4iP Council
Research Award Winner 2021
Third Place

Online Repositories, search costs and cumulative innovation

by Dr. Thomas Schaper

Postdoctoral researcher in economics at TUM School of Management

Rigorous empirical
research on
intellectual property

4iP Council is a European research council dedicated to developing high quality academic insight and empirical evidence on topics related to intellectual property and innovation. Our research is multi-industry, cross sector and technology focused. We work with academia, policy makers and regulators to facilitate a deeper understanding of the invention process and of technology investment decision-making.

www.4ipcouncil.com

Suggested citation

Schaper, Thomas. Online repositories search costs and cumulative innovation (February 2022). 4iP Council.

Online Repositories, Search Costs and Cumulative Innovation[1]

Thomas Schaper

Summary

Scientific advance is shaped by sequential and complementary intellectual efforts. Efficient access to existing information is, therefore, essential for technical progress. However, R&D spillovers, in particular those from applied research, tend to be clustered and internalized strongly within collaborative networks. Accordingly, access costs to external knowledge increase with the dispersion and disconnect between scientific communities. While the empirical literature studying the elasticity of access costs to existing knowledge on innovation outcomes has predominantly focused on variation in physical accessibility in contexts in which external information was scarce or available only at undue costs, questions regarding the efficiency of search and information retrieval have received significantly less attention.

This paper investigates whether universally accessible, topic-specific repositories of prior art can effectively decrease informational inefficiencies by reducing internal search costs for prior art. Such costs derive from challenges of absorbing and filtering most relevant information out of the sheer mass of scientific knowledge produced, *conditional* on accessibility. Specifically, this study contributes to the literature by disentangling the effect of access from the one of increased visibility of pieces of knowledge arising from the connection to a particular topic, established by the inclusion into a topic-targeted repository.

To investigate this, the paper studies the launch of the International AIDS Patents Database (*AIDS DB*, hereafter) in 1994, the historically first publicly accessible online repository of patent full-texts and images, covering of all inventions related to acquired immune deficiency syndrome. Leveraging the capacities of the new world wide web in the fight against the disease, this repository meant great improvement in the conditions of access and retrieval of information for researchers worldwide racing to develop effective technologies against HIV/AIDS.

This study empirically assesses the marginal impact of the AIDS DB on cumulative inventive search costs relying on publicly available citation data, tracing references to patents in the repository from follow-on inventions in the worldwide patent universe. In order to characterize inventor-level links, the analysis further relies on geo-localized addresses and assignment of inventors to scientific communities based on their prior collaborative activities

[1] JEL Codes: I23, O31, O32, O33. Keywords: knowledge diffusion, information technology, patents. The author thanks Sam Arts, Dirk Czarnitzki, Hanna Hottenrott, Stijn Kelchtermans, Adrián Kovács, Jeroen Mahieu, Tim Meyer, Maikel Pellens, Leonard Treuren, Dennis Verhoeven, Reinhilde Veugelers, Michael Ward, Martin Watzinger, Jesse Wursten and participants in seminars and presentations at KU Leuven, Max Planck, TU Eindhoven, TU Munich, VU Amsterdam, ZEW Mannheim, DRUID, EPIP and the SEI Doctoral Consortium for many useful comments and suggestions. Financial support from the Research Foundation Flanders (FWO) [11B5918N] and the German Research Foundation (DFG) [FOR5234/447967785] is gratefully acknowledged. An extended version of this paper is available from the author's website and at SSRN.

in both basic and applied science in the universe of all USPTO patents and all biomedical scientific articles indexed in PubMed.

To mitigate concerns of positive selection and endogenous treatment, the paper designs an empirical strategy relying on a within AIDS DB counterfactual: Exploiting idiosyncrasies of technology classes assigned to patents not being disease-specific, it estimates the elasticity of internal search costs on cumulative citations on AIDS DB patents for which the link to HIV/AIDS was, arguably, non-obvious to be detected through standard bibliographic search prior to the repository inclusion, against the baseline of a control group of equally indexed patents, comparable in timing, technical content, institutional and scientific prior art background, for which the disease-link was explicit already pre-AIDS DB from the content of their front-pages.

Results show that, after online deposit, cumulative citations to AIDS DB patents without explicit front-page reference to HIV/AIDS subjects increased by around 26% relative to citations to the control group. Effects are particularly pronounced for external spillovers and within citations from inventors working on HIV-related treatments, providing support for the effectiveness of the disease-specific repository in line with the policy objective. In support of these results being causally related to lower search costs, further analyses show that the marginal impact of database deposit was contingent on *how* visible the HIV/AIDS link was on a patent front page. Moreover, citations unlikely to reflect knowledge spillovers - from patents already under examination - were unaffected by database inclusion of cited patents and results are not driven by individual patent examiners. Differences in the rate of follow-up citations persisted for several years after online deposit, even as comprehensive non-disease-specific online patent databases became available. Estimates, additionally, reveal positive differential impact on cumulative citations to patents with intrinsically higher search costs: Private firm patents, recombinant patents, and patents introducing new medical subjects to technology classes. Furthermore, results imply significant second order effects on the visibility and increased patent citation rates to scientific references in patents without front-page links to HIV/AIDS. The paper provides robustness of these results using different estimations, time windows, stricter control group definitions, and impact weighted citation counts.

In a series of additional analyses, the study investigates repercussions on the intensive margin of knowledge spillovers following the establishment of the AIDS DB, comparing changes in the relation between HIV/AIDS patents and their follow-up inventions over time. Relative to non-indexed cited references, estimates provide indication for enhanced knowledge flows, evidenced by significantly higher rates of re-occurrence in citing patents of new words and novel scientific prior art references appearing in AIDS DB patents without front-page link to HIV, in particular among private firm citing inventors. Furthermore, international citations to patents without obvious HIV/AIDS link increased substantially, in particular from academic inventors and public research institutes. Finally, results provide evidence for a strong marginal impact of the AIDS DB on the diffusion of relevant knowledge across scientific community boundaries, in particular across previously disconnected communities, which was entirely driven by increased citations from private sector inventors to patents originating from distant network communities.

These findings primarily inform about how topic-specific repositories can enhance the cumulative impact of new scientific knowledge by reducing search and retrieval costs for researchers. In particular, the study contributes by providing new evidence for the effectiveness of topic-specific online repositories to decrease search costs for follow-up invention and show that these conditions for knowledge accumulation are analogous between open science and applied technology. This, further, speaks to the growing body of prior work on the importance of access to existing knowledge for scientific production, in particular on how access costs to information affect cumulative research impact. Finally, this paper has close antecedents in prior work regarding the role of modern information technologies on knowledge diffusion, spillovers and collaboration in research and development, as well as in the broader literature on the impact of information technologies on economic progress.

Contents

Online Repositories, Search Costs and Cumulative Innovation

Thomas Schaper

Abstract

Efficient access to existing knowledge is essential to technical advance, yet little is known about how access-enhancing institutions shape intertemporal knowledge spillovers. This paper investigates the cumulative technological impact of the CNIDR AIDS Database, the first, disease-targeted, online repository of electronic patent documents, launched in 1994. Tracing references from subsequent patents, results show that the marginal impact of the repository was largest (+30%) among patents for which the established disease-link was previously non-obvious to detect through standard bibliographic search, in line with predictions of stronger reduction of search costs. Further findings suggest that increased visibility and attention to more "hidden" prior art particularly benefited private sector HIV researchers, and was reflected in enhanced diffusion of technological knowledge across scientific community and geographic boundaries.

Keywords: knowledge diffusion, information technology, patents.

JEL Codes: I23, O31, O32, O33.

I. Introduction

Efficient access to existing information is essential for technical progress (Jones, 2003; Mokyr, 2005). This becomes particularly salient in times of acute intensification of disease-targeted research activities, as during pandemics or public health crises. Especially, *new* ideas necessitate the free flow of information and simultaneous experimentation in order to prevail (Rosenberg, 1976; Murray et al., 2016). However, R&D spillovers, in particular those from applied research, tend to be clustered and internalized within collaborative networks (Jaffe, Trajtenberg and Henderson, 1993; Cassiman and Veugelers, 2002; Singh, 2005; Belenzon and Schankerman, 2013; Akcigit, Hanley and Serrano-Velarde, 2020). Accordingly, access costs to external knowledge increase with the dispersion and disconnect between scientific communities.

Most of the empirical literature studying the elasticity of access costs to existing knowledge on innovation outcomes has, either explicitly or implicitly, focused on variation in physical accessibility in (historical) contexts in which external information was scarce or available only at undue costs.[2] Questions regarding the efficiency of search and information retrieval in the numerousness of accessible data points have received significantly less attention from the innovation literature. Furman and Stern (2011) show how specialized biological resource centres, providing access to certified bio-materials and lowering evaluation costs, can help amplifying cumulative impact of scientific discoveries. Similarly, Thompson and Hanley (2018) demonstrate in a randomized experiment that incorporating new scientific topics into Wikipedia articles enhances their diffusion in scientific literature. Zheng and Wang (2020) observe a decline in distant technological search for inventors located in China following the ban of Google's search engine in 2006.

This paper investigates whether universally accessible, topic-specific repositories of prior art can effectively decrease informational inefficiencies by reducing search costs for relevant prior art. Such costs derive from challenges of absorbing and filtering most relevant information out of the mass of scientific knowledge produced. Specifically, I contribute to the literature by disentangling the effect of access from the one of increased visibility of pieces of knowledge arising from the connection to a particular topic, established by the inclusion into a topic-targeted repository.

To investigate this, the paper studies the launch of the International AIDS Patents Database (*AIDS DB*, hereafter) in 1994, the historically first publicly accessible online repository of patent full-texts and images. This database, hosting initially 1,500 U.S. patents, meant great improvement in the conditions of access and retrieval of information for researchers worldwide racing to develop effective technologies against HIV/AIDS. The feature of being a centrally-maintained and expert-validated disease-targeted repository could decrease HIV inventors' search costs significantly for applicable technical prior art, which span a broad

[2] see the (extended) online version of the paper for a discussion of the scholarly literature on knowledge access and cumulative innovation.

range of technology classes and fields and many patent documents were, at first, not clearly recognizable as HIV-related from bibliographic searches.

To empirically assess the marginal impact of the AIDS DB on cumulative inventive search costs, the paper relies on publicly available citation data, tracing references to patents in the repository from follow-on inventions in the worldwide patent universe. The study, further, exploits data from USPTO examination procedures, patent front-page information as well as patent text to determine the specific technical content of inventions. In order to characterize inventor-level links, the analyses rely on geo-localized addresses and assignment of inventors to scientific communities based on their prior collaborative activities in both basic and applied science in the universe of all USPTO patents and all biomedical scientific articles indexed in PubMed.

To mitigate concerns of positive selection and endogenous treatment, the paper designs an empirical strategy relying on a within AIDS DB counterfactual: Exploiting idiosyncrasies of technology classes assigned to patents not being disease-specific, it estimates the elasticity of internal search costs on cumulative citations on AIDS DB patents for which the link to HIV/AIDS was, arguably, non-obvious to be detected through standard bibliographic search[3] prior to the repository inclusion. This approach compares the differential effects of database deposit for these patents to a control group of patents, also indexed in the AIDS DB and comparable in timing, technical content, institutional and scientific prior art background, for which the disease-link was explicit already pre-AIDS DB from the textual content of their frontpages.

Findings show that, after online deposit, cumulative citations to AIDS DB patents without explicit front-page reference to HIV/AIDS subjects increased by around 26% relative to citations to the control group. Effects are particularly pronounced for external spillovers, i.e. on the share of cumulative citations originating from outside applicants' organizations, and within citations from inventors working on HIV-related treatments, providing support for the effectiveness of the disease-specific repository in line with the policy objective. In support of these results being causally related to lower search costs, further results show that the marginal impact of database deposit was contingent on *how* visible the HIV/AIDS link was on a patent front page: For patents mentioning applicability to HIV/AIDS only in the patent abstract the effect was equally positive, compared to those mentioning this in the title, but significantly smaller in size compared to patents without explicit reference. Moreover, citations unlikely to reflect knowledge spillovers - those from patents already under examination - were unaffected by database inclusion of cited patents. Results are, further, not driven by individual patent examiners or within-changes in citation behavior of examiners over time.

