), and Chabé-Ferret (2015).

[6] An overview of how articles are submitted to PubMed Central is here: https://www.ncbi.nlm.nih.gov/pmc/about/submission-methods/

NIH articles, but sharply increases for NIH articles. Thus, the PAP pushes some articles that would not have otherwise been on PubMed Central – NIH articles accepted for publication after April 2008 – into the repository for anyone to access.

When the outcome is article-to-article citations – which measure the access of scientists to the scientific literature – IV estimates of the impact of going live on PubMed Central are always larger than non-IV estimates. Indeed, non-IV estimates suggest an increase of no more than about 7 percent, while the IV estimates suggest an increase of up to 50 percent. Perhaps surprisingly, the large IV estimates appear to be mostly driven by researchers at universities and in relatively rich countries. Future work will examine whether these effects are dominated by researchers at less prestigious institutions with less funding for library subscriptions.

When the outcome is patent-to-article citations – which measure the access of inventors to the scientific literature – non-IV estimates suggest an increase of no more than 10 percent, while the IV estimates suggest an increase of up to 190 percent. Undisclosed results using confidential data from the Census Bureau, which will be reported in future drafts, examine whether small/large, young/old, or high-/low- tech patent assignees drive these results.

Thus, for both article-to-article and patent-to-article citations, the effect on compliers – in this case, articles that go live on PubMed Central only because the PAP mandates it – is larger than the average treatment effect. This suggests that the gated articles pried open by the PAP may be particularly valuable to innovators.

When the outcome is Wiki-to-article citations – which measure the access of the general public to the scientific literature, the non-IV estimates of the impact of going live on PubMed Central are relatively large, suggesting increases of 16 to 45 percent. This is consistent with the public having less access, than scientists or inventors, to the typical article unless it is made freely available on PubMed Central. Unfortunately, the IV estimates are extremely noisy, making it difficult to draw any firm conclusions about the effect on compliers.

This paper distinguishes itself from previous work in three main ways. First, no other study has examined such a wide range of variables measuring access. Many studies have measured access using total article-to-article citations (the traditional measure)[7], a few studies

[7] Early studies in this area indicate that open access has a very large impact on total article-to-article citations. Lawrence (2001) found large effects for computer science, Antelman (2004) and Davis and Fromerth (2007) for mathematics, Antelman (2004) for philosophy, political science, and engineering, Schwarz and Kennicutt Jr (2004) and Metcalfe (2005, 2006) for astrophysics, Harnad et al. (2004) for physics, and Eysenbach (2006) for multidisciplinary science. See Craig et al. (2007) for a review of this early literature. These early studies simply compare, in one form or another, total article-to-article citations for open access and gated focal articles, making it difficult to draw causal conclusions. More recent articles, that attempt to account for the endogeneity of open access, have found much more modest effects. Using journal-level panel data, Evans and Reimer (2009) and McCabe and Snyder (2014) find that open access increases citations by approximately 8 percent. Using an instrumental variables strategy, Gaule and Maystre (2011) do not find a

have used article-to-article citations from authors in developing countries (Davis, 2011a; Evans and Reimer, 2009; Faber Frandsen, 2009; Gaulé, 2009) or at commercial enterprises (Staudt, forthcoming), and one study (Bryan and Ozcan, 2016) has measured inventor access using patent-to-article citations, focusing on 13 high-impact journals. However, due to the new linkages between a wide variety of high quality data, including confidential firm-level data at the Census Bureau, I am able to paint a much more comprehensive portrait of how open access affects specific groups of stakeholders in the scientific enterprise.

Second, no other work has used such a rich set of control variables to account for observed article heterogeneity in cross-sectional regressions, and more importantly, no other article has been able to use within article changes in open access status to compute citations received both before and after becoming freely available. Baseline citation data allow me to proxy for article quality in regressions in a way that was not possible in previous studies.

Finally, no other study has explicitly used the PAP as an instrument for the probability that an article is available on PubMed Central. Bryan and Ozcan (2016) eyeball a visual first stage and use it to scale reduced form estimates of the PAP's impact on patent-to-article citations, but they do not provide formal IV estimates with standard errors. Staudt (forthcoming) estimates the impact of the PAP on article-to-article citations (essentially the reduced form in this paper), but does not use it as an IV for open access.

2 Data

The availability of and linkages between a wide variety of high quality data make it possible to carry out the extensive empirical analysis in this paper. These include MEDLINE, PubMed Central, Web of Science, patent-article links from Marx and Fuegi (2019), MapAffil (Torvik, 2015), and confidential firm-level micro data with patent-firm links at the U.S. Census Bureau (Dreisigmeyer et al., 2018; Goldschlag and Perlman, 2017). The rest of this section describes how I combine these sources into my final analysis samples.

2.1 Focal Articles

The set of focal scientific articles is obtained from the 2016 MEDLINE baseline files, which are maintained by the National Institutes of Health (NIH) and index nearly the entire biomedical literature.[8] In this paper, I begin with the 2,846,178 articles published between April 2006

statistically significant impact of open access on citations, though the estimates are imprecise. Finally, using evidence from randomized controlled trials, Davis et al. (2008) and Davis (2011b) fail to find a statistically significant increase in the number of citations to open access articles.

[8] MEDLINE is freely available for bulk download from the National Institutes of Health (NIH): https://www.nlm.nih.gov/databases/download/pubmed_medline.html. Information on the elements for each

and April 2010 – 24 months before and after the implementation of the NIH Public Access Policy (PAP) in April 2008.[9]

In addition to providing a list of focal articles, MEDLINE provides a large amount of information about each article, including whether it received grant funding (in particular, NIH grant funding), the journal in which it is published, its publication date, and much more. This information allows me to construct a rich set of control variables, which are discussed in Section 2.5.

2.2 Treatment Articles: Going Live on PubMed Central

One of the main contributions of this paper the is estimation of how access to an article changes in response to that article becoming open access. In this context, becoming "open access" is synonymous with going live on PubMed Central, and is the treatment of interest. I identify articles that go live by first compiling a list of all articles available on PubMed Central.[10] To obtain the dates on which these articles went live, I acquired newly available data by querying the NIH's API.[11] Thus, in addition to knowing which MEDLINE articles ever go live on PubMed Central, I am also able to pinpoint the precise date on which they became available, enabling me to determine when they transitioned from gated to open access, and allowing me to compute both pre- and post-treatment citation outcomes for each treated article.

2.3 Control Articles: Assigning Pseudo-Live Dates

Since articles that *ever* go live on PubMed Central serve as the "treatment" group, the articles that *never* go live serve as the "control" group. Of course, there is no actual date on which these control articles go live on PubMed Central, preventing the computation of pre- and post-live citation counts. To get around this complication, I assign pseudo live dates to each control article. A contrived example best illustrates the process. Consider all articles (both treatment and control) published in June 2009. Suppose 25% of the treated articles go live on PubMed Central in August 2009 and 75% go live in December 2009. Based on these percentages, I randomly assign 25% of the control articles to a pseudo live date of August 2009 and 75% to a pseudo live date of December 2009. Using these pseudo

article is here: https://www.nlm.nih.gov/bsd/mms/medlineelements.html. Code for parsing and processing the 2016 baseline files is available here: https://github.com/EconJoe/medline2016-xmlparsers.

[9] All articles in MEDLINE have a publication year. Articles without a publication month are dropped from my sample.

[10] Bulk download available here: ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/PMC-ids.csv.gz

[11] See here: https://eutils.ncbi.nlm.nih.gov/entrez/. Code is available here:

live dates, I then compute the "pre-live" and "post-live" citation counts. This assignment process is then repeated for each publication month. In sum, for a given publication month, I randomly assign control articles to pseudo live months in a way that replicates the observed distribution of treated articles over live months.[12]

2.4 Measuring Access to the Biomedical Literature

Article-to-article citations, which measure researcher access to the scientific literature, are obtained from the Web of Science (WoS), maintained by Clarivate Analytics. MapAffil, developed by (Torvik, 2015), uses author affiliation information in MEDLINE to identify the type of institution at which the author is employed and the country in which that institution is located.[13] This allows me to determine whether an article that cites a focal article has authors employed at a commercial enterprise, university, or hospital as well as the income level of the country in which they are located.[14]

Patent-to-article citations, which measure inventor access to the scientific literature, are created by Marx and Fuegi (2019) using data from Microsoft Academic Graph (MAG).[15] In addition to documenting which patents cite MEDLINE articles, the data also indicate who inserted the citation into the patent application (e.g. the patent examiner or applicant) and whether the citation appeared in the front matter or body of the patent. One of the most important contributions of this paper is the linkages it establishes between patents and confidential firm-level data at the U.S. Census Bureau. Specifically, I use a patent-firm crosswalk, developed by Dreisigmeyer et al. (2018) and analyzed by Goldschlag and Perlman (2017), to link patents to firms in the Longitudinal Business Database (LBD), a comprehensive panel of business establishments in the United States (Jarmin and Miranda, 2002). This linkage allows me to determine whether a patent that cites a focal article is assigned to various firm types. In particular, I determine whether the patent is assigned to a young firm, a small firm, or a high-tech firm – firm types that disproportionately contribute to employment, output, and productivity growth (Haltiwanger et al., 2016).

[12] Appendix Figures A2.1 through A2.3 show how my main estimates in Tables 3, 6, and 9 vary over 100 replications of this random process. The main estimates are very stable, suggesting that a single idiosyncratic random assignment of control articles to pseudo-live dates is not driving my results.

[13] Until recently, MEDLINE only provided affiliation information for the first author listed on an article. Thus, fully characterizing the institutional affiliations of an article is not possible. However, in biomedicine, the first author position typically indicates the author who came up with the idea and wrote the manuscript, making institutional characterizations meaningful on the basis of the first author alone (Bhandari et al., 2004; Baerlocher et al., 2007; Packalen and Bhattacharya, 2019; Yu et al., 2020).

[14] To determine the income tercile of the country in which a first author is located, I use data from the UN National Accounts to classify each country into per capita GDP terciles by year. I then link this country-year level data to MapAffil, which enables me to link GDP tercile information to each MEDLINE article.

[15] The data are freely available for bulk download here: https://zenodo.org/record/3593486.Xhdya8R7mUk.

Wiki-to-article citations, which measure the general public's access to the scientific literature, are obtained from revision histories of Wikipedia entries that cite a MEDLINE article.[16] Each MEDLINE article has a unique identifier, called a PubMed ID (PMID), and the syntax for citing a PMID in a Wikipedia entry is relatively standard. This allows me to identify which entries cite a MEDLINE article and then parse these entries' revision histories to determine the date on which each entry cites the focal article.

The main outcomes of interest are citations (article, patent, Wiki) in the 6- and 12-month period *after* going live on PubMed Central (or after the pseudo-live date in the case of control articles). I also compute the number of citations an article receives in the 1-6 months, 7-12 months, and 13+ months *before* going live. These pre-treatment outcomes are used as covariates in regressions, and proxy for an article's underlying propensity to be cited.

2.5 Covariates and Instrument

As noted, in addition to pre-treatment outcomes, the data allow me to construct a rich set of covariates to control for observed article heterogeneity. These include an indicator for whether the article is NIH-supported, a set of indicators for publication month, an indicator for whether the article is a "Journal Article" (as opposed to an editorial, news article, etc.), an indicator for whether the article is written in English, the number of backward citations (i.e. references) and backward citations to articles published in an open access journal, the number of key words that tag the article, journal fixed effects, and field fixed effects. I also control for the age at which an article goes live on PubMed Central.

Finally, the data allow me to construct an instrument based on which articles were impacted by the PAP. In particular, I use the interaction between an indicator for NIH funding and an indicator for an article being published after April 2008.

2.6 Final Estimation Samples

As noted, I begin with the 2,846,178 focal articles published between April 2006 and April 2010. Since article-to-article citations are censored beginning in January 2011, and I want to allow all articles the same amount of time to accrue citations, I only keep articles published on or before January 2010, reducing the sample to 1,685,033 articles. Since it is necessary to control for pre-treatment outcomes, I only keep articles with at least 6 months between publication and going live on PubMed Central, which reduces the sample to 988,744 articles.

[16] Bulk downloads of Wikipedia articles and their revision histories are available here: https://dumps.wikimedia.org/enwiki/20200401/

Finally, it is possible that some articles non-randomly sorted across the PAP date. For instance, journals may have sped up the publication process for particular articles in order to avoid being subject to the PAP. To mitigate the effects of this behavior on the estimates, I produce two final estimation samples. The first eliminates articles published in April 2008, which reduces the sample to 958,597 articles. The second eliminates articles published 6 months before or after April 2008, which reduces the sample to 612,084 articles. This assumes that publication cannot be moved up 6 months to avoid the PAP requirements and authors will not be willing to wait 6 months to benefit from the PAP requirement.

2.7 Summary Statistics

Table 1 displays summary statistics for the main variables used in the analysis. Article-to-article, patent-to-article, and Wiki-to-article citations are all quite variable with large standard deviations relative to their means. Relative to control articles, treatment articles (i.e. those that ever go live on PubMed Central) typically receive about twice the number of both post-live citations (outcomes) and pre-live citations (covariates).

Unsurprisingly, article-to-article citations are much more common than patent-to-article citations, which are much more common than Wiki-to-article citations. For instance, the typical article receives about 1.2 article-to-article citations within six months of going live on PubMed Central, which is about 96 times the number of patent-to-article citations (0.013) and 959 times the number of Wiki-to-article citations (0.0013). As will be shown in Section 4.3, the rarity of Wiki-to-article citations causes problems with the precision of regression estimates, especially IV estimates.

