Online Repositories, Search Costs and Cumulative Innovation

By Thomas Schaper*

Draft: December 1, 2021

Efficient access to existing knowledge is essential to technical advance, yet little is known about how access-enhancing institutions shape intertemporal knowledge spillovers. In this paper, I investigate 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, I find 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.

JEL: I23, O31, O32, O33

Keywords: knowledge diffusion, information technology, patents

* TUM School of Management, Technical University of Munich, Arcisstraße 21, 80333 Munich, Germany. Contact: thomas.schaper@tum.de. Financial support from the Research Foundation Flanders (FWO) [11B5918N] and the German Research Foundation (DFG) [FOR5234] is gratefully acknowledged. I thank 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, TU Eindhoven, TU Munich, VU Amsterdam, ZEW Mannheim, DRUID, EPIP and the SEI Doctoral Consortium for many useful comments and suggestions.

I. Introduction

Scientific advance is shaped by sequential and complementary intellectual efforts. Efficient access to existing information is, therefore, 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 strongly 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.[1] [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

[1] Iaria, Schwarz and Waldinger (2018) study the disruption of international scientific cooperation induced by WWI and show how the exclusion from access to frontier research from opposite camps lastingly affected scientific productivity. Biasi and Moser (2021) and Bryan and Ozcan (2020) show that the authorized violation of copyrights to German scientific books in the U.S. during WWII and the 2008 open access mandate for NIH funded research publications enhanced the dissemination of scientific ideas, evidenced by citations from subsequent articles and patents. In recent working papers, Furman, Nagler and Watzinger (2021) and Berkes and Nencka (2020) document substantial increases in patenting activity following the opening of U.S. Patent Libraries 1975-1997 and the extensive roll-out of public 'Carnegie' libraries in the U.S. 1883-1991, respectively. Arts et al. (2020) provide concurring evidence due to the arrival of broadband internet to the U.K. in the early-2000s.

[2] A closely related stream of literature has investigated to what extent patent documents allow to disclose relevant scientific knowledge and affect follow-up inventions, providing ample confirming evidence: Graham and Hegde (2015); Hegde and Luo (2018); Baruffaldi and Simeth (2020); Hegde, Herkenhoff and Zhu (2020); Lück et al. (2020); de Rassenfosse, Pellegrino and Raiteri (2020)

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.

In this paper, I ask 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 sheer mass of scientific knowledge produced, *conditional* on accessibility. 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, I study 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. At the peak of the HIV/AIDS pandemic [3], the United States Patent and Trademark Office (USPTO) undertook an unprecedented effort to leverage the capacities of the new world wide web in the fight against the disease, by providing free of charge full-online access to all patents related to acquired immune deficiency syndrome, ranging from diagnostic testing to therapeutic treatments. This repository, 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. Before AIDS DB, interested inventors had comparably limited ways to screen, filter and efficiently rank weekly published new patents: Assessing the accurate technical content of a patent required to either inspect paper documents directly at the patent office, query computer terminals in patent depository libraries, or remote-order individual full-text copies via mail or fax. Moreover, adverse incentives in the drafting of patents, in particular of private firms, to inhibit effective disclosure of valuable information through the patent system, determine historically high and substantial efforts necessary for the

[3] Incidences of HIV/AIDS-related infections had exponentially spread since the early-1980s, reaching a peak of $>$ 20,000 cases in 1993 and a yearly mortality rate excess of almost 15,000 by 1995 in the U.S. alone (Source: U.S. Centres for Disease Control and Prevention (CDC)). As of 2020, about 38 million people worldwide are living with HIV, causing about 1.7% of deaths globally (Source: WHO).

retrieval and absorption of prior art from codified knowledge. The feature of being a centrally-maintained and expert-validated disease-targeted repository could decrease these search costs significantly, given that deposited patents span a broad range of technology classes and fields and many patent documents were, at first, not clearly recognizable as HIV-related from bibliographic searches. The AIDS DB was discontinued in March 1999, when the USPTO launched its comprehensive online database including full-text and images of all patents, as broader bandwidths became available.

I empirically assess 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. I, further, exploit 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, I rely on geolocalized addresses and assign 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, I design an empirical strategy relying on a within AIDS DB counterfactual: Exploiting idiosyncrasies of technology classes assigned to patents not being disease-specific, I estimate the elasticity of 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[4] prior to the repository inclusion. I compare the differential effects of database deposit for these patents to a control group of patents, also indexed in the AIDS DB and and comparable 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 front-pages, providing a baseline of online accessibility and the confounding factors relating to HIV-patents in the estimation

I find that, after online deposit, cumulative citations to AIDS DB patents without explicit

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

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, I further find 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. I further exclude that results are driven by individual patent examiners or within-changes in citation behaviour of examiners over time.

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, I find strongly 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, my 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. I show robustness of findings using different estimations, different time windows, stricter control group definitions, and impact weighted citation counts.

In a set of additional results, I investigate 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, I find 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 private firm citing inventors. I further estimate 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, I find 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, my 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, I contribute 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 applied technology.

My findings, 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, 2020; Furman, Nagler and Watzinger, 2021;

Biasi and Moser, 2021). Here, I confirm prior evidence that increasing accessibility to relevant prior art impacts subsequent invention and the diffusion of industrially applicable knowledge. Finally, my 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 S, 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, 2020; Biasi and Moser, 2021; Furman, Nagler and Watzinger, 2021). Such access costs, which constitute broader search costs from a point of view of microeconomic theory, have been historically substantial, but decreased drastically with the advent of modern information technologies, in particular broadband internet (Arts et al., 2020). Faster bandwidth, electronic file formats and online repositories have made scientific and technical information readily available and omni-accessible.

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), where each new patent document issued represents a closed box at the beginning of each period, containing a potential reward in form of a knowledge spillover. The costs to open a box and internalize its content are paid during the search, while the reward is revealed and collected only afterwards, and with some delay. After each stage, the inventor decides whether to incur the cost of another search round or to use the fallback option, i.e. rely on local search (e.g., March, 1991; Cohen and Levinthal, 1990) and the knowledge acquired in previous rounds. The accuracy with which the inventor can delineate the sample of prior art documents to inspect with regards to their utility determines the overall efficiency of the search, by optimizing the reward/cost ratio. Ceteris paribus, we would expect the amount of spillovers generated to

increase with the accuracy of up-front information provided.

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.[5] 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 (first, through patent library terminals, afterwards for their online repositories, see II.B). The usefulness of these for effectively detecting prior art 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.

Despite the advent of modern information technology in bibliographic search[6], including tools and algorithms based on latent semantic analysis, page rank or applications of artificial intelligence, efficiency in information retrieval remains a concern, as evidenced by several platform initiatives launched within the course of the COVID-19 pandemic: The U.S. Centres of Diseases Control and Prevention (CDC), in 2020, launched a topic-specific online access repository collecting comprehensive scientific evidence, clinical trials and gene sequence databases related to the novel coronavirus, its diagnostics and treatments.[7] Si-

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

[6] Modern web search engines based on LSA and page rank algorithms, such as Google Patents, might prove useful in lowering these costs through automated retrieval of related patents to inventors' searches. For instance, Zheng and Wang (2020) observe a relative decline distant technology search of inventors located in China following the Chinese Google-ban.