[3] i.e. would have required accessing the full-text patent document

Differences in the rate of follow-up citations gradually increased and persisted for several years after online deposit, even as comprehensive online patent databases became available at the end of the 1990s, strengthening my belief that the results are, indeed, attributable to reductions in search costs, provided by the disease-specific link, rather than online accessibility. In line with predictions, estimations reveal strong positive differential effects on cumulative citations to patents with intrinsically higher search costs: Private firm patents, recombinant patent, and patents introducing new medical subjects into technology classes. Furthermore, results imply significant second order effects on the visibility and increased patent citation rates to scientific references in patents without front-page links to HIV/AIDS. The paper provides robustness of findings using different estimations, different time windows, stricter control group definitions, and impact weighted citation counts.

In a series of additional analyses, the paper investigates repercussions on the intensive margin of knowledge spillovers generated among HIV researcher following the establishment of the AIDS DB, comparing changes in the relation between HIV/AIDS patents and their follow-up inventions over time. Relative to non-indexed cited references, results provide indication for enhanced knowledge flows, evidenced by significantly higher rates of re-occurrence in citing patents of new words and novel scientific prior art references appearing in AIDS DB patents without front-page link to HIV, following the launch of the database. These effects were strongest among citing private firm inventors. Further estimations investigate effects on the reach of spillovers generated across geographic boundaries and scientific collaboration networks. After AIDS DB deposit, international citations to indexed patents without previously obvious HIV/AIDS link increased substantially, in particular from academic inventors and public research institutes, while patents with HIV/AIDS references experienced a relative increase in domestic citations. Finally, based on changes in shortest path length between cited and citing inventors in the universe of (author-)inventors and their scientific collaborations, results show evidence for a strong marginal impact of the AIDS DB on the diffusion of relevant knowledge across scientific community boundaries, in particular across previously disconnected communities, which was entirely driven by increased citations from private sector inventors to patents without previously explicit link to HIV/AIDS originating from distant network communities.

This paper intends to make several contributions to the existing literature. Primarily, its findings inform about how topic-specific repositories can enhance the cumulative impact of new scientific knowledge by reducing search and retrieval costs for researchers, adding to findings of prior studies by Furman and Stern (2011) and Thompson and Hanley (2018). In particular, the present study contributes by providing new evidence for the effectiveness of topic-specific online repositories to decrease search costs for follow-up invention and show that these conditions for knowledge accumulations are analogous between open science and proprietary technology.

Findings reported herein, further, speak to the growing body of prior work on the importance of access to existing knowledge for scientific production (Moser and Voena, 2012; Murray et al., 2016; Iaria, Schwarz and Waldinger, 2018), in particular on how access costs to information affect cumulative research impact (Bryan and Ozcan, 2021; Furman, Nagler and Watzinger, 2021; Biasi and Moser, 2021). Here, the paper's results confirm prior evidence that increasing accessibility to relevant prior art impacts subsequent invention and the diffusion of industrially applicable knowledge. Finally, this paper has close antecedents in prior work regarding the role of modern information technologies on knowledge diffusion, spillovers and collaboration in research and development (Agrawal and Goldfarb, 2008; Ding et al., 2010; Forman and van Zeebroeck, 2012; Bertschek, Cerquera and Klein, 2013; Forman and van Zeebroeck, 2019; Zheng and Wang, 2020), as well as in the broader literature on the impact of information technologies on economic progress (e.g. Czernich et al., 2011; Dittmar, 2011).

II. Background

A. External Search and Prior Art Search Costs

The cumulativeness of R&D efforts is well documented in the innovation literature (e.g., Scotchmer, 1991; Galasso and Schankerman, 2015). Intertemporal spillovers from existing knowledge provide critical inputs for the direction of follow-up search, and spur the capacity of future advancement. Being non-rival in nature, these externalities generate social increasing returns to R&D investment (Griliches, 1991; Bloom, Schankerman and Van Reenen, 2013; Jones and Summers, 2020). In applied research, the primary channels, through which spillovers are internalized, rely on direct interaction, as knowledge flows tend to be intrinsically localized and strongly clustered among institutional networks (Jaffe, Trajtenberg and Henderson, 1993; Cassiman and Veugelers, 2002; Singh, 2005). When inventors conduct external search, i.e. attempt to source prior art information from outside their direct networks, important inputs are provided through patent documents and scientific publications. This is particularly given in fields in which these embody specific and valuable codified knowledge, such as chemical (including biomedical) technologies (e.g., Cohen, Nelson and Walsh, 2000; Jaffe, Trajtenberg and Fogarty, 2000; Giuri et al., 2007; Gambardella, Harhoff and Nagaoka, 2011). Recent studies provide ample empirical support for the effective disclosure function of the patent system (e.g., Hegde, Herkenhoff and Zhu, 2020; Baruffaldi and Simeth, 2020; Lück et al., 2020; de Rassenfosse, Pellegrino and Raiteri, 2020).

A precondition for the efficient absorption of external codified knowledge is posed by accessibility. The elasticity of access costs to physical copies of scientific and technical literature on cumulative innovation has been found to be large and significant in prior studies (e.g., Iaria, Schwarz and Waldinger, 2018; Bryan and Ozcan, 2021; Biasi and Moser, 2021; Furman, Nagler and Watzinger, 2021).

However, even conditional on full accessibility to prior art, inventors incur an additional and significant cost in capturing external knowledge spillovers: the search costs arising from the opportunity and mental effort necessary to screen the increasing bulk of information on new advances in a given technical domain, filter and rank these based on the relevance and usefulness for the inventor's specific inquiry, and find ways to integrate them in order to increase the value of a follow-up invention. For these characteristics, prior art search resembles a (non-stationary) sequential search problem with multiple periods (e.g., Pandora's problem in the model of Weitzman, 1979).

Patent systems provide several remedies for searching inventors to facilitate processing the information overload that comes with disclosure on the front page of patent documents. The most important of these are the technology classes an invention is assigned to.[4] However, patent classes have been frequently questioned in the literature with regards to accurately delineating narrow technological fields (e.g., Thompson and Fox-Kean, 2005; Benner and Waldfogel, 2008; Arts, Cassiman and Gomez, 2018). Next to the device of technology classes, most patent offices have, for some time, offered Boolean search facilities to their databases. The usefulness of these is, however, constrained by the fact that bibliographic patent text, in particular in U.S. patents, tends to be written in a highly abstract, legal jargon (e.g., Fromer, 2008; Ouelette, 2012; Lemley, 2012), making key word based searches prone to inaccuracy. These phenomena originate from private firms incentives in disclosing as little concrete information possible in patent documents, in order to conceal the nature of their inventions and protect from imitation (Risch, 2007; Devlin, 2009). Recent concurring evidence from computational linguistics by Kong et al. (2020) shows that private sector patents are significantly less readable than those of universities and public research institutions.

B. The CNIDR AIDS Patents Database Project

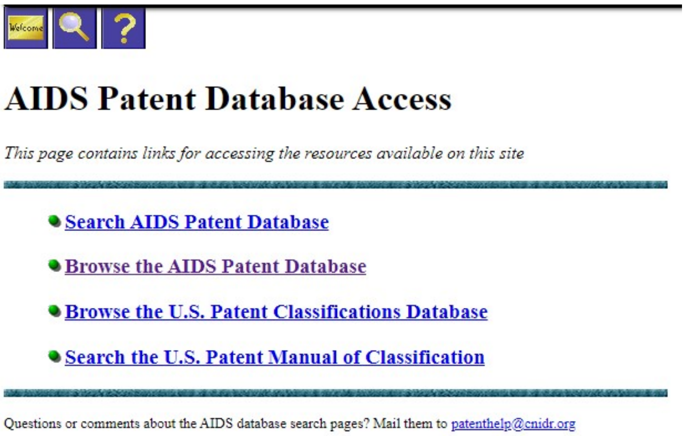
On October 26th 1994, the United States Department of Commerce announced the release of a new database allowing for immediate access to the full text and images of all U.S. patents related to the diagnostic testing and therapeutic treatment of *acquired immune deficiency syndrome (AIDS)*, the disease complex caused by infections with the *human immunodeficiency virus (HIV)*. The AIDS DB was created as a joint effort by the United States Patent and Trademark Office (USPTO), the National Science Foundation (NSF), and the Clearinghouse for Network Information Discovery and Retrieval (CNIDR).[5] Diagnosed HIV-1 infections had dramatically spread since the early-1980s. Following the identification of the new human retrovirus found to be the etiological agent of AIDS in 1983, by late-1994 about 1,500 patents had been issued by the USPTO on technologies relating to HIV/AIDS. These were included in the initial launch version of the database, which was periodically updated

[4] Most importantly the International Patent Classification (IPC), or the USPC and CPC for U.S. patents.

[5] Sources: States News Service, October 26, 1994; Federal Technology Report, McGraw-Hill, November 10, 1994; USPTO Press Release #98-12.

with new patents issued until February 4th of 1997, to host a final total of 2,916 patents. After 1995, the project page also included links to the full-text of HIV/AIDS-related patents issued by the European and Japanese Patent Offices. Figure 1 shows the access page to the AIDS DB which was provided through a link on the USPTO main website.

Figure 1: AIDS DB Access Page, Fall 1996



AIDS Patent Database Access

This page contains links for accessing the resources available on this site

- Search AIDS Patent Database
- Browse the AIDS Patent Database
- Browse the U.S. Patent Classifications Database
- Search the U.S. Patent Manual of Classification

Questions or comments about the AIDS database search pages? Mail them to patenthelp@cnidr.org

Notes: The figure shows a screenshot to the access page to the AIDS Patent Database hosted on the CNIDR server in December 1996. Web-links to the page were prominently included on the home page of the USPTO and the National Science Foundation (NSF). The database included a search form (allowing for keyword, class and boolean search) as well as a browse page, including the full list and links to all hosted patents. Worldwide access to AIDS DB pages was possible with a dial-in modem and a telephone line. The data base included full-text and high-resolution images and drawings of all patents related to HIV/AIDS. Download pages were optimized for small (56k) bandwidths.

Declared objective of the new online database was to connect and increase the informational efficiency between dispersed teams of researchers worldwide.[6] In fact, while all patents are by definition disclosed to the public, until then, researchers interested in technical information involving HIV/AIDS (or any other field) had to search paper files or local computer terminals at the patent office or the 78 depository libraries around the country [7], or rely on commercial services to conduct patent search surveys. The access to full patent documents from outside of library networks was even more difficult; The default remote delivery mode was ordering individual patent copies via mail or fax. With the new online repository, the external search costs to relevant prior art decreased suddenly for HIV/AIDS related knowledge: As discussed in section II.A, it is not-straightforward, even not for skilled inventors, to identify the applicability of a specific patent to a particular disease by inspecting the bibliographic information alone (e.g. provided on the patent front page or through bulletins/ newsletters). In fact, neither the USPC nor IPC patent classifications contain specific classes denoting HIV/AIDS (or other disease)-related inventions, which span a broad range of different technology fields and domains. Moreover, the majority of AIDS DB patents did not

[6] cf. Federal Technology Report, Nov. 10, 1994, p.4

[7]see Furman, Nagler and Watzinger, 2021 for an extensive discussion of the patent library system in the U.S.. In Europe, similar systems were in place in several countries, including the transnational PatLib library program from the European Patent Office.

include any textual reference to HIV/AIDS in title or abstract, which makes their retrieval through key word search comparably difficult.

One and a-half years after launch, the AIDS DB recorded about 2.2 thousand requests per day on average (a total of almost half a million request over a seven-month period), which originated from a large number of connecting points worldwide.[8] For more than four years, the AIDS DB remained the only online repository for full patent information available on the world wide web. It was not until late 1998 that the bulk volume of patents was made available online by the USPTO with their full text and images. Shortly after, in 1999, also the European Patent Office (EPO) launched its online platform.[9] The AIDS DB project was discontinued in March 1999, and all hosted patents were included into the main USPTO database.

