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

Joseph Staudt

Center for Economic Studies
U.S. Census Bureau

NBER Summer Institute
Science of Science Funding
July 16, 2020

Motivation

- Open Access has become much more common over time
- Main goals of open access:
  - Speed scientific discovery
  - Give access to public, who paid for research
- But does making articles open access achieve these goals?
- This paper estimates the effects of making an article freely available on:
  - Scientists – article-to-article cites
  - Inventors – patent-to-article cites
  - Public – Wiki-to-article cites

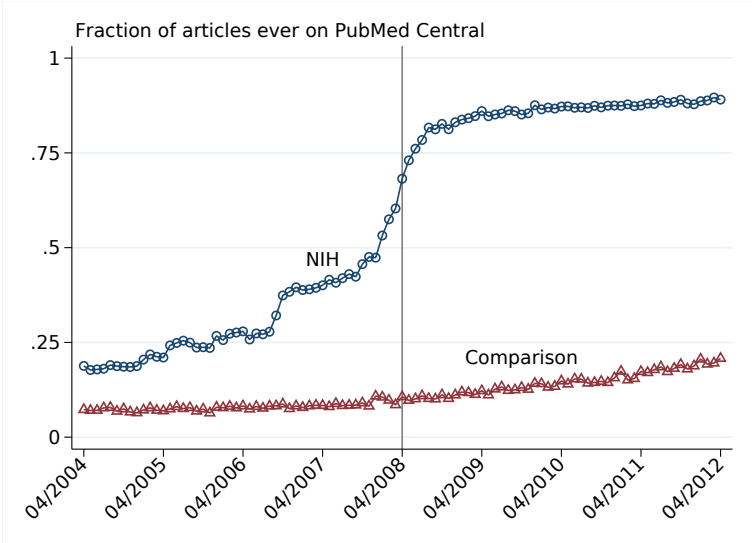
Open Access Goals

Open Access, PubMed Central, and the Public Access Policy (PAP)

> *"Open access (OA) literature is digital, online, free of charge, and free of most copyright and licensing restrictions."*
>
> — Peter Suber, Open Access

- PubMed Central is the the National Institutes of Health's (NIH) repository of open access full-text biomedical articles
- In April 2008, the NIH implemented the Public Access Policy (PAP) requiring all NIH-funded articles to be submitted to PubMed Central, *regardless of where the article is published*

Public Access Policy (PAP)



Data Sources

- Focal articles: MEDLINE
  - All 2.8 million biomedical articles published between April 2006 and April 2010.
  - Information on publication date, NIH funding, journal, and much more.
- PubMed Central Live dates: NIH API
- Article-to-article cites: Web of Science
- Patent-to-article cites: Marx and Fuegi (2020)
- Wiki-to-article cites: scraped from revision histories

Final Estimation Sample

Summary Statistics

Treatment and Control Articles

- Treated articles:
  - EVER goes live on PubMed Central
  - Compute pre- and post-outcomes using TRUE live date
- Control articles:
  - NEVER goes live on PubMed Central
  - Compute pre- and post-outcomes using PSEUDO live date

Schematic Figure

Research Design

- Diff-in-Diff
  - Obvious pre-trends, so not appropriate
- Regression adjustment
  - Assumes unconfoundedness – treatment is exogenous conditional on covariates
  - Covariates: pre-treatment outcomes, NIH indicator, publication month FEs, age at live FEs, and journal FEs
- Use NIH Public Access Policy (PAP) as IV
  - Operationalized as an interaction between an NIH dummy and a post-PAP dummy

DiD Pre-Trends | Regression Adjustment Research Design | IV Research Design | IV Assumptions | Covariate Balance

First Stages

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

Excluded Publication Months	Apr 08	Oct 07 - Oct 08
	(1)	(2)
NIH × Post-PAP	0.334***	0.327***
	(0.0260)	(0.0422)
F-Stat	165.0	60.0
Articles	958,597	612,084
Journal Clusters	4,836	4,745

