# Paywalls Constrain the Translation of Science into Technology

Paul-Emmanuel Anckaert[1]
SKEMA Business School - Université Cote d'Azur (GREDEG)
paul-emmanuel.anckaert@skema.edu

Sam Arts
Department of Management, Strategy and Innovation
Faculty of Economics and Business, KU Leuven
sam.arts@kuleuven.be

*Work in Progress – Do Not Distribute*

## ABSTRACT

For scientific discoveries to spur technological progress, inventors must be able to access and use them. Yet about 40% of scientific papers published in recent years remain behind paywalls. Using linked data on the near-universe of scientific publications since 2000 and the global patent record, we show that open access substantially increases inventors' use of science. Open-access papers are 7–19% more likely than comparable closed-access papers to be cited by patents, including within pairs of near-identical discoveries published contemporaneously by independent research teams. These effects are strongest for startups, small firms, and geographically distant inventors, consistent with paywalls disproportionately constraining those with more limited access to scientific knowledge. Together, the findings indicate that paywalls reduce and skew the translation of science into technology.

[1] Corresponding author. SKEMA Business School – Avenue Willy Brandt, 59777 Euralille, France, +33 3 59 82 22 38.

As famously argued by Vannevar Bush in his 1945 report to the U.S. President, public investment in scientific research is a central engine of technological progress and economic growth—a vision that led to the creation of the National Science Foundation (*1*). This vision rests on a critical assumption: that scientific discoveries are broadly accessible to those who build upon them to develop new technologies. When access is restricted, the innovation system may slow and become uneven, privileging actors with superior access to scientific knowledge.

Motivated by the belief that open access accelerates the diffusion of scientific ideas, broadens their use, and ultimately enhances their societal and economic impact, governments, research funders, and philanthropic organizations have launched major open-access initiatives, including Plan S and public-access mandates by agencies such as the U.S. National Institutes of Health. Yet, despite two decades of such efforts, a substantial share of the entire scientific record remains closed access and behind paywalls, limiting who can access and reuse frontier knowledge.

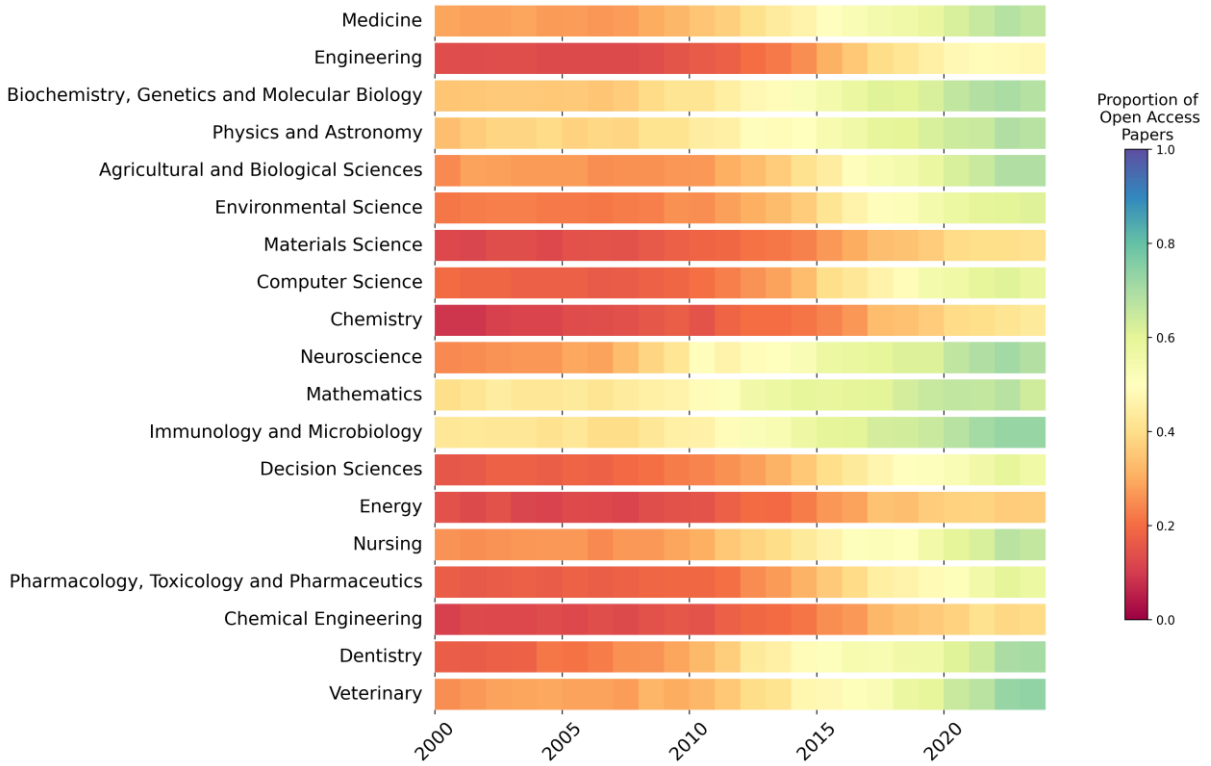
A large literature has examined whether open access increases scientific impact, typically measured by citations from scientific articles. The evidence is mixed, with estimates ranging from negligible to modest effects (*2*). However, scientific citations capture only one dimension of impact. For economic growth and societal welfare, a central outcome is whether scientific ideas are incorporated into new technologies (*3*). This distinction matters because the users of scientific knowledge differ sharply between academic and technological context. Academic scientists typically have institutional access to subscription journals, whereas inventors—often working in established firms, startups, or independent settings—frequently face higher barriers to accessing paywalled research. As a result, open access may matter relatively little for scientific impact, while being critical for the incorporation of scientific ideas into technological inventions. Prior work suggests that open access can increase the technological use of science, but existing studies focus on specific open-access mandates in selected scientific fields and short time windows (*4*, *5*), leaving open the question of whether open access systematically promotes technological progress across fields and over time.

In this paper, we first document the evolution of open-access publishing using the near-universe of scientific articles and conference proceedings published between 2000 and 2023 from OpenAlex—44 million articles across more than 44,000 journals. Next, we provide large-scale causal evidence on whether restricted access to science hinders innovation by linking the subset of 24 million articles published between 2000 and 2015 to the global patent record and examining the incorporation of scientific ideas into patented inventions. We measure technological impact in two complementary ways: through citations from patents to scientific articles and through direct traces of scientific ideas in patent texts. To identify the causal effect of open access, we use complementary research designs, including comparisons of near-identical scientific discoveries published by independent teams at roughly the same time, where only one article is open access. This approach separates access effects from confounding differences in scientific content and underlying technological potential. Finally, we examine who stands to benefit most from open access by analyzing how access affects the diffusion of scientific ideas across different types of organizations and inventors.

## A large share of the scientific record remains closed access

Figure 1 documents the evolution of open-access publishing across scientific fields by year of publication. While the share of open-access articles has increased substantially over time, from 22% in 2000 to 60% in 2023, a large fraction of the entire scientific record remains closed access—approximately 41% of all papers published between 2000 and 2023. Openness also varies widely across scientific fields. Averaged over time, certain fields, including mathematics and immunology and microbiology, display high levels of openness (56% and 55%, respectively), while others, including chemistry and materials science, remain substantially less open (23% and 27%). Trends also differ sharply within fields: some have consistently been open (mathematics), others have remained persistently closed (chemical engineering), and still others have undergone a major shift from closed to open access (such as agricultural and biological sciences).

**Figure 1: A large share of papers remains closed access**




*Notes:* Figure 1 shows the share of open-access papers by field of study (OpenAlex classification) over the period 2000-2023 ($n \approx 44$ million articles). Fields are top down ordered by publication volume. Open access status is derived from OpenAlex, which classifies articles as open access if they are currently freely available online without subscription or paywall, including both publisher-provided and repository-based open access. Following fields were excluded due to their limited relevance for technology development activity: Economics, Econometrics and Finance; Social Sciences; Arts and Humanities; Health Professions; Business, Management and Accounting; Earth and Planetary Sciences; and Psychology.

These temporal patterns align with major open-access policy initiatives. Biomedical fields such as immunology and microbiology, biochemistry, genetics and molecular biology, exhibit earlier increases in openness following the 2008 U.S. National Institutes of Health public-access mandate. A broad acceleration across nearly all fields after 2015 coincides with the launch of Plan S, a coordinated open-access mandate by major European research funders,

and subsequent national policies such as the 2020 U.S. directive requiring immediate public access to federally funded research.

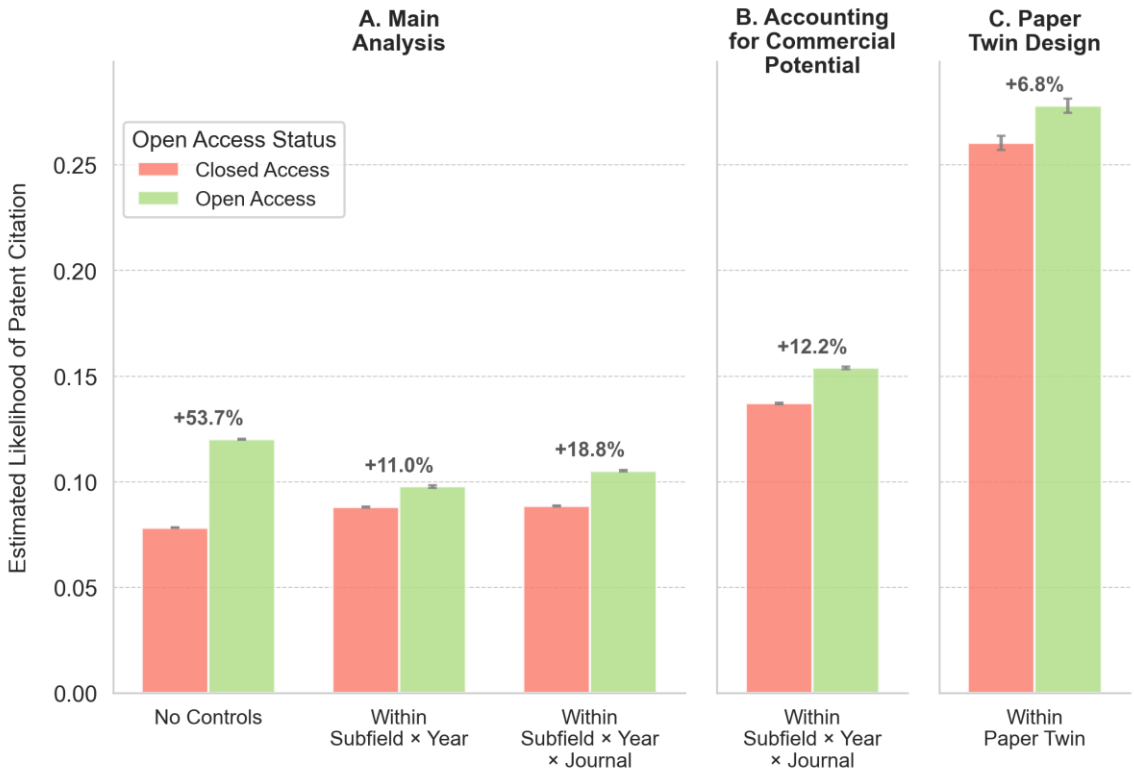
Yet, despite these initiatives, a substantial share of scientific knowledge remains behind paywalls: even in the most recent years, roughly 40% of newly published papers are not openly accessible. As a result, many potential users of scientific knowledge continue to face barriers to accessing frontier research.

**Closed access causally reduces the use of science in technology**

We next examine whether access restrictions affect the incorporation of scientific knowledge into patented inventions. Figure 2 summarizes our results and shows that open-access papers are substantially more likely to be cited by global patents than closed-access papers. Across specifications, open access is associated with a 6.8–18.8% higher probability of patent citation.

To isolate the causal effect of open access from confounding differences in scientific content, technological potential, and selective open-access publishing, we employ three complementary strategies. Detailed descriptions of the data sources and empirical methods are provided in the Supplementary Materials. First, in our main analyses, we compare open- and closed-access articles published in the same year and scientific subfield, then restrict the comparison set further to papers in the same journal, from the same author's institution, and ultimately by the same author (Table S2). The estimated effect of open access remains stable across these increasingly stringent specifications, indicating that the results are not driven by selective open-access publishing in particular scientific subfields or years, by more productive scientists, higher-quality journals, or better-resourced institutions. By showing that this effect is most pronounced for those prior art citations that more closely reflect inventors' actual use of scientific knowledge (i.e., "in-text" citations), we provide evidence that the observed effects can indeed be ascribed to inventor behavior (Fig. S3 and Table S7). Second, we separate open-access effects from inherent technological potential by controlling for each paper's predicted probability of future patent citation. This prediction is based solely on information available at the time of publication and is estimated using a machine-learning model trained on article abstracts (*6*). Conditioning on this measure leaves the open-access effect largely unchanged, indicating that access restrictions operate independently of the intrinsic technological potential of scientific ideas (Table S3). Third, we implement a paper-twins design that compares pairs of near-identical scientific discoveries published contemporaneously by independent research teams, of which only one is open access (*7*). By holding scientific content and intrinsic technological potential nearly constant, this quasi-experimental design strengthens the causal interpretation that restricted access to science reduces its incorporation into technology (Table S5).