III. Data

For the empirical study, I collect data from various sources. To retrieve the exact patents included in the AIDS DB, I web-scrape the archived historical pages of the CNIDR server. I recreate the full database content based on several snapshots of the AIDS DB browse pages containing the full list of US, EPO and JPO patents at different points in time between 1996 and 1998.[10] The archived snapshots also include links to the individual patent view pages, allowing to verify that the listed patents were indeed deposited with full text and images in the database (see Figures 8 and 9 in Appendix .B).

While I know for each patent the database status at a certain point in time, the exact inclusion date is not recorded. Based on the most recent patents in each recorded snapshot, however, I infer the average grant-to-database lag to be of 1-3 months, which is in line with historical information from the USPTO about the currentness of patents included in the database. I am able to retrieve the patent numbers of all U.S., European and Japanese patents that were deposited until February 1997.

I link the retrieved AIDS DB patent numbers to comprehensive information on the universe of patents worldwide in the EPO Patstat database (version spring 2018). Specifically, for each patent worldwide, this source provides information on filing, priority and publication date, documents part of the same international patent family, titles and abstracts, IPC technology classes and fields (based on Schmoch, 2008), raw inventor and assignee addresses, assignee sectors, as well as prior art references and citation links to all other patents. I supplement these data with specific information for U.S. patents concerning details on the patent prosecution process, namely examiners and examining art units (provided by Graham, Marco and Miller, 2015), and assigned USPC patent classes (Marco et al., 2015). To disambiguate

[8] Source: CNIDR Web Server Statistics Dec 5 1996, accessed online here on June 11th 2020.

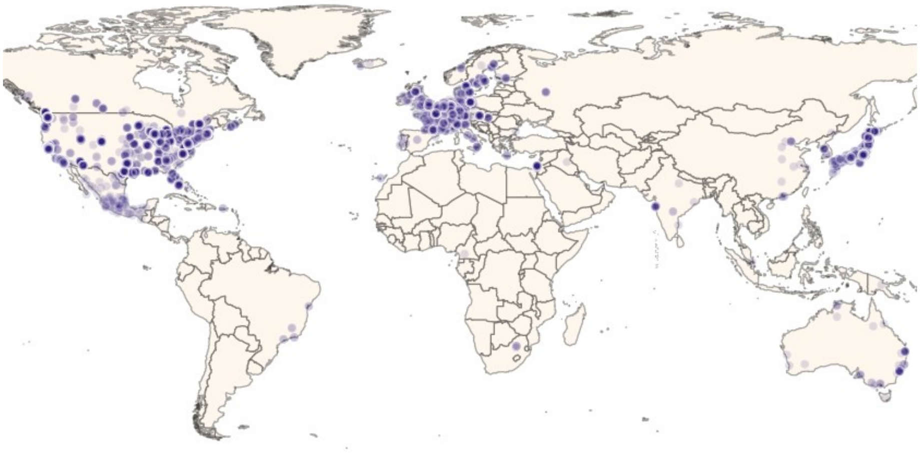
[9] Sources: Press releases/ historical archives of the USPTO and EPO websites.

[10] The chronologically first available snapshot containing the comprehensive database dates back to June 26th 1997 and was accessed here on June 11th 2020.

inventor identities and geo-locations for U.S. and European patents, I extensively rely on the data sets from Li et al. (2014) and Morrison, Riccaboni and Pammolli (2017). I assign patent locations to states, regions and metropolitan areas worldwide using geospatial boundary files provided by the United States Census Bureau, Eurostat and the OECD. Further, I rely on the disambiguation of Marx and Fuegi (2020) to identify links to scientific publications referenced in U.S. patents. Ex-ante indicators of technological novelty are obtained from Verhoeven, Bakker and Veugelers (2016). To retrieve knowledge flows associated with the re-use of new keywords, I use the list of stemmed keywords in U.S. patents from Arts, Cassiman and Gomez (2018). Finally, I determine firm self-citations based on Bureau Van Dijk *Orbis* Intellectual Property Data, linking patents worldwide to consolidated ultimate owners.

In order to detect the broader scientific communities in which AIDS DB inventors are embedded in, I trace their publishing activities in the universe of scientific articles in the MEDLINE database indexed in PubMed.[11] For this purpose, I start from the author name disambiguation of all authors in PubMed, provided by Smalheiser and Torvik (2009) and Torvik and Smalheiser (2009), and links to their U.S. patents from Torvik (2018). Subsequently, I establish the link between author and inventor identities using a within-patent probabilistic matching procedure based on author-inventor name strings.[12]

Figure 2: Geographic Dispersion of AIDS DB Inventors



Notes: Shown locations are individual inventor addresses from all AIDS DB patents deposited from Oct 1994 until Feb 1997. Opacity grades indicate intensity of patenting activity. Geo-coded inventor addresses are provided by Morrison, Riccaboni and Pammolli (2017). Coordinates are geo-mapped using QGIS.

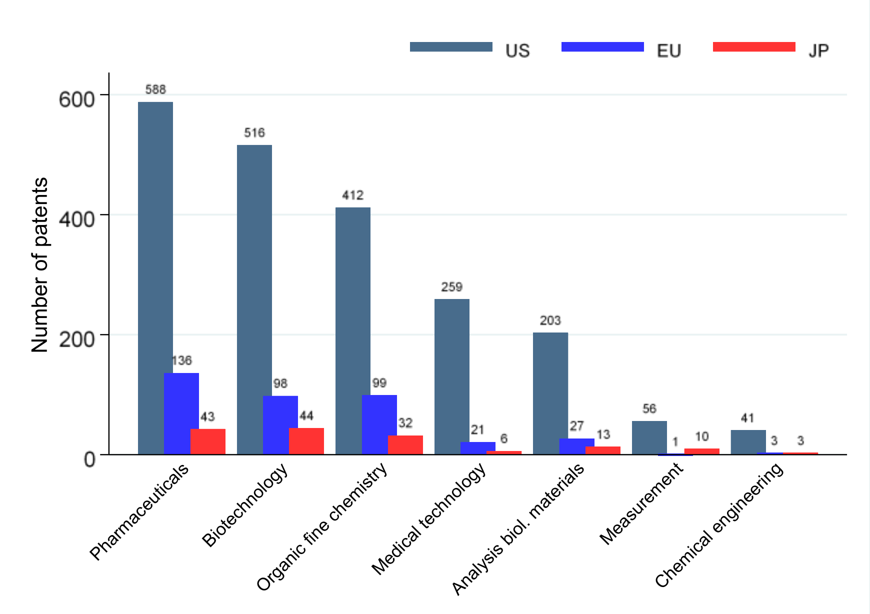
[11] PubMed data is publicly available and can be accessed here: https://pubmed.ncbi.nlm.nih.gov/

[12] I use a Jaro-Winkler similarity algorithm with varying acceptance thresholds. Random sample validation (N=200) of the matching approach yields a precision of 95,4% and recall of 97,8%.

The final assembled database contains detailed information about all patents deposited in the AIDS DB, and a large pool of comparison patents from the universe of similar technologies. It records all citation links to prior art and follow-up inventions, allowing to quantify the technological origins and cumulative impact of all patents included. Further, in order to evaluate the relation between cited and citing inventions in geographic and intellectual space, the data contains rich information about all inventors and assignees with their precise locations, prior patenting and publishing histories and embeddedness in the HIV/AIDS researching scientific communities.

Figure 2 shows that HIV/AIDS-related technologies were developed almost exclusively in the Western hemisphere. When looking at split counts of patents across main geographic areas and technology fields (see Figure 3), however, it becomes evident that the U.S. were by far the leading geographic area, accounting for the largest number of patents in each field. The AIDS DB hosted HIV/AIDS-related patents from a total of 14 distinct technology fields, covering a broad range of inventive domains.

Figure 3: Frequencies of AIDS DB Patents by Technology Fields and Geographic Origin



Notes: Bars show patent counts across technology fields by geographic areas. Counted are all patents deposited in the AIDS DB from Oct 1994 until Feb 1997. Fixed geographic areas are determined by the most represented geographical area among inventor locations of a patent (a random draw is taken in case of multiple). Technology fields are based on Schmoch (2008). Each patent is assigned to its most represented field (random draw in case of multiple). Seven further, less represented, fields are omitted in the graphic.

While geographically clustered, the HIV/AIDS inventor community was highly proliferated into small networks. Moreover, my data show that HIV/AIDS-technology research was strongly intertwined with advances in basic science; Many inventors listed on patents in the AIDS DB also ranked among the leading and most impactful fundamental science researchers in the area of the disease, as depicted in Table 8 in online Appendix .A.

IV. Effects on Cumulative Patent Citations

A. Empirical Strategy

The fundamental difficulty associated with identifying the marginal impact of an online repository, like AIDS DB, on cumulative invention arises from the need to isolate the intrinsic impact components of the embedded knowledge itself from the one of the access-enhancing institution. This study solves the endogenous link problem by adopting a *within* AIDS DB comparison, which accounts for all unobserved factors related to selection into the database. In order to disentangle the effects of increased visibility (attention effect) from the one of better access, it empirically estimates changes in cumulative impact across deposited patents as a function of additional up-front information on the technical content revealed by the explicit association with HIV/AIDS.

To determine the degree to which inclusion in the AIDS DB might have led to a shock in search costs, the analysis leverages knowledge about the conditions of external prior art search before the establishment of the AIDS DB. To detect whether a patent makes a clear reference to HIV/AIDS, I query all titles and abstract of AIDS DB patents for keywords of medical subject terms relating to HIV/AIDS, as defined by the National Library of Medicine.[13] The assumption behind this approach is that the likelihood that inventors would have detected a relevant HIV/AIDS-related patent out of a list of bibliographic information of numerous patents will be higher if a given patent makes a clear front-page reference to the disease. Accordingly, the inclusion of a patent into the AIDS DB likely entailed a stronger reduction in search costs for patents not making front-page references to HIV/AIDS, compared to those making them, increasing the visibility of the former for related prior art search, conditional on same (online) accessibility. I subsequently divide AIDS DB patents into two categories: With vs. without front-page reference.

While this within-comparison solves the positive selection of inclusion of patents into the repository, the criterion for unbiased inference requires, henceforth, these two groups to be comparable on all characteristics relating to cumulative diffusion except for the treatment status (”no reference”). To avoid comparing patents on different types of technologies within broader technological fields, which might have different dynamics of diffusion, I condition ”no reference” and control group (”with reference”) patents to be examined in the same art unit – the most granular inter-organizational units in the examination process at the USPTO, consisting of teams of patent examiners narrowly specializing in a particular technology.[14]

[13] see: https://meshb.nlm.nih.gov/. A full list of queried subject terms can be found in the extended online version of the paper.

[14] Given this constraint in data availability, I only consider U.S. patent family members. See Righi and Simcoe (2019) for an excellent discussion of the organization of art units at the USPTO

Table 1: Summary Statistics, Patents Without vs. With Front Page Reference to HIV/AIDS

Within AIDS DB sample	No reference		With reference		Diff
	Mean	SD	Mean	SD	p-val.
Yearly patent family citations at AIDS DB deposit	1.29	2.77	1.07	2.17	0.13
Generality index	0.46	0.25	0.44	0.25	0.16
Share breakthroughs (top-5%)	0.09		0.09		0.81
Share novel inventions	0.25		0.22		0.21
Share introducing new words	0.35		0.31		0.14
Share new-to-class medical subjects	0.37		0.34		0.26
Number of patent references	9.25	9.57	7.56	8.12	0.00
Share with scientific reference	0.90		0.90		0.95
Number of scientific references	13.90	22.91	11.48	13.66	0.03
Number of inventors	3.06	2.04	2.91	1.88	0.18
Share of new inventors	0.17		0.18		0.64
Share with author-inventors	0.94		0.95		0.53
Number of author-inventors	2.62	1.90	2.49	1.65	0.20
Patent family size	6.98	7.21	5.47	6.44	0.00
Share private firm patents	0.64		0.64		0.97
Assignee prior patent families	2.56k	11.15k	2.55k	5.29k	0.98
DB-to-publication lag (months)	19.07	21.46	18.74	21.23	0.79
DB-to-application lag (months)	55.81	27.94	56.15	28.47	0.83
Number of patents	870		497		
Number of technology fields	11		8		
Number of examining art units	33		33		

Notes: Row (1) reports the group mean and standard deviation for yearly patent family citations to AIDS DB patents without vs. with front-page reference to HIV/AIDS for year t0 relative to AIDS DB deposit. Inventor and applicant self-citations are removed from the counts. The following rows report ex-ante time-invariant characteristics. Control group patents consist of AIDS DB patents ”with reference” to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired ”no reference” patents. Technology fields are based on Schmoch (2008). Sample observations are weighted according to Iacus, King and Porro (2012). Column (6) reports p-values from two-sample t-tests with unequal variances for differences in sample means. The data were collected by the author and combine web-scraped information from the CNIDR server archive with data from the USPTO, PATSTAT, PubMed, BvD Orbis and several disambiguations (links) between them (see Section III for details).