3 Research Design

3.1 Institutional Details

3.1.1 Open Access

Open access literature is "digital, online, free of charge, and free of most copyright and licensing restrictions" (Suber, 2012, p. 4). There two main ways that an article can be made open access, and these are often referred to as the "gold" route and the "green" route (Guédon, 2004; Harnad et al., 2004, 2008; Rizor and Holley, 2014).[17] The gold route involves researchers publishing their articles in an open access journal – a journal that, in

[17] More recently, there has emerged a third route – the "black" or pirated route (Green, 2017), exemplified by SciHub. SciHub was launched in September 2011, after the end of my sample, and thus does not cause problems with my estimates of the effects of an article going live on PubMed Central.

contrast to traditional subscription-based journals, makes it contents freely available.[18] The green route involves researchers posting copies of their articles, regardless of the type of journal in which those articles are published, to personal webpages or depositing them in repositories like PubMed Central. It is worth noting that biomedicine (as well as related fields such as clinical medicine and health) is the field in which researchers are least likely to make their work available through green open access (Gargouri et al., 2012). Thus, if a biomedical article goes live on PubMed Central, especially one published in a subscription-based journal, it represents a meaningful increase in that article's availability to individuals without institutional library subscriptions.

3.1.2 PubMed Central, PubMed, and MEDLINE

As noted, the treatment of interest is an article going live on PubMed Central, which is distinct from (and often confused with) PubMed (no *Central*). Both PubMed Central and PubMed are searchable databases maintained by the NIH, but they serve different purposes. A search of PubMed yields references to articles – including information on title, authors, abstract, and more – whether they are freely available or not. In contrast, a search of PubMed Central only produces references to articles that are freely available in the repository, and thus can be read by anyone, in full. Thus, all articles in PubMed Central are contained in PubMed, but there are many articles in PubMed that are not in PubMed Central. In addition, articles in the two databases cross-reference each other. That is, if an article produced from a PubMed search is also indexed in PubMed Central, a link to the full text will be provided.

MEDLINE, the data I use to obtain the focal articles used in the empirical analysis, is the largest subset of PubMed. It contains PubMed articles published in journals that meet some minimal quality standards as set by the Literature Selection Technical Review Committee (LSTRC). In addition, each article in MEDLINE is reviewed by staff at the National Library of Medicine (NLM), and is tagged with keywords called Medical Subject Headings (MeSH). For additional information on PubMed Central, PubMed, and MEDLINE, as well as how they relate to each other, see Williamson and Minter (2019).

3.1.3 Public Access Policy (PAP)

On January 11, 2008, the NIH announced that the full text of all NIH-supported articles accepted for publication on or after April 7, 2008 were to be submitted, in final peer-reviewed

[18] Some journals are neither fully open nor fully subscription-based. For instance, some journals allow authors to pay a fee to make their article open access in an otherwise toll access journal. Other journals embargo articles for a time and then open them to all researchers.

form, to PubMed Central immediately upon acceptance for publication.[19] This mandate for submission to PubMed Central – the Public Access Policy (PAP) – applies to all NIH-funded articles, *regardless of where they are published*. Though journals, which officially retain the copyright to NIH articles they publish, have the right to delay an NIH article's availability on PubMed Central for up to one year, it is thereafter freely accessible to anyone.

3.2 Econometric Strategy

I use two key strategies to identify the impact, on citations, of an article becoming freely available on PubMed Central. First, I construct a rich set of covariates and estimate cross-sectional regressions. If, conditional on these covariates, the probability of an article going live on PubMed Central is unrelated to unobservable variables that also affect citations (unconfoundedness), then these estimates will yield the average treatment effect of going live. Key to this strategy is the measurement pre-treatment outcomes for each article, which proxy for the article's underlying propensity to be cited, and make the unconfoundedness assumption much more plausible.

Second, I use information on whether an article is NIH-funded and whether it is published before or after the PAP to construct an instrument for whether an article becomes available on PubMed Central. Intuitively, by mandating that NIH articles published after April 7, 2008 be made publicly available on PubMed Central, the PAP pushed some articles to become freely available when they would not be otherwise.

The identification assumptions for the IV approach are threefold. First, whether an article is published before or after the PAP has a causal effect on the probability that an article is made available on PubMed Central. This assumption is shown to be satisfied in Figure 1, and will be further verified by very large first-stage F-statistics in the next section. Second, whether an article is published before or after the PAP must be unrelated to omitted variables that affect the outcomes. Though this exogeneity condition is not directly testable, I do show that observable baseline characteristics (including pre-treatment outcomes) are very similar for articles published before and after the PAP suggesting unobservable characteristics are also likely to be similar. Indeed, Table 2 shows that no covariate has a standardized difference in means above 0.25 (a commonly used threshold), indicating that the covariates are similar

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

during the pre- and post-pap period (Rubin, 2001; Stuart, 2010). Finally, the effect of the PAP on citations can only arise through its effect on the the probability that an article is made available on PubMed Central (exclusion restriction). If these assumptions hold, then the IV estimates the local average treatment effect (LATE) – that is, the effect on compliers. In this context, the compliers are the set of articles that are made available on PubMed Central only because they are NIH articles published after April 2008 and the PAP applied to them.

More formally, I evaluate the impacts of making an article open access by estimating the following type of regression equation:

$$E[cites\_post_i|.] = g(\alpha + \beta pmc_i + \gamma cites\_pre_i + \tau X_i) \tag{1}$$

$cites\_post_i$ is a post-treatment (6 or 12 months) citation count for article i and $cites\_pre_i$ is a vector of pre-treatment citation counts for the 1-6 months, 7-12 months, and 13+ months before an article goes live on PubMed Central.[20] The variable of interest is pmc_i, an indicator for whether the article ever actually goes live on PubMed Central – that is, the "treatment" variable. X_i is a vector of additional covariates, including an indicator for whether the article is NIH-funded as well as sets of dummies indicating the month-year of the article's publication, the age (in months) at which the article went live, and the journal in which it is published.

If, conditional on the covariates, the treatment indicator (pmc_i) is exogenous, then the parameter β identifies the causal effect, on the outcomes of interest, of an article going live on PubMed Central. Though the set of covariates is rich, one can never rule out other unobserved variables that affect both the outcome and the treatment. Thus, as an alternative strategy, I use the NIH indicator interacted with an indicator for whether the article was published after the PAP as an instrument for pmc_i. As noted, Figure 1 shows that the first stage for this instrument is very strong. More formally, the first stage equation is estimated as:

$$pmc_i = \alpha^{fs} + \beta^{fs}(nih_i \times postpap_i) + \gamma^{fs} cites\_pre_i + \tau^{fs} X_i + \epsilon_i \tag{2}$$

nih_i is an indicator for whether article i is NIH-funded and $postpap_i$ is an indicator for whether the article is published after April 2008 (note that the main effects are included in X_i). If the instrument is valid, then β^{fs} is the impact of the instrument on the probability that an article goes live on PubMed Central, which can be used to scale the estimate of β from equation (1) to obtain the effect of going live on PubMed Central for the articles

[20] I use specifications in which these pre-treatment outcomes enter linearly as well as specifications in which they are discretized into percentile groups.

impacted by the PAP (i.e. the compliers).

For the main results, the conditional expectation function in equation (1), g, is modelled as exponential and the parameters are estimated using Poisson Pseudo Maximum Likelihood (PPML). For IV estimates of the Poisson model, I use a control function approach in which the residuals from the first stage equation (2) are included as an additional regressor in equation (1) (Cameron and Trivedi, 2010, p. 605-610).[21] All standard errors are clustered at the journal-level. In the appendix, I also present results using linear models estimated using OLS and 2SLS, which produce quite similar results.

4 Results

4.1 Article-to-Article Citations

Table 3 displays estimates of the effect, on article-to-article citations, of a focal article going live on PubMed Central. Panels A and B display estimates when the outcome variables are the number of citations received within 6 months and 12 of going live on PubMed Central (for treated articles) or of the pseudo live date (for control articles – see Section 2.3). The first three columns use all articles published between April 2006 and October 2009, except those published in April 2008. To mitigate the effects of articles possibly sorting across the PAP date, the last three columns further exclude articles published between October 2007 and October 2008 (six months before/after the PAP). Columns (1) and (4) display the non-IV PPML estimates and columns (2) and (5) display the PPML-IV estimates using the PAP as an IV, which is operationalized as the interaction between an indicator for NIH funding and an indicator for being published after April 2008. Columns (3) and (6) display the PPML reduced form estimates, obtained by regressing the outcomes on the instrument. Panel C displays the OLS first stage estimates, obtained by regressing the treatment on the instrument. All regressions include flexible functions of pre-treatment outcomes, include an NIH dummy, publication month fixed effects, journal fixed effects, and fixed effects for the age at which an article goes live on PubMed Central.

In all cases, the IV estimates are substantially larger than the non-IV estimates. For instance, when only articles published in April 2008 are excluded, the non-IV estimates (column 1) suggest that, relative to control articles, treated articles receive, on average, about 5.0% more citations within 6 months and about 4.7% more citations within 12 months of

[21] In practice, this control function approach produces estimates that are nearly identical to those obtained using PPML to produce reduced form estimates of the effect of the PAP on citation outcomes (replace pmc_i with $nih_i \times postpap_i$ in equation (1)) and then scaling by OLS first stage estimates from equation (2) to obtain the IV estimates.

going live on PubMed Central (see numbers in square brackets for implied percent changes). In contrast, the corresponding IV estimates (column 2) suggest relative citation increases of 18.1% and 15.7%, which are over three times larger than the non-IV estimates.

When articles published between October 2007 and October 2008 are excluded, similar patterns continue to hold. Non-IV estimates (column 4) suggests increases of 6.6 and 6.5 percent, while the PPML-IV estimates (column 5) suggest increases of 50.2 and 44.6 percent (approximately a 6-fold increase). Thus, the IV estimates remain much larger than the non-IV estimates.

Overall, two important patterns emerge from these estimates. First, IV estimates are much larger than non-IV estimates, ranging from about 3 to 6 times larger. This suggests that the average effect, on article-to-article citations, of going live on PubMed Central is larger for compliers than for the entire population of articles. That is, articles that are pushed onto PubMed Central only because they are affected by the PAP experience larger treatment effects than the typical article.

Second, estimates obtained using the sample that excludes articles published six months before/after the PAP are larger than estimates obtained using the sample that only excludes articles published in April 2008. The differences are fairly modest for the non-IV estimates, but are substantial for the IV estimates, with those obtained using the sample excluding articles published six months before/after the PAP over twice as large. The mechanism underlying these large differences for the IV estimates can be unpacked by examining the reduced form estimates along with the first stage estimates in Panel C.[22] First note that the first stages, with coefficients of 0.334 and 0.327, are very strong (first-stage F-statistics of 165.0 and 60.0) and suggest that, relative to non-NIH articles, NIH articles are about 33 percentage points more likely to go live on PubMed Central after the implementation of the PAP (This large increase is also suggested visually in Figure 1). Since these two first stage estimates are very similar across the two estimation samples, the differences in the IV estimates arise from the reduced form regressions of the outcomes on the instrument. Indeed, the reduced form estimates more than double when going from the sample that only excludes articles published in April 2008 to the sample that excludes articles published six months before/after the PAP.

The reason for this increase in the reduced form estimates can be seen in the top graph of Figure 2, which shows a visual version of the reduced form. First note that there are

[22] As noted, the PPML-IV estimates in columns (2) and (5) are obtained using a control function approach in which the residuals from the first stage equation (2) are included as an additional regressor in equation (1). Thus, unlike the linear 2SLS case, the IV estimates are not identical to the ratio of coefficients for the instrument from a PPML reduced form and a linear first stage (Panel C). In practice, they are nearly identical.

no obvious trends in the citation outcomes prior to the implementation of the PAP in April 2008. Further, we see that citation outcomes also do not change immediately after April 2008. Rather, they start to increase around six months later in October 2008. This delayed effect is not surprising for several reasons. First, in Figure 1, compliance with the PAP ramped up over a period of about six months after April 2008. Second, only articles *accepted*, not articles *published*, after April 7, 2008 are subject to the PAP. Thus, if an article is accepted on April 7, 2008, it is subject to the PAP, but it may not be published until several months later. Finally, journals are allowed to embargo articles for up to 12 months after publication. Thus, we would not necessarily expect citations to jump immediately after the implementation of the PAP. The trends in the graph also make it clear that, when the 12 month period around the PAP is excluded from the figure, the later periods, which have a larger estimated effects, receive more weight in the reduced form regressions, causing the reduced form estimates – and, by extension, the IV estimates – to increase.

Tables A1 and A4 show similar patterns when the outcomes are modelled as linear. In all cases, the 2SLS estimates are larger than the OLS estimates and estimates tend to be larger when the articles published six months before/after the PAP are excluded. Note that the Poisson models yield smaller implied percent increases than do the linear models. Indeed, estimates from Poisson models suggest percent increases that are about half the size. Since the citation outcomes are count variables, and a substantial number of articles receive zero citations, it is likely that the Poisson models more accurately model the conditional mean of citations.