**Figure 2: Open-access papers are more likely to be used in patented inventions**

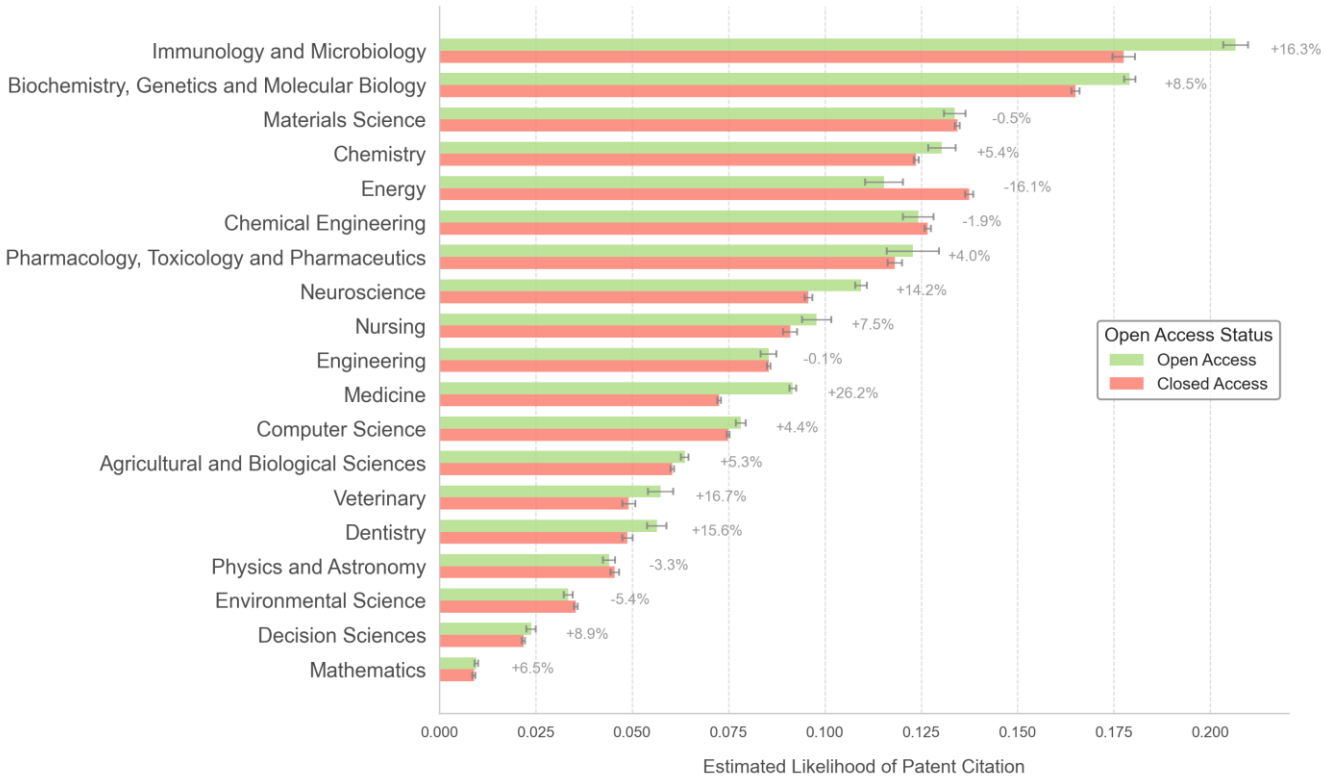



*Notes:* Figure 2 presents the estimated likelihood of patent citation for open- and closed-access papers, obtained using linear probability models. Panel (A) reports results from the main analysis with progressively added fixed effects ($n \approx 24.4$ million articles). Panel (B) reports results for the subsample of papers for which measures of commercial and scientific potential are available ($n \approx 10.4$ million articles). Panel (C) presents estimates from a paper-twin design that compares near-identical scientific discoveries, of which only one is open access ($n \approx 40{,}270$ articles). All specifications – unless explicitly stated otherwise – control for paper characteristics (i.e., references, conference proceeding status), journal commercial impact factor, and author team characteristics (i.e., size, prior publications, prior paper citations, and prior patent citations). Confidence intervals are shown at the 95% level.

In the Supplementary Materials, we show that open-access articles are not only more likely to be cited by patents relative to paywalled publications, but also yield a larger number of patent citations (Table S6) and are cited earlier in time, suggesting that open access accelerates the diffusion of scientific knowledge into patented inventions (Fig. S4 and Table S11). Moreover, conditional on being cited by a patent, open-access papers are semantically more similar to the citing patents than closed-access papers, indicating that open access improves inventors' use of scientific knowledge (Fig. S5 and Table S12). We further demonstrate that our results are robust when measuring the use of science in technology not only through patent citations, but also through direct traces of scientific ideas in patent texts, captured as the reuse of scientific concepts first introduced in the academic literature (e.g., "electroporation" or "monoclonal antibody") (*8*) (Table S10). Finally, we show that the positive effects of open access hold for both publisher-provided and repository-based open access (Table S9), and when excluding articles that may not have been open access at the time of publication (Table S8).

Figure 3 shows that the positive effect of open access on technological progress extends across a broad range of scientific fields. Fields are ordered by their baseline relevance for technological development, measured by the likelihood that their papers are cited by patents, with immunology and microbiology at the top and mathematics at the bottom. Open-access papers are markedly more likely to be incorporated into technology in fields such as medicine and immunology and microbiology. In contrast, estimated effects are smaller and in some cases statistically indistinguishable from zero in fields such as materials science and chemical engineering, which nonetheless play central roles in technological development. Together, these patterns suggest that access restrictions may shape the translation of science into technology as a broader feature of the innovation system, rather than solely a field-specific phenomenon.

**Figure 3: Open-access papers are more likely to be used in patented inventions in most scientific fields**



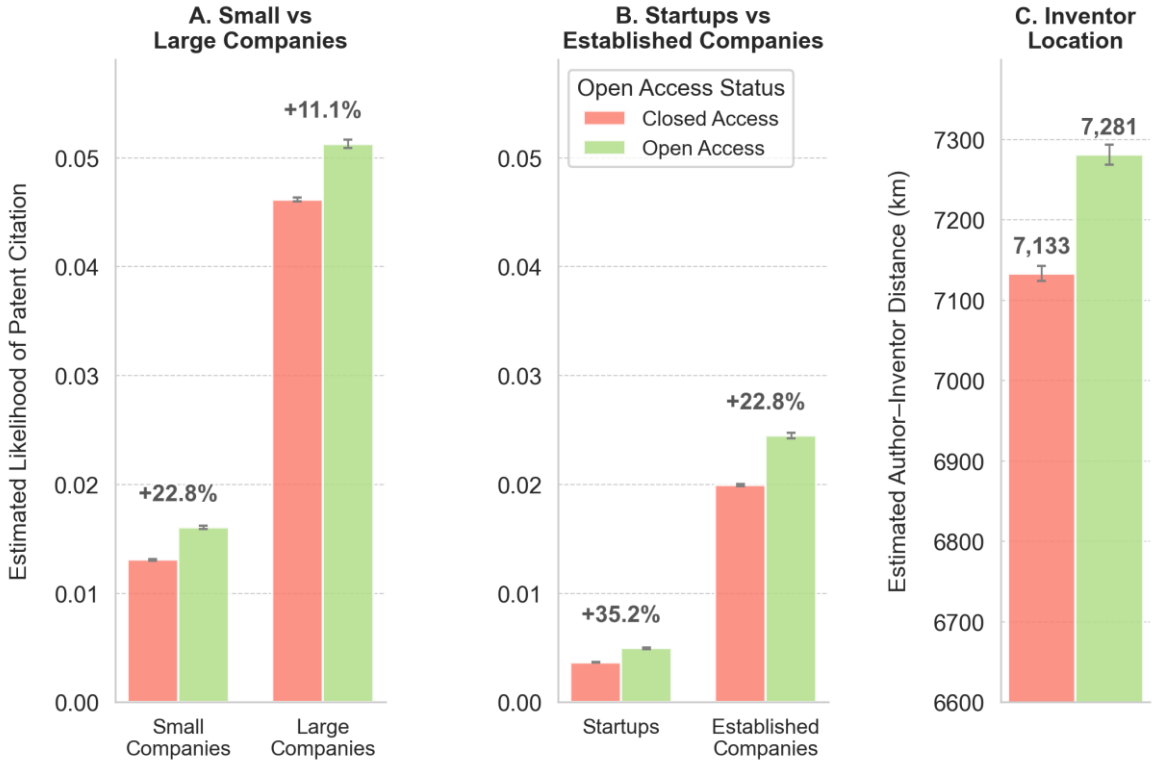

*Notes:* Figure 3 shows the estimated likelihood of patent citation for open- and closed-access papers by field of study (OpenAlex classification) over the period 2000–2015 (n ≈ 24 million articles), based on linear probability models. Fields are ordered from top to bottom by their baseline relevance for technological development. The specification includes subfield-by-year fixed effects and controls for paper characteristics (e.g., number of references, conference proceeding status), journal commercial impact factor, and author team characteristics (e.g., size, prior publications, prior paper citations, and prior patent citations). Confidence intervals are shown at the 95% level. Fig. S1 presents the same estimates without additional controls or fixed effects, while Fig. S2 reports the estimates using subfield-by-year-by-journal fixed effects.

## Access barriers shape who benefits from scientific knowledge

If access constraints drive the observed effects, the impact of open access should be largest for actors facing the greatest barriers to paywalled science. Figure 4 supports this prediction. While open access significantly benefits all firm types and inventors, the magnitude of the

effect varies systematically with access constraints. The open-access effect is more than twice as large for small firms than for large firms (23% versus 11%) and about 50% larger for startups than for established firms (35% versus 23%), consistent with smaller and younger firms facing greater barriers to accessing paywalled publications (Table S13 and Table S14).

**Figure 4: Open-access papers are disproportionately more likely to be used in patented inventions by small firms, startups, and geographically distant inventors**




*Notes:* In Figure 4, Panel (A) and Panel (B) present the estimated likelihood of patent citation for open- and closed-access papers, derived from linear probability models that distinguish between small and large firms, and between startups and established firms, respectively. Panel (C) reports the estimated geographic distance (in kilometers) between a paper's last author and the most distant inventor citing that paper, based on an OLS specification. In Panel (A), firms are classified annually as small or large based on the size of their patent portfolios over the preceding three years relative to the year-specific median. In Panel (B), firms are categorized as startups or established firms based on their founding year; firms are considered startups during their first three years of operation. Founding year information is obtained from the startup dataset developed by Ewens and Marx(*9*), which covers U.S.-based patent assignees. Accordingly, the analysis in Panel (B) is restricted to patent citations originating from U.S.-based firms. The specifications in Panels (A) and (B) include year-by-scientific-subfield fixed effects, while the specification in Panel (C) additionally includes last-author country fixed effects. All models control for paper characteristics, journal commercial impact factor, and author team characteristics. Confidence intervals are reported at the 95% level.

Alongside formal access to scientific publications, inventors can also obtain scientific knowledge through informal channels such as personal contacts and professional networks. Because personal and professional ties are often geographically constrained, inventors located farther from originating scientists face greater barriers to accessing scientific knowledge through these informal channels. Open access should therefore level the playing field and disproportionately benefit distant inventors. Consistent with this prediction, patents citing open-access papers originate from inventors located significantly farther from the originating

scientists than patents citing closed-access papers. Open access thus expands not only the likelihood that science is used in technology, but also its geographic reach (Table S15).

## Discussion

Our findings change how we understand the role of open access in technological progress. Open access not only increases the visibility of science, but causally accelerates and broadens the translation of scientific ideas into technology. By contrast, restricted access acts as a bottleneck that slows technological progress and skews innovation toward actors and locations with privileged access to scientific knowledge, disproportionately constraining startups, small firms, and geographically distant inventors.

From a policy perspective, these results imply that open-access policies influence not only scientific progress but also the rate, direction, and geographic reach of technological progress. When scientific knowledge is locked behind paywalls, the costs are borne not only by scientists but by society at large through slower and less inclusive technological development. By lowering barriers to the use of scientific ideas, open access helps realize the foundational vision articulated by Vannevar Bush—a system of publicly funded science that fuels innovation and economic growth well beyond the walls of academia.

## References


1. V. Bush, "Science, the Endless Frontier: A Report to the President" (Washington, DC, 1945).
2. A. Langham-Putrow, C. Bakker, A. Riegelman, Is the open access citation advantage real? A systematic review of the citation of open access and subscription-based articles. *PLoS One* **16** (2021).
3. M. Ahmadpoor, B. F. Jones, The dual frontier: Patented inventions and prior scientific advance. *Science (1979).* **357**, 583–587 (2017).
4. K. A. Bryan, Y. Ozcan, The impact of open access mandates on invention. *Review of Economics and Statistics* **103**, 954–967 (2021).
5. B. Probst, P. M. Lohmann, A. Kontoleon, L. D. Anadón, The impact of open access mandates on scientific research and technological development in the U.S. *iScience* **26** (2023).
6. R. Masclans, S. Hasan, W. M. Cohen, Measuring the Commercial Potential of Science. *Strategic Management Journal* **46**, 2199–2236 (2025).
7. R. K. Merton, Priorities in Scientific Discovery: A Chapter in the Sociology of Science. *Am. Sociol. Rev.* **22** (1957).
8. S. Arts, N. Melluso, "Reliance on scientific ideas in patenting" (2025); https://www.nber.org/conferences/innovation-information-initiative-technical-working-group-meeting-fall-2025.
9. M. Ewens, M. Marx, Firm Age and Invention: An Open Access Dataset. [Preprint] (2023).
10. J. Priem, H. Piwowar, R. Orr, OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. (2022).
11. Z. Lin, Y. Yin, L. Liu, D. Wang, SciSciNet: A large-scale open data lake for the science of science research. *Sci. Data* **10**, 1–22 (2023).
12. M. Bikard, Made in Academia: The Effect of Institutional Origin on Inventors' Attention to Science. *Organization Science* **29**, 818–836 (2018).