[7] Online link to CDC COVID-19 database here (accessed 28/07/2021).

multaneously, the USPTO launched its platform *Patents4Partnerships*, an initiative targeted towards markets for technologies regarding the prevention, diagnosis and treatment of COVID-19, assembling a complete listing of inventions currently available related to the on-going public health crisis.[8]

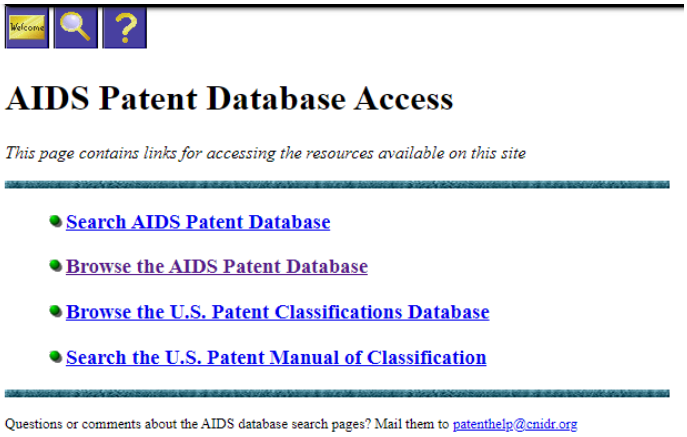
In sum, while there is consensus that patent documents can transfer valuable and specific technical information, inventors face substantial search costs to identify such information from bibliographic archives.

B. The International AIDS Patent 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 deficieny 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), by initiative of USPTO Commissioner Bruce Lehman.[9] Diagnosed HIV-1 infections had dramatically spread since the early-1980s, reaching a peak of $>$ 20,000 cases in 1993 and causing a yearly mortality rate excess due to AIDS of almost 15,000 by 1995 in the U.S. alone.[10] 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 with new HIV/AIDS-related patents issued until February 4th of 1997, to host a final total of 2,916 patents. Figure 1 shows the access page to the AIDS DB which was provided through a link on the USPTO main website (see also Figure 6 in Appendix .B).

[8] Online link to USPTO Patents4Partnerships platform here (accessed 28/07/2021).
[9] Sources: States News Service, October 26, 1994; Federal Technology Report, McGraw-Hill, November 10, 1994; USPTO Press Release #98-12.
[10] Source: U.S. Centres for Disease Control and Prevention (CDC).

FIGURE 1. AIDS DB ACCESS PAGE, FALL 1996



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.

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. The database system was designed and operated by CNIDR parent MCNC, a private, non-profit corporation located in Research Triangle Park, North Carolina. This included the development of a network-accessible search form and the optimization of electronic file formats allowing the inclusion of high-resolution images (complete with drawings, equations and diagrams) at enhanced compression.[11] Declared objective of the new online database was to connect and increase the informational efficiency between dispersed teams of researchers worldwide. Commerce Secretary Ronald Brown emphasized during the launch event on October 26th 1994:

> "This new online health care database is an excellent example of how our nation can best utilize the information superhighway to help improve people's lives, to help expand our knowledge and, most importantly, to connect ourselves with the resources and information previously beyond our reach." (Federal Technology

[11] Source: PR Newswire, Oct. 26 1994, Financial News

Report, Nov. 10, 1994, p.4)

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 [12], 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. The latter option was relatively faster, but also significantly more expensive, with delivery fees of several dozens of current USD per copy.[13] With the new online repository, the external search costs to relevant prior art decreased suddenly for HIV/AIDS related knowledge, as Commissioner Lehman explained at the release ceremony:

> "Anybody with a [personal computer] and a modem will now be able to hook into this database anywhere in the country, and almost all scientific researchers in this area have that capability.[..] Researchers will be able to immediately get into the files and access that nugget of information they have been trying to get for years to complete their work." (States News Service, Oct. 26, 1994; Federal Technology Report, Nov. 10, 1994, p.4)

Projections regarding intensive usage seemed to be rapidly fulfilled: One and a-half years after launch, the AIDS DB recorded about 2.2 thousand requests per day on average (a total of 484 thousand request over a seven-month period). Moreover, these requests originated from a large number of connecting points worldwide, with a total of about 24 thousand distinct hosts served over a seven-month period.[14]

Next to providing electronic access to full patents, by being a disease-targeted repository, the AIDS DB also bore the potential to significantly lower search costs for HIV/AIDS

[12] 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.

[13] Source: Historical website of the USPTO, accessed Feb 12 1997.

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

researchers: 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). This is particularly challenging when searching patents of inventors out of the searching inventor's intellectual network and community. In fact, neither the USPC nor IPC patent classifications contain specific classes denoting HIV/AIDS (or other disease)-related inventions, as these span a broad range of different technology fields and domains. Moreover, the majority of AIDS DB patents did not include any textual reference to HIV/AIDS in title or abstract, which makes their retrieval through key word search comparably difficult. Therefore, the labeling as 'possibly relevant patent', by inclusion in the database, brought a major information retrieval advantage to AIDS DB patents over similar technologies.

For more than four years, the AIDS DB remained the only online repository for full patent information available on the world wide web. The USPTO started to expand its online holdings in 1996, offering a comprehensive web-searchable catalogue of bibliographic front cover information for over 2 million patents, across all fields. But it was not until late 1998 that the bulk volume of patents was made available online with their full text and images. Shortly after, in 1999, also the European Patent Office (EPO) launched its online platform, including full text information for all patents worldwide.[15] The AIDS DB project was discontinued in March 1999, and all hosted patents were included into the main USPTO database.

III. Data

A. *Information about AIDS DB patents*

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

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

the full list of US, EPO and JPO patents at different points in time between 1996 and 1998 (Figure 8 provides an example of the layout of the scraped pages).[16] 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 Figure 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 included in each recorded snapshot, 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.[17] I am able to retrieve the patent numbers of all 2,916 U.S. patents that were deposited until February 1997 (1,668 of which were published at the time of the original launch in 1994, and 1,248 updated subsequently), as well as 695 European and 755 Japanese patents.

B. Patent universe data

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), which constitutes the main data source for my analysis. 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 inventor identities and geo-locations for U.S. and European patents, I extensively rely

[16] The chronologically first available snapshot containing the comprehensive database dates back to June 26th 1997 and was accessed here on June 11th 2020.
[17] See, e.g., here.

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 geo-spatial boundary files provided by the United States Census Bureau, Eurostat and the OECD.[18] 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.

C. *Scientific publication universe data*

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.[19] 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.[20] I further identify all publications of author-inventors relating to HIV/AIDS based on corresponding Medical Subject Headings (MeSH) terms assigned to the *human immune deficiency virus* and *acquired immunodeficiency syndrome* by the National Library of Medicine.[21] MeSH

[18] 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. Metropolitan areas are based on Combined Statistical Areas (CSAs) for the U.S. and OECD Functional Urban Areas (FUAs) for the rest of the world.

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

[20] 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%.

[21] see: https://meshb.nlm.nih.gov/. Identified MeSH terms are: *AIDS; Acquired Immune Deficiency Syndrome; Acquired Immuno-Deficiency Syndrome; Immunodeficiency Syndrome, Acquired;AIDS Arteritis, Central Nervous System; AIDS Dementia Complex;AIDS Serodiagnosis; HIV Seropositivity; HIV Seroprevalence; Lymphoma, AIDS-Related; HTLV-III; Human Immunodeficiency Virus; Human T Cell Lymphotropic Virus Type III; Human T Lymphotropic Virus Type III; Human T-Cell Leukemia Virus Type III; Human T-Cell Lymphotropic Virus Type III; Human T-Lymphotropic Virus Type III; Immunodeficiency Virus, Human; Immunodeficiency Viruses, Human;*

terms linked to publications are based on the MeSH tree version 2016.

D. *Final database*

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.