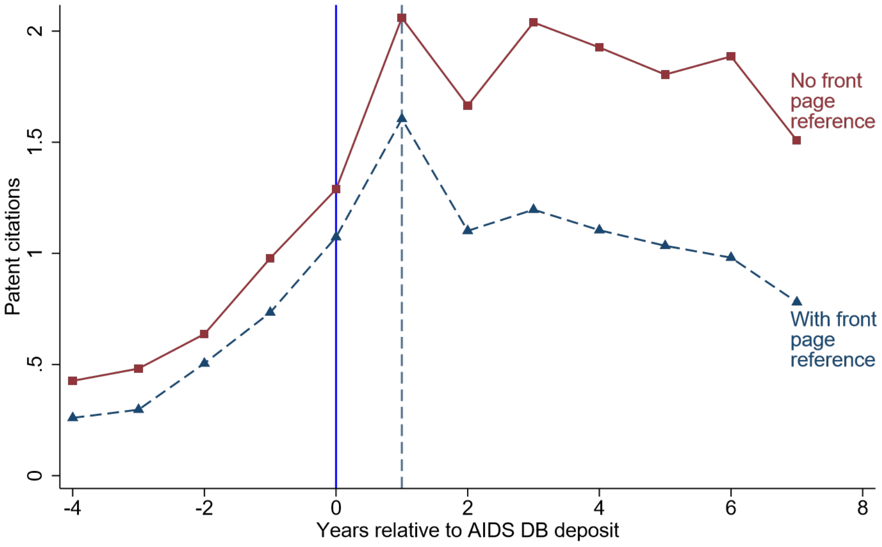
Next, within each AIDS DB patents - art units stratum, I further subdivide patent pairings based on whether they are assigned to a private firm or public institution, and whether they make prior art references to basic science, which have been widely shown to have significant influence on cumulative use and breadth of impact of technologies (Mansfield, 1995; Narin, Hamilton and Olivastro, 1997; Ahmadpoor and Jones, 2017). Finally, I pair patents, within these bins, based on coarsened invention filing and patent publication dates, and apply the weights of Iacus, King and Porro (2012) to ensure balanced estimation.[15] By this, all factors relating to the timing of invention, disclosure and online deposit are kept constant across the

[15] For ease of sample construction, I again assign a unique database deposit date to all patents in the same matched group. I based the unique date on the most frequent occurring, and earliest in case of multiple.

sample groups. Several examples of patent pairs with vs. without front-page references are discussed extensively in online Appendix .C.

To measure the realization of spillovers, the analysis predominantly relies on patent citations to AIDS DB patents as proxies. I use patent-level panel data to quantify the marginal effect of the AIDS DB on the cumulative rate of citations. Specifically, I create a data set with yearly observations of citation counts for each patent in all years following its initial filing date. In line with prior literature [16], I remove inventor and applicant self-citations from the counts, as those do not reflect spillovers from external search. I count a citation as a cumulative spillover with timing of the initial filing date of the citing patent. As initial filing date, I consider the priority date, for those patents with international priority, first or provisional filings, and the application filing date, for patents that are continuations or divisions of prior applications.

Figure 4: Group Means *Within* AIDS DB Comparison Yearly Patent Citations



Notes: The figure plots trends in group means across AIDS DB patents without vs. with front-page reference to HIV/AIDS. Control group patents consist of AIDS DB patents "with reference" to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired "no reference" patents. The *y*-axis scale reports levels of yearly patent family citations to patents in sample. Inventor self-citations are removed from the counts. The x-axis depicts years relative to online deposit (0). The dashed vertical line (1) indicates a 1-year lag of the database treatment, relative to deposit. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *PubMed*, *BvD Orbis* and several disambiguations (links) between them (see Section III for details).

Table 1 reports summary statistics for the within-AIDS DB matched sample. The strict applied selection criteria allow to pair 1,367 AIDS DB patents.[17] A total of 11 technology fields and 57 art units are represented in the sample, suggesting that the latter are significantly more granular in technological scope then 3-digit classes. Sample patents are highly comparable

[16] e.g., Jaffe, Trajtenberg and Henderson (1993); Thompson (2006); Singh and Marx (2013)

[17] By this, my estimation sample covers about half of all originally deposited patents in the AIDS DB. Inference is limited to this subset.

also on a broad range of relevant ex-ante patent-level characteristics, with no significant differences across the two groups.

Figure 4 plots almost perfectly parallel trends between ”no reference” and ”reference” patents in the relative periods until the AIDS DB deposit date, which is a necessary condition for inference of an average treatment effect on the treated, and suggests that the control group is well selected.

I then compare within-patent changes in differences in citation rates across groups after AIDS DB deposit in a generalized difference-in-differences framework by estimating the following regression equation:

$$(1) \qquad Y_{it} = \beta_1 * no\ reference_i \times post_{t(i)-1} + patentFE_i + yearFE_{t(i)} + \varphi_{y(i)} + \theta_{f(i)y(i)} + \epsilon_{it}$$

where *i* indexes patents, *t* indexes relative years to AIDS DB deposit, *y* indexes calendar years, and *f* indexes technology fields. The dependent variable measures the number of citations per relative year to deposit for each AIDS DB patent and, analogously, per relative year to deposit of the matched AIDS DB patent for each control patent. The coefficient β_1 measures changes in citation rates to AIDS DB patents without front-page reference to HIV/AIDS, after deposit, relative to the group of patents making such references, which are the excluded reference category. The interacted $post_{t(i)-1}$ indicator denotes the one-year lagged post-deposit status.[18] The regression model includes a full set of patent fixed effects. These control for all permanent differences across patents affecting the incoming citation patterns, for instance, the quality and complexity of an invention, its geographic origin, institutional context or technological field. Note, that a group fixed effect (e.g., an indicator for all ”no reference” patents) is omitted from the specification as it would be perfectly collinear with the sum of patent fixed effects of all patents in that group.

The model also includes fixed effects for relative years to the AIDS DB deposit date. These account for dynamic changes in the rate of citations over the life cycle common to all patents. Given that some patents enter the sample (i.e., are applied for and granted) several years before the AIDS DB launch, while others are deposited almost immediately after grant, this prevents results to be disproportionately driven by, e.g., more recently granted patents. Note, that the sum of pre-treatment and post-treatment relative year fixed effects is collinear to a $post_{t(i)-1}$ period indicator, which is therefore also omitted from the specification. To control for the confounding influence of shocks possibly affecting citation rates over time in the overall economy or the patent system (e.g., the enactment of the TRIPS Agreement in 1995), the regression further includes a full set of calendar year fixed effects (captured by the parameter $\varphi_{y(i)}$). Finally, I include linear field-year trends ($\theta_{f(i)y(i)}$) to control for idiosyncratic

[18] A one-year lag to assess the manifestation of technology spillovers is established also in prior work, e.g. Bloom, Schankerman and Van Reenen (2013).

variation in productivity of specific technology fields, for example, in up-rising biotechnology in the mid-1990s.

I estimate regression (1) on a symmetric sample window of five years preceding and five years following the switching of the $post_{t(i)-1}$ indicator. I report regression results of cumulative citation models, primarily, as OLS estimates.[19] For this, I standardize the number of yearly citations to mean zero and standard deviation one within technological fields.[20] This makes effect sizes on citation rates comparable across fields, despite the fixed functional form of the model, and avoids well-studied problems arising from the use of log-linearizations on distributions inflated with many zeros (e.g., Silva and Tenreyro, 2006). For comparison and robustness, however, I complement all results with estimates from Poisson pseudo-maximum likelihood models (Silva and Tenreyro, 2011). Given the panel structure of my data, and the common concern of possibly serially correlated regression residuals leading to deflated OLS standard errors and resulting over-rejections of the null hypothesis (Bertrand, Duflo and Mullainathan, 2004), I cluster all standard errors at the patent-level.

B. Effects on Patents Without vs. With Front-Page References to HIV/AIDS

Table 2 reports the econometric results from the estimation of regression (1) on the within AIDS DB sample. In column (1) of Table 2, I first estimate the model without field-specific linear trends. Starting from one year after deposit, patents without front-page reference to HIV/AIDS received on average .14 standard deviations more in cumulative citations relative to control group patents with front-page references (significant at the 1% level). Compared to the pre-deposit mean of citations, this implies a relative increase of .35 citations per year (about +29%) for the average ”no reference” patent in the sample. The effect is slightly smaller (+.12 standard deviations, +26%) when including field-specific time trends, in my preferred specification in column(2) (but equally significant at 1%), suggesting these to explain about one tenth of the dynamic differential.

I check the robustness of this finding across several alternative models: One caveat regarding the validity of these results might arise due to patents without specific references to HIV/AIDS covering more ’general’ technologies, which intrinsically experience a broader diffusion outside of the HIV-research community, possibly explaining the positive delta. On the other hand, effects could also be driven by new entry of inventors with more diverse backgrounds. To investigate this, in columns (3) of Table 2, I re-estimate the model considering only follow-up citations originating from the community of established HIV/AIDS inventors, identified as those appearing on patents indexed in the original AIDS DB. The point estimate of the treatment parameter in column (3) indicates that effects were larger for this sub-group (+28%, significant at the 1% level), however accounting for the vast majority of incoming

[19] as in Furman and Stern (2011); Galasso and Schankerman (2015); Biasi and Moser (2021)

[20] compare Iaria, Schwarz and Waldinger (2018) for a similar approach

citations. Another concern as to which extent my results capture knowledge spillovers from external search might arise from the fact that HIV-research was predominantly conducted by large institutions. Hence, a significant part of the observed effect could be due to within-firm spillovers resulting from, e.g., an intensification of efforts in HIV/AIDS research and not be due to online deposit. In column (4) of Table 2, next to inventor and applicant self-citations, I therefore also remove ultimate-owner level firm self-references from the dependent variable citation counts.[21] Estimates show that the effect is magnified (increase of .15 standard deviations, significant at 1%) when excluding these citations.[22]

Table 2: Effect for Patents Without HIV/AIDS Front-Page References, Within AIDS DB

Dependent variable:	OLS					
Number of patent citations	(1)	(2)	(3)	(4)	(5)	(6)
No reference x post$_{t-1}$	0.135*** (0.044)	0.123*** (0.044)	0.134*** (0.042)	0.150*** (0.044)	0.181*** (0.046)	-0.046 (0.044)
Abstr reference x post$_{t-1}$					0.101 (0.066)	
Patent fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Time/ year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Field time trends		Yes	Yes	Yes	Yes	Yes
DB inventor cites only			Yes			
Excl. firm self-cites				Yes		
In prosecution cites only						Yes
Observations	12,192	12,192	12,183	12,190	12,192	12,020
Number of patents	1,366	1,366	1,366	1,366	1,366	1,360
R^2	0.388	0.389	0.433	0.495	0.389	0.06
Mean at t_0	1.210	1.210	1.178	1.194	1.210	0.016
SD at t_0	2.570	2.570	2.498	2.554	2.570	0.152

Notes: Each column reports parameter estimates of regression (1) for the matched panel of U.S. AIDS DB patents without vs. with front-page reference to HIV/AIDS. The dependent variable measures the yearly number of family citations for years $t-4$ to $t+5$ relative to the one-year lagged online date. Inventor and applicant self-citations are removed from the counts. The reference category consists of AIDS DB patents "with reference" to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired "no reference" patents. The number of citations is standardized to mean zero and standard deviation one within technology fields (fields based on Schmoch (2008)). "DB inventor cites" are citations originating exclusively from HIV/AIDS inventors indexed in the AIDS DB. "Firm self cites" are self-citations at the ultimate owner-level. "In prosecution cites" are citations exclusively from patents already under examination at the time of DB inclusion of the cited patent. Sample observations are weighted based on Iacus, King and Porro (2012). Standard errors are clustered at the patent level. Significance levels: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *BvD Orbis* and several disambiguations (links) between them (see Section III for details).