13. M. Marx, D. H. Hsu, Revisiting the Entrepreneurial Commercialization of Academic Science: Evidence from "Twin" Discoveries. *Manage. Sci.* **68** (2022).
14. R. Sato, M. Yamada, H. Kashima, "Twin Papers : A Simple Framework of Causal Inference for Citations via Coupling" in *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17-21, 2022, Atlanta, GA, USA* (Association for Computing Machinery, 2022) vol. 1.
15. M. Bikard, Idea twins: Simultaneous discoveries as a research tool. *Strategic Management Journal* **41**, 1528–1543 (2020).
16. G. de Rassenfosse, J. Kozak, F. Seliger, Geocoding of worldwide patent data. *Sci. Data* **6** (2019).
17. M. Marx, A. Fuegi, Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal* **41**, 1572–1594 (2020).
18. J. Callaert, M. Pellens, B. Van Looy, Sources of inspiration? Making sense of scientific references in patents. *Scientometrics* **98**, 1617–1629 (2014).
19. M. Roach, W. M. Cohen, Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research. *Manage. Sci.* **59**, 504–525 (2013).
20. S. Arts, N. Melluso, R. Veugelers, Beyond Citations: Measuring Novel Scientific Ideas and their Impact in Publication Text. *Review of Economics and Statistics*, 1–33 (2025).
21. A. Cohan, S. Feldman, I. Beltagy, D. Downey, D. S. Weld, "SPECTER: Document-level representation learning using citation-informed transformers" in *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (2020).
22. S. Erhardt, M. Ghosh, E. Buunk, M. E. Rose, D. Harhoff, Logic Mill -- A Knowledge Navigation System. 25–35 (2022).
23. K. A. Bryan, Y. Ozcan, B. Sampat, In-text patent citations: A user's guide. *Res. Policy* **49** (2020).
24. M. Bikard, M. Marx, Bridging academia and industry: How geographic hubs connect university science and corporate technology. *Manage. Sci.* **66**, 3425–3443 (2020).

# Supplementary Materials for

## Paywalls Constrain the Translation of Science into Technology

### Materials and Methods

Data Sources

*Publication Data*

**OpenAlex:** All publication data are drawn from OpenAlex, an open bibliographic database covering the largest corpus of scientific work currently available (*10*). OpenAlex provides detailed publication-level metadata, including titles and abstracts, publication dates, open-access status, journal and article type, author names and identifiers, institutional affiliations and locations, reference lists, and field and subfield classifications.

**SciSciNet (v2):** To assess the scientific quality of the author team listed on a paper (see *Materials and Methods: Key Variables and Measures*), we use citation data on authors' prior scientific work. Specifically, we combine OpenAlex with citation information from SciSciNet v2, a large-scale dataset built on the OpenAlex snapshot that provides precomputed publication-level measures, including detailed citation trajectories (*11*).

**Technological and Scientific Potential:** In a subset of analyses, we further separate the effect of open access from differences in papers' underlying technological and scientific potential by controlling for each paper's *ex ante* technological and scientific potential, using the measures developed by Masclans et al. (*6*). These measures predict, respectively, the probability of future patent citation and of receiving above-median academic citations, using only information available at the time of publication and is estimated with a machine-learning model trained on article abstracts. They are available for more than 30 million university-affiliated articles published between 2000 and 2020 in the applied and natural sciences and engineering.

**Paper Twins:** As a complementary identification strategy, we exploit cases of simultaneous scientific discovery to further separate the effect of open access from differences in scientific content, technological potential, and selective open-access publishing. Specifically, we study "paper twins": near-identical findings published contemporaneously by independent research teams, as described by Merton (*7*) and used in prior work as an identification strategy (*12*, *13*). Identification comes from comparisons within paper-twin pairs in which one paper is open access and the other is closed access, allowing us to hold the underlying scientific discovery—and thus its technological potential—largely constant. We assemble these cases by combining three open-access datasets of paper twins identified from citation patterns, as developed by (i) Bikard, (ii) Marx and Hsu, and (iii) Sato et al. (*13–15*).

*Patent Data*

**PATSTAT:** All patent data are drawn from PATSTAT (Spring 2024 edition), the worldwide patent statistical database maintained by the European Patent Office (EPO). PATSTAT provides harmonized information on patent applications filed across major patent offices worldwide, including data on inventors and applicants, their addresses, filing and publication dates, patent families, technological classifications, and citation links.

**Startup Patenting:** To distinguish patents filed by startups from those filed by established firms, we use a recently developed startup patenting dataset by Ewens and Marx (*9*). This dataset reports the founding year of U.S.-based patent assignees for about 85% of patents filed between 1976 and 2021 and is limited to USPTO patents.

**Geolocation of Inventors:** To geolocate inventors listed on patents, we draw on two complementary data sources. First, we use PatentsView, a public database maintained by the U.S. Patent and Trademark Office (USPTO) that provides geolocation data for inventors based on address information reported in patent documents. Second, we use the dataset developed by de Rassenfosse et al. (*16*), which provides geographic coordinates for inventor locations in 18.8 million patent documents worldwide over more than 30 years. For patents not covered by either source, but for which inventor address information is available in PATSTAT, we geolocate inventors at the city level using the reported addresses.

*Paper-Patent Link*

**Patent-to-Paper Citations:** Following prior research, we use patent-to-paper citations as the traditional large-scale proxy for whether scientific papers are used as prior art in the development of patented technology. We draw these citations from the open-access dataset developed by Marx and Fuegi (*17*), which covers citations through 2023. The data include references from patents granted worldwide to scientific papers published since 1800. For each citation, the dataset identifies whether it originates from the patent applicant or the examiner, and whether it appears on the front page of the patent, in the body text, or in both locations.

**Reuse of Scientific Ideas:** As a complementary robustness check, we examine whether scientific ideas from papers later reappear in patent text. Our main analyses rely on patent-to-paper citations as a proxy for whether scientific papers are used as prior art in the development of patented technology. Prior inventor- and firm-level survey evidence, however, suggests that patents may build on scientific ideas without citing the relevant papers (*18*, *19*). We therefore draw on Arts et al. (*20*), who develop and validate a text-based approach to identify new scientific ideas in publication text, and use it to trace whether scientific ideas introduced in papers are later reused in the full text of U.S. patents (*8*). This measure complements formal patent citations by capturing the use of scientific ideas even when the patent does not cite the original paper.

Sample Construction – Main Analysis

We start from the January 2024 snapshot of OpenAlex, as processed and cleaned by Arts et al. (*20*), which contains 75,295,921 papers and conference proceedings published between 1901 and 2023. For simplicity, we refer to both document types as scientific articles or papers throughout the manuscript. We restrict the sample to articles published from 2000 onward, which coincides with the emergence and expansion of modern open-access publishing models and provides a consistent setting for studying variation in open-access status. Given our focus on the use of science in technology, we further exclude fields with limited relevance for technological development—Economics, Econometrics and Finance; Social Sciences; Arts and Humanities; Health Professions; Business, Management and Accounting; Earth and Planetary Sciences; and Psychology—based on the OpenAlex classification. After applying these restrictions, the baseline sample contains 44,431,576 articles. For analyses of the use of science in technology, we further restrict the sample to articles published by 2015 to allow a sufficiently long follow-up window for downstream technological use, yielding a final sample of 24,786,337 articles. In analyses that incorporate technological and scientific potential controls, the sample is further limited to articles for which these data are available, resulting in a subsample of 10,424,832 papers.

Sample Construction – Paper Twin Analysis

As mentioned in the *Materials and Methods: Data Sources* section, we combine three existing open-access datasets of paper twins: Bikard (*15*), with 10,927 paper pairs; Marx and Hsu (*13*), with 23,851 pairs; and Sato et al. (*14*), with 87,396 pairs. We remove duplicate pairs, pairs that cannot be matched to OpenAlex records, and pairs in which the two papers were published more than one calendar year apart. We further exclude pairs in fields with limited relevance for technological development, using the same criteria as in the main analysis: Economics, Econometrics and Finance; Social Sciences; Arts and Humanities; Health Professions; Business, Management and Accounting; Earth and Planetary Sciences; and Psychology. Finally, we restrict the sample to pairs with within-pair variation in open-access status, such that identification comes from comparing twin papers in which one paper is open access and the other is closed access. The resulting sample consists of 20,135 unique paper-pairs.

For robustness analyses, we further restrict the paper-twin sample to pairs with especially high semantic similarity in their titles and abstracts. Specifically, we represent the scientific content of each paper using PatSPECTER, a pre-trained document-level embedding model for scientific text, applied to the unprocessed title and abstract of each publication (*21*, *22*). This yields a 768-dimensional vector for each paper, which we use to compute pairwise cosine similarity within each twin pair. Applying a similarity threshold of 0.95, this yields a subsample of 7,236 pairs.

Key Variables and Measures

*Open-Access Status*

Open-access status is obtained from OpenAlex, which classifies articles as open access if they are freely available online without a subscription or paywall (see: https://help.openalex.org/hc/en-us/articles/24347035046295-Open-Access-OA). Accordingly, *Open Access* is coded as an indicator equal to 1 if an article is classified as open access in OpenAlex and 0 otherwise. This definition includes both articles that are freely available on the publisher's website and articles for which the publisher version remains paywalled but a free version is available through an online repository.

To assess whether our findings depend on how open access is defined, we use OpenAlex's more detailed open-access categories in two additional analyses (see *Supplementary Text: Types of Open Access*). First, we exclude bronze and green open-access articles, because these categories do not necessarily imply that the article was openly accessible on the publisher's website at the time of publication or that open access remained stable over time. This provides a more restrictive definition of open access, including diamond, gold and hybrid articles: those published in fully open-access journals (diamond or gold) or made openly accessible within otherwise subscription-based journals (hybrid). Second, we distinguish between *Publisher-Provided Open Access* and *Repository-Based Open Access*. Publisher-provided open access includes articles that are freely available on the publisher's website, whereas repository-based open access refers to articles for which access is provided through an online repository while the publisher version remains paywalled. As shown in *Supplementary Text: Types of Open Access*, the main results are robust to both alternative definitions, indicating that they are not driven by any single way of classifying open access.

*Dependent Variables*

**Patent Citation:** In accordance with prior work, we use patent-to-paper citations as the traditional large-scale proxy for whether scientific papers are used as prior art in the development of patented technology. Specifically, we define *Patent Citation* as an indicator variable that takes the value of 1 if a paper is cited by at least one patent within a ten-year window post-publication, and 0 otherwise.

**Patent Citation – Small versus Large Companies:** *Patent Citation by Small (Large) Company* is an indicator variable equal to 1 if a paper receives at least one patent citation from a small (large) company within ten years of publication, and 0 otherwise. We identify patent citations originating from companies using the sector allocation table in PATSTAT. We then classify companies as small or large based on the size of their patent portfolios over the preceding three years, using the year-specific median as the cutoff.

**Patent Citation – Startups versus Established Companies:** *Patent Citation by Startup (Established) Company* is an indicator variable equal to 1 if a paper receives at least one patent citation from a startup (established) company within ten years of publication, and 0 otherwise. Firms are classified as startups or established firms based on founding year, with firms

considered startups during their first three years of operation (five years in robustness analyses). Because founding year information is available only for U.S.-based assignees of USPTO patents (*9*), this analysis is restricted to citations originating from this subset of patents.

**Patent Citation – Author-Inventor Distance:** *Author-Inventor Distance* measures the maximum geographic distance, in kilometers, between a paper's last author and any inventor listed on the patents that cite that paper. This measure can therefore be constructed only for papers that receive at least one patent citation and for which the geographic locations of both the last author and the inventors on the citing patents can be identified.

**Patent Citation – Citation Origins:** *Patent Citation by Applicant (Examiner)* is an indicator variable equal to 1 if a paper receives at least one patent citation from a patent applicant (examiner) within ten years of publication, and 0 otherwise. This distinction is based on the dataset developed by Marx and Fuegi (*17*), which identifies whether each citation originates from the patent applicant or the examiner. Applicant citations more closely reflect inventors' own reliance on a scientific paper as prior art in the development of patented technology.

*In-Text (Front Page Only) Patent Citation* is an indicator variable equal to 1 if a paper receives at least one in-text (front-page-only) patent citation within ten years of publication, and 0 otherwise. This distinction is also provided by Marx and Fuegi (*17*). In-text citations more closely reflect inventors' direct use of a scientific paper as prior art in the development of patented technology (*23*).

**Patent Citations Count:** In accordance with prior work, we use patent-to-paper citations as the traditional large-scale proxy for whether scientific papers are used as prior art in the development of patented technology. To capture the intensity of such use, we define *Patent Citations Count* as the number of citations a paper receives from patents granted worldwide within a ten-year window after publication. To avoid double counting, citations are measured at the level of unique DOCDB patent families.

**Reuse of Scientific Ideas:** *Reuse in Patent* is an indicator variable equal to 1 if at least one scientific idea first introduced in a paper is later reused in the full text of a U.S. patent, and 0 otherwise. Because this measure can be constructed only for papers that introduce at least one new scientific idea, the analysis is restricted accordingly.

**Paper-Patent Similarity Score:** To assess how closely the content of a scientific paper aligns with the patent citing it, we measure the semantic similarity between a paper and its citing patents. Specifically, we compute cosine similarity between the text embeddings of the titles and abstracts of papers and their citing patents using the LogicMill tool developed by the Max Planck Institute (*22*) for a random sample of 100,000 distinct cited papers and all patents citing them. To focus on citations more likely to reflect inventors' direct use of scientific prior art, we restrict the analysis to in-text patent citations (see also *Supplementary Text: Citation Origins*) and exclude citations that appear only on the patent's front page. This yields 69,433 paper–patent similarity scores corresponding to 31,253 unique papers. Based on the distribution of these scores, we construct indicator variables for high similarity using three thresholds: *Paper–Patent Similarity Score* $\geq 0.90$, *Paper–Patent Similarity Score* $\geq 0.925$, and *Paper–Patent Similarity Score* $\geq 0.95$.