All Outcome Variables

Outcome: Article-to-Article Cites After 12 Months

$$E[artci\_post_i|.] = exp[\alpha + \beta pmc_i + \gamma artci\_pre_i + \delta X_i + \rho\hat{e}_i]$$

Excl. Pub. Months	Apr 08		Oct 07 - Oct 08	
	PPML (1)	PPML-IV (2)	PPML (3)	PPML-IV (4)
PMC Live	0.046***	0.146***	0.063***	0.369***
	(0.0148)	(0.0425)	(0.0194)	(0.0715)
%Δ	[4.7]	[15.7]	[6.5]	[44.6]
Articles	958,597		612,084	
Journal Clusters	4,836		4,745	

6 Month Cites | Linear Models | Residual Coefficients

Outcome: Patent-to-Article Cites After 12 Months

$$E[patci\_post_i|.] = exp[\alpha + \beta pmc_i + \gamma patci\_pre_i + \delta X_i + \rho\hat{\epsilon}_i]$$

Excl. Pub. Months	Apr 08		Oct 07 - Oct 08	
	PPML (1)	PPML-IV (2)	PPML (3)	PPML-IV (4)
PMC Live	0.023	0.318*	0.051	0.720***
	(0.0540)	(0.116)	(0.0687)	(0.274)
%Δ	[2.3]	[37.4]	[5.2]	[105.4]
Articles	958,597		612,084	
Journal Clusters	4,836		4,745	

6 Month Cites Linear Models Residual Coefficients

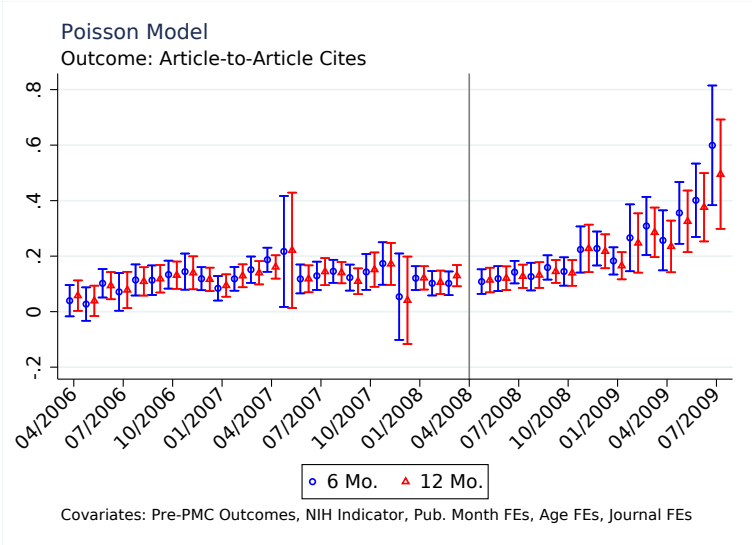
Outcome: Wiki-to-Article Cites After 12 Months

$$E[wikici\_post_i|.] = exp[\alpha + \beta pmc_i + \gamma wikici\_pre_i + \delta X_i + \rho\hat{\epsilon}_i]$$