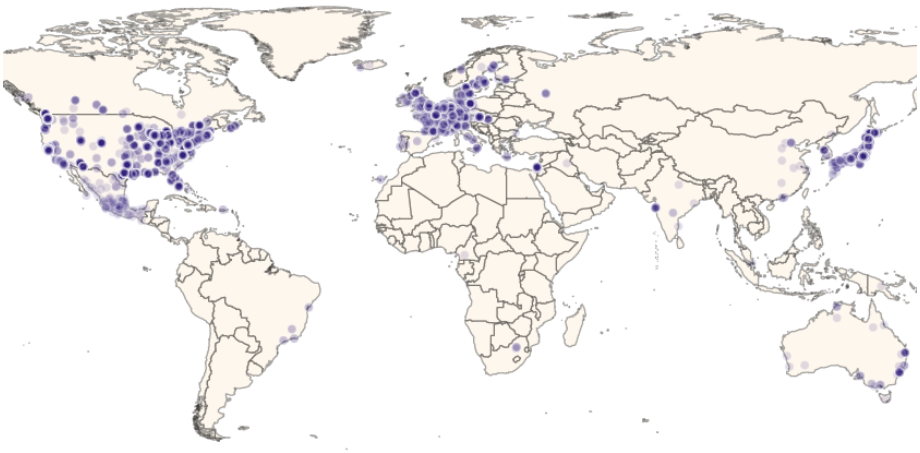
Figure 2 shows that HIV/AIDS-related technologies were developed almost exclusively in the Western hemisphere. Not surprisingly, the most active geographic hubs clustered in the North-East and West Coast of the U.S., Central and Western Europe and Japan. Further significant patenting activity originated from Israel and the East Coast of Australia.

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 in HIV/AIDS-related treatments, accounting for the largest number of patents in each field. Pharmaceutical technologies constitute the largest field share in the database, closely followed by patents in biotechnology and organic fine chemistry. Further well represented fields were medical technology, analysis of biological materials, measurement and chemical engineering. The AIDS DB hosted HIV/AIDS-related patents from a total of 14 distinct technology fields, covering a broad range of inventive domains.

While geographically clustered, the HIV/AIDS inventor community was highly proliferated into small networks. A community detection based on Louvain modularity maximization (Blondel et al., 2008) reveals a large central community cluster, spanning the research groups of later Nobel laureates Françoise Barré-Sinoussi and Luc Montagnier at Institute

LAV-HTLV-III; Lymphadenopathy-Associated Virus; Virus, Human Immunodeficiency; Human T-Cell Leukemia Virus

FIGURE 2. GEOGRAPHIC DISPERSION OF AIDSDB INVENTORS



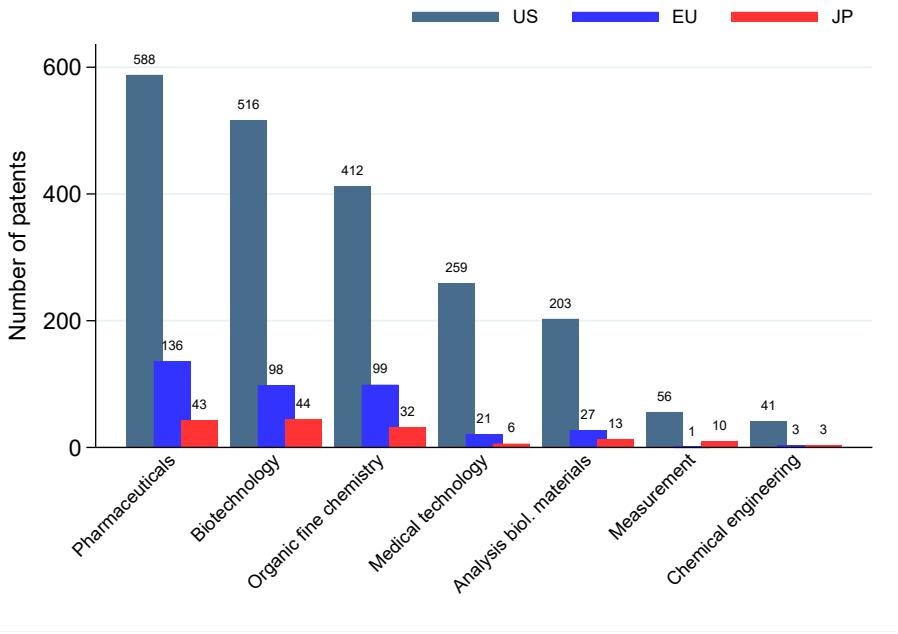
Notes: Shown locations are individual inventor addresses from all AIDS DB patents deposited from Oct 1994 until Feb 1997. Opacity grades indicate intensity of patenting activity. Geo-coded inventor addresses are provided by Morrison, Riccaboni and Pammolli (2017). Coordinates are geo-mapped using QGIS.

Pasteur in Paris and the laboratory of Robert Gallo at the National Institutes of Health in Bethesda, MD, in the U.S., and a large number (> 300) of small and unconnected clusters a few researchers each. Moreover, my assembled 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 Appendix .A; Several scientists (including Luc Montagnier, Samuel Broder, and William Haseltine) ranked both among the top-20 inventors and PubMed authors on HIV/AIDS-research by the end of 1996, and had large networks of collaborators (degree) in each sphere. This provides indication for a close connection between the technology and science frontier in the field.

IV. Effects on cumulative patent citations

This section outlines the empirical design and econometric results for the main analysis of the paper. Section IV.A presents the sample construction and empirical model, and IV.B the econometric results of the cumulative citation impact effects associated with

FIGURE 3. FREQUENCIES OF AIDSDB PATENTS OVER 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.

the launch of the AIDS DB. Heterogeneous impact effects, in line with the mechanism of reduced search costs, are reported in IV.D.

A. Empirical strategy

As addressed above, 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. As discussed in Section III.A, HIV/AIDS research originated a vibrant community of scientists, operating at the frontier of knowledge, and with a strong representation of public research institutions. Moreover, there was a concurring strong public interest in promoting and investment in HIV-related research. This suggests

a positive selection of patents in the AIDS DB based on intrinsic quality and social value, which directly reflect on the potential to cumulative impact.

I solve 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, I estimate 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, I leverage my knowledge about the conditions of external prior art search before the establishment of the AIDS DB. As discussed in Section II.A, until then, inventors interested in retrieving patents related to HIV/AIDS could most exclusively rely on the bibliographic front-page information, provided in library terminals, bulletins or newsgroups, in order to identify those out of the bulk volume of existing patents and order full-text copies to verify their actual content. While patent classes are not disease-specific, and associations from inventor names to HIV-research would typically require direct or indirect network ties, I focus particularly on the information conveyed in titles and abstracts; 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.[22] The assumption behind this approach is that - while inventors are, without doubt, highly-educated specialists in their respective domains and perfectly capable to judge the precise content of pertinent patents upon closer inspection - the likelihood that they will detect a relevant HIV/AIDS-related patent out of a list of bibliographic information of numerous patents will be higher if a

[22]see: https://meshb.nlm.nih.gov/. Queried keywords are: *AIDS; Acquired Immune Deficiency Syndrome; Acquired Immuno-Deficiency Syndrome; Immunodeficiency Syndrome, Acquired; AIDS Arteritis, Central Nervous System; AIDS Dementia Complex;AIDS Serodiagnosis; HIV Seropositivity; HIV Seroprevalence; Lymphoma, AIDS-Related; HTLV-III; Human Immunodeficiency Virus; Human T Cell Lymphotropic Virus Type III; Human T Lymphotropic Virus Type III; Human T-Cell Leukemia Virus Type III; Human T-Cell Lymphotropic Virus Type III; Human T-Lymphotropic Virus Type III; Immunodeficiency Virus, Human; Immunodeficiency Viruses, Human; LAV-HTLV-III; Lymphadenopathy-Associated Virus; Virus, Human Immunodeficiency; Human T-Cell Leukemia Virus* I use the same list of terms to retrieve scientific articles relating to HIV in PubMed.

given patent makes a clear front-page reference to the disease. Accordingly, by marking all deposited patents as disease-related, 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.[23]

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 exploit the richness of information regarding the examination process of U.S. patents, and condition "no reference" and control group ("with reference") patents to be examined in the same art unit.[24] Art units are the most granular inter-organizational units in the examination process of patent applications at the USPTO. Each art unit consists of a team of several patent examiners who specialize in a particular technology.[25] Using art units to detect technologically related patents has several advantages over the use of USPC patent classes in my study; First, given that patent applications are carefully screened and purposefully assigned to the competent art unit for examination, this avoids issues arising from randomness or misclassification that have been studied in the use of primary USPC (sub-)classes (Benner and Waldfogel, 2008; Arts, Cassiman and Gomez, 2018). Second, in particular for drugs and medical domains, art units provide a much more granular assessment device for specific technical content than the broader 3-digit classes, of which there are only eight in the USPC classification.