In column (5) of Table 2, I seek further support for lower search costs driving these results, by splitting up patents in the reference category based on how visible the HIV/AIDS reference

[21] Consolidated firm-level self-citations are based on the *BvD Orbis* firm-patent link.

[22] Unfortunately, the Orbis firm-patent link information is available only for about 60% of firm patents in my sample. I, therefore, do not rely on these in the preferred specification.

was on their front page. Precisely, I distinguish between patents making a textual HIV/AIDS reference in the abstract section only, and those making a reference already in the document title. The idea behind this is that retrieving information from patent abstracts required inventors to engage with a patent document already substantially more than just scrolling through lists of newly granted patents (including only titles, inventors, and classes) when searching for HIV-related prior art, implying somewhat higher search costs, and increasing the likelihood of overlooking a relevant patent making a reference only in the abstract. Hence, I add the category interaction of these patents to the model, comparing effects for lower search costs in cascading manner, relative to the background rate of patents with HIV/AIDS reference in the title. Results show the largest increase in citations for the ”no reference” category (.18 standard deviations more, significant at 1%), while effects are also positive, and about 45% smaller in size (although not significant below the 10% level in the OLS estimation). These patterns are widely in line with a reduction of search costs as mechanisms driving my results.

Another alternative explanation for the observed pattern might simply be that patent examiners became more likely to add references to certain AIDS DB patents, as their active involvement in the assembly process likely increased their attention to them as well.[23] I evaluate the severeness of this concern, in column (6) of Table 2, by re-estimating regression (1) *only* counting citations given from patents that were already under prosecution at the time of online deposit of the cited AIDS DB patent, i.e. filed before and granted after the AIDS DB deposit date.[24] These citations are very likely to be given by examiners rather than by the applicants.[25] They also cannot reflect knowledge spillovers from external search through online access to the AIDS DB, as patent applications were already pending and, accordingly, the inventive search process must have been terminated at the time of online deposit. The point estimate of β_1 in column (6) for changes in citations added during prosecution after database deposit across ”no reference” and ”with reference” patents tends slightly negative, but is highly insignificant. This suggests that the launch of the AIDS DB had no influence on citation practices of patent examiners, at least not within database indexed patents.

I further explore the possibility of confounding influences originating from patent examiner by excluding from yearly citation counts those references made by patents inspected by USPTO examiners who accounted for a large number of citations to AIDS DB patents after launch of the database. By this, I attempt to rule out the competing explanation that higher citation rates to AIDS DB patents without front-page references to HIV/AIDS could have been driven by a few very actively citing examiners whose attention was drawn towards these previously less visible inventions. The corresponding estimates are reported in the extended

[23] Examiner citations typically account for about 40% of citations included in U.S. patent documents. Unfortunately, the precise information about examiner-added citations is given only for U.S. patents granted after January 2001 and, therefore, not available for the vast majority of citing patents in my sample.

[24] In this case, deviant from my standard approach, I consider as citation date the grant date of a citing patent, which is arguably closest to the examination moment.

[25] see Arora, Belenzon and Lee (2018) for a similar approach

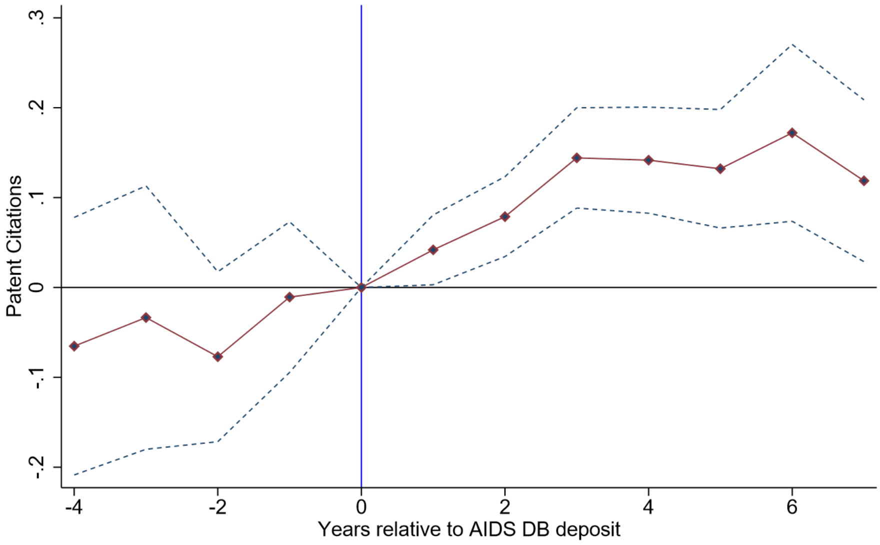
online version of the paper: When excluding citations from patents under review by the ex-post 10 most citing examiners (out of 1,016 total citing examiners), representing the top 1% and accounting for about 20% of total yearly citations to AIDS DB patents, the estimated citation premium is qualitatively robust and consistent with the main result in Table 2. The estimated yearly citation delta is even considerably higher when excluding the 50 most frequently citing examiners. Taken together, these sensitivity checks mitigate concerns that (changes in) examiner behavior would have significantly affected the observed increase in citation rates to ”no reference” patents in the AIDS DB, and support my interpretation of the estimated effect as elasticity of (a reduction in) search costs on the inventors’ side.

In online Appendix .A, I show further robustness of the entirety of the findings in this Section with quantitatively largely unchanged results: In Table 10 (online Appendix .A), I re-estimate all models with Poisson pseudo-maximum likelihood, yielding slightly larger effect sizes. In Table 11 I match and include only AIDS DB patents which are observed throughout all sample years (in order to form a balanced panel from $t - 4$ to $t + 5$), and show that results are robust and larger in magnitude for this sub-set. To address concerns that patent citations might exhibit exponential rather than linear cumulative growth rates, and accordingly small initial differences could result in large differences over time, fully or partially explaining the effect in the post-period, in Table 12 (online Appendix .A), I show robustness and substantially larger estimates for a sub-set of ”no reference” and ”with reference” patents additionally matched on yearly pre-period citation levels (and, accordingly, trends). Finally, in Table 13 (online Appendix .A), I provide alternative results for effects on impact weighted forward citations, suggesting real economic effects behind the observed increased knowledge flows.

C. Timing of Effects

To investigate the timing of the aggregate attention effects, I re-estimate regression (1) with yearly coefficients, by interacting the ”no reference” indicator with a set of individual year dummies for $t - 4$ to $t + 7$ relative to the database date (excluding the year of deposit as reference year). Figure 5 plots the corresponding point estimates within 95% confidence intervals. There are no significant differences estimated between citation trends of ”no reference” and ”with reference” group patents in the years prior to database inclusion, suggesting that differences in pre-trends cannot explain the results. On the other hand, the figure shows a steep relative increase in the rate of citations to ”no reference” patents in the years following their online availability, setting in highly significantly after one year, and reaching a plateau around the third year post-deposit and another subsequent peak after 6 years.

Figure 5: Yearly Effect for Patents Without HIV/AIDS Front-Page Reference



Notes: The figure plots parameter estimates from regression (1) with yearly coefficients for $t-4$ to $t+7$ relative to online deposit for the matched panel of U.S. AIDS DB patents without vs. with front-page reference to HIV/AIDS. The dependent variable measures the yearly number of family citations (inventor and applicant self-citations excluded). The year of deposit is omitted from the regression. The reference category consists of AIDS DB patents ”with reference” to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired ”no reference” patents. Sample observations are weighted based on Iacus, King and Porro (2012). 95% confidence intervals are based on clustered standard errors. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *BvD Orbis* and several disambiguations (links) between them (see Section III for details).

Figure 5 also indicates a stable and persistent effect over time, as estimated differences, first, gradually increase and seem decline only towards the very end of the sample (indicating the years 2000-2003), following the discontinuation of the AIDS DB. Noticeably, estimated yearly differentials seem to remain mostly unaffected by the launch, first, of the comprehensive bibliographic online database of the USPTO (in 1997, which corresponds to year $t+3$ for the initial cohort of patents uploaded in 1994) and, second, the full-text and images online catalogue including all U.S. patents (in 1998) and EPO Espacenet (1999). In fact, these newly launched databases mostly levelled out differences in external access costs, but only to a lesser extent search costs, as, unlike the AIDS DB, these were not disease-specific repositories. This observation provides further support for the believe that the observed effects are, indeed, caused by a unilateral shock to search costs for ”no reference” AIDS DB patents rather intrinsic quality differences or online electronic copy accessibility provided by means of the online repository.

D. *Differential effects for patents with intrinsically higher search costs*

The paper further investigates whether AIDS DB indexing particularly benefited the cumulative diffusion of patents that are associated with intrinsically higher search costs: Private firm patents, recombinant patents, and relying on medical subjects new to a technology class.

Table 3 shows estimation results for regression (1) as triple-differences for heterogeneity of marginal impact of the online repository for split-samples of these patents.[26] Columns (1)-(2) of Table 3 report differential effects for the group of private firm patents. As discussed in Section II.A, these patents are subjected to adverse incentives of private firms against information disclosure towards rivals and particular prone to attempt to conceal the nature of the underlying inventions. Accordingly, I expect the shock to search costs from the disease-specific link to have been disproportionately higher for these patents. Results in Table 3 indicate that corporate assignee patents without front-pages references to HIV/AIDS received .17 standard deviations in citations more after AIDS DB deposit than ”no reference” patents from public research institutions (e.g. universities, government research institutes, hospitals, etc.). Relative to the average patent in the sample this implies an additional increase of 35% (significant at the 10% level). Column (2) of Table 3 investigates robustness of this finding counting only incoming citations from the group of HIV/AIDS inventors, indexed in the AIDS DB. The estimated coefficient is equally positive, yet smaller in size and not significant below the 10% level. For both cases of citation counts, the relative changes after database deposit for private firm patents with front-page references to HIV/AIDS are close to zero and not statistically significant.

Next, I compare differential effects for patents departing from existing trajectories of search, making them ceteris paribus more difficult to retrieve, e.g. if inventors aim to assess the relevance of new advances by inspecting the prior art cited.[27] First, I evaluate marginal impacts on recombinant (novel) patents (e.g., Fleming, 2001). I identify these as making new combinations of previously uncombined technology classes in the prior art they cite, using the measure suggested by Verhoeven, Bakker and Veugelers (2016) at the IPC-group level (IPC-6). Columns (3)-(4) of Table 3 show that effects on novel patents without front-page references to HIV/AIDS were .25 standard deviations larger compared to non-novel ”no reference” patents (significant at the 5% level, + 54% relative to baseline) and that this pattern was robust for citations incoming from HIV/AIDS inventors (significant at the 10% level, given slightly reduced effect size).

Novel patents *with* frontpage reference to HIV/AIDS, at the same time, did not exhibit a significantly different change in citations relative to the background rate of non-novel ”with reference” patents. Finally, I assess heterogeneity in impact for patents making scientific prior art references to articles in PubMed which were indexed in MeSH terms previously not linked to the technology class of the citing patent. Coefficient estimates in columns (5)-(6) of Table 3 show similar patterns for ”no reference” patents making such new connections to scientific underpinnings, significant at the 10% level in both models, while the differential effect of database deposit for this group was opposite in patents with front-page HIV/AIDS references.

[26] This econometric specification compares changes in cumulative citations of, e.g., *private firm* patents with frontpage references to HIV/AIDS to *private firm* patents without such references, and analogously for the sub-groups of novel patents and patents linking to new medical subjects.