Tables 4 and 5 attempt to determine the sources of the citation increases observed, in Table 3, for articles that go live on PubMed Central. Table 4 examines article-to-article citations from researchers with different types of institutional affiliations. The outcomes in columns (1)-(4), (5)-(8), and (9)-(12) are, respectively, the number of citations from articles with a first author affiliated with 1) a commercial enterprise, 2) a university, and 3) a hospital. The non-IV estimates tend to be largest for commercial citations and smallest for university citations, though the differences are modest. Overall, the non-IV estimates range from 6.4 to 11.2 percent – all a bit larger than the main non-IV estimates in Table 3. Thus, the effect, on the typical article, of going live on PubMed Central seems to be driven by citations from all three institutional affiliation types. As with the main estimates, the PPML-IV estimates in Table 4 tend to be larger than the non-IV estimates. The only exception is for commercial citations using the sample excluding articles published in April 2008. Estimates for commercial citations are also noisy and are much smaller than estimates for university citations or hospital citations. The IV estimates for university citations are larger than the main estimates (ranging from 23.7 to 61.3 percent), while estimates for hospital citations tend to be smaller or close to the main estimates (ranging from 16.0 to 42.8 percent). This suggests that citations from researchers affiliated with universities disproportionately contribute to the citation increases for complier articles observed in Table 3.[23]

Table 5 examines article-to-article citations from researchers located in countries with different income levels. The outcomes in columns (1)-(4), (5)-(8), and (9)-(12) are, respectively, the number of citations from articles with a first author located in a country in the first, second, and third terciles of GDP per capita (for a given year). Once again, the main patterns revealed in Table 3 – IV estimates larger than non-IV estimates and larger IV estimates when using the sample that excludes articles published six months before/after the PAP – continue to hold, in most cases, in Table 5. The more striking result is that the main citation results appear to be predominantly driven by citations from researchers located in countries in the the third tercile of the GDP per capita distribution. Estimates – both non-IV and IV – tend to be smaller (and statistically insignificant) for citations from researchers in the first and second terciles. In contrast, the citations from researchers in the third tercile are always statistically significant and larger than the main estimates. Thus, the main citation effects in Table 3, for both the typical article and complier articles, appear to be mainly driven by researchers located in relatively rich countries.

It is perhaps surprising that the results in Tables 4 and 5 suggest that the citation increases observed in Table 3, especially for complier articles, seem to be driven mainly by citations from researchers at universities and researchers in relatively rich countries – groups we might expect to have broad access to the biomedical literature regardless of how much of it is freely available on PubMed Central. Several possible explanations and directions for future work will be discussed in the next section.

4.2 Patent-to-Article Citations

Table 6 displays estimates analogous to those in Table 3, but with article-to-article citation outcomes replaced with patent-to-article citation outcomes. Whereas the former reflect changes in access for academic researchers, the later reflect changes in access for inventors.

As with article-to-article citations, the IV estimates are substantially larger than the non-IV estimates. For instance, when only articles published in April 2008 are excluded, the non-IV estimates of 0.0499 and 0.0227 (column 1) suggest that, relative to control articles, treated articles receive, on average, about 5.1 and 2.3 percent more patent citations within 6

[23] Also similar to the main estimates in Table 3, the IV-estimates in Table 4 tend to be larger using the sample that excludes articles published six months before/after the PAP. All first stages (not reported) are very strong and around 0.33, so the differences across samples are again attributable to differences in the reduced form estimates.

and 12 months of going live on PubMed Central. In contrast, the corresponding IV estimates of 0.533 and 0.318 (column 2) suggest much larger increases of 70.4 and 37.4 percent. When articles published six months before/after the PAP are excluded, a similar pattern holds. Non-IV estimates (column 4) suggest modest percent increases of 10.0 and 5.2 percent while IV estimates (column 5) suggest much larger increases of 188.3 and 105.4 percent. These results suggest that, as with article-to-article citations, the average effect, on patent-to-article citations, of going live on PubMed Central is larger for compliers than for the entire population of articles.

Also similar to article-to-article citations, estimates for patent-to-article citations tend to be larger when estimates are obtained using the sample that excludes articles published six months before/after the PAP. Again, this is especially true for the IV estimates which more than double. The first stages, with coefficients of 0.337 and 0.328, are again very strong (first-stage F-statistics of 165.4 and 60.1) and suggest a 33 percentage point increase in the likelihood of an NIH article going live on PubMed Central after the PAP (relative to a non-NIH article).[24] The similarity of these two estimates suggests that the IV differences across the samples again arises from the reduced form differences. Indeed, the reduced form estimates become much larger when going from the sample that only excludes articles published in April 2008 to the sample that excludes articles published six months before/after the PAP.

The middle graph of Figure 2 suggests that, as with article-to-article citations, there is no evidence of trends in patent-to-article citations prior to the implementation of the PAP in April 2008. Moreover, there is again no immediate citation increase after the PAP, but an increase starting around 9 months later in January 2009. Again, this delayed response is not surprising given the gradual increase in compliance with the PAP, the fact that the date of acceptance for publication (not the date of publication itself) determines whether an article is subject to the PAP, and journals' ability to embargo articles for up to 12 months. However, it means that, when the 12 month period around the PAP is excluded from the figure, the later periods receive more weight in the reduced form regressions, causing the reduced form and IV estimates to increase.

Thus, the two main trends for article-to-article citations also hold for patent-to-article citations. Specifically, IV estimates are much larger than non-IV estimates and estimates obtained using the sample that excludes articles published six months before/after the PAP are larger than those obtained using the sample that only excludes articles published in April

[24] The first stage coefficients and F-statistics can differ between Tables 3, 4, 5, 6, 7, and 8 because the regressions control for pre-treatment outcome variables, and thus are different for article-to-article, patent-to-article, and Wiki-to-article citation outcomes. In practice, the differences are negligible.

2008.

Tables A2 and A5 show similar patterns emerge when the outcomes are modelled as linear. In all cases, the 2SLS estimates are larger than the OLS estimates and estimates tend to be larger when the articles published six months before/after the PAP are excluded. However, unlike for article-to-article citations Poisson models tend to imply larger percent changes than linear models. Again, since patent citation outcomes are counts, the Poisson models likely yield more accurate estimates of the conditional mean function. In fact, since patent-to-article citations are much rarer than article-to-article citations (see Table 1), the conditional means of these outcomes are even less likely to be well-approximated using a linear function.

Similar to Tables 4 and 5 for article-to-article citations, Table 7 attempts to determine the sources of the patent-to-article citation increase observed in Table 6. Specifically, it examines patent-to-article citations that are located in the of body patent versus citations that are only located in the front-matter. Citations located in the body better represent knowledge actually drawn upon when creating the invention, as opposed to citations inserted in the front-matter for legal reasons. All estimates are larger for citations located in the body of the patent, though none of the PPML estimates are statistically significant. The PPML-IV estimates tend to be substantially larger for citations located in the body, suggesting that the rate at which inventors incorporate information from an article into their invention substantially increases for complier articles.

4.3 Wiki-to-Article Citations

Table 8 displays estimates analogous to those in Tables 3 and 6, but with Wiki-to-article citations as the outcomes. In contrast to article-to-article citations or patent-to-article citations, which measure the access of innovators to an article, Wiki-to-article citations measure the access of the general public.

Unlike when the outcomes are article-to-article or patent-to-article citations, the non-IV estimates of the effect of an article going live on PubMed Central are quite large for Wiki-to-article citations. For instance, when only articles published in April 2008 are excluded, the non-IV estimates (column 1) imply increases of 45.1 and 37.9 percent for citations received within 6 and 12 months of going live. When articles published six months before/after the PAP are excluded, the implied percent changes remain relatively large at 16.3 and 46.2 percent for 6 and 12 month citations (column 4). Thus, if the unconfoundedness assumption holds, these estimates suggest that, for the typical article, going live on PubMed Central has a larger effect on access for the general public than for innovators.

Unfortunately, the IV estimates are extremely imprecise. Taken at face value, the point estimates suggest that the effect of going live on PubMed Central for compliers is negative and typically large.[25] Panel C shows that the first stages are again very strong (first-stage F-statistics of 165.4 and 60.1) with NIH articles 33 percentage points more likely to go live on PubMed Central after the PAP (relative to non-NIH articles), but the reduced form estimates are negative and very imprecise. When scaled by the first stage, these produce IV estimates that are large, negative, and noisy. Given the imprecision of the estimates, I would not feel comfortable drawing any strong conclusions from these IV estimates.

Tables A3 and A6 show similar patterns when the outcomes are modelled as linear. Specifically, the OLS estimates tend to imply large percent increases while the 2SLS estimates are very noisy.

5 Discussion

The main estimates, for article-to-article (Table 3), patent-to-article (Table 6), and Wiki-to-article (Table 8) citations, of the impact of an article going live on PubMed Central suggest positive effects for all broadly defined groups analyzed in this paper – scientists, inventors, and the public. However, the estimates also suggest substantially different magnitudes across stakeholders and article type (e.g. average articles versus complier articles).

The modest non-IV estimates for article-to-article (4.7-6.6%) and patent-to-article (2.3-10.0%) citations suggest that scientists and inventors do not substantially increase their use of the *typical* article after it goes live on PubMed Central. This is consistent with these groups having easy access to the typical biomedical article and thus not increasing their use of such articles after the articles become freely available on PubMed Central. In contrast, the much larger non-IV estimates for Wiki-to-article citations (16.3-46.2%) suggest that individuals who edit Wikipedia entries, and thus propagate knowledge to the public, do substantially increase their use of the typical biomedical article after it goes live. It is certainly intuitive that innovators – scientists and inventors – would have broader access than editors of Wikipedia entries to the typical biomedical article in the absence of it being freely available on PubMed Central. In sum, the non-IV estimates suggest that, while all three groups of stakeholders increase their use of the average article after it is made freely available, the public increases its use the most.

Though scientists and inventors do not seem to substantially increase their use of the *typical* article after it is made freely available, the relatively large PPML-IV estimates (15.7-

[25] The large negative estimates are entirely driven by the journal fixed effects. Dropping these, the estimates are always large and positive, but still extremely noisy.

50.2% for article-to-article citations and 37.4-188.3% for patent-to-article citations) suggest that they do substantially increase their use of a particular subset of articles after these articles go live. In particular, innovators appear to considerably increase their use of complier articles – those articles actually affected by the PAP. These complier articles are only available on PubMed Central because they are forced by the PAP, and, in the absence of the PAP, would not be on PubMed Central. The large effects for compliers may be due to journals more zealously protecting valuable high quality articles behind a paywall, causing such articles to experience large citation increases after they are pried open by the PAP. In addition, high quality journals may have been more likely than lower quality journals to closely guard their entire corpus, only allowing some of it into the open after being forced by the PAP, causing large citation increases from innovators eager to access the content in these high-quality journals. Future versions of this paper will examine this latter mechanism by measuring whether higher quality journals were more likely to refrain from posting articles on PubMed Central prior to the PAP.

That the increase in article-to-article citations, especially for complier articles, is driven by citations from researchers at universities and located in rich countries is, at first glance, surprising. Indeed, we might expect these groups of researchers to have extensive access to most biomedical articles, regardless of whether those articles are posted on PubMed Central. However, there are several possible explanations. First, it is possible that, while researchers at universities and in relatively rich countries have broad access to the biomedical literature through institutional subscriptions, placing an article on PubMed Central makes it easier to find, and thus cite. This is especially true given the centrality of PubMed and PubMed Central in searches for literature in biomedicine and related fields (see footnote 4). Second, educational institutions, even in the rich world, are highly unequal in terms of resources. It is possible that the increase in citations is driven by researchers at less prestigious institutions with limited funds for extensive journal subscriptions. Future versions of this paper will examine this possibility by linking measures of university quality to MapAffil.

6 Conclusion

In this paper, I analyzed the effects of making articles freely available on the use of those articles. In my setting, making an article freely available is synonymous with the article going live on PubMed Central, the NIH's repository of free full-text articles. Use of these focal articles was measured with citations from three different groups: scientists (article-to-article), inventors (patent-to-article), and the public (Wiki-to-article). Using an unconfoundedness approach, I find that making a typical article freely available does not substantially increase the rate at which scientists or inventors use it, but does substantially increase the rate at which the general public is exposed to it. Using the NIH's Public Access Policy (PAP) as an instrument, I find that scientists and inventors substantially increased their use of complier articles – those articles that are live on PubMed Central only because they are required to be by the PAP.

References

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

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

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

Baerlocher, Mark Otto, Marshall Newton, Tina Gautam, George Tomlinson, and Allan S Detsky (2007), "The meaning of author order in medical research." *Journal of Investigative Medicine*, 55, 174–180.

Bhandari, Mohit, Jason W Busse, Abhaya V Kulkarni, P J Devereaux, Pamela Leece, and Gordon H Guyatt (2004), "Interpreting authorship order and corresponding authorship." *Epidemiology*, 15, 125–126.

Bryan, Kevin A and Yasin Ozcan (2016), "The impact of open access mandates on invention."

Cameron, A.C. and P.K. Trivedi (2010), *Microeconometrics Using Stata, Revised Edition*. Stata Press, URL https://books.google.com/books?id=UkKQRAAACAAJ.

Chabé-Ferret, Sylvain (2015), "Analysis of the bias of matching and difference-in-difference under alternative earnings and selection processes." *Journal of Econometrics*, 185, 110–123.

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

Davis, Philip M (2011a), "Do discounted journal access programs help researchers in sub-saharan africa? a bibliometric analysis." *Learned Publishing*, 24, 287–298.

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

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

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

De Groote, Sandra L and Josephine L Dorsch (2003), "Measuring use patterns of online journals and databases." *Journal of the Medical Library Association*, 91, 231.

De Groote, Sandra L, Mary Shultz, and Deborah D Blecic (2014), "Information-seeking behavior and the use of online resources: a snapshot of current health sciences faculty." *Journal of the Medical Library Association: JMLA*, 102, 169.

Dreisigmeyer, David, Nathan Goldschlag, Marina Krylova, Wei Ouyang, and Elisabeth Perlman (2018), "Building a Better Bridge: Improving Patent Assignee-Firm Links." CES Technical Notes Series 18-01, Center for Economic Studies, U.S. Census Bureau, URL https://ideas.repec.org/p/cen/tnotes/18-01.html.