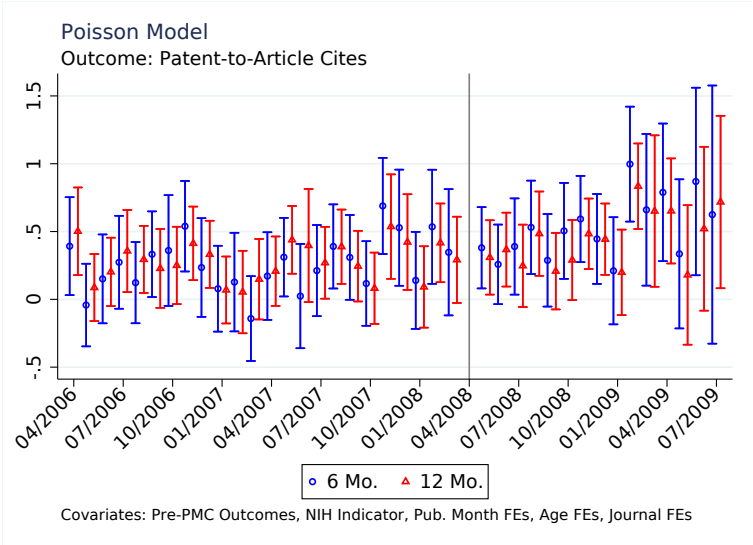
Excl. Pub. Months	Apr 08		Oct 07 - Oct 08	
	PPML (1)	PPML-IV (2)	PPML (3)	PPML-IV (4)
PMC Live	0.321**	-0.292	0.380**	-0.454
	(0.153)	(0.421)	(0.191)	(0.694)
%Δ	[37.9]	[-25.3]	[46.2]	[-36.5]
Articles	958,597		612,084	
Journal Clusters	4,836		4,745	

6 Month Cites Linear Models Residual Coefficients

Reduced Form: Article-to-Article Cites



Reduced Form: Patent-to-Article Cites



Additional Results in Paper

- Where are extra cites coming from?
  - Article-to-article cites:
    - From scientists at different institutions (e.g. university, hospital, commercial enterprise)
    - From scientists located in poor/developing/rich countries
  - Patent-to-article cites:
    - From applicants or examiners
    - Located in the front-matter or body of patent

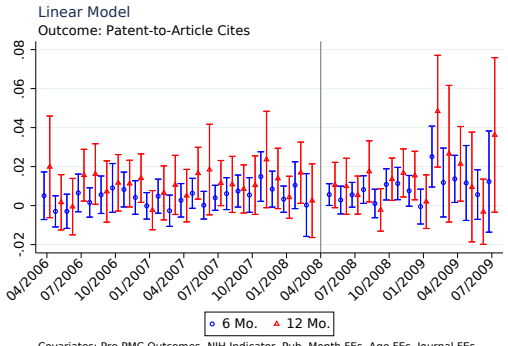
Conclusions

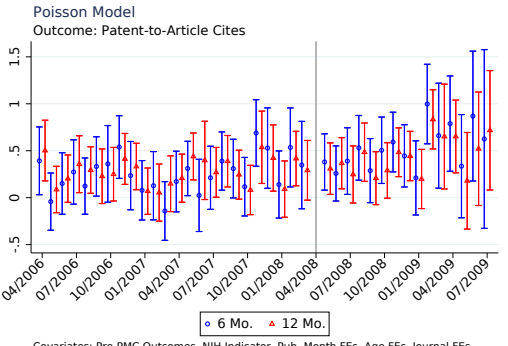
- Open access (via PubMed Central) may not increase access to the *typical* article
- Access to articles pried open by the PAP (compliers) does increase substantially
  - These articles are only on PubMed Central because of the PAP
  - Perhaps journals/publishers more zealously protect, behind a paywall, high quality articles.

Future Work

- Which articles benefit? Interact PubMed Central indicator with:
  - Pre-treatment outcomes
  - Journal quality
- Patent-to-article cites from different types of patent assignee firms:
  - Small/Large
  - Young/Old
  - High-Tech/Low-Tech

Reduced Form: Patent-to-Article Cites





Main Estimates

Goals of Public Access Policy

"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."