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

[23] Patents with front-page reference to HIV/AIDS accounted for about one third of AIDS DB patents (N=745).
[24] Given this constraint in data availability, I only consider patent family members filed at the USPTO.
[25] see Righi and Simcoe (2019) for an excellent discussion of the organization of art units at the USPTO.

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 balance in the estimation.[26] By this, I keep all factors relating to the timing of invention, disclosure and online deposit constant across the sample groups. Several examples of patent pairs with vs. without front-page references are discussed extensively in Appendix .C.

To measure the realization of spillovers I predominantly rely on patent citations to AIDS DB patents as proxies. While these measures are widely established, they are known to be imperfect and noisy. In section V.B I, therefore, employ alternative metrics to identify knowledge flows. 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 [27], I remove inventor and applicant self-citations from the counts, as those do not reflect spillovers from external search. I am interested in determining the date closest in time to the inventive effort leading to the future patent. Therefore, 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.

Table 1 reports summary statistics for the within-AIDS DB matched sample. My strict selection criteria allow to pair 1,367 AIDS DB patents.[28] 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. Most noticeable, the small difference

[26] 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. Note, that the exact database deposit dates are only approximated.

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

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

Table 1. Summary statistics, patents with vs. without 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	.13
Generality index	.46	.25	.44	.25	.16
Share breakthroughs (top-5%)	.09		.09		.81
Share novel technologies	.25		.22		.21
Share introducing new words	.35		.31		.14
Share new-to-class medical subjects	.37		.34		.26
Number of patent references	9.25	9.57	7.56	8.12	.00
Share with scientific reference	.90		.90		.95
Number of scientific references	13.90	22.91	11.48	13.66	.03
Number of inventors	3.06	2.04	2.91	1.88	.18
Share of new inventors	.17		.18		.64
Share with author-inventors	.94		.95		.53
Number of author-inventors	2.62	1.90	2.49	1.65	.20
Patent family size	6.98	7.21	5.47	6.44	.00
Share private firm patents	.64		.64		.97
Assignee prior patent families	2.56k	11.15k	2.55k	5.29k	.98
DB-to-publication lag (m)	19.07	21.46	18.74	21.23	.79
DB-to-application lag (m)	55.81	27.94	56.15	28.47	.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 t_0 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).

in the level of citations received in the year pre-AIDS DB deposit across "no reference" and "reference" groups (see row (1), Table 1) is not statistically significant. Sample patents are highly comparable also on a broad range of relevant ex-ante patent-level characteristics,

with no significant differences across the two groups, except for the small deviations regarding the mean number of prior art references and the scope of geographic patent protection (family size). In particular, patents without front-page references to HIV/AIDS do not appear to be more general in applicability, are not more often 'hit' patents (top-5% most highly cited in a given technology class-year cohort), nor more often recombinant (novel) patents.[29] They do not introduce more new keywords and do not reference scientific prior art in previously unconnected medical subjects to the technology class.[30] Patents across sample groups do also not significantly differ in terms of inventor team or assignee characteristics. Table 1 further shows the high science-intensity of inventions in the sample: About 90% of included patents make prior art references to scientific publications, while the average within-patent share of inventors with prior scientific publications is even at 95%. Summary statistics also reveal strong involvement of public research institutions, accounting for one third of patent assignees. Average lags between deposit and patent publication (18 months) and application (55 months) indicate that I observe a large share of the sample patents for a significant time span before database inclusion.[31]

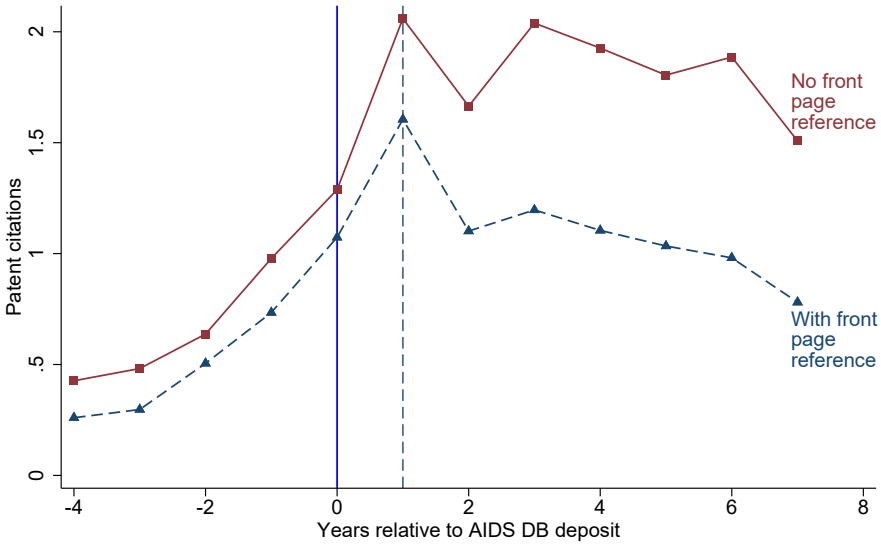
In order to evaluate whether small differences between the two sample groups might cause them to diverge dynamically, i.e. in pre-period trends of citations, I inspect group means over time prior to inclusion in the online database. Figure 4 plots almost perfectly parallel trends between "no reference" 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. Figure 4 also reveals that starting from $t + 1$ differences in group means can be seen to substantially increase, while trends still follow largely parallel patterns. Note, that the spike in patent citations observed in the first period after deposit date is due to the drastic increase in overall patent

[29] "Generality" is based on the index proposed by Trajtenberg, Henderson and Jaffe (1997). "Novel patents" are those making unprecedented combinations in technological prior art cited at the IPC-6-level, based on Verhoeven, Bakker and Veugelers (2016))

[30] New keywords are based on Arts, Cassiman and Gomez (2018). Medical subjects are retrieved based on MeSH terms assigned to scientific prior art references in PubMed (MeSH tree version 2018). SNPR disambiguation comes from Marx and Fuegi (2020)

[31] Online deposit of AIDS DB patents is assumed to take place on average one month after patent publication, compare Section III.A

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).

filings at the USPTO immediately prior to the enactment of reforms to the U.S. patent system to implement provisions of the TRIPS agreement, that became effective on June 8$^{\text{th}}$ 1995 (see Figure 10 in Appendix .B).[32]

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:

$$(1) \qquad Y_{it} = \beta_1 * no\ reference_i \times post_{t-1} + patentFE_i + yearFE_t + \phi_y + \theta_{fy} + \epsilon_{it},$$

[32]These included, among others, a change in maximum patent term (from 17 years post-grant to 20 years post-filing), as well as the establishment of provisional applications, in line with the TRIPS agreement.

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. Control patents have the function to provide a reference level of citations that would have been received by a matched AIDS DB patent from database deposit in the absence of a shock to search costs, e.g. if the database would have been a non-disease specific online repository. 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 $post_{t-1}$ indicator denotes the one-year lagged post-deposit status.[33] 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. This allows to keep constant a broad range of patent-level characteristics in a flexible and non-parametric manner. 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 $post_{t-1}$ period indicator, which is therefore also omitted from the specification. To control for the confounding influence of shocks possibly affecting

[33] Allocating the treatment to set in one year after database deposit seems a conservative lower bound for intertemporal spillovers from newly accessed prior art information to plausibly affect follow-up inventions. 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).