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

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

Faber Frandsen, Tove (2009), "Attracted to open access journals: a bibliometric author analysis in the field of biology." *Journal of documentation*, 65, 58–82.

Gargouri, Yassine, Vincent Larivière, Yves Gingras, Les Carr, and Stevan Harnad (2012), "Green and gold open access percentages and growth, by discipline." *arXiv preprint arXiv:1206.3664*.

Gaulé, Patrick (2009), "Access to scientific literature in india." *Journal of the American Society for Information Science and Technology*, 60, 2548–2553.

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

Goldschlag, Nathan and Elisabeth Perlman (2017), "Business Dynamic Statistics of Innovative Firms." Working Papers 17-72, Center for Economic Studies, U.S. Census Bureau, URL https://ideas.repec.org/p/cen/wpaper/17-72.html.

Green, Toby (2017), "We've failed: Pirate black open access is trumping green and gold and we must change our approach." *Learned Publishing*, 30.

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

Haines, Laura L, Jeanene Light, Donna O'Malley, and Frances A Delwiche (2010), "Information-seeking behavior of basic science researchers: implications for library services." *Journal of the Medical Library Association: JMLA*, 98, 73.

Haltiwanger, John, Ron S Jarmin, Robert B Kulick, and Javier Miranda (2016), "High growth young firms: Contribution to job, output and productivity growth." *US Census Bureau Center for Economic Studies Paper No. CES-WP-16-49*.

Harnad, Stevan, Tim Brody, François Vallières, Les Carr, Steve Hitchcock, Yves Gingras, Charles Oppenheim, Chawki Hajjem, and Eberhard R Hilf (2008), "The access/impact problem and the green and gold roads to open access: An update." *Serials review*, 34, 36–40.

Harnad, Stevan, Tim Brody, François Vallières, Les Carr, Steve Hitchcock, Yves Gingras, Charles Oppenheim, Heinrich Stamerjohanns, and Eberhard R Hilf (2004), "The access/impact problem and the green and gold roads to open access." *Serials review*, 30, 310–314.

Heilman, James M and Andrew G West (2015), "Wikipedia and medicine: quantifying readership, editors, and the significance of natural language." *Journal of medical Internet research*, 17, e62.

Houghton, John, Alma Swan, and Sheridan Brown (2011), "Access to research and technical information in denmark."

Imbens, Guido W and Jeffrey M Wooldridge (2009), "Recent developments in the econometrics of program evaluation." *Journal of economic literature*, 47, 5–86.

Islamaj Dogan, Rezarta, G Craig Murray, Aurélie Névéol, and Zhiyong Lu (2009), "Understanding pubmed® user search behavior through log analysis." *Database*, 2009.

Jarmin, Ron S and Javier Miranda (2002), "The longitudinal business database." *Available at SSRN 2128793*.

Laurent, Michaël R and Tim J Vickers (2009), "Seeking health information online: does wikipedia matter?" *Journal of the American Medical Informatics Association*, 16, 471–479.

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

Lechner, Michael (2011), “The estimation of causal effects by difference-in-difference methods.” *Foundations and Trends® in Econometrics*, 4, 165–224.

Marx, Matt and Aaron Fuegi (2019), “Reliance on science: Worldwide front-page patent citations to scientific articles.” Working Paper 3331686, Boston University Questrom School of Business, URL https://ssrn.com/abstract=3331686orhttp://dx.doi.org/10.2139/ssrn.3331686.

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

McKenzie, David (2012), “Beyond baseline and follow-up: The case for more t in experiments.” *Journal of development Economics*, 99, 210–221.

Mesgari, Mostafa, Chitu Okoli, Mohamad Mehdi, Finn Årup Nielsen, and Arto Lanamäki (2015), “the sum of all human knowledge: A systematic review of scholarly research on the content of w ikipedia.” *Journal of the Association for Information Science and Technology*, 66, 219–245.

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

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

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

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

Packalen, Mikko and Jay Bhattacharya (2019), “Age and the trying out of new ideas.” *Journal of Human Capital*, 13, 341–373.

Rizor, Sara L and Robert P Holley (2014), “Open access goals revisited: How green and gold open access are meeting (or not) their original goals.” *Journal of Scholarly Publishing*, 45, 321–335.

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

Rubin, Donald B (2001), “Using propensity scores to help design observational studies: application to the tobacco litigation.” *Health Services and Outcomes Research Methodology*, 2, 169–188.

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

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

Spoerri, Anselm (2007), “What is popular on wikipedia and why?” *First Monday*, 12.

Staudt, Joseph (forthcoming), “Mandating access: Assessing the nih’s public access policy.” *Economic Policy.*

Stuart, Elizabeth A (2010), “Matching methods for causal inference: A review and a look forward.” *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 25, 1.

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

Tennant, Jonathan P, François Waldner, Damien C Jacques, Paola Masuzzo, Lauren B Collister, and Chris HJ Hartgerink (2016), “The academic, economic and societal impacts of open access: an evidence-based review.” *F1000Research*, 5.

Tenopir, Carol, Donald W King, and Amy Bush (2004), “Medical faculty’s use of print and electronic journals: changes over time and in comparison with scientists.” *Journal of the Medical Library Association*, 92, 233.

Tenopir, Carol, Donald W King, Michael T Clarke, Kyoungsik Na, and Xiang Zhou (2007), “Journal reading patterns and preferences of pediatricians.” *Journal of the Medical Library Association*, 95, 56.

Torvik, Vetle I. (2015), “Mapaffil: A bibliographic tool for mapping author affiliation strings to cities and their geocodes worldwide.” *D-Lib Magazine*, 21.

Ware, Mark and Michael Mabe (2015), “The stm report: An overview of scientific and scholarly journal publishing.”

Williamson, Peace Ossom and Christian IJ Minter (2019), "Exploring pubmed as a reliable resource for scholarly communications services." *Journal of the Medical Library Association: JMLA*, 107, 16.

Yu, Huifeng, Gerald Marschke, Matthew B Ross, Joseph Staudt, and Bruce A Weinberg (2020), "Publish or perish: Selective attrition as a unifying explanation for patterns in innovation over the career." *Working Paper*.

Zerhouni, Elias (2008), "Congressional testimony." *H.R.6845 - Fair Copyright in Research Works Act*.

Figure 1: Fraction of NIH and comparison articles ever on PubMed Central.

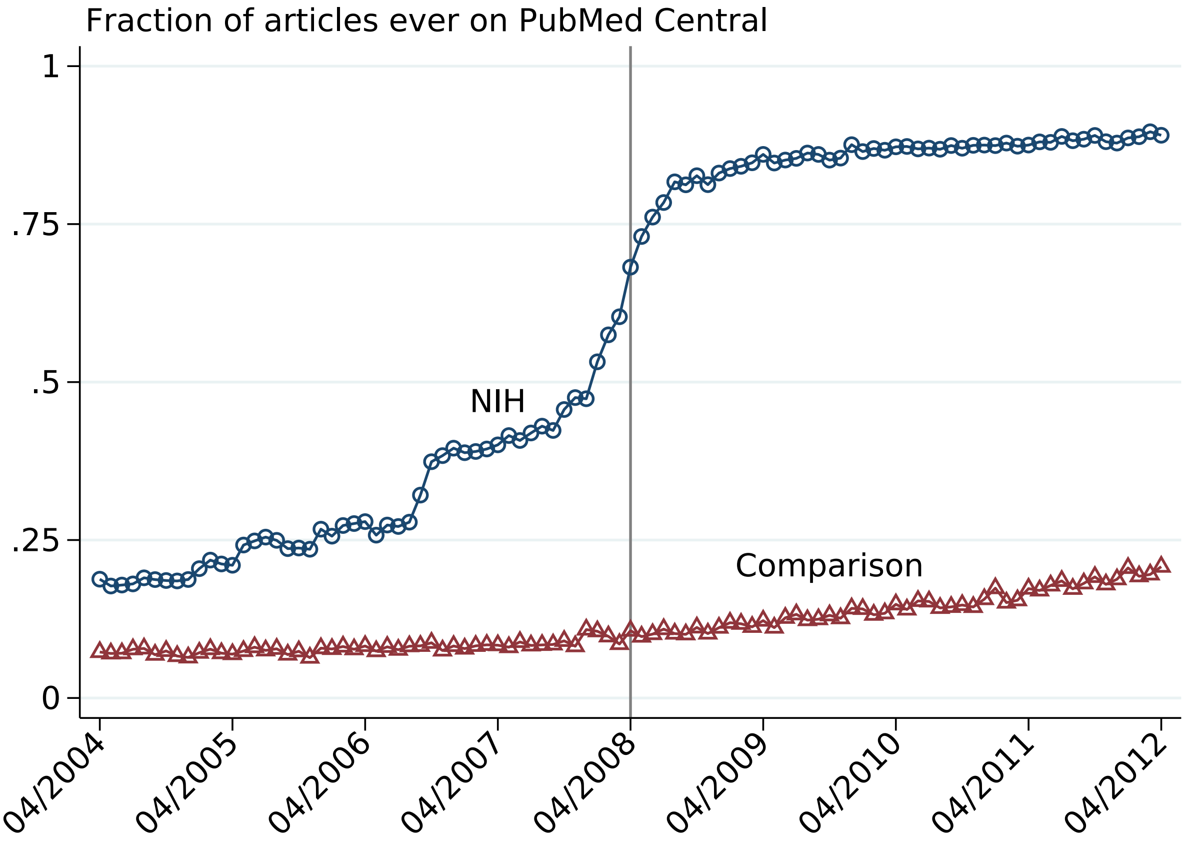
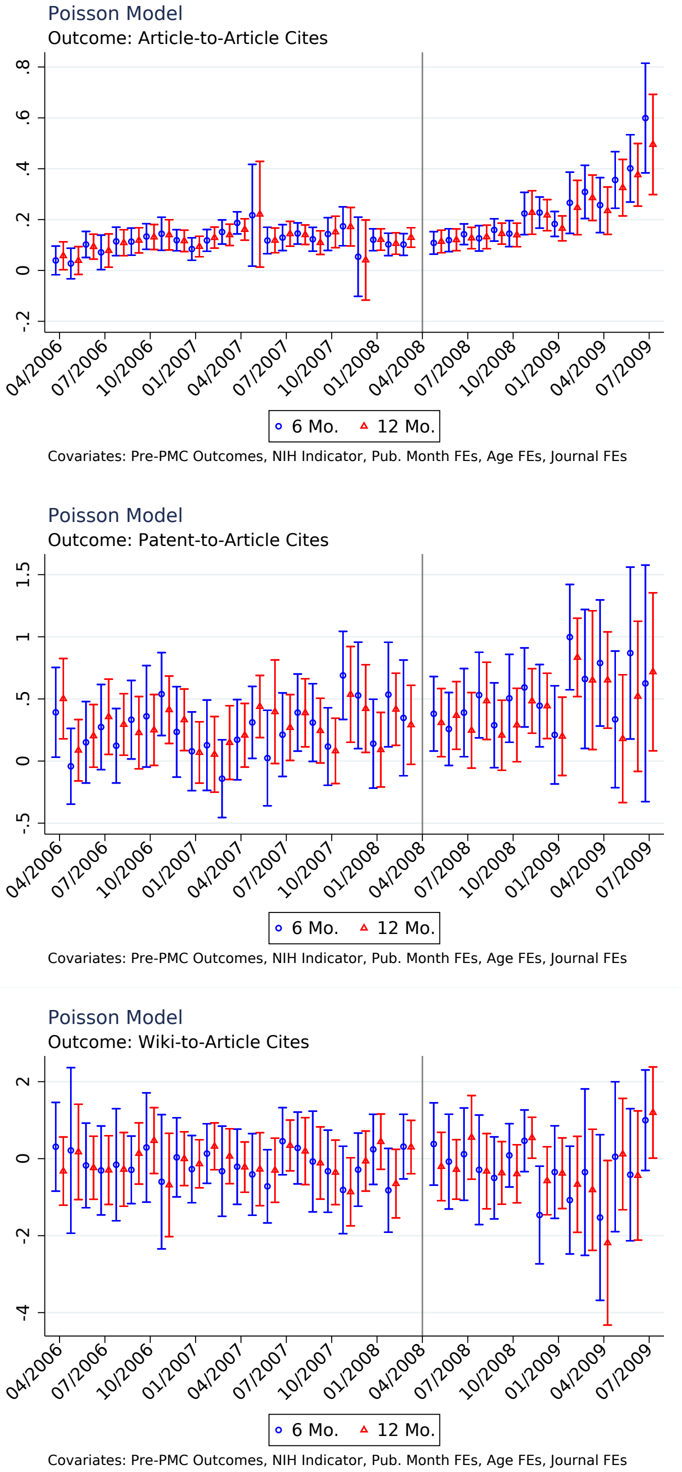


Figure 2: Visual Reduced Form for Citation Outcomes.



Notes – These graphs show visual reduced form regressions – regressions of a citation outcome (article-to-article, patent-to-article, or Wiki-to-article) on the instrument. The instrument is the NIH's Public Access Policy (PAP), which applies to NIH articles accepted for publication after April 2008. Specifically, these graphs show coefficients and 95% confidence intervals from regressions of citations on publication month dummies interacted with an indicator for an article being NIH funded (non-interacted publication month dummies are also included in regressions, but are not displayed in the graphs). The coefficients are interpreted as the citation ratio between NIH and non-NIH articles published in a particular month. All regressions control for pre-treatment outcomes and include, publication month fixed effects, age at which an article goes live fixed effects, and journal fixed effects. Confidence intervals are computed using standard errors clustered at the journal level.

Table 1: Summary Statistics.