— Elias Zerhouni, NIH Director

Motivation

Final Estimation Sample

- 2.8 million articles published between April 2006 and April 2010
- Article-to-article cites are censored beginning in January 2011
  - Retain 1.7 million articles published before January 2010
- To control for pre-treatment outcomes
  - Retain 988,744 articles with at least 6 months between publication and live dates
- To prevent possible sorting across PAP date
  - Drop articles published in April 2008 – 958,677 left
  - Drop articles published between October 2007 and October 2008 (6 months before/after PAP) – 612,174 left

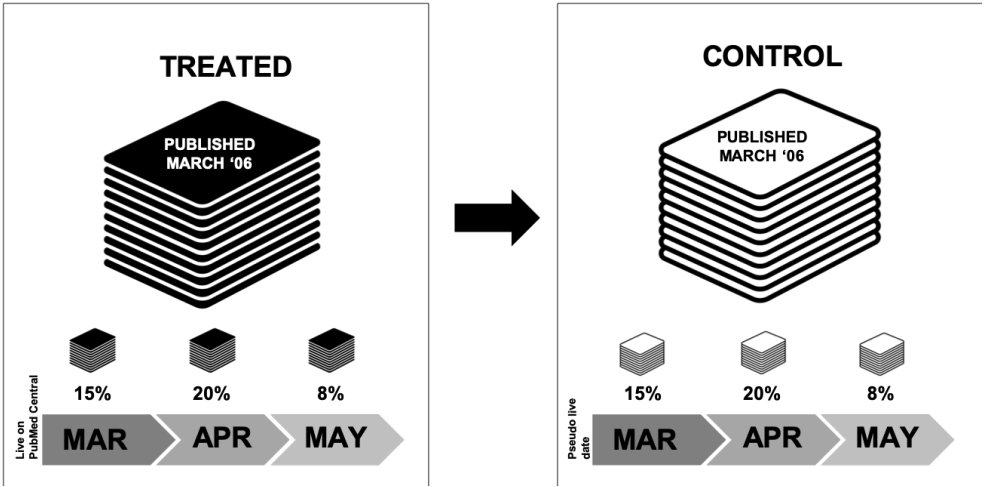
Data Sources

Summary Statistics

	All		Treated		Control	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Post-Treatment Outcomes						
6 Month Article Cites	1.247	4.22	2.388	4.193	1.053	4.200
12 Month Article Cites	2.455	8.29	4.691	7.919	2.075	8.292
6 Month Patent Cites	0.013	0.216	0.025	0.365	0.011	0.179
12 Month Patent Cites	0.026	0.354	0.050	0.518	0.022	0.318
Articles	988,744		144,269		844,475	

Data Sources

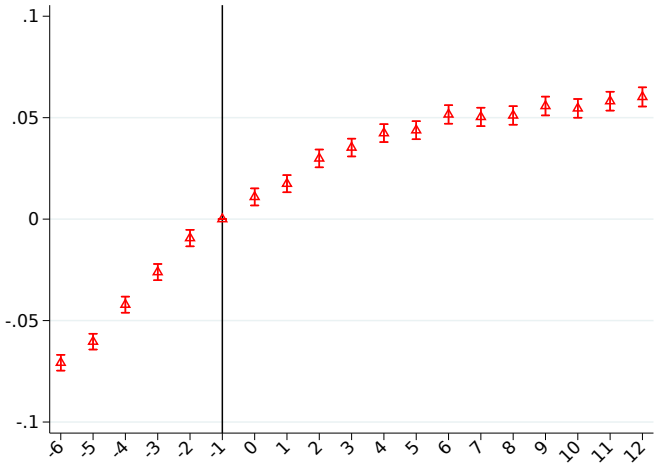
Pseudo Live Dates for Control Articles



Description Slide

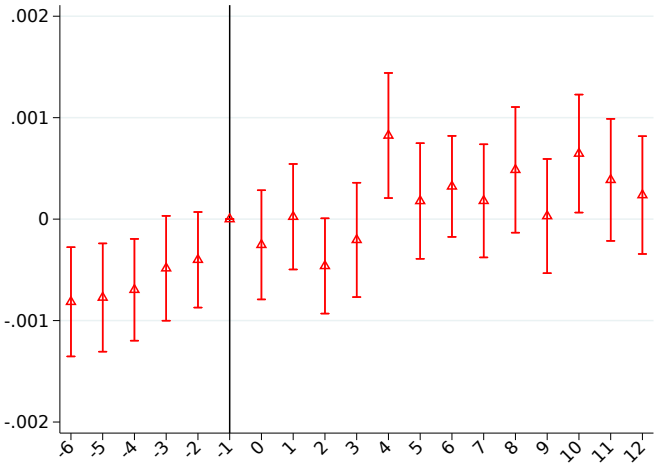
Diff-in-Diff Pre-Trends: Article-to-Article Cites