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 ϕ_y). Finally, in the preferred specification, I include linear field-year trends (θ_{fy}) to control for idiosyncratic variation in productivity of specific technology fields, for example, in up-rising biotechnology in the mid-1990s.[34]

I estimate regression (1) on a symmetric sample window of five years preceding and five years following the switching of the $post_{t-1}$ indicator, i.e., for example, ranging from October 1991 to October 2000 for patents deposited in the initial launch of the AIDS DB database on October 26$^{\text{th}}$ 1994. The fact that some patents are represented only for later years in the sample window can be considered innocuous, as it is accounted for by the pairing (and weighting) of patents across groups, based on same filing/grant timing and individual year fixed effects. I report regression results of cumulative citation models, primarily, as OLS estimates.[35] For this, I standardize the number of yearly citations to mean zero and standard deviation one within technological fields.[36] 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 or the commonly used ln(n+1) transformation of the data (e.g., Silva and Tenreyro, 2006). For comparison and robustness, however, I complement all results with estimates from Poisson pseudo-maximum likelihood models.[37] Given the underlying count distribution, these are likely to be the most efficient estimator and model the conditional mean of citations most accurately. Moreover, they provide asymptotically correct standard errors even with over-dispersion, as citation counts are likely to exhibit

[34] Recent contributions in the treatment effects literature have expressed concerns regarding the naive use of two-way fixed effects estimators with staggered treatment adoption arising from bias induced by unequal weighting of individual two-by-two estimators due to variance in treatment effects over time (Goodman-Bacon, 2021; De Chaisemartin and d'Haultfoeuille, 2020). In my specification, these concerns are mitigated in several ways: First, treated and control group units are fixed over time, without within-unit variation ("switching") in group status at any point of the sample. Second, to account for variance in treatment effects over time, next to event years, my model includes year dummies to capture absolute time-varying effects (as suggested by Goodman-Bacon (2021)) and units across groups are matched pairwise on timing. Third, the test suggested by De Chaisemartin and d'Haultfoeuille (2020) formally excludes the presence of any negative weights in the sum of all average treatment effects in my sample.

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

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

[37] similar to Bryan and Ozcan, 2020; Biasi and Moser, 2021

(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. Within AIDS DB comparison: 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 reveived 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.[38]

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

[38] In unreported results, I do not find any significant heterogeneity of effects conditional on pre-DB citation levels. In particular, I also don't find that inclusion into the AIDS DB would have impacted the likelihood for "no reference" patents of entering the top-impact ranks of a given field-year distribution. This is consistent with the view that patent that already had a strong visibility in a field, evidenced by prior received citations, would have not disproportionately benefited from online indexing.

Table 2. Effect for patents without HIV/AIDS front page reference, within AIDS DB

Dependent variable:	OLS					
Number of patent citations	(1)	(2)	(3)	(4)	(5)	(6)
No reference × post$_{t-1}$	0.135***	0.123***	0.134***	0.150***	0.181***	-0.046
	(0.044)	(0.044)	(0.042)	(0.044)	(0.046)	(0.044)
Abstr reference × 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	.388	.389	.433	.495	.389	.060
Mean at t_0	1.210	1.210	1.178	1.194	1.210	.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).

larger for this sub-group (+28%, significant at the 1% level), however accounting for the vast majority of incoming 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.[39] Estimates show that the effect is magnified (increase of .15 standard deviations, significant at 1%) when excluding these citations.[40]

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 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 to certain AIDS DB patents, as their active involvement in the assembly process likely increased their attention to them as well.[41] I evaluate the severeness of this concern, in column (6) of Table 2, by re-estimating regres-

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

[40] 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.

[41] Examiner citations typically account for about 40% of citations included in U.S. patent documents. As examiner added citation cannot reflect knowledge flows among inventors, they have the potential to introduce significant noise in these type of analyses (Alcacer and Gittelman, 2006; Thompson, 2006). 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. Note, however, that as long as their share does not unilaterally change over time for either HIV/AIDS or control group patents, the presence of examiner added citations is innocuous to my estimation given the specification I employ.

sion (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.[42] These citations are very likely to be given by examiners rather than by the applicants.[43] 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 behaviour in two ways: First, I re-estimate the main specification in Table 2 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. Table 3 provides the corresponding estimates: 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 with relatively +32% even slightly larger in size (column (1) of Table 3, significant at the 1% level). Accounting for variation in citations across fields over time in column (2) of Table 3 yields an adjustment of the effect to +29% (point estimate 0.12, significant at the 1% level), which is perfectly consistent with the main result in Table 2.

[42] 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.

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

TABLE 3. EFFECT FOR PATENTS WITHOUT HIV/AIDS FRONT PAGE REFERENCE, EXCLUDING TOP-CITING EXAMINERS

Dependent variable:	*Excl. top 1% citing examiners*		*Excl. top 5% citing examiners*	
Number of patent citations	(1)	(2)	(3)	(4)
No reference × post$_{t-1}$	0.132***	0.120**	0.146***	0.131**
	(0.047)	(0.047)	(0.053)	(0.051)
Patent fixed effects	Yes	Yes	Yes	Yes
Time/ year fixed effects	Yes	Yes	Yes	Yes
Field time trends		Yes		Yes
Observations	12,192	12,192	12,192	12,192
Number of patents	1,366	1,366	1,366	1,366
R^2	.412	.414	.359	.361
Mean at t_0	.962	.962	.709	.709
SD at t_0	2.345	2.345	2.190	2.190

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, excluding citations from patents inspected by the top 1% (n=10) and top 5% (n=50) of examiners in number of patents citing the AIDS DB. 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)). 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).

The estimated yearly citation delta is even considerably higher (+ 40%) when excluding the 50 most frequently citing examiners, as reported in columns (3)-(4) in Table 3.

Second, I investigate changes in the individual citing behaviour in response to the AIDS DB launch across all patent examiners at the USPTO. For this, I construct a data set where each observation provides the count of citations from patents inspected by a given examiner in a given year (based on publication year) to a patent in the AIDS DB sample.[44] Including a full set of citing examiner fixed effects, results reported in Table 9 in Appedix .A show

[44]To ensure a minimal relatedness to contents, I restrict the examiner sample to only examiners who inspected at least one patent with a prior art citation - either before or after database deposit - to an AIDS DB patent during the sample period.

no significant deviation for examiners to cite "no reference" compared to "with reference" AIDS DB patents post online-deposit, relative to their individual sample mean.[45] This is robust for citations from only examiners who were previously to online-deposit citing HIV patents, as well as from only those citing in the same technology field and art unit (columns (2)-(4) in Table 9). Taken together, these sensitivity checks mitigate concerns that (changes in) examiner behaviour 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 Appendix .A is show further robustness of the entirety of the findings in this Section with quantitatively largely unchanged results: In Table 10 in Appendix .A, I re-estimate all models with Poisson pseudo-maximum likelihood, which is the more efficient estimator given the count nature of the citation data, 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 of 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 of Appendix .A, I provide alternative results for effects on impact weighted forward citations, suggesting real economic effects behind the observed increased knowledge flows.