[27] Prior art references are also included on the front page of patents documents.

These patterns are in line with my predictions of these groups of patents experiencing larger marginal impact from the disease-specific link established by AIDS DB indexing, given previously higher retrieval costs.

Table 3:Differential Effects for Patents with Intrinsically High Search Costs, Within AIDS DB

Dependent variable:	*Private firm patents*		*Novel patents*		*New-to-class MeSH*	
Number of patent citations	(1)	(2)	(3)	(4)	(5)	(6)
No reference x post$_{t-1}$ x *cat*	0.166* (0.086)	0.119 (0.086)	0.252** (0.109)	0.214* (0.109)	0.291* (0.176)	0.291* (0.156)
Post$_{t-1}$ x *cat*	-0.036 (0.072)	-0.008 (0.068)	-0.101 (0.070)	-0.092 (0.076)	-0.304* (0.148)	-0.278** (0.129)
Main category interactions	Incl	Incl	Incl	Incl	Incl	Incl
Non-new MeSH interactions					Incl	Incl
Patent fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Time/ year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Field time trends	Yes	Yes	Yes	Yes	Yes	Yes
DB inventor cites only		Yes		Yes		Yes
Observations	12,192	12,183	12,192	12,183	12,192	12,183
Number of patents	1,366	1,366	1,366	1,366	1,366	1,366
R^2	0.441	0.433	0.441	0.434	0.442	0.434
Mean at t_0	1.210	1.178	1.210	1.178	1.210	1.178
SD at t_0	2.570	2.498	2.570	2.498	2.570	2.498

Notes: Each column reports parameter estimates of regression (1) split up as triple-differences for heterogeneity of effects on patents associated with intrinsically higher search costs in the matched panel of U.S. AIDS DB patents without vs. with front-page reference to HIV/AIDS. The $post_{t-1}$ parameter captures relative changes in citations to patents with front-page reference to HIV/AIDS in each split-sample category. Main category × $post_{t-1}$ interactions are included in all models. The dependent variable measures the yearly number of family citations for years $t-4$ to $t+5$ relative to the one-year lagged online date. The number of citations is standardized to mean zero and standard deviation one within technology fields (fields based on Schmoch (2008)). Inventor and applicant self-citations are removed from the counts. The reference category consists of AIDS DB patents ”with reference” to HIV/AIDS, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired ”no reference” patents. In columns (1)-(2) report differential effect estimates for the sub-sample of private firm patents. Columns (3)-(4) report corresponding estimates for novel patents. As ”novel” are considered patents making novel combinations of technological prior art classes (IPC-6 level), following Verhoeven, Bakker and Veugelers 2016. Columns (5)-(6) show heterogeneous effects for the split-sample of patents referencing scientific prior art in medical subject terms (MeSH) that have not been previously linked to their respective technology class. A full set of interactions for patents making non-new-to-class medical subjects references are included. ”DB inventor cites” are citations originating from HIV/AIDS inventors indexed in the AIDS DB. Sample observations are weighted based on Iacus, King and Porro (2012). Standard errors are clustered at the patent level. Significance levels: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, *BvD Orbis* and several disambiguations (links) between them (see Section III for details).

Also the recency since availability of the online deposited patents may affect search costs associated with their retrieval. As age of knowledge is intrinsically linked to higher diffusion levels, patents that had already been granted several years before their AIDS DB inclusion might have experienced relatively lower rates of excess citations from online indexing, ceteris paribus. On the other hand, due to the short-term higher visibility of new inventions associated with their recent publication and announcement in the *USPTO Patent Gazzette*, it is also thinkable that more dated patents would have benefited relatively more from the additional attention drawn to them. In line with these contradicting predictions, I do not find any differential effects conditional on recency since first publication of the patented knowledge. As reported in the results of Table 14 in the online Appendix A, the citation premium to ”no reference” patents relative to patents with front-page reference to HIV/AIDS is estimated to be homogeneous across the patent age distribution.

The investigation of heterogeneity of effects also does not yield any significant differences based individual countries or aggregated geographic areas of origin of patents. Obviously, variation in our sample on this dimension is relatively contained, as HIV research was strongly clustered and concentrated mostly in the U.S., and to a much lesser extent in Europe and Japan (see Figure 3). This makes it difficult to meaningfully estimate differential impact effects for sub-samples of locations outside of these hubs. I further do not find heterogeneity based on individual level features of pre-database centrality and degree of connectedness of inventors within the HIV-research community (see Table 8 in online Appendix .A), in particular no disproportional gains from higher visibility of patents from more peripheral inventors to the community.[28] This motivates the further inquiry of differences in impact of the AIDS DB at the receiving end of the knowledge spillover, conducted in Section V.C.

E. Second Order Effects on Citations to Referenced Scientific Publications

Finally, I investigate whether the higher visibility associated with inclusion in the database of patents with previously higher search costs generated second order effects on visibility of the scientific prior art applied in these patents. The showcasing of ”hidden” technologies linked to the treatment and diagnostics of HIV/AIDS may have further led to a socially desirable display of useful scientific knowledge and revealed potential for new modes of application of fundamental insights. This seems particularly standing to reason, given the closeness and strong reliance on science of inventions in the fields relevant to HIV/AIDS research.

To estimate second order effects on the subsequent use in technology of papers included among scientific references in patents in the AIDS DB, I compute yearly cumulative patent citation rates to each PubMed article cited by a patent in the sample.[29] The corresponding estimates are reported in the extended online version of the paper. Results indicate a strong and robust second order effect of the AIDS DB establishment on the visibility and subsequent use of the scientific knowledge components linked to patents with previously non-obvious link to HIV/AIDS, which provides further evidence in support of the effectiveness of the online repository in line with the policy objective.

[28] Estimation results unreported.

[29] Patent-to-article citations are sourced from Marx and Fuegi (2020)

V. Changes in the Quality and Reach of Knowledge Spillovers

In this section, the paper scrutinizes whether the increase in informational efficiency due to the establishment of the AIDS DB had repercussions on the intensive margin of knowledge spillovers generated among researchers, which was a declared policy objective behind the database project. Precisely, I evaluate changes in likelihood of transfer of new knowledge elements from indexed patents to citing follow-up applications (Section V.B), as well as in the reach of citation links across geographic distances and HIV-researcher community boundaries (Section V.C). In each analysis, I further assess heterogeneity of effects between private sector and public research institute inventors, in order to investigate to which degree they benefited (differently) from access to the online repository. The distinction between corporate vs. academic inventors is in this case particularly interesting given that private sector researchers faced substantially higher access barriers to external patent documents, while inventors from public research institutes were embedded in more sophisticated and far reaching information systems and communication channels other than AIDS DB (e.g. BITNET, patent libraries, etc.), as discussed in Section II.A.

A. Citation-level Estimation Model

The estimation of treatment effects on diffusion patterns is challenging insofar as it, primarily, requires to isolate the natural (intrinsic) diffusion of a specific piece of knowledge in time from the influence of the institution, or policy, assessed. In the following, I move my inquiry from a patent-level to a citation link-level analysis *within* citing applications (patents), thereby holding constant all factors impacting the context and dynamics at the receiving end of knowledge spillovers. Specifically, I determine differences between AIDS DB patents and other references of same timing cited within the same citing patent, and compare changes in these differences over time across AIDS DB patents without vs. with front-page reference to HIV/AIDS, to evaluate the marginal impact of the online repository. For this, I create a citation-level data set containing one observation for each cited patent-citing patent link made from any follow-up patent to each AIDS DB patent during a five-year period before and after the one-year lagged launch of the database, i.e. between 1991 and 2000.[30] For each citing patent and cited AIDS DB patent, I further include one observation for each non-AIDS DB patent, referenced as prior art in the citing patent, that was granted and first published in the same year as the paired AIDS DB patent. I again remove inventor and applicant self-citations between cited and citing patents. For each within citing patent-cited year group, I assign equal weight of .5 to both the sum of all AIDS DB and all non-AIDS DB cited patent observations, in order to give each cited year the same weight within citing patents.[31] Finally, I balance the

[30] To ensure results are based on the same sample of patents, I only include citations links to AIDS DB patents included in the external matched control sample utilized in Section IV.

[31]So, e.g., if a citing patent - cited year group contains 1 AIDS DB and 2 non-AIDS DB cited patents, the assigned individual weights are .5, for the AIDS DB patent, and, respectively, .25 for each non-AIDS DB patent.

data set by giving each citing patent a weight of one, in order not to overweight the importance of citing applications with many references to prior art.

I then investigate how the relation to prior AIDS DB vs. non-AIDS DB patents, cited within same applications, changed after the establishment of the online repository by estimating the following type of regressions:

$$(2) \quad Y_{ij} = \beta_1 * no\ reference_j \times post1994_{t-1} + \beta_2 * with\ reference_j \times post1994_{t-1} + \\ + \beta_3 * no\ reference_j + \beta_4 * with\ reference_j + \\ + citingPatent_i \times citedYear_{y(j)}FE + \theta_{tr(i)} + \epsilon_{ij}$$

where *i* indexes citing patents, *t* indexes years, *y* indexes cited years and *r* indexes citing geographic regions. Y_{ij} is the generic dependent variable quantifying the quality and reach of knowledge flow associated with a citation. The coefficients β_1 and β_2 measure the change in outcomes after one-year lagged online deposit (i.e. after 1995) between AIDS DB and non-AIDS DB control patents for patents without vs. with front-page references to HIV/AIDS, while β_3 and β_4 capture the respective pre-AIDS DB differences.

The regression model includes a fixed effect for each citing patent × cited year pair. These fixed effects control for all differences regarding the context of invention of the citing patent that might affect knowledge flows, for instance, the identity of the citing inventor, the citing institution, their scientific networks and quality. Specifically, they also account for all unobserved shocks affecting knowledge flows and information access channels of citing inventors, for example, increased resources for specific research lines, as these are held constant within a citing patent. Similarly, the fixed effects control for all permanent differences in access to knowledge across geographic regions, or permanent differences in citation patterns across technological fields. Moreover, they account for age differences across cited prior art, by holding constant all changes due to the natural diffusion of knowledge over time within a citing patent. Note, that the sum of all citing patents fixed effects would be collinear to a $post_{t-1}$ period indicator, which is therefore omitted from the specification. The regressions further include region-specific time trends, absorbing changing intrinsic components of knowledge agglomeration in specific geographic areas over time, for example, Maryland in the U.S. becoming more central to the global HIV-researcher community over the years.[32] To account for potential correlations of regression residuals regarding the presence of unobserved random shocks to knowledge production (e.g., national R&D policies or related specific developments), I cluster standard errors at the citing patent country level. Summary statistics for the cited-citing level sample are reported in Table 15 in the online Appendix .A.

[32] Regions are aggregated to federal states in the U.S., Mexico and Australia, NUTS-1 regions in the E.U., prefectures for Japan, provinces for Korea and Canada, and districts in Israel.

B. Quality of Knowledge Flows

To evaluate changes in the intensity of knowledge spillovers associated with citations, I proxy the quality of knowledge flow by the re-occurrence in subsequently citing patents of new knowledge elements originally appearing on cited patent documents: New words in patent text, and novel scientific references. Corresponding estimates are reported in the extended online version of the paper. Results indicate significantly higher rates of re-occurrence in citing patents of new words and novel scientific prior art references appearing in AIDS DB patents without front-page link to HIV, in particular among private firm citing inventors.