*Control Variables*

**Paper Level:** At the paper level, we control for the total number of references cited by the paper (*Paper References (ln)*) and whether the paper is a conference proceeding (*Conference Proceeding*). We also include publication year fixed effects (*Year*) and fixed effects for the primary scientific subfield assigned to each paper (*Subfield*), as classified by OpenAlex. OpenAlex assigns each paper to one of 252 subfields nested within 26 broader fields, using subfield labels derived from Scopus. A machine-learning model determines the appropriate subfield for each paper based on the text of its title and abstract, as well as co-citation patterns (https://help.openalex.org/hc/en-us/articles/24736129405719-Topics).

In a subset of analyses, we also control for each paper's *ex ante* technological and scientific potential (*Technological and Scientific Potential*), limited to papers in our main sample for which these data are available (*6*). Controlling for *scientific* potential, in addition to *technological* potential, is relevant since the scientific promise of an idea or discovery may also correspond to technological applicability (e.g., CRISPR).

**Author-Team Level:** At the author-team level, we control for the number of authors listed on the paper (*Team Size (ln)*), as well as the team's prior scientific productivity and its scientific and technological impact. Scientific productivity is measured as the average number of publications authored by team members before the focal paper's publication (*Team Papers (ln)*). Scientific impact is measured as the average citation impact of team members' prior publications, based on the number of citations those publications received in the five years preceding the focal paper's publication (*Team Paper Citations (ln)*). Technological impact is measured as the average number of patent-to-paper citations received by team members' prior publications, based on the number of patent citations those publications received in the five years preceding the focal paper's publication (*Team Patent Citations (ln)*). In the most stringent specifications, we include fixed effects for the last author's institution (*Institution*) and the last author (*Author*).

**Journal Level:** At the journal level, we control for the technological relevance of the journal in which the article is published using the *Journal Commercial Impact Factor* introduced by Bikard and Marx (*24*). This measure is analogous to the traditional journal impact factor, but replaces citations from papers to papers with citations from patents to papers. In more stringent specifications, we instead include journal fixed effects.

**Supplementary Text**

Main Analysis Regressions

**Table S1** reports paper-level summary statistics for the variables used in these analyses. **Table S2** reports the regression results underlying panel A of Figure 2, and **Table S3** reports the corresponding results for panel B. Relative to **Table S2**, the specifications in **Table S3** are identical, except that they additionally control for papers' *ex ante* scientific and technological potential at the time of publication.

In **Table S2** and **Table S3**, we estimate the likelihood that a paper receives at least one patent citation within ten years of publication (*Patent Citation*) using linear probability models implemented with the *reghdfe* routine in Stata, which allows us to absorb high-dimensional interacted fixed effects. Model 1 includes neither controls nor fixed effects and therefore reports the raw descriptive difference between open- and closed-access papers. In this specification, open-access papers are 54% more likely to receive a patent citation. Model 2 includes publication year × scientific subfield fixed effects, such that open- and closed-access papers are compared within the same subfield and publication year. Model 3 further adds journal fixed effects, so identification comes from comparing open- and closed-access papers published in the same journal, subfield, and year. Model 4 adds last-author affiliation fixed effects (for example, Harvard University or KU Leuven), restricting comparisons to papers published in the same journal, subfield, and year and originating from the same last-author institution. Model 5 adds last-author fixed effects (for example, Robert Langer) on top of this specification, such that the most stringent comparison is between open- and closed-access papers published in the same journal, subfield, and year by the same last author from the same institution. Together with the control variables, these increasingly stringent specifications absorb a large amount of time-, field-, journal-, institutional-, author-, and team-level heterogeneity, thereby reducing concern that the estimated open-access effect is driven by selective open-access publishing across subfields, journals, institutions, or scientists. Across all specifications, the estimated effect of open access on the likelihood of patent citation remains positive, statistically significant at $p < 0.001$, and economically meaningful: in **Table S2**, Models 2–5, open access is associated with an 11–21% higher probability of patent citation.

While the fixed-effects specifications in **Table S2** account for selective open-access publishing across subfields, journals, institutions, and scientists, papers with greater ex ante scientific or technological potential may still be more likely both to be made open access and to be cited later in patents. **Table S3** therefore reports the same specifications as **Table S2**, but additionally controls for papers' *ex ante* scientific and technological potential, measured at the time of publication using a machine-learning model trained on article abstracts (*6*). Because these measures are available only for a subset of the sample, the number of observations falls from 24,786,337 to 10,424,832. Even in this more demanding subsample and conditional on these additional controls, the estimated effect of open access remains positive, statistically significant at $p < 0.001$, and economically meaningful: in **Table S3**, Models 2–5, open access is associated with a 12–16% higher probability of patent citation. These results indicate that the estimates in **Table S2** are unlikely to be explained by the possibility that papers with greater

underlying scientific or technological potential are both more likely to be made open access and more likely to be cited by patents.

To further strengthen identification, panel C of Figure 2 in the main text exploits paper twins—near-identical scientific discoveries published contemporaneously by independent teams, of which one paper is open access and the other is closed access. While the fixed-effects specifications in **Table S2** account for selective open-access publishing across scientific subfields, journals, institutions, and scientists, and **Table S3** further controls for papers' ex ante scientific and technological potential at the time of publication, these approaches cannot fully eliminate concern that open- and closed-access papers may still differ in underlying scientific content and thus technological potential. The paper-twin design addresses this concern by holding the underlying scientific content, and thus much of its technological potential, largely constant. As shown in the corresponding results reported in **Table S5** and discussed in greater detail below (see *Supplementary Text: Paper Twin Analysis Regressions*), the estimated effect of open access remains positive, statistically significant at $p < 0.001$, and economically meaningful: in **Table S5**, Models 1-3, open access is associated with a 6–13% higher probability of patent citation.

In **Table S6**, we examine whether open access is associated not only with the likelihood that a paper is cited at least once by patents, but also with the intensity of such use, measured by the total number of patent citations a paper receives within ten years of publication (*Patent Citations Count*). To avoid double counting, citations are measured at the level of unique DOCDB patent families. Given the count nature of the dependent variable, and to maintain consistent with the high-dimensional fixed-effects framework used above, we estimate these models using the *ppmlhdfe* routine in Stata. The results confirm that open-access papers are not only more likely to be cited by patents than closed-access papers, but also receive more patent citations overall, with estimated increases of 12–25% ($p < 0.001$ across all specifications).

Citation Origins

Our main interpretation is that open access increases patent citations because inventors may otherwise lack access to relevant scientific prior art when developing patented technology. If this interpretation is correct, the open-access effect should be strongest for citation types that most closely reflect inventors' use of scientific papers as prior art.

Patent citations to scientific papers differ both in who adds them and in where they appear within the patent document. Some are added by patent applicants, whereas others are introduced by examiners during the examination process. Applicant citations more closely reflect inventors' own reliance on scientific prior art, whereas examiner citations may more often reflect references added for legal or disclosure purposes. Examiners should also be less constrained by access barriers to paywalled research than inventors. Likewise, citations appearing in the body text of the patent more closely reflect inventors' use of scientific papers in the development of patented technology than citations appearing only on the front page (*23*).

We therefore test whether the open-access effect is stronger for applicant citations than for examiner citations, and stronger for in-text citations than for front-page-only citations. The results, reported in **Fig. S3** and **Table S7**, support this prediction. Open access is associated with a larger increase in the likelihood of applicant citations than examiner citations (22% versus 6%), and with a larger increase in the likelihood of in-text citations than front-page-only citations (35% versus 9%). Taken together, these patterns support our interpretation that open access matters in part because it relaxes inventors' access constraints, thereby increasing their use of scientific papers as prior art in the development of patented technology.

Types of Open Access

Open access exists in several forms that differ in how articles are made freely available, when that access begins and how long it lasts, and what reuse rights are granted. Broadly, these forms can be grouped into two main routes: publisher-provided open access and repository-based open access.

Publisher-provided open access includes diamond, gold, hybrid, and bronze open access. Both diamond and gold open access refer to articles published in fully open-access journals, where all content is made freely available by the publisher. However, whereas gold open access typically requires the author – or a funder or institution on the author's behalf – to pay article processing charges (APCs), these APCs are covered by the journal itself in case of diamond open access, usually through institutional or community support. Hybrid open access refers to articles published in otherwise subscription-based journals, where a specific article is made freely available by the publisher under an open license. Like gold open access, the APCs for hybrid open access are typically paid by the author. Thus, diamond, gold and hybrid open access all provide access via the publisher's platform, but hybrid open access occurs within journals that otherwise remain paywalled. All three typically provide immediate free availability upon publication, and usually include clearly defined reuse rights through an open license such as a Creative Commons (CC) license. By contrast, bronze open access refers to articles that are free to read on the publisher's website but do not have a clearly identifiable open license. As a result, bronze open access does not guarantee clear reuse rights, may not ensure long-term open accessibility, and is often temporary.

Repository-based open access corresponds to green open access. Under this model, the article remains paywalled on the publisher's website, but a free version is available in an open repository, such as an institutional repository or subject-specific archive. These repository versions may consist of a preprint, an accepted manuscript, or a postprint, depending on journal policies, and they do not necessarily correspond to the final publisher-formatted version. Green open access typically does not require payment of an article processing charge, because it relies on self-archiving rather than publisher-mediated open dissemination. However, access through repositories may be subject to embargo periods, so that an article may become openly available only after a delay, for example six or twelve months after publication. In addition, reuse rights may be more limited and depend on publisher policies.

These distinctions matter because not all open-access forms guarantee the same timing, stability, or legal clarity of access. Diamond, gold and hybrid open access generally imply immediate publisher-provided access with explicit reuse rights. By contrast, bronze and green open access may be less stable over time: access may be withdrawn or may begin only some time after publication, for example after an embargo period has passed. Consequently, a paper classified as bronze or green open access at the time of data collection was not necessarily open access at the time of publication, nor will it necessarily remain open access in the future.

To assess whether our findings depend on the specific type of open access considered, we conduct two additional analyses.

First, we adopt a more restrictive definition of open access by including only diamond, gold and hybrid articles and excluding bronze and green articles. This restriction focuses on forms of open access that are typically immediately available on the publisher's platform and accompanied by explicit reuse rights. We then compare this subset of open-access articles with closed-access publications. The results, reported in **Table S8**, confirm the robustness of the findings presented in **Table S2**, with estimates indicating an even larger open-access effect: open access is now associated with an 12–29% higher probability of patent citation (Models 2–5).

Second, we examine whether the positive relationship between open-access publishing and patent citation holds across the two main routes to open access: publisher-provided open access and repository-based open access. The analyses reported in **Table S9** distinguish between these two routes and show that findings are robust for both types, with the open-access effect ranging from 8 to 24% for publisher-provided open access and 15 to 20% for repository-based open access (Models 2–5).

Taken together, these results indicate that our findings are not driven by any single type of open access. The positive association between open access and patent citation holds when attention is restricted to diamond, gold and hybrid open access, and it also holds for both publisher-provided and repository-based open access.

Reuse of Scientific Ideas

Our main analyses use patent-to-paper citations to measure whether scientific papers are used as prior art in the development of patented technology. Because patents may build on scientific ideas without formally citing the relevant papers (*18*, *19*), we complement this citation-based measure with a text-based measure of whether scientific ideas introduced in papers later reappear in patent text.

Specifically, we use the approach developed by Arts et al. (*20*) and extended by Arts and Melluso (*8*) to trace whether scientific ideas first introduced in the scientific literature (as measured by new noun phrases that appear for the first time in the title or abstract of papers) are later reused in the full text of U.S. patents. The dependent variable, *Reuse in Patent*, equals 1 if at least one scientific idea introduced in a paper is later reused in a patent, and 0 otherwise. Consistent with the main analysis, we estimate linear probability models using the *reghdfe*

routine in Stata. In addition to the controls used in the main analysis, we include the log-transformed number of scientific ideas introduced in the paper (*Scientific Ideas*), because papers that introduce more such ideas are mechanically more likely to have at least one reused in a patent. Because this measure can be constructed only for papers that introduce at least one new scientific idea, the sample is restricted to 6,918,397 papers, compared with 24,786,337 papers in the main analysis.

The results, reported in **Table S10**, show that scientific ideas introduced in open-access papers are significantly more likely to be reused in subsequent patents than those introduced in closed-access papers. Across Models 2–5, open access is associated with a 2% to 7% higher probability of reuse. This finding shows that the main results are not limited to formal patent citations, but also extend to a text-based measure of the use of scientific ideas in patented technology.

Time to First Patent Citation

The main text shows that open-access papers are substantially more likely to be cited by patents than comparable closed-access papers. Here, we examine whether open-access papers are also cited earlier in time. If open access reduces inventors' barriers to scientific prior art, its effects should appear not only in whether papers are cited by patents, but also in how quickly those citations occur.

**Fig. S4** shows Kaplan–Meier survival curves for the time from publication to the first patent citation, separately for open- and closed-access papers. The curves represent the probability that a paper has not yet been cited in a patent at each point in time, providing a nonparametric comparison of time-to-first-citation across access types. Overall, the survival curves indicate that open-access papers are cited sooner and more frequently than closed-access papers. For example, the cumulative probability of receiving at least one patent citation reaches 5% after about 2.4 years for open-access papers, compared with about 3.7 years for closed-access papers.

To assess whether this difference persists after conditioning on observable characteristics, we estimate stratified Cox proportional hazards models using the same paper-, journal-, and team-level controls as in the main regressions. The results, reported in **Table S11**, show that open access is associated with an 8–18% higher hazard of first patent citation across specifications (Models 2–5). In other words, at a given point in time, open-access papers are more likely than otherwise comparable closed-access papers to receive their first patent citation.

Taken together, the descriptive and model-based evidence indicates that open access is associated not only with a higher likelihood of patent citation, but also with earlier use of scientific papers in patented technology.

Paper-Patent Similarity

In the main text, we show that open-access papers are more likely to be cited by patents than closed-access papers. Here, we examine a further question: conditional on being cited, is an open-access paper also more closely related to the technical content of the citing patent? If so, this would suggest a closer relationship between the cited paper and the patented invention, providing a complementary indication of more substantive use of the underlying scientific knowledge.