$$article\_cites_{ia} = \beta pmc_{ia} + \gamma_i + \delta_a + \epsilon_{ia}$$



Diff-in-Diff Pre-Trends: Patent-to-Article Cites

$$patent\_cites_{ia} = \beta pmc_{ia} + \gamma_i + \delta_a + \epsilon_{ia}$$



Research Design: Regression Adjustment

- Unconfoundedness: Conditional on covariates, an article going live on PubMed Central is exogenous (unrelated to potential outcomes)
- If true, we can estimate average treatment effects
- Pre-treatment outcomes proxy for underlying article quality
- Other covariates:
  - NIH dummy
  - Publication month-year FEs
  - Age at live FEs
  - Journal FEs
  - Many others possible...

Research Design

Research Design: IV using Public Access Policy (PAP)

- April 2008: Public Access Policy (PAP) goes into effect
  - NIH articles accepted after April 2008 must be submitted to PubMed Central
  - Does *not* apply to non-NIH articles or NIH articles accepted before April 2008
- Use PAP as IV for an article going live on PubMed Central
  - Operationalized as an interaction between an NIH dummy and a post-PAP dummy

Research Design

Research Design: IV using Public Access Policy (PAP)

- Three core assumptions underlying the use of the PAP as an IV for an article going live on PubMed Central:
  - First Stage: 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
  - Exogeneity: whether an article is published before or after the PAP must be unrelated to omitted variables that affect the outcomes.
  - Exclusion Restriction: 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

Research Design

Covariate Balance

	Pre-PAP Mean	Post-PAP Mean	Std. Diff.
Pre-Treatment Article Cite Outcomes			
1-6 Month Pre-PMC	0.79	0.68	-0.034
7-12 Month Pre-PMC	0.48	0.23	-0.108
Pre-Treatment Patent Cite Outcomes			
1-6 Month Pre-PMC	0.01	0.01	-0.017
7-12 Month Pre-PMC	0.01	0.00	-0.025
Other Covariates			
NIH Article	0.13	0.14	0.024
Unique N-Grams in Text	100.22	101.10	0.016
English Language	0.91	0.92	0.021
Journal Article	0.92	0.91	-0.016
Non-NIH Grants (Count)	0.03	0.06	0.069
MeSH Descriptors (Count)	11.16	11.12	-0.007
Authors (Count)	4.65	4.78	0.019

First Stages

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

Excl. Pub. Months	Apr 08			Oct 07 - Oct 08		
	Article (1)	Patent (2)	Wiki (3)	Article (4)	Patent (5)	Wiki (6)
NIH × Post-PAP	0.334***	0.337***	0.337***	0.327***	0.328***	0.328***
	(0.0260)	(0.0262)	(0.0262)	(0.0422)	(0.0423)	(0.0423)
F-Stat	165.0	165.4	165.4	60.0	60.1	60.1
Articles		958,597			612,084	
Journal Clusters		4,836			4,745	

Main Results

Outcome: Article-to-Article Cites After 6 Months

$$E[artci\_post_i|.] = exp[\alpha + \beta pmc_i + \gamma artci\_pre_i + \delta X_i + \rho \hat{e}_i]$$

Excl. Pub. Months	Apr 08		Oct 07 - Oct 08	
	PPML (1)	PPML-IV (2)	PPML (3)	PPML-IV (4)
PMC Live	0.049***	0.166***	0.0643***	0.407***
	(0.0152)	(0.0435)	(0.0200)	(0.0752)
%Δ	[5.0]	[18.1]	[6.6]	[50.2]
Residual		-0.126***		-0.362***
		(0.0416)		(0.0726)
Articles	958,597		612,084	
Journal Clusters	4,836		4,745	