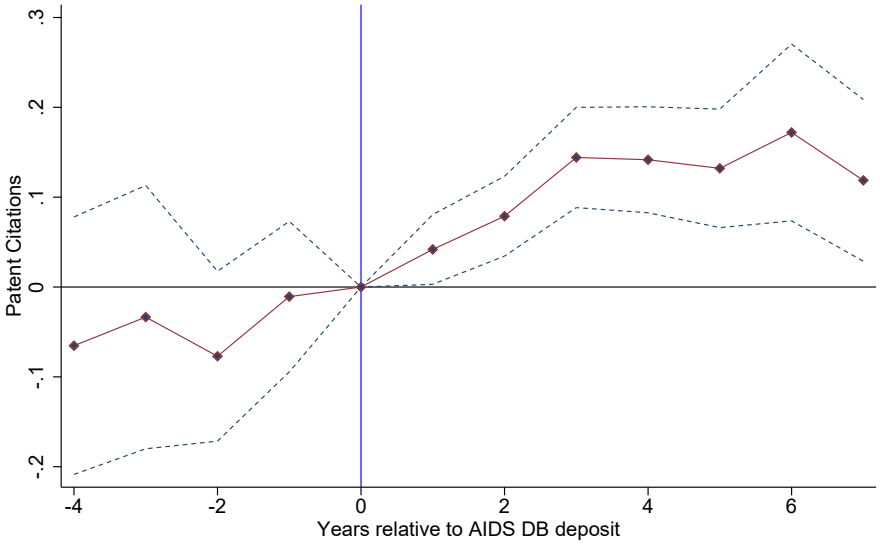
C. Timing of effects

To investigate the timing of the aggregate attention effects reported in Section IV.B, I re-estimate regression (1) with yearly coefficients, by interacting the "no reference" indicator

[45] The idea behind this approach is that, as shown by prior literature, the assignment of patents to individual examiners can be considered as quasi-random within art units. Therefore, a large share of the variation stemming from increased applicant citations to "no reference" AIDS DB patents should be absorbed be the examiner fixed effects, unless there would be a significant change in citation behaviour of the examiners themselves.

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 IV.B also indicates a stable and persistent effect over time, as estimated differ-

ences, 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. Mechanism: Differential effects for patents with intrinsically higher search costs

I further investigate 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 4 shows estimation results for regression (1) as triple-differences for heterogeneity of marginal impact of the online repository for split-samples of these patents.[46] Columns (1)-(2) of Table 4 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 4 indicate that corporate assignee patents without front-pages references to HIV/AIDS received .17 standard deviations in citations more after AIDS DB deposit

[46] This econometric specification compares changes in cumulative citations of, e.g., *private firm* patents with front-page 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.

TABLE 4. DIFFERENTIAL EFFECTS FOR PATENTS WITH INTRINSICALLY HIGHER 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 × post$_{t-1}$ × *cat*	0.166*	0.119	0.252**	0.214*	0.291*	0.291*
	(0.086)	(0.086)	(0.109)	(0.109)	(0.176)	(0.156)
Post$_{t-1}$ × *cat*	-0.036	-0.008	-0.101	-0.092	-0.304**	-0.278**
	(0.072)	(0.068)	(0.070)	(0.076)	(0.148)	(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	.441	.433	.441	.434	.442	.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).

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

4 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.[47]

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.[48] 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 4 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* front-page 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 4 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

[47] In unreported additional estimation results, I do not find significant heterogeneity of effects for other institutional categories of patent applicants, namely universities, hospitals and government research organizations. In particular government institutions, such as e.g. the National Institutes of Health in the U.S. or Institut Pasteur in France, held a leading and exposed role in the development of HIV-related treatments. Therefore, their patents enjoyed high visibility irrespective of the AIDS DB. These observations are consistent with the view that database inclusion would have disproportionately benefited the diffusion of private sector invention. Note however, that for some of the non-corporate assignee categories my sample size is very small, e.g. hospitals (n = 18), which makes it statistically difficult to estimate differential effects.

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

references. 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.

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 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.

My 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 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 Appendix .A), in particular no disproportional gains from higher visibility of patents from more peripheral inventors to the community.[49] 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.

[49] Estimation results unreported.

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.[50] In a next step, I link these reference citations back to the characteristics and timing of AIDS DB deposit of the focal patent(s) and construct an SNPR panel, where each observation is a PubMed article referenced in an AIDS DB patent cited in a given year relative to the online deposit (of the AIDS DB patent).[51]

Table 5 reports results for the estimation of equation 1 on this SNPR-level panel. The coefficient estimate in column (1) of Table 5 implies that scientific articles referenced in a AIDS DB patent without front-page reference to HIV/AIDS experience an additional .071 standard deviations in yearly citations after online deposit of the referencing patent compared to patents with obvious front-page links to HIV/AIDS (signifcant at the 1% level). This corresponds to an increase of + 17.9% in patent references to these articles relative to the pre-treatment level. When accounting for technology field trends, in column (2) of Table 5, this effect is estimated to be even larger, at + 18.6% (significant at the 1% level). In column (4) of Table 5, I include scientific publication fixed effects in the model,

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

[51] Note that in this case, unlike in the main estimation, it is possible that scientific articles appear both in the treated and control group, as they may be referenced by multiple patents. In the case of articles cited simultaneously in both a "no reference" and "with reference" patents, inference will be limited to variation stemming from different timing and/ or references unique to one of the two groups.

TABLE 5. SECOND ORDER EFFECT ON SNPRS IN PATENTS WITHOUT HIV/AIDS FRONT PAGE REFERENCE

Dependent variable:	OLS			
Number of patent citations to scientific references	(1)	(2)	(3)	(4)
No reference × post$_{t-1}$	0.071***	0.074***	0.074***	0.069**
	(0.026)	(0.026)	(0.028)	(0.027)
Patent fixed effects	Yes	Yes	Yes	Yes
Time/ year fixed effects	Yes	Yes	Yes	Yes
Field time trends		Yes	Yes	Yes
Referenced paper fixed effects			Yes	Yes
Paper year time trends				Yes
Observations	96,532	96,532	96,532	96,472
Number of patents	1,086	1,086	1,086	1,086
Number of scientific papers	7,848	7,848	7,848	7,843
R^2	.318	.318	.690	.692
Mean at t_0	2.464	2.464	2.464	2.464
SD at t_0	6.197	6.197	6.197	6.197

Notes: Each column reports parameter estimates of regression (1) for the matched panel of scientific references to articles in PubMed in U.S. AIDS DB patents without vs. with front-page reference to HIV/AIDS. The dependent variable measures the yearly number of family citations to scientific articles referenced in AIDS DB patents 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)). Sample observations are weighted based on Iacus, King and Porro (2012) and 1/ number of SNPRs in each patent. 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 order to account for idiosyncatic, time-invariant differences in potential of technological applicability across scientific discoveries. The coefficient estimate is robust and significant at the 1% level. We additionally controlling for variation of citation rates over the life-cycle of publications, with publication year × citing year fixed effects, in column (4) of Table 5, the relative effect size is minimally reduced to + 17.4% (significant at the 1% level). This 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.

V. Changes in the quality and reach of knowledge spillovers

In this section, I investigate 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.[52] 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.[53] 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.[54] Finally, I balance the 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:

$$\begin{aligned} Y_{ijy} =& \beta_1 * no\ reference_j \times post1994_{t-1} \\ &+ \beta_2 * with\ reference_j \times post1994_{t-1} \\ &+ \beta_3 * no\ reference_j + \beta_4 * with\ reference_j \\ &+ citPatent \times citedYearFE_{i\times y} + \theta_{tr} + \epsilon_{ij}, \end{aligned} \tag{2}$$

where i indexes citing patents, t indexes years, y indexes cited years and r indexes citing geographic regions. Y_{ijy} is the generic dependent variable quantifying the quality and

[52] Another possible approach to estimating this would be to compare changes in citation-level quality and reach *within* cited patents over time, similar to the approach used in Section IV), thus controlling for all time invariant heterogeneity across AIDS DB patents. Empirically however, given that by construction it conditions on receiving citations in both periods of analysis, which leads to very few patents, and accordingly observations (N $<$ 100), in several of the split-samples studied in this section, this would make statistical inference difficult due to very large confidence intervals estimated.

[53] 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.

[54] 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.

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 $\times$ 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.[55] 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.[56] Summary statistics for the cited-citing level sample are reported in Table 15 in Appendix .A.

[55] 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.

[56] I assign each patent to a unique country, based on the most frequent occurring inventor country-location listed on the patent. In the rare case of multiple equally frequent countries, I chose a random location.

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.