	All		**PMC Live**		**Control**	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Post-Treatment Outcomes						
Citations 6 Months After Going Live						
Article-to-Article	1.247	4.22	2.388	4.193	1.053	4.200
Patent-to-Article	0.013	0.216	0.025	0.365	0.011	0.179
Wiki-to-Article	0.0013	0.0826	0.0026	0.0665	0.0011	0.0851
Citations 12 Months After Going Live						
Article-to-Article	2.455	8.29	4.691	7.919	2.075	8.292
Patent-to-Article	0.026	0.354	0.050	0.518	0.022	0.318
Wiki-to-Article	0.0023	0.0982	0.0046	0.0969	0.0019	0.0984
Pre-Treatment Outcomes (Controls)						
Citations 1-6 Months Before Going Live						
Article-to-Article	0.757	3.383	1.371	2.807	0.652	3.461
Patent-to-Article	0.009	0.161	0.016	0.221	0.007	0.149
Wiki-to-Article	0.0012	0.0587	0.0021	0.0699	0.0011	0.0566
Citations 7-12 Months Before Going Live						
Article-to-Article	0.409	2.684	0.608	1.881	0.375	2.796
Patent-to-Article	0.005	0.128	0.009	0.192	0.005	0.114
Wiki-to-Article	0.0008	0.0344	0.0011	0.0392	0.0008	0.0335
Other Covariates						
PMC Live	0.145	0.352	1.000	0.000	0.000	0.000
NIH Article	0.133	0.339	0.524	0.500	0.066	0.249
Post-PAP	0.287	0.452	0.383	0.486	0.271	0.444
Age at Live	14.5	8.354	13.8	7.711	14.6	8.45
Observations	958,677		138,974		819,703	

Table 2: Covariate Balance: Means and Standardized Differences.

	NIH			non-NIH			All		
	Pre-PAP	Post-PAP	Std. Diff.	Pre-PAP	Post-PAP	Std. Diff.	Pre-PAP	Post-PAP	Std. Diff.
Citations 1-6 Months Before Going Live									
Article-to-Article	1.78	1.72	-0.013	0.64	0.52	-0.044	0.79	0.68	-0.034
Patent-to-Article	0.02	0.02	-0.012	0.01	0.00	-0.020	0.01	0.01	-0.017
Wiki-to-Article	0.00	0.00	0.012	0.00	0.00	0.007	0.00	0.00	0.008
Citations 7-12 Months Before Going Live									
Article-to-Article	1.06	0.58	-0.166	0.39	0.18	-0.099	0.48	0.23	-0.108
Patent-to-Article	0.01	0.01	-0.038	0.01	0.00	-0.024	0.01	0.00	-0.025
Wiki-to-Article	0.00	0.00	-0.003	0.00	0.00	0.002	0.00	0.00	0.001
Other Covariates									
NIH Article	1.00	1.00	-	0.00	0.00	-	0.13	0.14	0.024
Unique N-Grams in Text	127.14	127.43	0.007	96.19	96.86	0.012	100.22	101.10	0.016
English Language	1.00	1.00	0.001	0.90	0.91	0.019	0.91	0.92	0.021
Journal Article	0.98	0.98	-0.041	0.91	0.90	-0.017	0.92	0.91	-0.016
Non-NIH Grants (Count)	0.08	0.13	0.107	0.03	0.04	0.059	0.03	0.06	0.069
MeSH Descriptors (Count)	9.07	9.28	0.035	11.16	11.12	-0.007	10.79	10.71	-0.014
Authors (Count)	5.50	5.76	0.061	4.52	4.62	0.014	4.65	4.78	0.019
Observations	86,064	38,150	127,214	594,370	237,093	831,463	683,434	275,243	958,677

Notes – The values under "Pre-PAP" and 'Post-PAP" are means of the variables. The values under "Std. Diff." are the standardized differences between the means in the pre- and post-PAP periods. These are computed as $(\bar{X}_{pre} - \bar{X}_{post})/\sqrt{(v_{pre} + v_{post})/2}$, where $\bar{X}_{pre}$ and v_{pre} are the mean and variance of covariate X during the pre-PAP period and $\bar{X}_{post}$ and v_{post} are the same quantities for articles in the post-PAP period. No covariate has a standardized difference above 0.25 (a commonly used threshold), indicating that the covariates are similar during the pre- and post-pap period. (Rubin, 2001; Stuart, 2010).

Table 3: Effect, on Article-to-Article Citations from All Researchers, of Going Live on PubMed Central.

Excluded Publication Months		Apr 08			Oct 07 - Oct 08	
	PPML	PPML-IV	PPML-RF	PPML	PPML-IV	PPML-RF
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 6 Month Cites						
PMC Live	0.0492***	0.166***		0.0643***	0.407***	
	(0.0152)	(0.0435)		(0.0200)	(0.0752)	
	[5]	[18.1]		[6.6]	[50.2]	
NIH × Post-PAP (Red. Form)			0.0561***			0.134***
			(0.0149)			(0.0262)
			[5.8]			[14.3]
Panel B: 12 Month Cites						
PMC Live	0.0462***	0.146***		0.0626***	0.369***	
	(0.0148)	(0.0425)		(0.0194)	(0.0715)	
	[4.7]	[15.7]		[6.5]	[44.6]	
NIH × Post-PAP (Red. Form)			0.0492***			0.121***
			(0.0144)			(0.0249)
			[5]			[12.9]
Panel C: First Stage						
NIH × Post-PAP		0.334***			0.327***	
		(0.0260)			(0.0422)	
Observations (Articles)	958,597	958,597	958,597	612,084	612,084	612,084
Journal Clusters	4,836	4,836	4,836	4,745	4,745	4,745
First Stage F-Stat		165.0			60.0	

Notes – This table displays the cross-sectional estimates of the effect, on article-to-article citations, of a focal article going live on PubMed Central. For treated articles (i.e. those that actually go live on PubMed Central), article-to-article citations are measured 6 (Panel A) and 12 (Panel B) months after an article goes live. For control articles that never go live on PubMed Central, citations are measured after the assigned pseudo-live date (see Section 2.3 for details). The NIH's Public Access Policy (PAP), which applies to NIH articles accepted for publication after April 2008, is used as an instrument for going live on PubMed Central (columns (2) and (5)), and is operationalized using the interaction between an indicator for an article being NIH funded and an indicator for an article being published after the April 2008. Columns (1)-(3) exclude articles published in April 2008 (the month the PAP was implemented) and columns (4)-(6) exclude articles published between October 2007 and October 2008 (6 months before and after the PAP). April 2006 is the earliest month articles in the sample are published and July 2009 is the latest month. All regressions control for pre-treatment outcomes and include an indicator for whether the article is NIH funded, publication month fixed effects, age at which an article goes live fixed effects, and journal fixed effects. The numbers next to "PMC Live" are estimates of the coefficient for the treatment indicator (i.e. an indicator for actually going live on PubMed Central). *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels. The numbers in parentheses are standard errors, clustered at the journal-level. The numbers in square brackets are implied percent changes, which are computed as $100 * (e^{\hat{\beta}} - 1)$. The PPML-IV estimates in columns (2) and (5) are obtained using a control function approach in which the residuals from the first stage equation (2) are included as an additional regressor in equation (1). Thus, they are not identical to the ratio of coefficients for the instrument from a PPML reduced form and a linear first stage (Panel C).

Table 4: Effect, on Article-to-Article Citations from Researchers with Different Institutional Affiliations, of Going Live on PubMed Central.

	Commercial Affiliations				University Affiliation				Hospital Affiliation			
Excluded Publication Months	Apr 08		Oct 07 - Oct 08		Apr 08		Oct 07 - Oct 08		Apr 08		Oct 07 - Oct 08	
	PPML	PPML-IV	PPML	PPML-IV	PPML	PPML-IV	PPML	PPML-IV	PPML	PPML-IV	PPML	PPML-IV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: 6 Month Cites												
PMC Live	0.0665*	0.000294	0.0828*	0.205	0.0662***	0.243***	0.0814***	0.473***	0.0766***	0.151**	0.0647*	0.359***
	(0.0372)	(0.158)	(0.0497)	(0.268)	(0.0158)	(0.0569)	(0.0217)	(0.0932)	(0.0273)	(0.0713)	(0.0346)	(0.121)
	[6.9]	[0]	[8.6]	[22.8]	[6.8]	[27.5]	[8.5]	[60.5]	[8]	[16.3]	[6.7]	[43.2]
Panel B: 12 Month Cites												
PMC Live	0.0835**	0.0732	0.0980**	0.368	0.0657***	0.213***	0.0833***	0.430***	0.0762***	0.163**	0.0696**	0.383***
	(0.0342)	(0.138)	(0.0440)	(0.248)	(0.0150)	(0.0550)	(0.0210)	(0.0878)	(0.0284)	(0.0698)	(0.0348)	(0.115)
	[8.7]	[7.6]	[10.3]	[44.5]	[6.8]	[23.7]	[8.7]	[53.7]	[7.9]	[17.7]	[7.2]	[46.7]
Observations (Articles)	958,597	958,597	612,084	612,084	958,597	958,597	612,084	612,084	958,597	958,597	612,084	612,084
Journal Clusters	4,836	4,836	4,745	4,745	4,836	4,836	4,745	4,745	4,836	4,836	4,745	4,745

Notes – This table displays the cross-sectional estimates of the effect, on article-to-article citations from researchers at different institutional types, of a focal article going live on PubMed Central. For treated articles (i.e. those that actually go live on PubMed Central), article-to-article citations are measured 6 (Panel A) and 12 (Panel B) months after an article goes live. For control articles that never go live on PubMed Central, citations are measured after the assigned pseudo-live date (see Section 2.3 for details). The NIH's Public Access Policy (PAP), which applies to NIH articles accepted for publication after April 2008, is used as an instrument for going live on PubMed Central (columns (2), (4), (6), (8), (10), and (12)), and is operationalized using the interaction between an indicator for an article being NIH funded and an indicator for an article being published after April 2008. Columns (1)-(2), (5)-(6), and (9)-(10) exclude articles published in April 2008 (the month the PAP was implemented) and columns (3)-(4), (7)-(8), and (11)-(12) exclude articles published between October 2007 and October 2008 (6 months before and after the PAP). April 2006 is the earliest month articles in the sample are published and July 2009 is the latest month. All regressions control for pre-treatment outcomes and include an indicator for whether the article is NIH funded, publication month fixed effects, age at which an article goes live fixed effects, and journal fixed effects. The numbers next to "PMC Live" are estimates of the coefficient for the treatment indicator (i.e. an indicator for actually going live on PubMed Central). *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels. The numbers in parentheses are standard errors, clustered at the journal-level. The numbers in square brackets are implied percent changes, which are computed as $100 * (e^{\hat{\beta}} - 1)$. The PPML-IV estimates are obtained using a control function approach in which the residuals from the first stage equation (2) are included as an additional regressor in equation (1).

Table 5: Effect, on Article-to-Article Citations from Researchers Located in Countries with Different Income Levels, of Going Live on PubMed Central.

	1st Tercile GDP/Capita				2nd Tercile GDP/Capita				3rd Tercile GDP/Capita			
Excluded Publication Months	Apr 08		Oct 07 - Oct 08		Apr 08		Oct 07 - Oct 08		Apr 08		Oct 07 - Oct 08	
	PPML	PPML-IV	PPML	PPML-IV	PPML	PPML-IV	PPML	PPML-IV	PPML	PPML-IV	PPML	PPML-IV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: 6 Month Cites												
PMC Live	0.0421	0.155	0.00840	0.0286	0.00922	0.0618	-0.0130	0.177	0.0572***	0.170***	0.0763***	0.444***
	(0.0526)	(0.176)	(0.0649)	(0.266)	(0.0326)	(0.108)	(0.0394)	(0.164)	(0.0145)	(0.0393)	(0.0189)	(0.0721)
	[4.3]	[16.8]	[.8]	[2.9]	[.9]	[6.4]	[-1.3]	[19.4]	[5.9]	[18.5]	[7.9]	[55.9]
Panel B: 12 Month Cites												
PMC Live	0.0840*	0.0816	0.0467	0.0544	0.0257	0.0968	0.0180	0.236*	0.0532***	0.148***	0.0735***	0.402***
	(0.0453)	(0.142)	(0.0574)	(0.214)	(0.0296)	(0.0963)	(0.0353)	(0.134)	(0.0140)	(0.0381)	(0.0184)	(0.0691)
	[8.8]	[8.5]	[4.8]	[5.6]	[2.6]	[10.2]	[1.8]	[26.6]	[5.5]	[16]	[7.6]	[49.5]
Observations (Articles)	958,597	958,597	612,084	612,084	958,597	958,597	612,084	612,084	958,597	958,597	612,084	612,084
Journal Clusters	4,836	4,836	4,745	4,745	4,836	4,836	4,745	4,745	4,836	4,836	4,745	4,745

Notes – This table displays the cross-sectional estimates of the effect, on article-to-article citations from researchers located in countries with different levels of GDP per capita, of a focal article going live on PubMed Central. For treated articles (i.e. those that actually go live on PubMed Central), article-to-article citations are measured 6 (Panel A) and 12 (Panel B) months after an article goes live. For control articles that never go live on PubMed Central, citations are measured after the assigned pseudo-live date (see Section 2.3 for details). The NIH's Public Access Policy (PAP), which applies to NIH articles accepted for publication after April 2008, is used as an instrument for going live on PubMed Central (columns (2), (4), (6), (8), (10), and (12)), and is operationalized using the interaction between an indicator for an article being NIH funded and an indicator for an article being published after April 2008. Columns (1)-(2), (5)-(6), and (9)-(10) exclude articles published in April 2008 (the month the PAP was implemented) and columns (3)-(4), (7)-(8), and (11)-(12) exclude articles published between October 2007 and October 2008 (6 months before and after the PAP). April 2006 is the earliest month articles in the sample are published and July 2009 is the latest month. All regressions control for pre-treatment outcomes and include an indicator for whether the article is NIH funded, publication month fixed effects, age at which an article goes live fixed effects, and journal fixed effects. The numbers next to "PMC Live" are estimates of the coefficient for the treatment indicator (i.e. an indicator for actually going live on PubMed Central). *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels. The numbers in parentheses are standard errors, clustered at the journal-level. The numbers in square brackets are implied percent changes, which are computed as $100 * (e^{\hat{\beta}} - 1)$. The PPML-IV estimates are obtained using a control function approach in which the residuals from the first stage equation (2) are included as an additional regressor in equation (1).