As described in the *Materials and Methods: Key Variables and Measures* section, we measure semantic similarity between cited papers and the patents citing them using cosine similarity based on text embeddings of the titles and abstracts of both documents. Similarity scores are calculated using the LogicMill tool (*22*) for a random sample of 100,000 distinct cited papers and all patents citing them. Restricting the analysis to in-text patent citations, which are more likely to reflect inventors' direct use of scientific prior art, yields 69,433 paper–patent similarity scores corresponding to 31,253 unique papers.

**Fig. S5** shows the kernel-density distribution of paper–patent similarity scores separately for paper–patent pairs involving open-access and closed-access papers. The distribution indicates that paper–patent dyads involving open-access papers tend to exhibit higher semantic similarity than paper–patent pairs involving closed-access papers. **Table S12** reports linear probability models estimating whether a cited paper–patent dyad exceeds increasingly stringent similarity thresholds (*Paper–Patent Similarity Score* $\geq 0.90$, $\geq 0.925$, and $\geq 0.95$). The specifications progressively add (i) publication year × subfield fixed effects (Models 1, 4, and 7), such that paper–patent dyads involving open- and closed-access papers are compared within the same scientific subfield and publication year; (ii) citing-patent fixed effects (Models 2, 5, and 8), such that comparisons are made across different papers cited within the same citing patent; and (iii) both sets of fixed effects simultaneously (Models 3, 6, and 9). In the most stringent specification, we compare paper–patent dyads in which the cited papers were published in the same scientific subfield and year and are cited by the same patent, but differ in open-access status, allowing us to test whether the open-access paper is more semantically similar to the citing patent than the corresponding closed-access paper. Across specifications, the results show that open-access papers are significantly more likely to exhibit high content similarity to the citing patent than comparable closed-access papers.

Taken together, the descriptive and regression evidence shows that, conditional on being cited by a patent, open-access papers are more semantically similar to the citing patents than closed-access papers. This pattern suggests a closer relationship between the cited paper and the patented invention, providing a complementary indication that open access is associated not only with a higher likelihood of patent citation, but also with more substantive use of the underlying scientific knowledge.

Paper Twin Analysis Regression

While the fixed-effects specifications in Table S2 account for selective open-access publishing across scientific subfields, journals, institutions, and scientists, and **Table S3** further controls for papers' *ex ante* scientific and technological potential at the time of publication, these approaches cannot fully eliminate concern that open- and closed-access papers may still differ in underlying scientific content and thus technological potential. To further strengthen identification, we therefore exploit paper twins: near-identical scientific discoveries published contemporaneously by independent research teams, of which one paper is open access and the other is closed access. This design allows us to compare papers built around nearly the same underlying scientific discovery, while reducing concern that the estimated open-access effect is driven by differences in scientific content, intrinsic technological potential, or selective open-access publishing. Further details on the motivation, data sources, and sample construction for this analysis are provided in *Materials and Methods: Data Sources and Sample Construction – Paper-Twin Analysis*.

**Table S4** reports summary statistics for the variables used in this analysis, and **Table S5** reports the regression results. As in the main analysis, we estimate the likelihood that a paper receives at least one patent citation within ten years of publication (*Patent Citation*) using linear probability models implemented with the *reghdfe* routine in Stata. Model 1 includes paper-twin-pair fixed effects (*Simultaneous Discovery*) and publication-year fixed effects (*Year*), such that identification comes from within-twin comparisons between papers reporting nearly the same discovery but differing in open-access status. Model 2 adds the same set of control variables used in the main specification, as well as a dummy variable, *First Published*, indicating which paper within a twin pair was published earlier when publication dates differ. Panel C of Figure 2 in the main text is based on the estimated outcomes from Model 2. Following prior work using paper twins (*12*, *24*), Model 3 further adds journal-level and team-level controls tailored to this design. Specifically, at the journal level, we control for the journal's conventional impact factor (*Journal Impact Factor*). At the author-team level, we additionally control for the share of authors based in the United States (*Share of US-Based Authors*), whether the paper includes at least one author affiliated with industry (Collaboration with Industry), and whether the last author is based in the United States (*Main Author Is US-Based*).

Across Models 1–3, the estimated effect of open access remains positive, statistically significant at $p < 0.001$, and economically meaningful, with open access papers being 6–13% more likely to cited by a patent than their closed-access twin counterparts. Thus, even among near-identical simultaneous discoveries, the open-access paper is more likely to be cited by patents than the corresponding closed-access twin. This pattern strengthens the interpretation that the main results are not driven by differences in the underlying scientific content or intrinsic technological potential.

Models 4–6 further tighten the design by restricting the sample to twin pairs with especially high semantic similarity, using a cosine-similarity thresholds of 0.95. This restriction further limits the analysis to the most similar paper pairs and therefore further reduces residual differences in how the same underlying discovery is described. Despite the significant drop in the number of observations, the estimated open-access effect remains robust in this subsample.

Taken together, these results show that the positive association between open access and patent citation persists even under a stringent within-discovery research design.

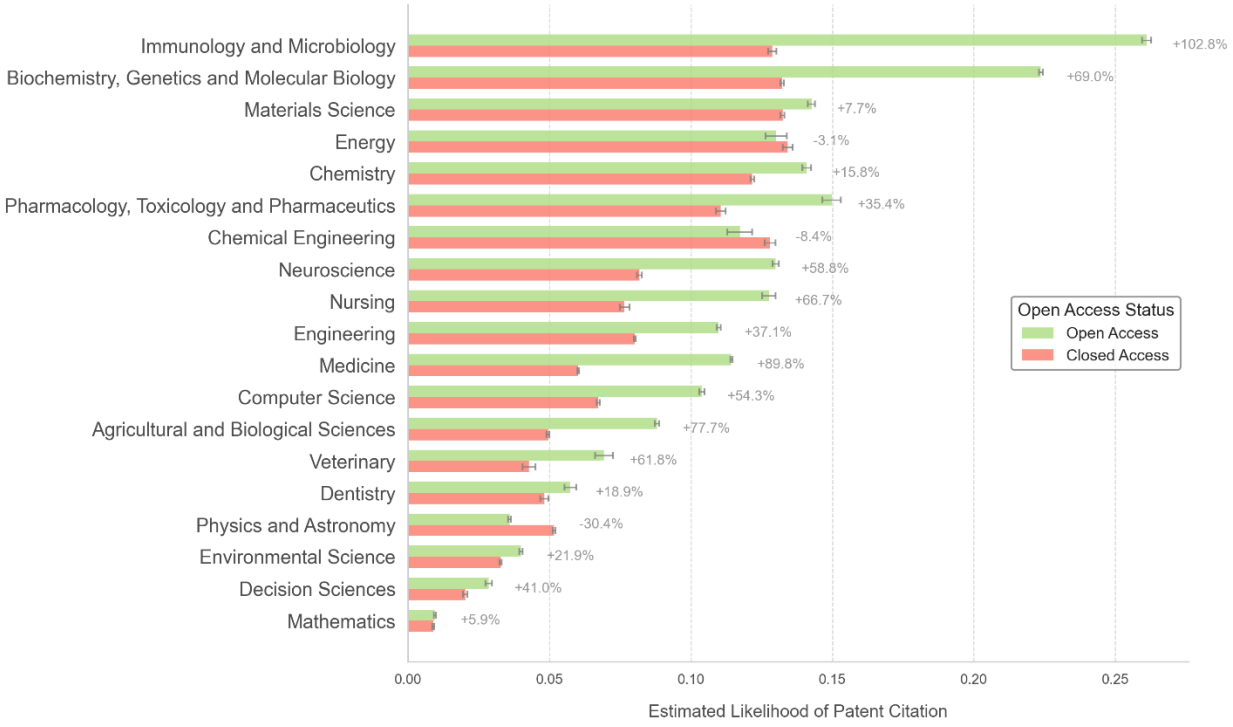



**Fig. S1.** Shows the descriptive patterns of patent citation likelihood for open- and closed-access papers by field of study (OpenAlex classification) over 2000–2015 (n ≈ 24 million articles). Fields are ordered from top to bottom by their baseline relevance for technological development. Confidence intervals are shown at the 95% level.

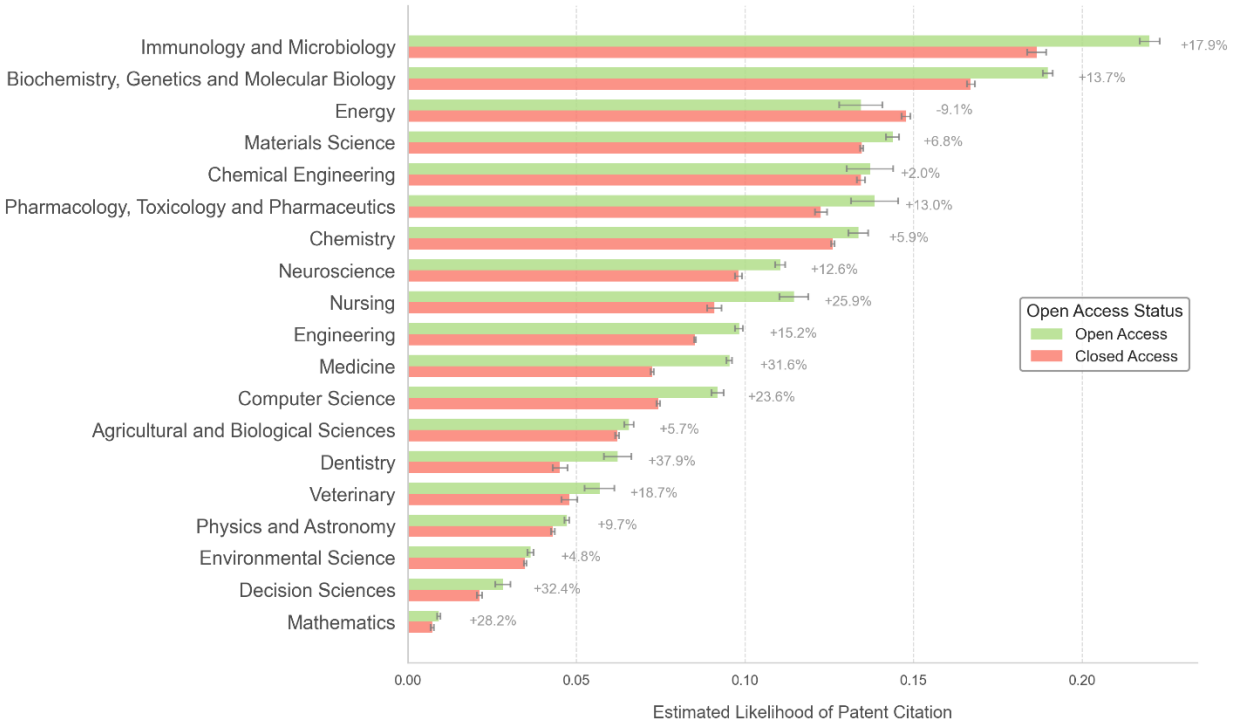



**Fig. S2.** shows the estimated likelihood of patent citation for open- and closed-access papers by field of study (OpenAlex classification) over the period 2000–2015 (n ≈ 24 million articles), based on linear probability models. Fields are ordered from top to bottom by their baseline relevance for technological development. The specification includes subfield-by-year-by-journal fixed effects and controls for paper characteristics (e.g., number of references, conference proceeding status), journal commercial impact factor, and author team characteristics (e.g., size, prior publications, prior paper citations, and prior patent citations). Confidence intervals are shown at the 95% level.

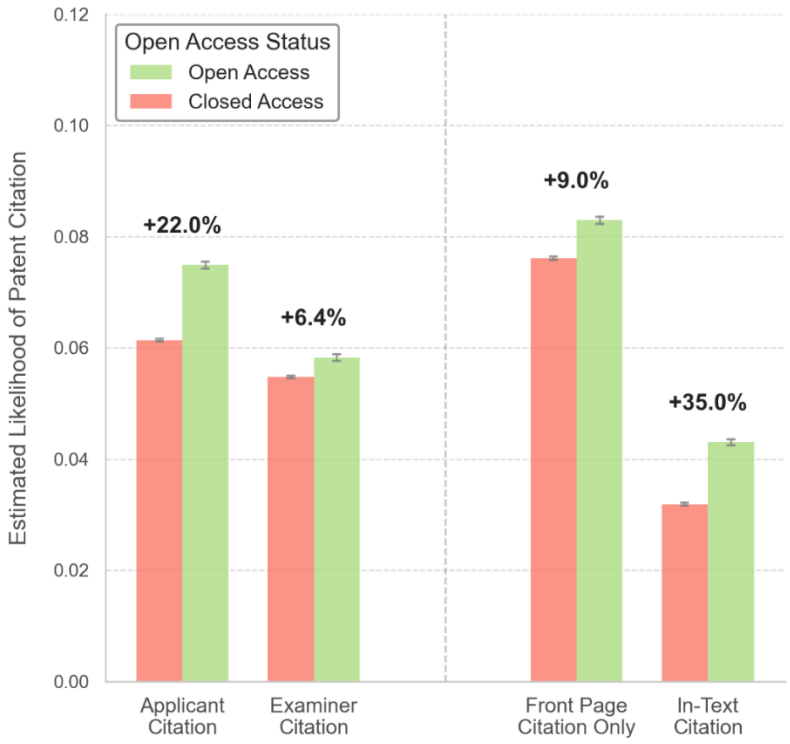



**Fig. S3.** Open access and the likelihood of patent citation, distinguishing applicant versus examiner citations and front-page versus in-text citations. Estimates based on Table S7, Models 1–4.