Table 6 reports the results from the estimation of regression (2) for the re-occurrence of new words and novel SNPRs in citing patents (columns (1) and (3)), split up as triple-differences for citations from private firm patents in columns (2) and (4).[57] I identify "new words" as unique keywords on a specific patent that appear for the first time in the universe of all U.S. patents since 1976, using the patent text data provided by Arts, Cassiman and Gomez (2018). New keywords are a suitable measure to proxy the transfer of unique knowledge elements, as they occur comparatively often in patents, in about 30% of patents in my sample (compare Table 1). Furthermore, new scientific and technical words have been used as an alternative measure to patent citations for tracing knowledge flows in prior literature (e.g., Iaria, Schwarz and Waldinger, 2018; de Rassenfosse, Pellegrino and Raiteri, 2020; Baruffaldi and Pöge, 2020). The likelihood of re-use of new words introduced increased for both AIDS DB patents without and with front-page reference after database deposit, as shown in column (1) of Table 6. This appears to have been much more pronounced for the "with reference" group of patents, suggesting gains from online accessibility to likely be driving these results. Indeed, the split-sample differences estimated in column (2) reveal that effects on "with reference" patents were large and significant (at the 1% level) only for the sub-set of citing corporate patents, while not different from the control group for public research institutions' citations (captured by the main category interactions in column (2)). On the other hand, the latter seemed to have benefited more from enhanced access to external patents more difficult to identify as HIV-related. I re-investigate this pattern for the likelihood of re-occurrence of novel

[57] All regression models control for the number of new words introduced and scientific references made by each cited patents.

Table 6. Effects on quality of generated spillovers

Dependent variable:	*New word*		*Novel SNPR*	
Probability of re-occurrence	(1)	(2)	(3)	(4)
No reference × post1994$_{t-1}$	0.012*	0.027**	0.031***	-0.021**
	(0.007)	(0.012)	(0.006)	(0.008)
With reference × post1994$_{t-1}$	0.068***	0.016	-0.011	-0.139***
	(0.006)	(0.019)	(0.030)	(0.028)
No reference × post1994$_{t-1}$		-0.008		0.084***
× private firm citing		(0.068)		(0.014)
With reference × post1994$_{t-1}$		0.075***		0.188***
× private firm citing		(0.025)		(0.028)
Main category interactions	Incl	Incl	Incl	Incl
Non-new/novel interactions	Incl	Incl	Incl	Incl
Citing patent × 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	.261	.263	.359	.368
Conditional mean at t_0	.028	.015	.095	.060

Notes: Each column reports parameter estimates of regression (2) on citing patents-cited year pairs for the subset of cited patents introducing new words (columns (1)-(2)) and referencing novel scientific publications (columns (3)-(4)). "New words" are unique keywords appearing for the first time on a patent in the universe of U.S. patents since 1976. "Novel SNPR" are non-patent references to scientific publications in PubMed which are in the top-5% of the distribution of new medical subject term combinations in a their respective year and field of publication, following Boudreau et al. (2016). The dependent variable measures the probability of re-occurrence of a new word and novel scientific reference in citing U.S. patents filed between 1991 and 2000. Inventor and applicant self-citations are excluded. Displayed are parameter estimates for patents introducing new words and referencing novel science only. Full triple interactions for patents with no new words and non-novel science are included. Controls for the number of new words and the number of scientific references are included. The reference category consists of cited non-AIDS DB patents, published in the same year, within the same application. Sample observations are weighted in order to give equal weight to each citing patent. Standard errors are clustered at citing country-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*, *PubMed* and several disambiguations (links) between them. Patent key words are obtained from Arts, Cassiman and Gomez (2018) (see Section III for details).

scientific references. I measure "novel SNPRs" as references to scientific publications that score among the top-5% of publications in a scientific field-year cohort regarding the share of unprecedented combinations of MeSH terms assigned in the universe of all biomedical

research articles.[58] Technological usefulness and application potential of such recombining scientific advances is, arguably, harder to realize and less visible outside of the sphere of basic science and fundamental research. Findings reported in columns (3)-(4) of Table 6 support this prediction, as estimates point to a disproportionate increase in likelihood of re-citation of novel scientific advances in patents without front-page reference to HIV/AIDS, and large and positive effects concentrated in the sub-set of private sector citations which, in the case of "no reference" patents, are about 4 times larger than for academic citations.

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.[59] The share of international citation links is then simply given by the inverse of the overlap, e.g. 0% if all citing and cited inventors are located in the same country, and vice versa.[60] 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.[61] 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

[58] Compare, e.g., Boudreau et al. (2016) for a similar measure of scientific novelty.

[59] for each patent-inventor instance, I use all disambiguated locations for the same inventor listed on patent documents for the patent family (see Section III.A for details)

[60] 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: $(1 \times .5 + 1 \times .5) \times .5 + (1 \times .5 + 0 \times .5) \times .5 = .75$

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

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. patents and > 16 million authors indexed in PubMed and their realm-transcending collaborative ties.[62]

Table 7 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 $= \infty$, no finite shortest path) as outcome variables. In the main effect specifications (columns (1) and (3)) of Table 16, 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-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.[63]

When looking at heterogeneity at the receiving end, estimates in column (2) of Table 7 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

[62]Inventor and author identities are disambiguated based on Li et al. (2014); Morrison, Riccaboni and Pammolli (2017); Smalheiser and Torvik (2009); Torvik and Smalheiser (2009). Inventor information covers years 1976-2011, author information years 1858-2009. For details, see Section III.

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

TABLE 7. EFFECTS ON THE REACH OF GENERATED SPILLOVERS

Dependent variable:	*International*		*Detached community*	
Probability of distant citation	(1)	(2)	(3)	(4)
No reference × post1994$_{t-1}$	0.020**	0.058***	0.007*	-0.012**
	(0.009)	(0.006)	(0.004)	(0.006)
With reference × post1994$_{t-1}$	-0.026***	-0.110***	-0.021**	-0.036***
	(0.009)	(0.022)	(0.008)	(0.010)
No reference × post1994$_{t-1}$ × private firm citing		-0.049***		0.027***
		(0.016)		(0.008)
With reference × post1994$_{t-1}$ × private firm citing		0.119***		0.023
		(0.043)		(0.015)
Main category interactions		Incl		Incl
Citing patent × 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	.556	.556	.731	.730
Mean at t_0	.352	.352	.201	.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 country-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*, 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).

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 7 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 Appendix .A. I further show robustness of findings for social distance metrics based exclusively on inventor network graphs in Table 18 in 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. This can be seen also from summary statistics in Table 15 in Appendix .A: Shortest paths between cited and citing patents in the inventor graph are, on average, 6.8 degrees long for AIDS DB patents prior to AIDS DB launch (6.9 for control patents). When considering prior collaborations on basic research articles, however, this distance decreases to only 3.5 degrees on average (3.6 for control group patents). 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.

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 my 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.

My findings, 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 my 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.

Arts, Sam, Mattia Nardotto, Thomas Schaper, and Jesse Wursten. 2020. "The Impact of Broadband Internet on Invention: Evidence from the U.K." *Working Paper*.

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 patent-based 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. 2020. "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, Clément, 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 fédérale 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, Jurriën 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.