Main Estimates

Outcome: Article-to-Article Cites After 12 Months

$$E[artci\_post_i|.] = exp[\alpha + \beta pmc_i + \gamma artci\_pre_i + \delta X_i + \rho \hat{e}_i]$$

Excl. Pub. Months	Apr 08		Oct 07 - Oct 08	
	PPML (1)	PPML-IV (2)	PPML (3)	PPML-IV (4)
PMC Live	0.046***	0.146***	0.063***	0.369***
	(0.0148)	(0.0425)	(0.0194)	(0.0715)
%Δ	[4.7]	[15.7]	[6.5]	[44.6]
Residual		-0.108***		-0.323***
		(0.0408)		(0.0682)
Articles	958,597		612,084	
Journal Clusters	4,836		4,745	

Main Estimates

Outcome: Article-to-Article Cites After 12 Months

$$artci\_post_i = \alpha + \beta pmc_i + \gamma artci\_pre_i + \delta X_i + \rho \hat{\epsilon}_i + e_i$$

Excl. Pub. Months	Apr 08		Oct 07 - Oct 08	
	OLS (1)	2SLS (2)	OLS (3)	2SLS (4)
PMC Live	0.231**	1.237***	0.324**	0.2.925***
	(0.102)	(0.282)	(0.134)	(0.603)
Articles	958,597		612,084	
Journal Clusters	4,836		4,745	

Main Estimates

Outcome: Article-to-Article Cites After 6 Months

$$artci\_post_i = \alpha + \beta pmc_i + \gamma artci\_pre_i + \delta X_i + \rho \hat{\epsilon}_i + e_i$$

Excl. Pub. Months	Apr 08		Oct 07 - Oct 08	
	OLS (1)	2SLS (2)	OLS (3)	2SLS (4)
PMC Live	0.133**	0.642***	0.172**	0.1.448***
	(0.0519)	(0.138)	(0.0689)	(0.293)
Articles	958,597		612,084	
Journal Clusters	4,836		4,745	

Main Estimates

Outcome: Patent-to-Article Cites After 6 Months

$$E[patci\_post_i|.] = exp[\alpha + \beta pmc_i + \gamma patci\_pre_i + \delta X_i + \rho\hat{\epsilon}_i]$$

Excl. Pub. Months	Apr 08		Oct 07 - Oct 08	
	PPML (1)	PPML-IV (2)	PPML (3)	PPML-IV (4)
PMC Live	0.0499	0.533***	0.0951	1.059***
	(0.0611)	(0.191)	(0.0780)	(0.287)
%Δ	[5.1]	[70.4]	[10.0]	[188.3]
Residual		-0.516***		-1.004***
		(0.189)		(0.282)
Articles	958,597		612,084	
Journal Clusters	4,836		4,745	

Main Estimates

Outcome: Patent-to-Article Cites After 12 Months

$$E[patci\_post_i|.] = exp[\alpha + \beta pmc_i + \gamma patci\_pre_i + \delta X_i + \rho\hat{\epsilon}_i]$$

Excl. Pub. Months	Apr 08		Oct 07 - Oct 08	
	PPML (1)	PPML-IV (2)	PPML (3)	PPML-IV (4)
PMC Live	0.023	0.318*	0.051	0.720***
	(0.0540)	(0.116)	(0.0687)	(0.274)
%Δ	[2.3]	[37.4]	[5.2]	[105.4]
Residual		-0.316*		-0.696***
		(0.164)		(0.262)
Articles	958,597		612,084	
Journal Clusters	4,836		4,745	

Main Estimates

Outcome: Patent-to-Article Cites After 12 Months

$$patci\_post_i = \alpha + \beta pmc_i + \gamma patci\_pre_i + \delta X_i + \rho \hat{e}_i + e_i$$