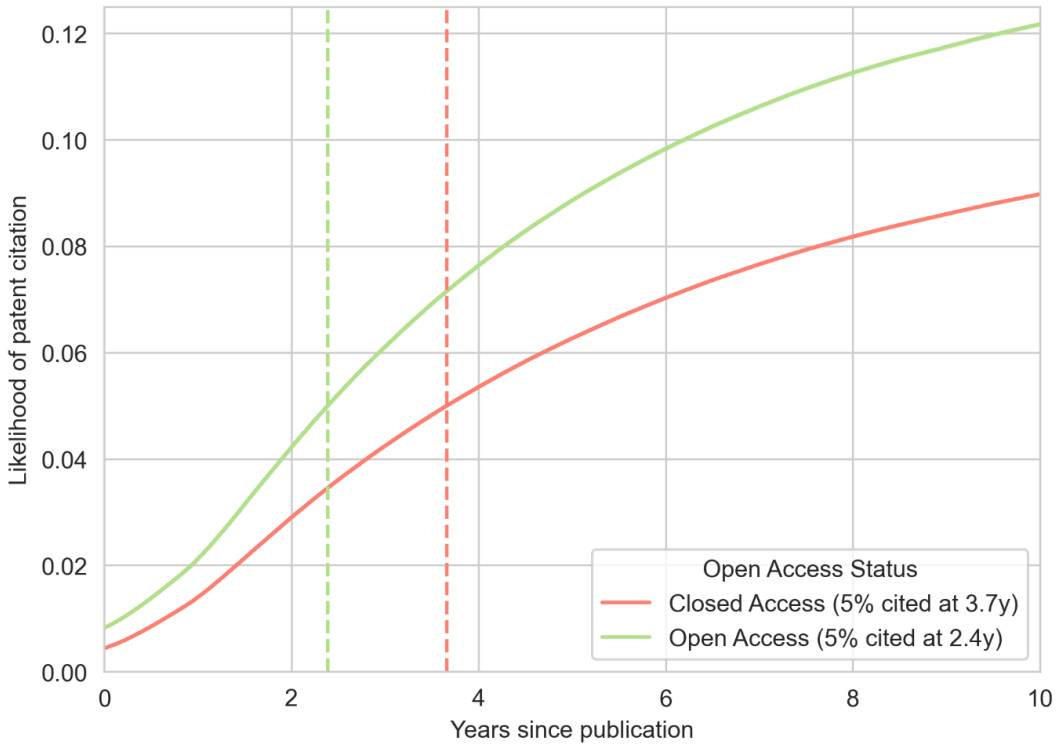



**Fig. S4.** Kaplan–Meier survival curves showing the time from publication to the first patent citation for open- and closed-access papers. The curves represent the probability that a paper has not yet been cited in a patent at each point in time, providing a nonparametric comparison of time to first citation across access types.

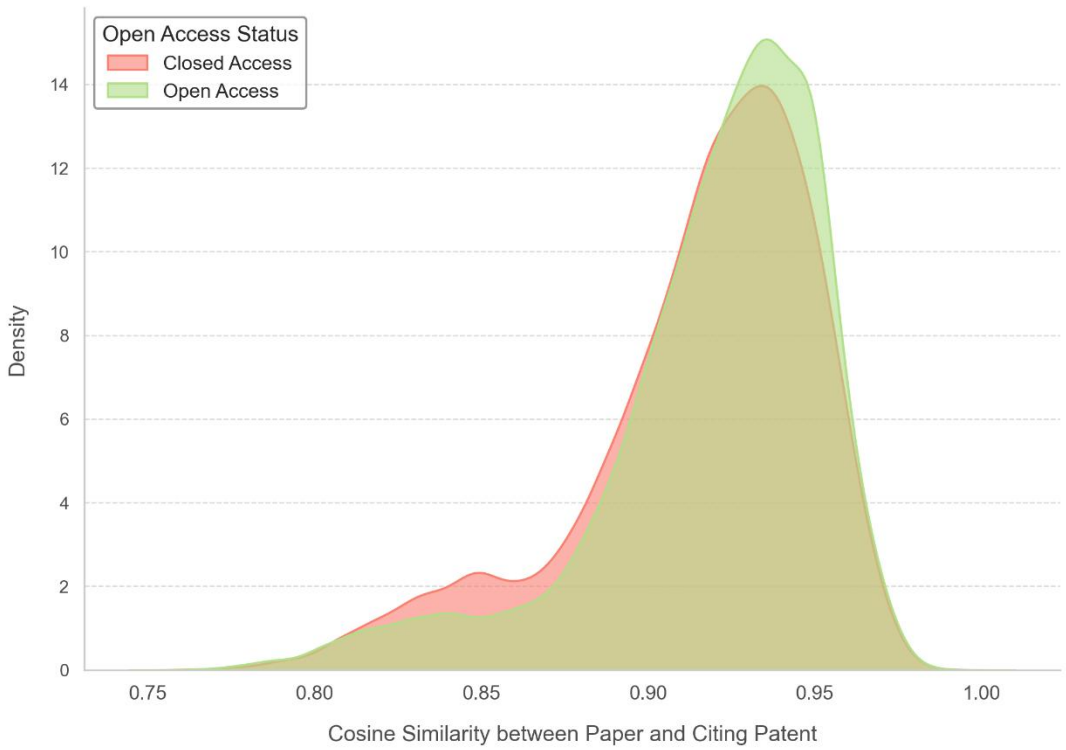



**Fig. S5.** Kernel density distribution of paper–patent semantic similarity scores for open- and closed-access papers. Semantic similarity is measured using cosine similarity based on text embeddings of titles and abstracts of cited papers and their citing patents. The analysis uses a random sample of 100,000 cited papers and their full set of citing patents, restricted to in-text citations, yielding 69,433 paper–patent similarity scores for 31,253 unique papers.

**Table S1.** Summary statistics – Main analysis.

| | *N* | Mean | Standard Dev. | Min. | Max. |
|---|---|---|---|---|---|
| Patent Citation | 24,786,337 | 0.091 | 0.288 | 0.000 | 1.000 |
| Patent Citation by Small Company | 24,786,337 | 0.014 | 0.117 | 0.000 | 1.000 |
| Patent Citation by Large Company | 24,786,337 | 0.048 | 0.213 | 0.000 | 1.000 |
| Patent Citation by Startup (<3y) | 24,786,337 | 0.004 | 0.064 | 0.000 | 1.000 |
| Patent Citation by Est. Comp. (>3y) | 24,786,337 | 0.021 | 0.144 | 0.000 | 1.000 |
| Patent Citation by Startup (<5y) | 24,786,337 | 0.006 | 0.077 | 0.000 | 1.000 |
| Patent Citation by Est. Comp. (>5y) | 24,786,337 | 0.020 | 0.141 | 0.000 | 1.000 |
| Author-Inventor Distance | 1,618,216 | 7194.292 | 4536.368 | 0.000 | 20001.78 |
| Applicant Citation | 24,786,337 | 0.065 | 0.247 | 0.000 | 1.000 |
| Examiner Citation | 24,786,337 | 0.056 | 0.230 | 0.000 | 1.000 |
| Front Page Citation Only | 24,786,337 | 0.078 | 0.268 | 0.000 | 1.000 |
| Reuse in Patent | 6,918,397 | 0.126 | 0. 331 | 0.000 | 1.000 |
| Patent Citations Count | 24,786,337 | 0.295 | 4.104 | 0.000 | 2662.000 |
| Open Access | 24,786,337 | 0.306 | 0.461 | 0.000 | 1.000 |
| Conference Proceeding | 24,786,337 | 0.002 | 0.045 | 0.000 | 1.000 |
| Scientific Potential | 10,424,832 | 0.576 | 0.269 | 0.002 | 0.998 |
| Technological Potential | 10,424,832 | 0.447 | 0.324 | 0.001 | 0.999 |
| Paper References (ln) | 24,786,337 | 2.005 | 1.562 | 0.000 | 8.326 |
| Journal Commercial Impact Factor | 24,786,337 | 0.017 | 0.042 | 0.000 | 7.375 |
| Team Size (ln) | 24,786,337 | 1.436 | 0.559 | 0.000 | 8.569 |
| Team Papers (ln) | 24,786,337 | 2.893 | 1.411 | 0.000 | 11.622 |
| Team Paper Citations (ln) | 24,786,337 | 3.742 | 1.816 | 0.000 | 10.737 |
| Team Patent Citations (ln) | 24,786,337 | 1.495 | 1.595 | 0.000 | 9.520 |

**Table S2.** Open-access and the likelihood of patent citation. Main analysis.

| | (1) Patent Citation | (2) Patent Citation | (3) Patent Citation | (4) Patent Citation | (5) Patent Citation |
|---|---|---|---|---|---|
| Open Access | 0.0420*** | 0.0097*** | 0.0166*** | 0.0240*** | 0.0185*** |
| | (0.0001) | (0.0005) | (0.0004) | (0.0007) | (0.0013) |
| Conference Proceeding | | 0.0024 | | | |
| | | (0.0032) | | | |
| Paper References (ln) | | 0.0296*** | 0.0320*** | 0.0358*** | 0.0371*** |
| | | (0.0005) | (0.0002) | (0.0003) | (0.0005) |
| Journal Commercial Impact Factor | | 0.9774*** | | | |
| | | (0.0367) | | | |
| Team Size (ln) | | 0.0114*** | 0.0160*** | 0.0203*** | 0.0197*** |
| | | (0.0006) | (0.0002) | (0.0006) | (0.0013) |
| Team Papers (ln) | | -0.0325*** | -0.0285*** | -0.0370*** | -0.0298*** |
| | | (0.0005) | (0.0002) | (0.0004) | (0.0014) |
| Team Paper Citations (ln) | | 0.0001 | 0.0011*** | -0.0025*** | 0.0017 |
| | | (0.0003) | (0.0001) | (0.0003) | (0.0011) |
| Team Patent Citations (ln) | | 0.0527*** | 0.0422*** | 0.0456*** | 0.0296*** |
| | | (0.0008) | (0.0002) | (0.0003) | (0.0007) |
| Mean(Dep. Var.) | 0.0910 | 0.0910 | 0.0937 | 0.1224 | 0.1203 |
| Year x Subfield FE | | x | | | |
| Year x Subfield x Journal FE | | | x | | |
| Year x Subfield x Journal x Inst. FE | | | | x | |
| Year x Subfield x Journal x Inst. x Auth. FE | | | | | x |
| N | 24,786,337 | 24,786,334 | 22,844,381 | 4,291,667 | 1,535,904 |
| Full Information Available | 24,786,337 | 24,786,337 | 24,639,920 | 15,423,932 | 15,423,932 |
| Dropped | 0 | 3 | 1,795,539 | 11,132,265 | 13,888,028 |

*Note:* This table reports estimates from linear probability models (STATA REGHDFE). Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

**Table S3.** Open-access and the likelihood of patent citation. Main analysis, accounting for technological and scientific potential.

| | (1) Patent Citation | (2) Patent Citation | (3) Patent Citation | (4) Patent Citation | (5) Patent Citation |
|---|---|---|---|---|---|
| Open Access | 0.0316*** | 0.0161*** | 0.0167*** | 0.0240*** | 0.0204*** |
| | (0.0002) | (0.0006) | (0.0005) | (0.0011) | (0.0020) |
| Scientific Potential | | -0.0101*** | -0.0062*** | -0.0084*** | 0.0004 |
| | | (0.0017) | (0.0006) | (0.0018) | (0.0032) |
| Technological Potential | | 0.1574*** | 0.1485*** | 0.1408*** | 0.1109*** |
| | | (0.0047) | (0.0009) | (0.0016) | (0.0030) |
| Conference Proceeding | | -0.0205*** | | | |
| | | (0.0042) | | | |
| Paper References (ln) | | 0.0292*** | 0.0318*** | 0.0399*** | 0.0426*** |
| | | (0.0006) | (0.0003) | (0.0005) | (0.0009) |
| Journal Commercial Impact Factor | | 0.8410*** | | | |
| | | (0.0392) | | | |
| Team Size (ln) | | 0.0133*** | 0.0165*** | 0.0210*** | 0.0215*** |
| | | (0.0009) | (0.0003) | (0.0009) | (0.0020) |
| Team Papers (ln) | | -0.0426*** | -0.0392*** | -0.0450*** | -0.0328*** |
| | | (0.0008) | (0.0002) | (0.0007) | (0.0023) |
| Team Paper Citations (ln) | | 0.0003 | 0.0001 | -0.0049*** | 0.0005 |
| | | (0.0006) | (0.0002) | (0.0007) | (0.0021) |
| Team Patent Citations (ln) | | 0.0565*** | 0.0499*** | 0.0541*** | 0.0338*** |
| | | (0.0010) | (0.0002) | (0.0004) | (0.0012) |
| Mean(Dep. Var.) | 0.1411 | 0.1411 | 0.1439 | 0.1613 | 0.1573 |
| Year x SubField FE | | x | | | |
| Year x Subfield x Journal FE | | | x | | |
| Year x Subfield x Journal x Inst. FE | | | | x | |
| Year x Subfield x Journal x Inst. x Auth. FE | | | | | x |
| N | 10,424,832 | 10,424821 | 9,571,492 | 2,024,062 | 715,948 |
| Full Information Available | 10,424,832 | 10,424,832 | 1,0422,218 | 8,606,166 | 8,606,166 |
| Dropped | 0 | 11 | 850,726 | 658,2104 | 7,890,218 |

*Note:* This table reports estimates from linear probability models (STATA REGHDFE). Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