Table 6: Effect, on Patent-to-Article Citations, of Going Live on PubMed Central.

Excluded Publication Months		Apr 08			Oct 07 - Oct 08	
	PPML	PPML-IV	PPML-RF	PPML	PPML-IV	PPML-RF
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 6 Month Cites						
PMC Live	0.0499	0.533***		0.0951	1.059***	
	(0.0611)	(0.191)		(0.0780)	(0.287)	
	[5.1]	[70.4]		[10]	[188.3]	
NIH × Post-PAP (Red. Form)			0.179***			0.345***
			(0.0643)			(0.0945)
			[19.6]			[41.2]
Panel B: 12 Month Cites						
PMC Live	0.0227	0.318*		0.0506	0.720***	
	(0.0540)	(0.166)		(0.0687)	(0.274)	
	[2.3]	[37.4]		[5.2]	[105.4]	
NIH × Post-PAP (Red. Form)			0.107*			0.235***
			(0.0557)			(0.0893)
			[11.3]			[26.5]
Panel C: First Stage						
NIH × Post-PAP		0.337***			0.328***	
		(0.0262)			(0.0423)	
Observations (Articles)	958,597	958,597	958,597	612,084	612,084	612,084
Journal Clusters	4,836	4,836	4,836	4,745	4,745	4,745
First Stage F-Stat		165.4			60.1	

Notes – This table displays the cross-sectional estimates of the effect, on patent-to-article citations, of a focal article going live on PubMed Central. For treated articles (i.e. those that actually go live on PubMed Central), patent-to-article citations are measured 6 (Panel A) and 12 (Panel B) months after an article goes live. For control articles that never go live on PubMed Central, citations are measured after the assigned pseudo-live date (see Section 2.3 for details). The NIH's Public Access Policy (PAP), which applies to NIH articles accepted for publication after April 2008, is used as an instrument for going live on PubMed Central (columns (2) and (5)), and is operationalized using the interaction between an indicator for an article being NIH funded and an indicator for an article being published after the April 2008. Columns (1)-(3) exclude articles published in April 2008 (the month the PAP was implemented) and columns (4)-(6) exclude articles published between October 2007 and October 2008 (6 months before and after the PAP). April 2006 is the earliest month articles in the sample are published and July 2009 is the latest month. All regressions control for pre-treatment outcomes and include an indicator for whether the article is NIH funded, publication month fixed effects, age at which an article goes live fixed effects, and journal fixed effects. The numbers next to "PMC Live" are estimates of the coefficient for the treatment indicator (i.e. an indicator for actually going live on PubMed Central). *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels. The numbers in parentheses are standard errors, clustered at the journal-level. The numbers in square brackets are implied percent changes, which are computed as $100 * (e^{\hat{\beta}} - 1)$. The PPML-IV estimates in columns (2) and (5) are obtained using a control function approach in which the residuals from the first stage equation (2) are included as an additional regressor in equation (1). Thus, they are not identical to the ratio of coefficients for the instrument from a PPML reduced form and a linear first stage (Panel C).

Table 7: Effect, on Patent-to-Article Citations located in the Body and Front Matter, of Going Live on PubMed Central.

	Patent Body				Patent Front-Matter Only			
Excluded Publication Months	Apr 08		Oct 07 - Oct 08		Apr 08		Oct 07 - Oct 08	
	PPML	PPML-IV	PPML	PPML-IV	PPML	PPML-IV	PPML	PPML-IV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: 6 Month Cites								
PMC Live	0.0850	0.843***	0.140	1.215**	0.0401	0.300	0.0850	0.805***
	(0.0798)	(0.0798)	(0.118)	(0.543)	(0.0722)	(0.219)	(0.0862)	(0.300)
	[8.9]	[132.3]	[15]	[237]	[4.1]	[35]	[8.9]	[123.7]
Panel B: 12 Month Cites								
PMC Live	0.0460	0.484*	0.130	0.858*	0.0203	0.129	0.0326	0.467*
	(0.0610)	(0.0610)	(0.0943)	(0.472)	(0.0651)	(0.188)	(0.0744)	(0.270)
	[4.7]	[62.3]	[13.9]	[135.8]	[2.1]	[13.8]	[3.3]	[59.5]
Observations	958,597	958,597	612,084	612,084	958,597	958,597	612,084	612,084
Journal Clusters	4,836	4,836	4,745	4,745	4,836	4,836	4,745	4,745
First Stage F-Stat		X		X		X		X

Notes – This table displays the cross-sectional estimates of the effect, on patent-to-article citations located in the patent body and located only in the front-matter of the patent, of a focal article going live on PubMed Central. For treated articles (i.e. those that actually go live on PubMed Central), patent-to-article citations are measured 6 (Panel A) and 12 (Panel B) months after an article goes live. For control articles that never go live on PubMed Central, citations are measured after the assigned pseudo-live date (see Section 2.3 for details). The NIH's Public Access Policy (PAP), which applies to NIH articles accepted for publication after April 2008, is used as an instrument for going live on PubMed Central (columns (2), (4), (6), (8), (10), and (12)), and is operationalized using the interaction between an indicator for an article being NIH funded and an indicator for an article being published after April 2008. Columns (1)-(2), (5)-(6), and (9)-(10) exclude articles published in April 2008 (the month the PAP was implemented) and columns (3)-(4), (7)-(8), and (11)-(12) exclude articles published between October 2007 and October 2008 (6 months before and after the PAP). April 2006 is the earliest month articles in the sample are published and July 2009 is the latest month. All regressions control for pre-treatment outcomes and include an indicator for whether the article is NIH funded, publication month fixed effects, age at which an article goes live fixed effects, and journal fixed effects. The numbers next to "PMC Live" are estimates of the coefficient for the treatment indicator (i.e. an indicator for actually going live on PubMed Central). *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels. The numbers in parentheses are standard errors, clustered at the journal-level. The numbers in square brackets are implied percent changes, which are computed as $100 * (e^{\hat{\beta}} - 1)$. The PPML-IV estimates are obtained using a control function approach in which the residuals from the first stage equation (2) are included as an additional regressor in equation (1).

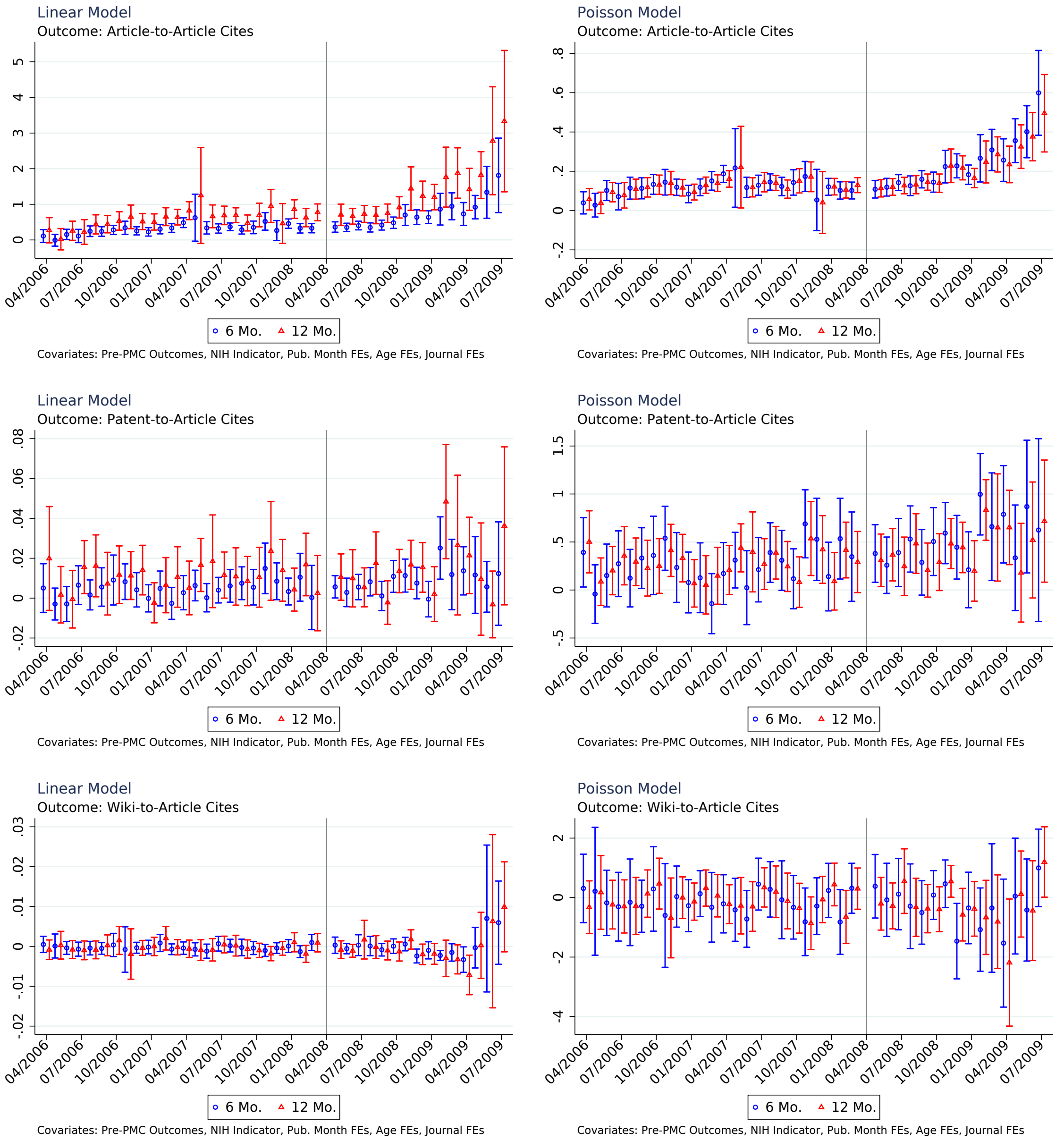
Table 8: Effect, on Wiki-to-Article Citations, of Going Live on PubMed Central.

Excluded Publication Months		Apr 08			Oct 07 - Oct 08	
	PPML	PPML-IV	PPML-RF	PPML	PPML-IV	PPML-RF
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 6 Month Cites						
PMC Live	0.372*	-0.0831		0.151	-0.927	
	(0.215)	(0.534)		(0.285)	(0.908)	
	[45.1]	[-8]		[16.3]	[-60.4]	
NIH × Post-PAP (Red. Form)			-0.0288			-0.310
			(0.185)			(0.304)
			[-2.8]			[-26.7]
Panel B: 12 Month Cites						
PMC Live	0.321**	-0.292		0.380**	-0.454	
	(0.153)	(0.421)		(0.191)	(0.694)	
	[37.9]	[-25.3]		[46.2]	[-36.5]	
NIH × Post-PAP (Red. Form)			-0.105			-0.178
			(0.151)			(0.244)
			[-10]			[-16.3]
Panel C: First Stage						
NIH × Post-PAP		0.337***			0.328***	
		(0.0262)			(0.0423)	
Observations (Articles)	958,597	958,597	958,597	612,084	612,084	612,084
Journal Clusters	4,836	4,836	4,836	4,745	4,745	4,745
First Stage F-Stat		165.4			60.1	

Notes – This table displays the cross-sectional estimates of the effect, on Wiki-to-article citations, of a focal article going live on PubMed Central. For treated articles (i.e. those that actually go live on PubMed Central), article-to-article citations are measured 6 (Panel A) and 12 (Panel B) months after an article goes live. For control articles that never go live on PubMed Central, citations are measured after the assigned pseudo-live date (see Section 2.3 for details). The NIH's Public Access Policy (PAP), which applies to NIH articles accepted for publication after April 2008, is used as an instrument for going live on PubMed Central (columns (2) and (5)), and is operationalized using the interaction between an indicator for an article being NIH funded and an indicator for an article being published after the April 2008. Columns (1)-(3) exclude articles published in April 2008 (the month the PAP was implemented) and columns (4)-(6) exclude articles published between October 2007 and October 2008 (6 months before and after the PAP). April 2006 is the earliest month articles in the sample are published and July 2009 is the latest month. All regressions control for pre-treatment outcomes and include an indicator for whether the article is NIH funded, publication month fixed effects, age at which an article goes live fixed effects, and journal fixed effects. The numbers next to "PMC Live" are estimates of the coefficient for the treatment indicator (i.e. an indicator for actually going live on PubMed Central). *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels. The numbers in parentheses are standard errors, clustered at the journal-level. The numbers in square brackets are implied percent changes, which are computed as $100 * (e^{\hat{\beta}} - 1)$. The PPML-IV estimates in columns (2) and (5) are obtained using a control function approach in which the residuals from the first stage equation (2) are included as an additional regressor in equation (1). Thus, they are not identical to the ratio of coefficients for the instrument from a PPML reduced form and a linear first stage (Panel C).