C. Geographic and Social Distance of Spillovers

I further investigate changes in international citations and knowledge flows across network boundaries of scientific communities following the launch of the AIDS DB, which were particularly emphasized as leading objectives behind the repository (compare Section II.B). To measure the geographic spillover distance, for each pair of cited and citing patents, I determine the share of overlap between geographic locations of all citing inventors and all cited inventors. The share of international citation links is then simply given by the inverse of the overlap.[33]

To determine the social distance between inventor communities, I consider each patenting inventor as a node in a dynamic, undirected social network of researchers, whose edges (connections) are based on observed prior collaborations between these inventors at a given point in time.[34] Subsequently, I determine, for each pair of citing and cited inventors in the data, the shortest path in the network graph of all > 5 million inventors of USPTO patents and their existing collaborative ties (as evidenced by co-appearance on prior patents) at the moment of filing of the citing patent. I consider as *minimal social distance* between a citing and cited patent, the shortest of all paths between any inventor pair involved. To account for the fact that, in any finite network, the existence of a network tie between a citing and cited patent increases stochastically in the number of inventors, I control for the count of inventors on the cited patent in all specifications. Given that AIDS DB inventors were strongly intertwined with the community of basic science authors (as shown in Section III.D), I further consider their existing collaborative ties in the universe of fundamental science, based on prior scientific co-authorships on biomedical publications, and determine the minimal social distance between any pair of cited and citing patents' inventors based on the comprehensive author-inventor network graph consisting of the union of all > 5 million inventors on U.S.

[33] Following the same reasoning, if a patent with two inventors, one located in the U.S. and the other in France, cites a prior patent with equally two inventors, one located in France and the other in Japan, the share of international citations will be: (1 × .5 + 1 × .5) × .5 + (1 × .5 + 0 × .5) × .5 = .75

[34] Knowledge flows are found to be naturally clustered alongside these collaborative network ties (e.g., Singh, 2005).

patents and > 16 million authors indexed in PubMed and their realm-transcending collaborative ties.[35]

Table 4: Effects on the Reach of Generated Spillovers

Dependent variable:	*International*		*Detached community*	
Probability of distant citation	(1)	(2)	(3)	(4)
No reference x post1994$_{t-1}$	0.020* (0.009)	0.058*** (0.006)	0.007* (0.004)	-0.012** (0.006)
With reference x post1994$_{t-1}$	-0.026*** (0.009)	-0.110*** (0.022)	-0.021** (0.008)	-0.036*** (0.010)
No reference x post1994$_{t-1}$ x private firm citing		-0.049*** (0.016)		0.027*** (0.008)
With reference x post1994$_{t-1}$ x private firm citing		0.119*** (0.043)		0.023 (0.015)
Main category interactions		Incl		Incl
Citing patent x cited year FE	Yes	Yes	Yes	Yes
Citing region time trends	Yes	Yes	Yes	Yes
Observations	36,690	36,618	36,684	36,612
Number of citing clusters	8,573	8,554	8,570	8,551
R^2	0.556	0.556	0.731	0.730
Mean at t_0	0.352	0.352	0.201	0.201

Notes: Each column reports parameter estimates of regression (2) on citing patents-cited year pairs. The main category parameter is included. The dependent variable in columns (1)-(2) measures shares of international citations between all pairwise links of citing and cited inventor locations for AIDS DB and control group patents in citing patents between 1991 and 2000. The dependent variable in columns (3)-(4) measures the probability that a citation originates from a research team which is entirely unconnected to the networks of direct and indirect collaborators (social distance =) of any cited inventor at the time of filing of the citing patent. Displayed are parameter estimates for the post-period only. Main category parameters and full sets of interactions are included. Inventor and applicant self-citations are excluded. The reference category consists of non-AIDS DB patents, published in the same year, cited within the same application. Sample observations are weighted in order to give equal weight to each citing patent. Standard errors are clustered at the citing countrylevel. Significance levels: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT*, and *BvD Orbis*. Geo-coordinates and inventor/ author identities are disambiguated based on input data from Li et al. (2014); Morrison, Riccaboni and Pammolli (2017); Smalheiser and Torvik (2009); Torvik and Smalheiser (2009) (see Section III for details).

Table 4 compares results for the estimation of regression (2) with the share of international citations as well as the likelihood of citation to an entirely unconnected community (social distance = ∞, no finite shortest path) as outcome variables. In the main effect specifications (columns (1) and (3)), AIDS DB patents with front-page references to HIV/AIDS received relatively less citations from outside of geographic and social network boundaries after database inclusion (significant at the 1% level). This suggests a thickening of citation clusters within these boundaries following online accessibility. For AIDS DB patents without front-

[35] Inventor information covers years 1976-2011, author information years 1858-2009. For details, see Section III.

page references, on the other hand, I observe opposite patterns, suggesting a positive influence of DB indexing for patents previously more difficult to detect as HIV-related to be referenced across geographic and scientific network boundaries. Effect magnitudes indicate a relative average increase of + 20% (for international citations) and + 58% (for across-community citations) compared to pre-AIDS DB levels.[36]

When looking at heterogeneity at the receiving end, estimates in column (2) of Table 4 show that the impact on international spillovers from ”no reference” patents was much smaller for private firm inventors (-.05 percentage points, - 85%), suggesting that gains in enhanced retrieval of HIV-relevant prior art with higher search costs from abroad were particularly driven and internalized by academic inventors. Private sector researchers exhibited a strong and positive heterogeneous increase in foreign citations to patents with front-page references to HIV/AIDS, corroborating the prediction of stronger benefits of online accessibility for this category.

Split-sample results in column (4) of Table 4 show, on the other hand, that positive effects on citations to detached scientific communities were strongly driven by corporate inventors; Their likelihood of citation to external patents without HIV/AIDS front-page references outside of the network of direct or indirect collaborators increased by .03 percentage points compared to the background rate of control group references, cited by the same patent, which was about three times larger than the corresponding effect for non-firm inventors (significant at the 1% level). For citations to prior AIDS DB patents with HIV/AIDS front-page reference, instead, there were no significant differential effects for private firm inventors.

For the findings on the reach of spillovers generated, I provide more results on different sub-level of geographic and social distance in Tables 16 and 17 in the online Appendix .A. I further show robustness of findings for social distance metrics based exclusively on inventor network graphs in Table 18 in online Appendix .A. While results are qualitatively robust, these show that shortest paths based exclusively on inventor networks drastically overstate the true distances between researchers in strongly science-intensive environments, like in this case, where patents are only a partial indicator of research output. Depending on the exact conceptualization of social distance, this might have important implications for estimating the true extent of distance of knowledge flows.

Taken together, these results suggest a strong positive impact of the institution of the AIDS DB on facilitating the flow and diffusion of, HIV/AIDS related, technical knowledge across dispersed communities of inventors, in particular for those technologies that were more difficult to detect as HIV/AIDS-related in pre-AIDS DB external search efforts. Combined with the results in Section IV, these findings provide credible support that part of the cumulative impact of the AIDS DB online repository can be attributed to broader diffusion across distant teams of researchers, both in geographic and social terms.

[36] Pre-AIDS DB level estimates not reported in the table.

VI. Conclusion

Access to existing knowledge is a crucial input for technical progress and economic growth. However, due to constraints of bounded rationality, the costs to filter out relevant knowledge inputs in light of a growing abundance and general availability of information increase for inventors and other scientists alike. Many examples of public and private sector institutions and devices have emerged over the past three decades, in the form of search engines, structured databases and platforms, but we still know very little about their repercussions on scientific production, and the underlying mechanisms governing these. Yet, the problem of ’too-many-giants’, on whose shoulders to stand, on has relevance and important implications far beyond prior art search, but becomes salient also with regards to questions like media literacy and public political opinion-forming (e.g., Bimber 2001; Gavazza, Nardotto and Valletti 2019).

The case of the 1994 AIDS Patent Database, as an early modern-era information-enhancing institution, enables me to study these two concurring mechanisms separately: On the one hand, the online repository provided broad accessibility at minimal cost to the full body of technical prior art related to the deadly infectious disease behind the HIV-pandemic. Patent documents, despite their abstract jargon and strategic motives of patent holders to ’conceal’ the nature of the underlying invention, have been shown by prior literature to be important carriers of codified knowledge and relevant channels of knowledge transfer between distant inventors (e.g., Furman, Nagler and Watzinger, 2021; Hegde, Herkenhoff and Zhu, 2020).

The main stand-alone contribution of this paper arises with regards the design of such institutions. The disease-specific connection, established by inclusion in the AIDS Patent Database, appears to have disproportionately benefited the visibility and subsequent diffusion of technical advances that were more difficult to identify as related to HIV/AIDS with the previous capabilities of external prior art search, based most exclusively on bibliographic information. The stronger reduction in search costs explains 30% of the variation in cumulative diffusion between these patents and those making clear front-page references to the disease. This speaks to prior findings by Thompson and Hanley (2018), who show a causal increase in follow-up citations to scientific articles appearing in topic-specific pages in Wikipedia. The catalytic effect of the topic-connection in the online repository, in my analysis is strongest for the cumulative impact of technologies embodying new ideas and novel concepts. These are particularly vulnerable to barriers affecting knowledge flows and, at the same time, need often parallel experimentation in order to prevail (Murray et al., 2016). In my analyses it shows that, not only did patents with higher up-front retrieval costs experience the relatively strongest increase in cumulative impact, but the effects on scientific and geographic community-crossing citations were also strongly concentrated in these patents, and disproportionately benefited private firm inventors. Considering this remarkable effectiveness of a comparatively low-cost policy measure, that is an online database, this has important and corroborating implications for public and private sector decision makers regarding the imperative of free access to prior art for the productivity of researchers and makes a powerful argument for the establishment of access-providing institutions.

Findings reported herein, therefore, speak in particular to the effective organization and design of patent search devices, which are historically structured based on technology classes. Complementary categorizations, such as use-indexed headings (based on, for instance, medical subjects or specific diseases), could provide useful tools for prior art searching inventors to process and condense the thousands of patents granted every year even in the most narrow technology classes. Nevertheless, for the interpretation and evaluation of transferability of these findings it should, obviously, be taken into account that HIV/AIDS research constituted a very particular and dynamic domain, spanning the frontier of many sub-disciplines both of basic science and technological knowledge, especially at the time it is observed in the empirical setting. Similar to other studies focusing on nascent and highly-innovative domains, it should, therefore, be subject to further discussion to which extent these findings can be transferred to other contexts and different circumstances of inventive search.

REFERENCES

Agrawal, Ajay, and Avi Goldfarb. 2008. "Restructuring research: Communication costs and the democratization of university innovation." *American Economic Review*, 98(4): 1578–90.

Ahmadpoor, Mohammad, and Benjamin F Jones. 2017. "The dual frontier: Patented inventions and prior scientific advance." *Science*, 357(6351): 583–587.

Akcigit, Ufuk, Douglas Hanley, and Nicolas Serrano-Velarde. 2020. "Back to Basics: Basic Research Spillovers, Innovation Policy and Growth." *The Review of Economic Studies*.

Alcacer, Juan, and Michelle Gittelman. 2006. "Patent citations as a measure of knowledge flows: The influence of examiner citations." *The Review of Economics and Statistics*, 88(4): 774–779.

Arora, Ashish, Sharon Belenzon, and Honggi Lee. 2018. "Reversed citations and the localization of knowledge spillovers." *Journal of Economic Geography*, 18(3): 495–521.

Arts, Sam, Bruno Cassiman, and Juan Carlos Gomez. 2018. "Text matching to measure patent similarity." *Strategic Management Journal*, 39(1): 62–84.

Baruffaldi, Stefano, and Felix Pöge. 2020. "A Firm Scientific Community: Industry Participation and Knowledge Diffusion." *IZA Discussion Paper No. 13419, Available at SSRN: https://ssrn.com/abstract=3643183*.

Baruffaldi, Stefano H, and Markus Simeth. 2020. "Patents and knowledge diffusion: The effect of early disclosure." *Research Policy*, 49(4): 103927.

Belenzon, Sharon, and Mark Schankerman. 2013. "Spreading the word: Geography, policy, and knowledge spillovers." *Review of Economics and Statistics*, 95(3): 884–903.

Benner, Mary, and Joel Waldfogel. 2008. "Close to you? Bias and precision in patentbased measures of technological proximity." *Research Policy*, 37(9): 1556–1567.

Berkes, Enrico, and Peter Nencka. 2020. "Knowledge Access: The Effects of Carnegie Libraries on Innovation." *Working Paper*.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How much should we trust differences-in-differences estimates?" *The Quarterly Journal of Economics*, 119(1): 249–275.

Bertschek, Irene, Daniel Cerquera, and Gordon J Klein. 2013. "More bits–more bucks? Measuring the impact of broadband internet on firm performance." *Information Economics and Policy*, 25(3): 190–203.