**Table S4.** Summary statistics – Paper twin design.

| | *N* | Mean | Standard Dev. | Min. | Max. |
|---|---|---|---|---|---|
| Patent Citation | 40,270 | 0.269 | 0.444 | 0.000 | 1.000 |
| Open Access | 40,270 | 0.500 | 0.500 | 0.000 | 1.000 |
| Conference Proceeding | 40,270 | 0.050 | 0.217 | 0.000 | 1.000 |
| Paper References | 40,270 | 3.287 | 0.737 | 0.000 | 7.619 |
| Journal Commercial Impact Factor | 40,270 | 0.024 | 0.076 | 0.000 | 6.000 |
| Journal Impact Factor | 40,270 | 1.986 | 3.095 | 0.000 | 35.616 |
| Team Size (ln) | 40,270 | 1.434 | 0.484 | 0.000 | 5.442 |
| Team Papers (ln) | 40,270 | 3.766 | 1.011 | 0.000 | 8.174 |
| Team Paper Citations (ln) | 40,270 | 4.542 | 1.097 | 0.000 | 9.236 |
| Team Patent Citations (ln) | 40,270 | 3.056 | 2.380 | 0.000 | 10.615 |
| Published First | 40,270 | 0.412 | 0.492 | 0.000 | 1.000 |
| Main Author is US-Based | 40,270 | 0.320 | 0.467 | 0.000 | 1.000 |
| Share of US-Based Authors | 40,270 | 0.354 | 0.441 | 0.000 | 1.000 |
| Collaboration with Industry | 40,270 | 0.102 | 0.303 | 0.000 | 1.000 |

**Table S5.** Open access and the likelihood of patent citation. Paper twin design.

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | All Twins | | | Twin Similarity Score >= 0.95 | | |
| Open Access | 0.0339*** | 0.0177*** | 0.0147*** | 0.0307*** | 0.0156** | 0.0133* |
| | (0.0033) | (0.0034) | (0.0034) | (0.0051) | (0.0052) | (0.0052) |
| Conference Proceeding | | -0.0061 | 0.0038 | | -0.0421* | -0.0219 |
| | | (0.0121) | (0.0122) | | (0.0208) | (0.0211) |
| Paper References (ln) | | 0.0544*** | 0.0491*** | | 0.0555*** | 0.0510*** |
| | | (0.0040) | (0.0040) | | (0.0070) | (0.0070) |
| Journal Comm. Impact Factor | | 0.3425*** | 0.1719*** | | 0.2681*** | 0.1419*** |
| | | (0.0523) | (0.0364) | | (0.0521) | (0.0422) |
| Team Size (ln) | | 0.0131 | 0.0107 | | 0.0212 | 0.0203 |
| | | (0.0075) | (0.0076) | | (0.0111) | (0.0113) |
| Team Papers (ln) | | -0.0322*** | -0.0285*** | | -0.0290*** | -0.0267*** |
| | | (0.0050) | (0.0049) | | (0.0071) | (0.0071) |
| Team Paper Citations (ln) | | 0.0172*** | 0.0130** | | 0.0212** | 0.0190** |
| | | (0.0048) | (0.0048) | | (0.0073) | (0.0073) |
| Team Patent Citations (ln) | | 0.0351*** | 0.0337*** | | 0.0311*** | 0.0302*** |
| | | (0.0020) | (0.0020) | | (0.0030) | (0.0030) |
| Published First | | 0.0057 | 0.0081 | | 0.0224** | 0.0218** |
| | | (0.0043) | (0.0042) | | (0.0071) | (0.0071) |
| Journal Impact Factor | | | 0.0112*** | | | 0.0094*** |
| | | | (0.0014) | | | (0.0017) |
| Main Author is US-Based | | | -0.0128 | | | -0.0066 |
| | | | (0.0100) | | | (0.0151) |
| Share of US-Based Authors | | | 0.0311** | | | 0.0116 |
| | | | (0.0120) | | | (0.0174) |
| Collaboration with Industry | | | 0.0448*** | | | 0.0107 |
| | | | (0.0104) | | | (0.0167) |
| Mean(Dep. Var.) | 0.2692 | 0.2692 | 0.2692 | 0.2938 | 0.2938 | 0.2938 |
| Simultaneous Discovery FE | x | x | x | x | x | x |
| Year FE | x | x | x | x | x | x |
| N | 40,266 | 40,266 | 40,266 | 14,468 | 14,468 | 14,468 |

*Note:* This table reports estimates from linear probability models (STATA REGHDFE). Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

**Table S6.** Open-access and the number of patent citations (DOCDB-level). Main analysis.

| | (1)<br>Patent Citations Count | (2)<br>Patent Citations Count | (3)<br>Patent Citations Count | (4)<br>Patent Citations Count | (5)<br>Patent Citations Count |
|---|---|---|---|---|---|
| Open Access | 0.6120***<br>(0.0055) | 0.1133***<br>(0.0097) | 0.1564***<br>(0.0089) | 0.2235***<br>(0.0234) | 0.2148***<br>(0.0336) |
| Conference Proceeding | | -0.3431***<br>(0.0538) | | | |
| Paper References (ln) | | 0.4982***<br>(0.0047) | 0.4562***<br>(0.0047) | 0.4137***<br>(0.0092) | 0.4203***<br>(0.0167) |
| Journal Commercial Impact Factor | | 0.9327***<br>(0.0570) | | | |
| Team Size (ln) | | 0.4004***<br>(0.0132) | 0.2979***<br>(0.0085) | 0.3217***<br>(0.0256) | 0.2800***<br>(0.0543) |
| Team Papers (ln) | | -0.6879***<br>(0.0101) | -0.5664***<br>(0.0060) | -0.5056***<br>(0.0190) | -0.3659***<br>(0.0584) |
| Team Paper Citations (ln) | | 0.0613***<br>(0.0098) | -0.0574***<br>(0.0061) | -0.1460***<br>(0.0215) | -0.0059<br>(0.0630) |
| Team Patent Citations (ln) | | 0.8203***<br>(0.0092) | 0.6683***<br>(0.0056) | 0.6486***<br>(0.0190) | 0.5085***<br>(0.0600) |
| Mean(Dep. Var.) | 0.2952 | 0.2953 | 0.5017 | 1.3926 | 1.7605 |
| Year x SubField FE | | x | | | |
| Year x Subfield x Journal FE | | | x | | |
| Year x Subfield x Journal x Inst. FE | | | | x | |
| Year x Subfield x Journal x Inst. x Auth. FE | | | | | x |
| N | 24,786,337 | 24,778,700 | 13,923,210 | 1,255,743 | 347,175 |
| Full Information Available | 24,786,337 | 24,786,337 | 24,639,920 | 15,423,932 | 15,423,932 |
| Dropped | 0 | 7,637 | 10,716,710 | 14,168,189 | 15,076,757 |

*Note:* This table reports estimates from Poisson models (STATA PPMLHDFE). Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

**Table S7.** Open access and the likelihood of patent citation, distinguishing applicant versus examiner citations and front-page versus in-text citations.

| | (1) Applicant Citation | (2) Examiner Citation | (3) Front Page Citation Only | (4) In-Text Citation | (5) Applicant Citation | (6) Examiner Citation | (7) Front Page Citation Only | (8) In-Text Citation |
|---|---|---|---|---|---|---|---|---|
| Open Access | 0.0135*** | 0.0035*** | 0.0068*** | 0.0111*** | 0.0178*** | 0.0081*** | 0.0143*** | 0.0111*** |
| | (0.0005) | (0.0004) | (0.0005) | (0.0004) | (0.0003) | (0.0003) | (0.0003) | (0.0003) |
| Conference Proceeding | 0.0055* | -0.0008 | 0.0052 | -0.0044*** | | | | |
| | (0.0024) | (0.0027) | (0.0033) | (0.0009) | | | | |
| Paper References (ln) | 0.0207*** | 0.0193*** | 0.0262*** | 0.0114*** | 0.0257*** | 0.0205*** | 0.0278*** | 0.0166*** |
| | (0.0004) | (0.0004) | (0.0004) | (0.0003) | (0.0002) | (0.0001) | (0.0002) | (0.0001) |
| Journal Commercial Impact Factor | 0.8874*** | 0.6500*** | 0.8677*** | 0.5905*** | | | | |
| | (0.0326) | (0.0265) | (0.0335) | (0.0229) | | | | |
| Team Size (ln) | 0.0021*** | 0.0124*** | 0.0128*** | -0.0013*** | 0.0089*** | 0.0156*** | 0.0166*** | 0.0057*** |
| | (0.0003) | (0.0005) | (0.0006) | (0.0002) | (0.0002) | (0.0002) | (0.0002) | (0.0002) |
| Team Papers (ln) | -0.0276*** | -0.0196*** | -0.0282*** | -0.0165*** | -0.0246*** | -0.0171*** | -0.0251*** | -0.0144*** |
| | (0.0005) | (0.0004) | (0.0005) | (0.0005) | (0.0001) | (0.0001) | (0.0001) | (0.0001) |
| Team Paper Citations (ln) | 0.0021*** | -0.0020*** | -0.0011*** | 0.0018*** | 0.0029*** | -0.0006*** | 0.0005*** | 0.0022*** |
| | (0.0002) | (0.0003) | (0.0003) | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) |
| Team Patent Citations (ln) | 0.0424*** | 0.0332*** | 0.0463*** | 0.0245*** | 0.0346*** | 0.0259*** | 0.0370*** | 0.0200*** |
| | (0.0007) | (0.0006) | (0.0006) | (0.0006) | (0.0002) | (0.0001) | (0.0002) | (0.0001) |
| Mean(Dep. Var.) | 0.0655 | 0.0558 | 0.0782 | 0.0353 | 0.0674 | 0.0576 | 0.0806 | 0.0363 |
| Eydx(Open Access) | 0.2201 | 0.0643 | 0.0897 | 0.3498 | 0.2866 | 0.1479 | 0.1871 | 0.3375 |
| Year x Subfield FE | x | x | x | x | | | | |
| Year x Subfield x Journal FE | | | | | x | x | x | x |
| N | 24,786,334 | 24,786,334 | 24,786,334 | 24,786,334 | 22,844,381 | 22,844,381 | 22,844,381 | 22,844,381 |

*Note:* This table reports estimates from linear probability models (STATA REGHDFE). Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

**Table S8.** Open access and the likelihood of patent citation, excluding articles that may not have been open at the time of publication (i.e., bronze and green open access).

| | (1) Patent Citation | (2) Patent Citation | (3) Patent Citation | (4) Patent Citation | (5) Patent Citation |
|---|---|---|---|---|---|
| Open Access | 0.0309*** | 0.0101*** | 0.0240*** | 0.0288*** | 0.0216*** |
| | (0.0002) | (0.0006) | (0.0008) | (0.0026) | (0.0045) |
| Conference Proceeding | | 0.0073 | | | |
| | | (0.0039) | | | |
| Paper References (ln) | | 0.0283*** | 0.0265*** | 0.0250*** | 0.0263*** |
| | | (0.0005) | (0.0002) | (0.0003) | (0.0006) |
| Journal Commercial Impact Factor | | 0.9123*** | | | |
| | | (0.0385) | | | |
| Team Size (ln) | | 0.0088*** | 0.0124*** | 0.0176*** | 0.0158*** |
| | | (0.0004) | (0.0002) | (0.0007) | (0.0016) |
| Team Papers (ln) | | -0.0292*** | -0.0256*** | -0.0357*** | -0.0310*** |
| | | (0.0005) | (0.0002) | (0.0005) | (0.0016) |
| Team Paper Citations (ln) | | 0.0003 | 0.0010*** | -0.0022*** | 0.0017 |
| | | (0.0003) | (0.0001) | (0.0004) | (0.0013) |
| Team Patent Citations (ln) | | 0.0488*** | 0.0391*** | 0.0438*** | 0.0300*** |
| | | (0.0007) | (0.0002) | (0.0003) | (0.0009) |
| Mean(Dep. Var.) | 0.0821 | 0.0821 | 0.0846 | 0.1174 | 0.1177 |
| Year x SubField FE | | x | | | |
| Year x Subfield x Journal FE | | | x | | |
| Year x Subfield x Journal x Inst. FE | | | | x | |
| Year x Subfield x Journal x Inst. x Auth. FE | | | | | x |
| N | 19,726,456 | 19,726,453 | 17,950,881 | 3,059,642 | 1,084,843 |
| Full Information Available | 19,726,456 | 19,726,456 | 19,581,070 | 11,623,279 | 11,623,279 |
| Dropped | 0 | 3 | 1,630,189 | 8,563,637 | 10,538,436 |

*Note:* This table reports estimates from linear probability models (STATA REGHDFE). Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

**Table S9.** Open access and the likelihood of patent citation, distinguishing between publisher-provided and repository-based open access.

| | (1)<br>Patent Citation | (2)<br>Patent Citation | (3)<br>Patent Citation | (4)<br>Patent Citation | (5)<br>Patent Citation |
|---|---|---|---|---|---|
| Publisher-Provided Open Access | 0.0324***<br>(0.0001) | 0.0072***<br>(0.0005) | 0.0157***<br>(0.0005) | 0.0272***<br>(0.0009) | 0.0192***<br>(0.0014) |
| Repository-Based Open Access | 0.0709***<br>(0.0002) | 0.0176***<br>(0.0008) | 0.0175***<br>(0.0004) | 0.0206***<br>(0.0009) | 0.0175***<br>(0.0018) |
| Conference Proceeding | | 0.0031<br>(0.0031) | | | |
| Paper References (ln) | | 0.0294***<br>(0.0005) | 0.0320***<br>(0.0002) | 0.0358***<br>(0.0003) | 0.0371***<br>(0.0005) |
| Journal Commercial Impact Factor | | 0.9763***<br>(0.0366) | | | |
| Team Size (ln) | | 0.0115***<br>(0.0006) | 0.0160***<br>(0.0002) | 0.0203***<br>(0.0006) | 0.0197***<br>(0.0013) |
| Team Papers (ln) | | -0.0325***<br>(0.0005) | -0.0285***<br>(0.0002) | -0.0370***<br>(0.0004) | -0.0298***<br>(0.0014) |
| Team Paper Citations (ln) | | 0.0000<br>(0.0003) | 0.0011***<br>(0.0001) | -0.0025***<br>(0.0003) | 0.0017<br>(0.0011) |
| Team Patent Citations (ln) | | 0.0527***<br>(0.0008) | 0.0422***<br>(0.0002) | 0.0456***<br>(0.0003) | 0.0296***<br>(0.0007) |
| Mean(Dep. Var.) | 0.0910 | 0.0910 | 0.0937 | 0.1224 | 0.1203 |
| Year x SubField FE | | x | | | |
| Year x Subfield x Journal FE | | | x | | |
| Year x Subfield x Journal x Inst. FE | | | | x | |
| Year x Subfield x Journal x Inst. x Auth. FE | | | | | x |
| N | 24,786,337 | 24,786,334 | 22,844,381 | 4,291,667 | 1,535,904 |
| Full Information Available | 24,786,337 | 24,786,337 | 24,639,920 | 15,423,932 | 15,423,932 |
| Dropped | 0 | 3 | 1,795,539 | 11,132,265 | 13,888,028 |