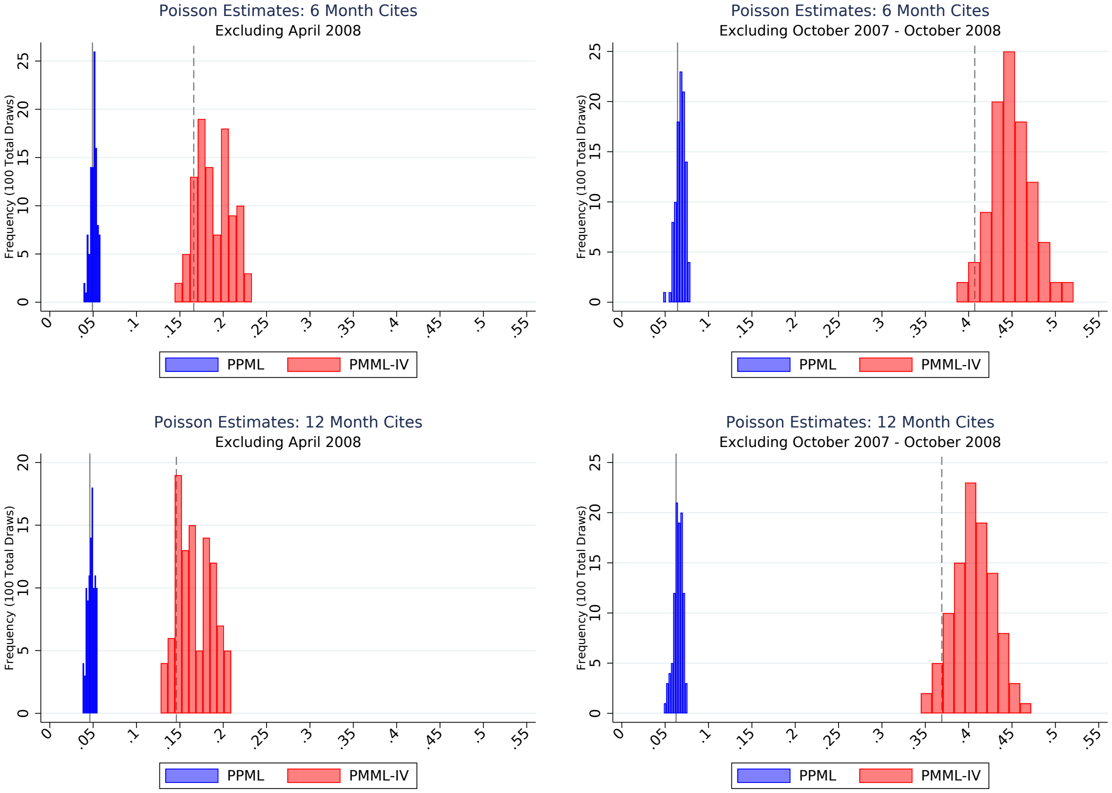
A Appendix

Figure A1: Visual Reduced Form for Citation Outcomes.



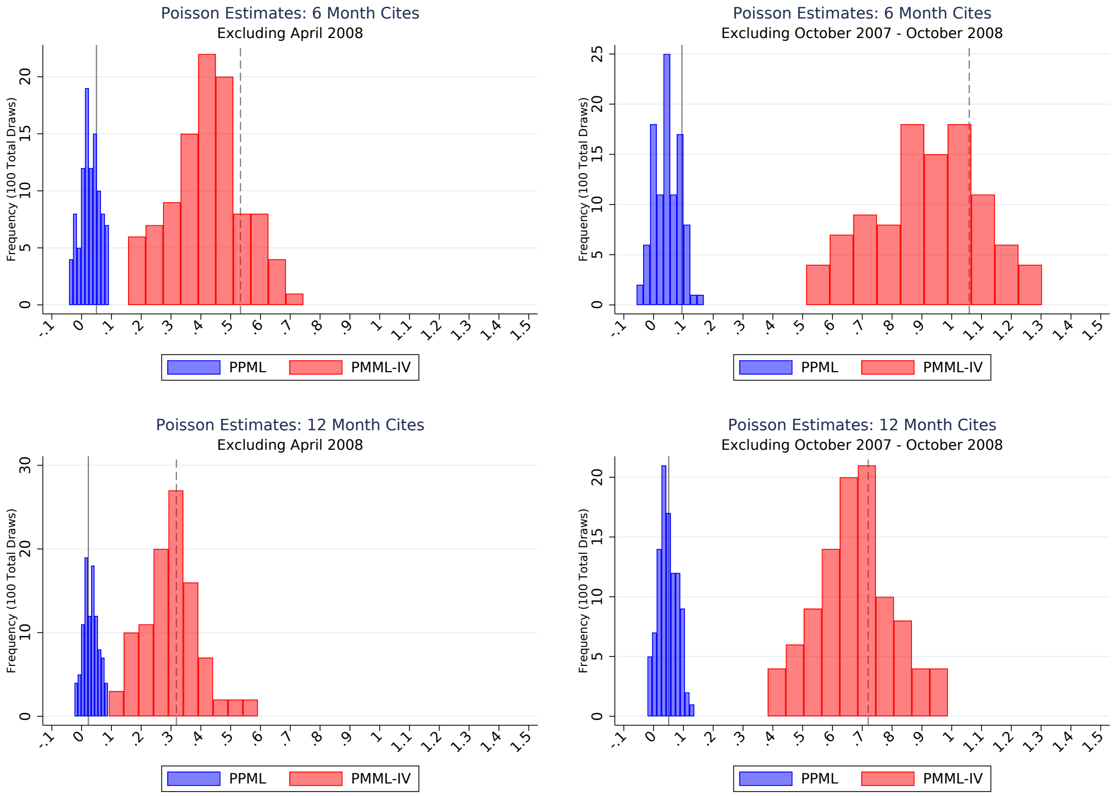
Notes – These graphs show visual reduced form regressions – regressions of a citation outcome (article-to-article, patent-to-article, or Wiki-to-article) on the instrument. The instrument is the NIH's Public Access Policy (PAP), which applies to NIH articles accepted for publication after April 2008. Specifically, these graphs show coefficients and 95% confidence intervals from regressions of citations on publication month dummies interacted with an indicator for an article being NIH funded (non-interacted publication month dummies are also included in regressions, but are not displayed in the graphs). For linear (Poisson) models, the coefficients are interpreted as the citation difference (ratio) between NIH and non-NIH articles published in a particular month. All regressions control for pre-treatment outcomes and include, publication month fixed effects, age at which an article goes live fixed effects, and journal fixed effects. Confidence intervals are computed using standard errors clustered at the journal level.

Figure A2.1: Distribution, over 100 random assignments of control articles to pseudo live dates, of the the effect, on article-to-article citations from all researchers, of going live on PubMed Central.



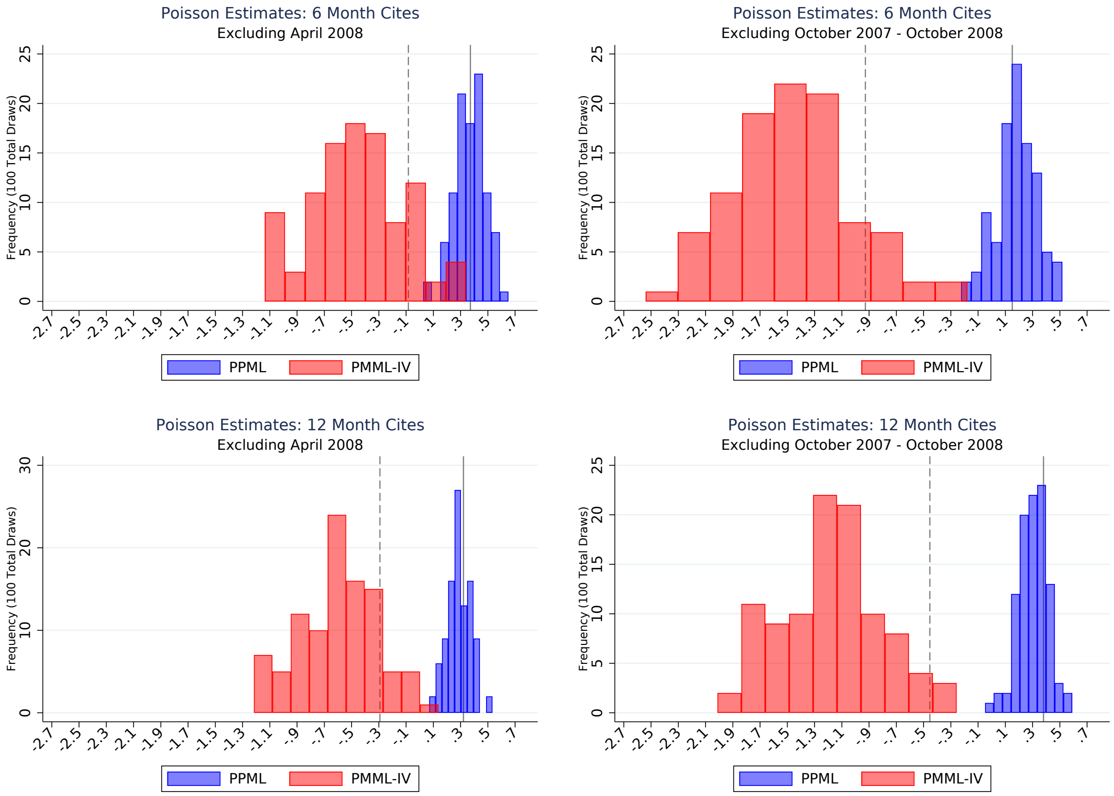
Notes – These graphs show .

Figure A2.2: Distribution, over 100 random assignments of control articles to pseudo live dates, of the the effect, on patent-to-article citations from all inventors, of going live on PubMed Central.



Notes – These graphs show .

Figure A2.3: Distribution, over 100 random assignments of control articles to pseudo live dates, of the the effect, on Wiki-to-article citations from all inventors, of going live on PubMed Central.



Notes – These graphs show .

Table A1: Effect, on Article-to-Article Citations from All Researchers, of Going Live on PubMed Central.

Excluded Publication Months	Apr 08				Oct 07 - Oct 08			
	Linear		Poisson		Linear		Poisson	
	Non-IV	IV	Non-IV	IV	Non-IV	IV	Non-IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: 6 Month Cites								
PMC Live	0.127**	0.631***	0.0492***	0.166***	0.166**	1.426***	0.0643***	0.407***
	(0.0509)	(0.1364)	(0.0152)	(0.0435)	(0.0675)	(0.2882)	(0.0200)	(0.0752)
	[10.2]	[50.6]	[5.0]	[18.1]	[13.5]	[116.1]	[6.6]	[50.2]
Panel B: 12 Month Cites								
PMC Live	0.217**	1.204***	0.0462***	0.146***	0.308**	2.875***	0.0626***	0.369***
	(0.1012)	(0.2812)	(0.0148)	(0.0425)	(0.1328)	(0.5971)	(0.0194)	(0.0715)
	[8.8]	[49.1]	[4.7]	[15.7]	[12.7]	[118.8]	[6.5]	[44.6]
Observations	958,597	958,597	958,597	958,597	612,084	612,084	612,084	612,084
Journal Clusters	4,836	4,836	4,836	4,836	4,745	4,745	4,745	4,745

Notes – This table displays the cross-sectional estimates of the effect, on article-to-article citations, of a focal article going live on PubMed Central. For treated articles (i.e. those that actually go live on PubMed Central), article-to-article citations are measured 6 (Panel A) and 12 (Panel B) months after an article goes live. For control articles that never go live on PubMed Central, citations are measured after the assigned pseudo-live date (see Section 2.3 for details). The NIH's Public Access Policy (PAP), which applies to NIH articles accepted for publication after April 2008, is used as an instrument for going live on PubMed Central (columns (2), (4), (6), and (8)), and is operationalized using the interaction between an indicator for an article being NIH funded and an indicator for an article being published after the April 2008. Columns (1)-(4) exclude articles published in April 2008 (the month the PAP was implemented) and columns (5)-(8) exclude articles published between October 2007 and October 2008 (6 months before and after the PAP). April 2006 is the earliest month articles in the sample are published and July 2009 is the latest month. All regressions control for pre-treatment outcomes and include an indicator for whether the article is NIH funded, publication month fixed effects, age at which an article goes live fixed effects, and journal fixed effects. The numbers next to "PMC Live" are estimates of the coefficient for the treatment indicator (i.e. an indicator for actually going live on PubMed Central). *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels. The numbers in parentheses are standard errors, clustered at the journal-level. The numbers in square brackets are implied percent changes, which are computed as $100 * (e^{\hat{\beta}} - 1)$ for the PPML estimates and $100 * (\hat{\beta}/\bar{y})$ for the linear estimates ($\bar{y}$ is the mean of the outcome variable). The Poisson IV estimates in columns (4) and (8) are obtained indirectly, by computing the ratio of coefficients for the instrument from a Poisson reduced form and a linear first stage. The standard errors for these estimates are bootstrapped using 100 replications.

Table A2: Effect, on Patent-to-Article Citations, of Going Live on PubMed Central.