Excl. Pub. Months	Apr 08		Oct 07 - Oct 08	
	OLS (1)	2SLS (2)	OLS (3)	2SLS (4)
PMC Live	0.00221	0.00444	0.00319	0.0192
	(0.00239)	(0.00836)	(0.00323)	(0.0138)
Articles	958,597		612,084	
Journal Clusters	4,836		4,745	

Main Estimates

Outcome: Patent-to-Article Cites After 6 Months

$$patci\_post_i = \alpha + \beta pmc_i + \gamma patci\_pre_i + \delta X_i + \rho \hat{\epsilon}_i + e_i$$

Excl. Pub. Months	Apr 08		Oct 07 - Oct 08	
	OLS (1)	2SLS (2)	OLS (3)	2SLS (4)
PMC Live	0.00205	0.00711	0.00259	0.0164**
	(0.00144)	(0.00525)	(0.00186)	(0.00789)
Articles	958,597		612,084	
Journal Clusters	4,836		4,745	

Main Estimates

Outcome: Wiki-to-Article Cites After 6 Months

$$E[wikici\_post_i|.] = exp[\alpha + \beta pmc_i + \gamma wikici\_pre_i + \delta X_i + \rho \hat{e}_i]$$

Excl. Pub. Months	Apr 08		Oct 07 - Oct 08	
	PPML (1)	PPML-IV (2)	PPML (3)	PPML-IV (4)
PMC Live	0.372*	-0.0831	0.151	-0.927
	(0.215)	(0.534)	(0.285)	(0.908)
%Δ	[45.1]	[-7.8]	[16.3]	[-60.4]
Residual		0.490		1.137
		(0.592)		(1.065)
Articles	958,597		612,084	
Journal Clusters	4,836		4,745	

Main Estimates

Outcome: Wiki-to-Article Cites After 12 Months

$$E[wikici\_post_i|.] = exp[\alpha + \beta pmc_i + \gamma wikici\_pre_i + \delta X_i + \rho \hat{e}_i]$$

Excl. Pub. Months	Apr 08		Oct 07 - Oct 08	
	PPML (1)	PPML-IV (2)	PPML (3)	PPML-IV (4)
PMC Live	0.321**	-0.292	0.380**	-0.454
	(0.153)	(0.421)	(0.191)	(0.694)
%Δ	[37.9]	[-25.3]	[46.2]	[-36.5]
Residual		0.658		0.871
		(0.467)		(0.770)
Articles	958,597		612,084	
Journal Clusters	4,836		4,745	

Main Estimates

Outcome: Wiki-to-Article Cites After 12 Months

$$wikici\_post_i = \alpha + \beta pmc_i + \gamma wikici\_pre_i + \delta X_i + \rho \hat{\epsilon}_i + e_i$$

Excl. Pub. Months	Apr 08		Oct 07 - Oct 08	
	OLS	2SLS	OLS	2SLS
	(1)	(2)	(3)	(4)
PMC Live	0.00119**	-0.00108	0.00158**	-0.00310
	(0.000498)	(0.00149)	(0.000693)	(0.00257)
Articles	958,597		612,084	
Journal Clusters	4,836		4,745	

Main Estimates

Outcome: Wiki-to-Article Cites After 6 Months

$$wikici\_post_i = \alpha + \beta pmc_i + \gamma wikici\_pre_i + \delta X_i + \rho \hat{\epsilon}_i + e_i$$

Excl. Pub. Months	Apr 08		Oct 07 - Oct 08	
	OLS	2SLS	OLS	2SLS
	(1)	(2)	(3)	(4)
PMC Live	0.000799**	-0.000303	0.000511	-0.00244*
	(0.000378)	(0.000939)	(0.000529)	(0.00147)
Articles	958,597		612,084	
Journal Clusters	4,836		4,745	

Main Estimates