Biasi, Barbara, and Petra Moser. 2021. "Effects of Copyrights on Science - Evidence from the US Book Republication Program." *American Economic Journal: Microeconomics*, 13(4): 218–260.

Bimber, Bruce. 2001. “Information and political engagement in America: The search for effects of information technology at the individual level.” *Political Research Quarterly*, 54(1): 53–67.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. “Fast unfolding of communities in large networks.” *Journal of statistical mechanics: theory and experiment*, 2008(10): P10008.

Bloom, Nicholas, Mark Schankerman, and John Van Reenen. 2013. “Identifying technology spillovers and product market rivalry.” *Econometrica*, 81(4): 1347–1393.

Boudreau, Kevin J, Eva C Guinan, Karim R Lakhani, and Christoph Riedl. 2016. “Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science.” *Management Science*, 62(10): 2765–2783.

Bryan, Kevin A, and Yasin Ozcan. 2021. “The impact of open access mandates on invention.” *The Review of Economics and Statistics*, forthcoming.

Cassiman, Bruno, and Reinhilde Veugelers. 2002. “R&D Cooperation and Spillovers: Some Empirical Evidence from Belgium.” *American Economic Review*, 92(4): 1169–1184.

Cohen, Wesley M, and Daniel A Levinthal. 1990. “Absorptive capacity: A new perspective on learning and innovation.” *Administrative science quarterly*, 128–152.

Cohen, Wesley M, Richard R Nelson, and John P Walsh. 2000. “Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not).” National Bureau of Economic Research Working Paper 7552.

Czernich, Nina, Oliver Falck, Tobias Kretschmer, and Ludger Woessmann. 2011. “Broadband infrastructure and economic growth.” *The Economic Journal*, 121(552): 505–532.

De Chaisemartin, Cl´ement, and Xavier d’Haultfoeuille. 2020. “Two-way fixed effects estimators with heterogeneous treatment effects.” *American Economic Review*, 110(9): 2964–96.

de Rassenfosse, G., G. Pellegrino, and E. Raiteri. 2020. “Do patents enable disclosure? Evidence from the 1951 Invention Secrecy Act.” Ecole polytechnique f´ed´erale de Lausanne.

Devlin, Alan. 2009. “The misunderstood function of disclosure in patent law.” *Harvard Journal of Law & Technology*, 23: 401.

Ding, Waverly W, Sharon G Levin, Paula E Stephan, and Anne E Winkler. 2010. “The impact of information technology on academic scientists’ productivity and collaboration patterns.” *Management Science*, 56(9): 1439–1461.

Dittmar, Jeremiah E. 2011. “Information technology and economic change: the impact of the printing press.” *The Quarterly Journal of Economics*, 126(3): 1133–1172.

Fleming, Lee. 2001. “Recombinant uncertainty in technological search.” *Management science*, 47(1): 117–132.

Forman, Chris, and Nicolas van Zeebroeck. 2012. “From wires to partners: How the Internet has fostered R&D collaborations within firms.” *Management science*, 58(8): 1549–1568.

Forman, Chris, and Nicolas van Zeebroeck. 2019. “Digital technology adoption and knowledge flows within firms: Can the Internet overcome geographic and technological distance?” *Research Policy*, 48(8): 103697.

Fromer, Jeanne C. 2008. “Patent disclosure.” *Iowa L. Rev.*, 94: 539.

Furman, Jeffrey L., and Scott Stern. 2011. “Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research.” *American Economic Review*, 101(5): 1933–63.

Furman, Jeffrey L, Markus Nagler, and Martin Watzinger. 2021. “Disclosure and subsequent innovation: Evidence from the patent depository library program.” *American Economic Journal: Economic Policy*, 13(4): 239–270.

Galasso, Alberto, and Mark Schankerman. 2015. “Patents and cumulative innovation: Causal evidence from the courts.” *The Quarterly Journal of Economics*, 130(1): 317–369.

Gambardella, A, D Harhoff, and Nagaoka S. 2011. “The Social Value of Patent Disclosure.” LMU Munich.

Gavazza, Alessandro, Mattia Nardotto, and Tommaso Valletti. 2019. “Internet and politics: Evidence from UK local elections and local government policies.” *The Review of Economic Studies*, 86(5): 2092–2135.

Giuri, Paola, Myriam Mariani, Stefano Brusoni, Gustavo Crespi, Dominique Francoz, Alfonso Gambardella, Walter Garcia-Fontes, Aldo Geuna, Raul Gonzales, Dietmar Harhoff, and Karin Hoisl. 2007. “Inventors and invention processes in Europe: Results from the PatVal-EU survey.” *Research Policy*, 36(8): 1107–1127.

Goodman-Bacon, Andrew. 2021. “Difference-in-differences with variation in treatment timing.” *Journal of Econometrics*.

Graham, S, and D Hegde. 2015. “Disclosing patents’ secrets.” *Science*, 347(6219): 236– 237.

Graham, Stuart JH, Alan C Marco, and Richard Miller. 2015. “The USPTO patent examination research dataset: A window on the process of patent examination.” *Georgia Tech Scheller College of Business Research Paper No. WP*, 43.

Griliches, Zvi. 1991. “The search for R&D spillovers.” National Bureau of Economic Research.

Hegde, Deepak, and Hong Luo. 2018. “Patent publication and the market for ideas.” *Management Science*, 64(2): 652–672.

Hegde, Deepak, Kyle Herkenhoff, and Chenqi Zhu. 2020. "Patent disclosure and innovation." *Available at SSRN 3158031*.

Iacus, Stefano M, Gary King, and Giuseppe Porro. 2012. "Causal inference without balance checking: Coarsened exact matching." *Political analysis*, 1–24.

Iaria, Alessandro, Carlo Schwarz, and Fabian Waldinger. 2018. "Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science*." *The Quarterly Journal of Economics*, 133(2): 927–991.

Jaffe, Adam B, Manuel Trajtenberg, and Michael S Fogarty. 2000. "Knowledge spillovers and patent citations: Evidence from a survey of inventors." *American Economic Review*, 90(2): 215–218.

Jaffe, Adam B, Manuel Trajtenberg, and Rebecca Henderson. 1993. "Geographic localization of knowledge spillovers as evidenced by patent citations." *The Quarterly Journal of Economics*, 108(3): 577–598.

Jones, Benjamin F, and Lawrence H Summers. 2020. "A Calculation of the Social Returns to Innovation." National Bureau of Economic Research Working Paper 27863.

Jones, Eric. 2003. *The European miracle: environments, economies and geopolitics in the history of Europe and Asia.* Cambridge University Press.

Kong, Nancy, Uwe Dulleck, Adam B Jaffe, Sowmya Vajjala, et al. 2020. "Linguistic Metrics for Patent Disclosure: Evidence from University Versus Corporate Patents." *NBER Working Paper*, , (w27803).

Lemley, Mark A. 2012. "The myth of the sole inventor." *Mich. L. Rev.*, 110(5): 709.

Li, Guan-Cheng, Ronald Lai, Alexander D'Amour, David M Doolin, Ye Sun, Vetle I Torvik, Z Yu Amy, and Lee Fleming. 2014. "Disambiguation and co-authorship networks of the US patent inventor database (1975–2010)." *Research Policy*, 43(6): 941–955.

Lück, Sonja, Benjamin Balsmeier, Florian Seliger, and Lee Fleming. 2020. "Early disclosure of invention and reduced duplication: An empirical test." *Management Science*, 66(6): 2677–2685.

Mansfield, Edwin. 1995. "Academic research underlying industrial innovations: sources, characteristics, and financing." *The review of Economics and Statistics*, 55–65.

March, James G. 1991. "Exploration and exploitation in organizational learning." *Organization science*, 2(1): 71–87.

Marco, Alan C, Michael Carley, Steven Jackson, and Amanda Myers. 2015. "The uspto historical patent data files: Two centuries of innovation." *Available at SSRN 2616724*.

Marx, Matt, and Aaron Fuegi. 2020. "Reliance on science: Worldwide front-page patent citations to scientific articles." *Strategic Management Journal*, 41(9): 1572–1594.

Mokyr, Joel. 2005. "Long-term economic growth and the history of technology." In *Handbook of economic growth*. Vol. 1, 1113–1180. Elsevier.

Morrison, Greg, Massimo Riccaboni, and Fabio Pammolli. 2017. "Disambiguation of patent inventors and assignees using high-resolution geolocation data." *Scientific data*, 4: 170064.

Moser, Petra, and Alessandra Voena. 2012. "Compulsory Licensing: Evidence from the Trading with the Enemy Act." *American Economic Review*, 102(1): 396–427.

Murray, Fiona, Philippe Aghion, Mathias Dewatripont, Julian Kolev, and Scott Stern. 2016. "Of Mice and Academics: Examining the Effect of Openness on Innovation." *American Economic Journal: Economic Policy*, 8(1): 212–52.

Narin, Francis, Kimberly S Hamilton, and Dominic Olivastro. 1997. "The increasing linkage between US technology and public science." *Research policy*, 26(3): 317–330.

Ouelette, Lisa L. 2012. "Do patents disclose useful information?" *Harvard Journal of Law & Technology*, 25(2): 545–608.

Righi, Cesare, and Timothy Simcoe. 2019. "Patent examiner specialization." *Research Policy*, 48(1): 137–148.

Risch, Michael. 2007. "The Failure of Public Notice in Patent Prosecution." *Harv. JL & Tech.*, 21: 179.

Rosenberg, Nathan. 1976. *Perspectives on technology.* CUP Archive.

Schmoch, Ulrich. 2008. "Concept of a technology classification for country comparisons." *Final report to the world intellectual property organisation (wipo), WIPO*.

Scotchmer, Suzanne. 1991. "Standing on the shoulders of giants: cumulative research and the patent law." *Journal of economic perspectives*, 5(1): 29–41.

Silva, JMC Santos, and Silvana Tenreyro. 2006. "The log of gravity." *The Review of Economics and statistics*, 88(4): 641–658.

Silva, JMC Santos, and Silvana Tenreyro. 2011. "Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator." *Economics Letters*, 112(2): 220–222.

Singh, Jasjit. 2005. "Collaborative networks as determinants of knowledge diffusion patterns." *Management science*, 51(5): 756–770.

Singh, Jasjit, and Matt Marx. 2013. "Geographic constraints on knowledge spillovers: Political borders vs. spatial proximity." *Management Science*, 59(9): 2056–2078.

Smalheiser, Neil R, and Vetle I Torvik. 2009. "Author name disambiguation." *Annual review of information science and technology*, 43(1): 1.

Thompson, Neil C., and Douglas Hanley. 2018. "Science is Shaped by Wikipedia: Evidence From a Randomized Control Trial." *Available at SSRN 3039505*.

Thompson, Peter. 2006. "Patent citations and the geography of knowledge spillovers: evidence from inventor-and examiner-added citations." *The Review of Economics and Statistics*, 88(2): 383–388.

Thompson, Peter, and Melanie Fox-Kean. 2005. "Patent citations and the geography of knowledge spillovers: A reassessment." *American Economic Review*, 95(1): 450–460.

Torvik, Vetle I. 2018. "Author-Linked data for Author-ity 2009."

Torvik, Vetle I, and Neil R Smalheiser. 2009. "Author name disambiguation in MEDLINE." *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(3): 1–29.

Trajtenberg, Manuel, Rebecca Henderson, and Adam Jaffe. 1997. "University versus corporate patents: A window on the basicness of invention." *Economics of Innovation and new technology*, 5(1): 19–50.

Verhoeven, Dennis, Jurri¨en Bakker, and Reinhilde Veugelers. 2016. "Measuring technological novelty with patent-based indicators." *Research Policy*, 45(3): 707–723.

Weitzman, Martin L. 1979. "Optimal search for the best alternative." *Econometrica: Journal of the Econometric Society*, 641–654.

Zheng, Yanfeng, and Qinyu Wang. 2020. "Shadow of the great firewall: The impact of Google blockade on innovation in China." *Strategic Management Journal*, 41(12): 2234–2260.