Excluded Publication Months	Apr 08				Oct 07 - Oct 08			
	Linear		Poisson		Linear		Poisson	
	Non-IV	IV	Non-IV	IV	Non-IV	IV	Non-IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: 6 Month Cites								
PMC Live	0.002	0.008	0.0499	0.533***	0.002	0.017**	0.0951	1.059***
	(0.0014)	(0.0052)	(0.0611)	(0.191)	(0.0018)	(0.0078)	(0.0780)	(0.287)
	[15.9]	[60.3]	[5.1]	[70.4]	[18.0]	[128.4]	[10.0]	[188.3]
Panel B: 12 Month Cites								
PMC Live	0.002	0.004	0.0227	0.318*	0.003	0.018	0.0506	0.720***
	(0.0024)	(0.0084)	(0.0540)	(0.166)	(0.0032)	(0.0136)	(0.0687)	(0.274)
	[7.8]	[15.9]	[2.3]	[37.4]	[10.1]	[69.6]	[5.2]	[105.4]
Observations	958,597	958,597	958,597	958,597	612,084	612,084	612,084	612,084
Journal Clusters	4,836	4,836	4,836	4,836	4,745	4,745	4,745	4,745

Notes – This table displays the cross-sectional estimates of the effect, on patent-to-article citations, of a focal article going live on PubMed Central. For treated articles (i.e. those that actually go live on PubMed Central), patent-to-article citations are measured 6 (Panel A) and 12 (Panel B) months after an article goes live. For control articles that never go live on PubMed Central, citations are measured after the assigned pseudo-live date (see Section 2.3 for details). The NIH's Public Access Policy (PAP), which applies to NIH articles accepted for publication after April 2008, is used as an instrument for going live on PubMed Central (columns (2), (4), (6), and (8)), and is operationalized using the interaction between an indicator for an article being NIH funded and an indicator for an article being published after the April 2008. Columns (1)-(4) exclude articles published in April 2008 (the month the PAP was implemented) and columns (5)-(8) exclude articles published between October 2007 and October 2008 (6 months before and after the PAP). April 2006 is the earliest month articles in the sample are published and July 2009 is the latest month. All regressions control for pre-treatment outcomes and include an indicator for whether the article is NIH funded, publication month fixed effects, age at which an article goes live fixed effects, and journal fixed effects. The numbers next to "PMC Live" are estimates of the coefficient for the treatment indicator (i.e. an indicator for actually going live on PubMed Central). *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels. The numbers in parentheses are standard errors, clustered at the journal-level. The numbers in square brackets are implied percent changes, which are computed as $100 * (e^{\hat{\beta}} - 1)$ for the PPML estimates and $100 * (\hat{\beta}/\bar{y})$ for the linear estimates ($\bar{y}$ is the mean of the outcome variable). The Poisson IV estimates in columns (4) and (8) are obtained indirectly, by computing the ratio of coefficients for the instrument from a Poisson reduced form and a linear first stage. The standard errors for these estimates are bootstrapped using 100 replications.

Table A3: Effect, on Wiki-to-Article Citations, of Going Live on PubMed Central.

Excluded Publication Months	Apr 08				Oct 07 - Oct 08			
	Linear		Poisson		Linear		Poisson	
	Non-IV	IV	Non-IV	IV	Non-IV	IV	Non-IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: 6 Month Cites								
PMC Live	0.00080**	-0.00020	0.372*	-0.0831	0.00056	-0.0024*	0.151	-0.927
	(0.00038)	(0.00094)	(0.215)	(0.534)	(0.00053)	(0.00147)	(0.285)	(0.908)
	[62.1]	[-15.4]	[45.1]	[-8.0]	[41.1]	[-179.3]	[16.3]	[-60.4]
Panel B: 12 Month Cites								
PMC Live	0.0012**	-0.00095	0.321**	-0.292	0.0017**	-0.0032	0.380**	-0.454
	(0.00050)	(0.00149)	(0.153)	(0.421)	(0.00069)	(0.00256)	(0.191)	(0.694)
	[53.1]	[-41.6]	[37.9]	[-25.3]	[67.4]	[-129.1]	[46.2]	[-36.5]
Observations	958,597	958,597	958,597	958,597	612,084	612,084	612,084	612,084
Journal Clusters	4,836	4,836	4,836	4,836	4,745	4,745	4,745	4,745

Notes – This table displays the cross-sectional estimates of the effect, on Wiki-to-article citations, of a focal article going live on PubMed Central. For treated articles (i.e. those that actually go live on PubMed Central), article-to-article citations are measured 6 (Panel A) and 12 (Panel B) months after an article goes live. For control articles that never go live on PubMed Central, citations are measured after the assigned pseudo-live date (see Section 2.3 for details). The NIH's Public Access Policy (PAP), which applies to NIH articles accepted for publication after April 2008, is used as an instrument for going live on PubMed Central (columns (2), (4), (6), and (8)), and is operationalized using the interaction between an indicator for an article being NIH funded and an indicator for an article being published after the April 2008. Columns (1)-(4) exclude articles published in April 2008 (the month the PAP was implemented) and columns (5)-(8) exclude articles published between October 2007 and October 2008 (6 months before and after the PAP). April 2006 is the earliest month articles in the sample are published and July 2009 is the latest month. All regressions control for pre-treatment outcomes and include an indicator for whether the article is NIH funded, publication month fixed effects, age at which an article goes live fixed effects, and journal fixed effects. The numbers next to "PMC Live" are estimates of the coefficient for the treatment indicator (i.e. an indicator for actually going live on PubMed Central). *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels. The numbers in parentheses are standard errors, clustered at the journal-level. The numbers in square brackets are implied percent changes, which are computed as $100 * (e^{\hat{\beta}} - 1)$ for the PPML estimates and $100 * (\hat{\beta}/\bar{y})$ for the linear estimates ($\bar{y}$ is the mean of the outcome variable). The Poisson IV estimates in columns (4) and (8) are obtained indirectly, by computing the ratio of coefficients for the instrument from a Poisson reduced form and a linear first stage. The standard errors for these estimates are bootstrapped using 100 replications.

Table A4: Article-to-Article Reduced Forms and First Stages.

	Apr 08		Oct 07 - Oct 08	
	Linear	Poisson	Linear	Poisson
	(1)	(2)	(3)	(4)
Panel A: 6 Month Cites				
NIH × Post-PAP	0.211***	0.0561***	0.466***	0.134***
	(0.0456)	(0.0149)	(0.0959)	(0.0262)
	[16.9]	[5.8]	[37.9]	[14.3]
Panel B: 12 Month Cites				
NIH × Post-PAP	0.402***	0.0492***	0.940***	0.121***
	(0.0932)	(0.0144)	(0.1984)	(0.0249)
	[16.4]	[5.0]	[38.8]	[12.9]
Panel C: PMC Live				
NIH × Post-PAP	0.334***		0.327***	
	(0.0260)		(0.0422)	
F-Stat	165.0		60.0	
Observations	958,597	958,597	612,0847	612,084
Journal Clusters	4,836	4,836	4,745	4,745

Notes – This table displays the reduced forms (Panels A and B) and first stages (Panel C) for the IV estimates presented in Table 2. Columns (1), (2), (3), and (4) of this table correspond to columns (2), (4), (6), and (8) of Table 2. For treated articles (i.e. those that actually go live on PubMed Central), article-to-article citations are measured 6 (Panel A) and 12 (Panel B) months after an article goes live. For treated articles (i.e. those that actually go live on PubMed Central), article-to-article citations are measured 6 (Panel A) and 12 (Panel B) months after an article goes live. For control articles that never go live on PubMed Central, citations are measured after the assigned pseudo-live date (see Section 2.3 for details). Columns (1)-(2) exclude articles published in April 2008 (the month the PAP was implemented) and columns (3)-(4) exclude articles published between October 2007 and October 2008 (6 months before and after the PAP). April 2006 is the earliest month articles in the sample are published and July 2009 is the latest month. All regressions control for pre-treatment outcomes and include an indicator for whether the article is NIH funded, publication month fixed effects, age at which an article goes live fixed effects, and journal fixed effects. The NIH's Public Access Policy (PAP), which applies to NIH articles accepted for publication after April 2008, is used as an instrument for going live on PubMed Central, and is operationalized using the interaction between an indicator for an article being NIH funded and an indicator for an article being published after the April 2008. Thus, the numbers next to "NIH × Post-PAP" are estimates of the coefficient for this interaction on citations (Panels A and B) and the probability of going live on PubMed Central (Panel C). *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels. The numbers in parentheses are standard errors, clustered at the journal-level. The numbers in square brackets are implied percent changes, which are computed as $100 * (e^{\hat{\beta}} - 1)$ for the Poisson estimates and $100 * (\hat{\beta}/\bar{y})$ for the linear estimates ($\bar{y}$ is the mean of the outcome variable).

Table A5: Patent-to-Article Reduced Forms and First Stages.

	Apr 08		Oct 07 - Oct 08	
	Linear	Poisson	Linear	Poisson
	(1)	(2)	(3)	(4)
Panel A: 6 Month Cites				
NIH × Post-PAP	0.003	0.179***	0.006**	0.345***
	(0.0018)	(0.0643)	(0.0025)	(0.0945)
	[20.3]	[19.6]	[42.2]	[41.2]
Panel B: 12 Month Cites				
NIH × Post-PAP	0.001	0.107*	0.006	0.235***
	(0.0028)	(0.0557)	(0.0044)	(0.0893)
	[5.3]	[11.3]	[22.8]	[26.5]
Panel C: PMC Live				
NIH × Post-PAP	0.337***		0.328***	
	(0.0261)		(0.0423)	
F-Stat	165.7		60.2	
Observations	958,597	958,597	612,084	612,084
Journal Clusters	4,836	4,836	4,745	4,745

Notes – This table displays the reduced forms (Panels A and B) and first stages (Panel C) for the IV estimates presented in Table 3. Columns (1), (2), (3), and (4) of this table correspond to columns (2), (4), (6), and (8) of Table 3. For treated articles (i.e. those that actually go live on PubMed Central), patent-to-article citations are measured 6 (Panel A) and 12 (Panel B) months after an article goes live. For treated articles (i.e. those that actually go live on PubMed Central), article-to-article citations are measured 6 (Panel A) and 12 (Panel B) months after an article goes live. For control articles that never go live on PubMed Central, citations are measured after the assigned pseudo-live date (see Section 2.3 for details). Columns (1)-(2) exclude articles published in April 2008 (the month the PAP was implemented) and columns (3)-(4) exclude articles published between October 2007 and October 2008 (6 months before and after the PAP). April 2006 is the earliest month articles in the sample are published and July 2009 is the latest month. All regressions control for pre-treatment outcomes and include an indicator for whether the article is NIH funded, publication month fixed effects, age at which an article goes live fixed effects, and journal fixed effects. The NIH's Public Access Policy (PAP), which applies to NIH articles accepted for publication after April 2008, is used as an instrument for going live on PubMed Central, and is operationalized using the interaction between an indicator for an article being NIH funded and an indicator for an article being published after the April 2008. Thus, the numbers next to "NIH × Post-PAP" are estimates of the coefficient for this interaction on citations (Panels A and B) and the probability of going live on PubMed Central (Panel C). *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels. The numbers in parentheses are standard errors, clustered at the journal-level. The numbers in square brackets are implied percent changes, which are computed as $100 * (e^{\hat{\beta}} - 1)$ for the Poisson estimates and $100 * (\hat{\beta}/\bar{y})$ for the linear estimates ($\bar{y}$ is the mean of the outcome variable).

Table A6: Wiki-to-Article Reduced Forms and First Stages.

	Apr 08		Oct 07 - Oct 08	
	Linear	Poisson	Linear	Poisson
	(1)	(2)	(3)	(4)
Panel A: 6 Month Cites				
NIH × Post-PAP	-0.00007	-0.0288	-0.00080*	-0.310
	(0.000314)	(0.185)	(0.000479)	(0.304)
	[-5.2]	[-2.8]	[-58.9]	[-26.7]
Panel B: 12 Month Cites				
NIH × Post-PAP	-0.00032	-0.105	-0.00104	-0.178
	(0.000491)	(0.151)	(0.000784)	(0.244)
	[-14.0]	[-10.0]	[-42.4]	[-16.3]
Panel C: PMC Live				
NIH × Post-PAP	0.337***		0.328***	
	(0.0261)		(0.0423)	
F-Stat	165.7		60.2	
Observations	958,597	958,597	612,084	612,084
Journal Clusters	4,836	4,836	4,745	4,745

Notes – This table displays the reduced forms (Panels A and B) and first stages (Panel C) for the IV estimates presented in Table 3. Columns (1), (2), (3), and (4) of this table correspond to columns (2), (4), (6), and (8) of Table 3. For treated articles (i.e. those that actually go live on PubMed Central), Wiki-to-article citations are measured 6 (Panel A) and 12 (Panel B) months after an article goes live. For treated articles (i.e. those that actually go live on PubMed Central), article-to-article citations are measured 6 (Panel A) and 12 (Panel B) months after an article goes live. For control articles that never go live on PubMed Central, citations are measured after the assigned pseudo-live date (see Section 2.3 for details). Columns (1)-(2) exclude articles published in April 2008 (the month the PAP was implemented) and columns (3)-(4) exclude articles published between October 2007 and October 2008 (6 months before and after the PAP). April 2006 is the earliest month articles in the sample are published and July 2009 is the latest month. All regressions control for pre-treatment outcomes and include an indicator for whether the article is NIH funded, publication month fixed effects, age at which an article goes live fixed effects, and journal fixed effects. The NIH's Public Access Policy (PAP), which applies to NIH articles accepted for publication after April 2008, is used as an instrument for going live on PubMed Central, and is operationalized using the interaction between an indicator for an article being NIH funded and an indicator for an article being published after the April 2008. Thus, the numbers next to "NIH × Post-PAP" are estimates of the coefficient for this interaction on citations (Panels A and B) and the probability of going live on PubMed Central (Panel C). *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels. The numbers in parentheses are standard errors, clustered at the journal-level. The numbers in square brackets are implied percent changes, which are computed as $100 * (e^{\hat{\beta}} - 1)$ for the Poisson estimates and $100 * (\hat{\beta}/\bar{y})$ for the linear estimates ($\bar{y}$ is the mean of the outcome variable).