*Note:* This table reports estimates from linear probability models (STATA REGHDFE). Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

**Table S10.** Open-access and the likelihood that scientific ideas are reused in patents.

| | (1)<br>Reuse in Patent | (2)<br>Reuse in Patent | (3)<br>Reuse in Patent | (4)<br>Reuse in Patent | (5)<br>Reuse in Patent |
|---|---|---|---|---|---|
| Open Access | 0.0063***<br>(0.0003) | 0.0047***<br>(0.0004) | 0.0039***<br>(0.0005) | 0.0032+<br>(0.0018) | 0.0105**<br>(0.0037) |
| Conference Proceeding | | 0.0094***<br>(0.0027) | | | |
| Paper References (ln) | | -0.0019***<br>(0.0002) | -0.0007***<br>(0.0002) | -0.0009<br>(0.0007) | -0.0011<br>(0.0014) |
| Journal Commercial Impact Factor | | 0.2752***<br>(0.0128) | | | |
| Team Size (ln) | | -0.0019***<br>(0.0004) | -0.0015***<br>(0.0004) | 0.0028*<br>(0.0013) | 0.0032<br>(0.0036) |
| Team Papers (ln) | | -0.0197***<br>(0.0004) | -0.0176***<br>(0.0003) | -0.0215***<br>(0.0011) | -0.0088*<br>(0.0041) |
| Team Paper Citations (ln) | | 0.0061***<br>(0.0002) | 0.0056***<br>(0.0002) | 0.0042***<br>(0.0010) | 0.0023<br>(0.0037) |
| Team Patent Citations (ln) | | 0.0249***<br>(0.0004) | 0.0190***<br>(0.0002) | 0.0194***<br>(0.0007) | 0.0091***<br>(0.0024) |
| Scientific Ideas | | 0.1476***<br>(0.0024) | 0.1492***<br>(0.0006) | 0.1545***<br>(0.0012) | 0.1520***<br>(0.0023) |
| Mean(Dep. Var.) | 0.1256 | 0.1256 | 0.1287 | 0.1429 | 0.1483 |
| Year x SubField FE | | x | | | |
| Year x Subfield x Journal FE | | | x | | |
| Year x Subfield x Journal x Inst. FE | | | | x | |
| Year x Subfield x Journal x Inst. x Auth. FE | | | | | x |
| N | 6,918,397 | 6,918,390 | 5,924,638 | 835,498 | 254,916 |
| Full Information Available | 6,918,397 | 6,918,397 | 6,883,433 | 4,834,443 | 4,834,443 |
| Dropped | 0 | 7 | 958,795 | 3,998,945 | 4,579,527 |

*Note:* This table reports estimates from linear probability models (STATA REGHDFE). Robust standard errors in parentheses. $^{+}$ $p < 0.10$, $^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$

**Table S11.** Open-access and the hazard of receiving a first patent citation.

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Open Access | 1.560*** | 1.078*** | 1.123*** | 1.1831*** | 1.1307*** |
| | (0.002) | (0.002) | (0.003) | (0.0072) | (0.0122) |
| Conference Proceeding | | 0.997 | | | |
| | | (0.017) | | | |
| Paper References (ln) | | 1.622*** | 1.507*** | 1.4436*** | 1.4425*** |
| | | (0.001) | (0.002) | (0.0036) | (0.0064) |
| Journal Commercial Impact Factor | | 2.420*** | | | |
| | | (0.058) | | | |
| Team Size (ln) | | 1.420*** | 1.254*** | 1.2603*** | 1.2549*** |
| | | (0.002) | (0.002) | (0.0060) | (0.0130) |
| Team Papers (ln) | | 0.620*** | 0.662*** | 0.6544*** | 0.6618*** |
| | | (0.001) | (0.001) | (0.0025) | (0.0083) |
| Team Paper Citations (ln) | | 1.060*** | 0.967*** | 0.8997*** | 0.9616** |
| | | (0.001) | (0.001) | (0.0036) | (0.0116) |
| Team Patent Citations (ln) | | 1.805*** | 1.631*** | 1.6686*** | 1.5901*** |
| | | (0.001) | (0.001) | (0.0045) | (0.0155) |
| **Strata:** | | | | | |
| Year x Subfield | | x | | | |
| Year x Subfield x Journal | | | x | | |
| Year x Subfield x Journal x Inst. FE | | | | x | |
| Year x Subfield x Journal x Inst. x Auth. FE | | | | | x |
| N | 24,660,186 | 24,660,186 | 24,513,953 | 15,321,803 | 15,321,803 |

*Note:* This table reports exponentiated coefficients (hazard ratios) from stratified Cox proportional hazards regression models. Robust standard errors are in parentheses; $^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$

**Table S12.** Open access and semantic similarity between paper and citing patent (in-text citations only).

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|
| | Paper-Patent Similarity Score >= 0.90 | | | Paper-Patent Similarity Score >= 0.925 | | | Paper-Patent Similarity Score >= 0.95 | | |
| Open Access | 0.0388*** | 0.0823** | 0.0318*** | 0.0337*** | 0.0685* | 0.0183+ | 0.0133** | 0.0233** | 0.0105+ |
| | (0.0064) | (0.0267) | (0.0082) | (0.0080) | (0.0270) | (0.0095) | (0.0050) | (0.0081) | (0.0059) |
| Mean(Dep. Var.) | 0.7617 | 0.7772 | 0.7785 | 0.4936 | 0.4930 | 0.4939 | 0.1417 | 0.1348 | 0.1347 |
| Year x Subfield FE | x | | x | x | | x | x | | x |
| Patent FE | | x | x | | x | x | | x | x |
| N | 69,186 | 31,421 | 30,823 | 69,186 | 31,421 | 30,823 | 69,186 | 31,421 | 30,823 |

*Note:* This table reports estimates from linear probability models (STATA REGHDFE). Robust standard errors in parentheses, clustered at the paper level. $^{+}$ $p < 0.10$, $^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$

**Table S13.** Open access and the likelihood of patent citation, distinguishing between patent citations originating from small versus large companies.

| | (1) Patent Citation by Small Company | (2) Patent Citation by Large Company | (3) Patent Citation by Small Company | (4) Patent Citation by Large Company |
|---|---|---|---|---|
| Open Access | 0.0030*** | 0.0051*** | 0.0037*** | 0.0094*** |
| | (0.0001) | (0.0003) | (0.0001) | (0.0002) |
| Conference Proceeding | -0.0012** | 0.0133*** | | |
| | (0.0004) | (0.0033) | | |
| Paper References (ln) | 0.0049*** | 0.0142*** | 0.0066*** | 0.0169*** |
| | (0.0001) | (0.0003) | (0.0001) | (0.0001) |
| Journal Commercial Impact Factor | 0.2161*** | 0.6408*** | | |
| | (0.0085) | (0.0229) | | |
| Team Size (ln) | 0.0008*** | 0.0048*** | 0.0030*** | 0.0091*** |
| | (0.0001) | (0.0003) | (0.0001) | (0.0002) |
| Team Papers (ln) | -0.0059*** | -0.0200*** | -0.0055*** | -0.0182*** |
| | (0.0001) | (0.0004) | (0.0001) | (0.0001) |
| Team Paper Citations (ln) | -0.0000 | 0.0005* | 0.0004*** | 0.0012*** |
| | (0.0001) | (0.0002) | (0.0000) | (0.0001) |
| Team Patent Citations (ln) | 0.0091*** | 0.0310*** | 0.0078*** | 0.0262*** |
| | (0.0002) | (0.0006) | (0.0001) | (0.0001) |
| Mean(Dep. Var.) | 0.0140 | 0.0477 | 0.0143 | 0.0491 |
| Eydx(Open Access) | 0.2282 | 0.1110 | 0.2824 | 0.2026 |
| Year x Subfield FE | x | x | | |
| Year x Subfield x Journal FE | | | x | x |
| N | 24,786,334 | 24,786,334 | 22,844,381 | 22,844,381 |

*Note:* This table reports estimates from linear probability models (STATA REGHDFE). Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

**Table S14.** Open access and the likelihood of patent citation, distinguishing between patent citations originating from startups versus established companies.

| | (1) Patent Citation by Startup (≤3y) | (2) Patent Citation by Est. Comp. (>3y) | (3) Patent Citation by Startup (≤5y) | (4) Patent Citation by Est. Comp. (>5y) | (5) Patent Citation by Startup (≤3y) | (6) Patent Citation by Est. Comp. (>3y) | (7) Patent Citation by Startup (≤5y) | (8) Patent Citation by Est. Comp. (>5y) |
|---|---|---|---|---|---|---|---|---|
| Open Access | 0.0013*** | 0.0046*** | 0.0019*** | 0.0042*** | 0.0014*** | 0.0061*** | 0.0021*** | 0.0057*** |
| | (0.0001) | (0.0002) | (0.0001) | (0.0002) | (0.0001) | (0.0002) | (0.0001) | (0.0002) |
| Conference Proceeding | -0.0004 | 0.0057*** | -0.0004 | 0.0057*** | | | | |
| | (0.0002) | (0.0015) | (0.0003) | (0.0015) | | | | |
| Paper References (ln) | 0.0013*** | 0.0060*** | 0.0019*** | 0.0056*** | 0.0022*** | 0.0089*** | 0.0031*** | 0.0083*** |
| | (0.0000) | (0.0002) | (0.0001) | (0.0002) | (0.0000) | (0.0001) | (0.0000) | (0.0001) |
| Journal Comm. Impact Factor | 0.0893*** | 0.4038*** | 0.1269*** | 0.3874*** | | | | |
| | (0.0037) | (0.0140) | (0.0053) | (0.0135) | | | | |
| Team Size (ln) | -0.0003*** | -0.0000 | -0.0005*** | 0.0000 | 0.0008*** | 0.0039*** | 0.0011*** | 0.0038*** |
| | (0.0000) | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) |
| Team Papers (ln) | -0.0022*** | -0.0114*** | -0.0031*** | -0.0109*** | -0.0022*** | -0.0106*** | -0.0031*** | -0.0102*** |
| | (0.0001) | (0.0003) | (0.0001) | (0.0003) | (0.0000) | (0.0001) | (0.0000) | (0.0001) |
| Team Paper Citations (ln) | 0.0002*** | 0.0013*** | 0.0002*** | 0.0013*** | 0.0004*** | 0.0017*** | 0.0005*** | 0.0017*** |
| | (0.0000) | (0.0001) | (0.0000) | (0.0001) | (0.0000) | (0.0000) | (0.0000) | (0.0000) |
| Team Patent Citations (ln) | 0.0031*** | 0.0159*** | 0.0046*** | 0.0151*** | 0.0029*** | 0.0139*** | 0.0042*** | 0.0132*** |
| | (0.0001) | (0.0004) | (0.0002) | (0.0004) | (0.0000) | (0.0001) | (0.0000) | (0.0001) |
| Mean(Dep. Var.) | 0.0041 | 0.0213 | 0.0059 | 0.0201 | 0.0042 | 0.0220 | 0.0061 | 0.0208 |
| Eydx(Open Access) | 0.3521 | 0.2285 | 0.3456 | 0.2230 | 0.3836 | 0.3052 | 0.3897 | 0.2999 |
| Year x Subfield FE | x | x | x | x | | | | |
| Year x Subfield x Journal FE | | | | | x | x | x | x |
| N | 24,786,334 | 24,786,334 | 24,786,334 | 24,786,334 | 22,844,381 | 22,844,381 | 22,844,381 | 22,844,381 |

*Note:* This table reports estimates from linear probability models (STATA REGHDFE). Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

**Table S15.** Open access and author-inventor distance: estimated geographic distance (in km) between a paper's last author location and the location of the most distant inventor citing that paper.

| | (1)<br>Author-Inventor Distance | (2)<br>Author-Inventor Distance | (3)<br>Author-Inventor Distance | (4)<br>Author-Inventor Distance |
|---|---|---|---|---|
| Open Access | 296.43***<br>(11.16) | 147.87***<br>(11.19) | 240.88***<br>(13.72) | 151.59***<br>(13.11) |
| Conference Proceeding | | -149.64<br>(99.81) | | |
| Paper References (ln) | | 234.53***<br>(5.35) | | 266.17***<br>(6.86) |
| Journal Commercial Impact Factor | | 3676.07***<br>(150.46) | | |
| Team Size (ln) | | 97.26***<br>(10.34) | | 145.78***<br>(10.58) |
| Team Papers (ln) | | -252.18***<br>(8.09) | | -228.82***<br>(7.71) |
| Team Paper Citations (ln) | | 57.29***<br>(8.02) | | 5.58<br>(8.34) |
| Team Patent Citations (ln) | | 322.31***<br>(7.56) | | 322.83***<br>(5.65) |
| Mean(Dep. Var.) | 7195.6 | 7195.6 | 7224.1 | 7224.1 |
| Year x Subfield FE | x | x | | |
| Year x Subfield x Journal FE | | | x | x |
| Last Author Institution Country FE | x | x | x | x |
| N | 1,599,308 | 1,599,308 | 1,310,055 | 1,310,055 |

*Note:* This table reports estimates from OLS models (STATA REGHDFE). Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.