A Tables

Table 8. Top-20 HIV/AIDS inventors vs. authors, by the end of 1996

Top-20 AIDSDB inventors

Name	Location	↓Patents	Citations	Degree	Centrality
Montagnier, Luc	Paris, France	36	557	52	.091
Mueller, Richard A.	Chicago, IL, U.S.	28	378	22	.005
Schinazi, Raymond F.	Atlanta, GA, U.S.	20	379	19	.008
Gallo, Robert C.	Bethesda, MD, U.S.	20	626	42	.254
Sessler, Jonathan L.	San Francisco, CA, U.S.	18	853	18	.001
Paessens, Arnold	Duesseldorf, Germany	17	129	33	.004
Hargrave, Karl D.	Ridgefield, CT, U.S.	16	93	12	.000
Chu, Chung K.	Atlanta, GA, U.S.	16	234	11	.004
Norbeck, Daniel W.	Chicago, IL, U.S.	15	498	19	.000
Carter, William A.	Philadelphia, PA, U.S.	15	157	0	-
Fleet, George W.	Oxford, U.K.	14	81	10	.005
Clavel, Francois	Paris, France	13	174	13	.000
Geutard, Denise	Paris, France	13	174	13	.000
Hoshino, Hiroo	Takasaki, Japan	13	68	28	.004
Krenitsky, Thomas A.	Chapel Hill, NC, U.S.	13	136	9	.000
Hemmi, Gregory W.	San Francisco, CA, U.S.	12	653	7	.000
Broder, Samuel	Bethesda, MD, U.S.	12	214	18	.248
Alizon, Marc	Paris, France	12	202	14	.001
Rideout, Janet L.	Chapel Hill, NC, U.S.	12	175	13	.001
Haseltine, William A.	Boston, MA, U.S.	12	203	19	.002

Top-20 AIDSDB PubMed authors

Name	Location	↓HIV articles	Citations	Degree	Centrality
Levy, Jay A.	San Francisco, CA, U.S.	154	9,417	271	.087
Ho, David D. A.	New York, NY, U.S.	107	16,748	289	.125
Montagnier, Luc	Paris, France	106	3,752	307	.121
Sodroski, Joseph G.	Boston, MA, U.S.	105	9,146	216	.047
Mitsuya, Hiroaki	Bethesda, MD, U.S.	103	5,089	220	.051
Baba, Masanori	Fukushima, Japan	86	3,748	165	.023
Broder, Samuel	Bethesda, MD, U.S.	84	5,003	183	.040
Matthews, Thomas J.	Durham, NC, U.S.	82	4,796	266	.072
Haseltine, William A.	Boston, MA, U.S.	81	4,862	128	.027
Nakashima, Hideki	Tokyo, Japan	77	1,700	211	.022
Nara, Peter L.	Frederick, MD, U.S.	75	3,590	244	.069
Bolognesi, Dani P.	Durham, NC, U.S.	73	3,827	164	.022
Chermann, Jean-Claude	Paris, France	71	1,335	177	.061
Guertler, Lutz G.	Munich, Germany	70	813	175	.039
Weiss, Robert A.	London, U.K.	67	5,452	176	.047
Yarchoan, Robert	Bethesda, MD, U.S.	67	3,605	182	.035
Berzofsky, Jay A.	Bethesda, MD, U.S.	64	4,089	148	.015
Busch, Michael P.	San Francisco, CA, U.S.	63	2,708	188	.047
Lane, H. Clifford	Bethesda, MD, U.S.	62	3,018	194	.026
De Rossi, Anita	Padua, Italy	62	1,494	170	.057

Notes: Inventor betweenness centrality values multiplied by ×100, author centrality by by ×10

TABLE 9. RESPONSIVENESS TO AIDS DB DEPOSIT OF CITATION RATES FROM PATENTS WITHIN INDIVIDUAL EXAMINERS

Dependent variable:	OLS			
Number of patent citations	(1)	(2)	(3)	(4)
No reference × post$_{t-1}$	0.002	-0.002	-0.015	-0.049
	(0.003)	(0.010)	(0.021)	(0.053)
Patent/ year/ field fixed effects	Yes	Yes	Yes	Yes
Citing examiner fixed effects	Yes	Yes	Yes	Yes
Pre-deposit DB cites		Yes	Yes	Yes
Pre-deposit same IPC cites			Yes	
Pre-deposit same art unit cites				Yes
Observations	10,501,376	2,377,280	928,913	296,069
Number of citing examiners	1,016	230	230	230
R^2	.005	.009	.018	.041
Mean at t_0	.0001	.0007	.0017	.0053
SD at t_0	.014	.029	.046	.081

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 by individual examiners at the USPTO for years $t-4$ to $t+5$ relative to the one-year lagged online date. I have no information whether citations are given by the examiners or applicants. 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)). "Pre-deposit DB cites" indicates a sample restriction to exclusively citations of examiners with prior citations to patents in the AIDS DB before DB deposit. "Same IPC" indicates a sample restriction to citations of examiners pre-DB citing AIDS DB patents in the same technological field as a cited patent. "Same art unit" indicates a sample restriction to citations of examiners pre-DB citing AIDS DB patents in the art unit. 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).

TABLE 10. EFFECT FOR PATENTS WITHOUT HIV/AIDS FRONT PAGE REFERENCE, WITHIN AIDS DB, PPML ESTIMATES

Dependent variable:	Poisson					
Number of patent citations	(1)	(2)	(3)	(4)	(5)	(6)
No reference × post_{t-1}	0.330***	0.301***	0.297***	0.304***	0.521***	-0.925**
	(0.109)	(0.112)	(0.110)	(0.111)	(0.142)	(0.413)
Abstr reference × post_{t-1}					0.352*	
					(0.189)	
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	10,727	10,717	10,624	10,684	10,717	2,761
Number of patents	1,196	1,195	1,185	1,191	1,195	309
Mean at t_0	1.210	1.210	1.178	1.194	1.210	.016

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 from a Poisson pseudo-maximum likelihood estimation (correct for over-dispersion by construction) to address the count nature of citations. 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. "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).

TABLE 11. EFFECT FOR PATENTS WITHOUT HIV/AIDS FRONT PAGE REFERENCE, WITHIN AIDS DB, ONLY PATENTS OBSERVED IN ALL SAMPLE YEARS

Dependent variable:	OLS					
Number of patent citations	(1)	(2)	(3)	(4)	(5)	(6)
No reference × post$_{t-1}$	0.154**	0.140**	0.158***	0.169***	0.212***	-0.142**
	(0.062)	(0.062)	(0.060)	(0.063)	(0.066)	(0.056)
Abstr reference × post$_{t-1}$					0.122	
					(0.091)	
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	6,535	6,535	6,514	6,525	6,535	6,240
Number of patents	655	655	653	654	655	624
R^2	.428	.431	.477	.469	.431	.052
Mean at t_0	1.707	1.707	1.640	1.680	1.707	.038
SD at t_0	3.271	3.271	3.168	3.251	3.271	.232

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 only for patents observed in each year of the sample period. 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).

TABLE 12. EFFECT FOR PATENTS WITHOUT HIV/AIDS FRONT PAGE REFERENCE, WITHIN AIDS DB, MATCHED ON PRE-TRENDS

Dependent variable:	OLS					
Number of patent citations	(1)	(2)	(3)	(4)	(5)	(6)
No reference × post$_{t-1}$	0.155***	0.147**	0.124**	0.147***	0.239***	0.030
	(0.059)	(0.058)	(0.057)	(0.056)	(0.053)	(0.047)
Abstr reference × post$_{t-1}$					0.166**	
					(0.081)	
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	6,774	6,774	6,818	6,829	6,774	6,661
Number of patents	775	775	781	781	775	768
R^2	.225	.233	.261	.278	.232	.022
Mean at t_0	.492	.492	.493	.499	.492	.006
SD at t_0	1.035	1.035	1.040	1.038	1.035	.080

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, additionally matched on pre-period yearly citation levels. 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).

TABLE 13. EFFECT FOR PATENTS WITHOUT HIV/AIDS FRONT PAGE REFERENCE, WITHIN AIDS DB, IMPACT WEIGHTED CITATIONS

Dependent variable:	OLS					
Log(impact weighted citations)	(1)	(2)	(3)	(4)	(5)	(6)
No reference × post_{t-1}	0.159**	0.138*	0.137*	0.142*	0.226**	-0.014
	(0.078)	(0.078)	(0.076)	(0.078)	(0.099)	(0.018)
Abstr reference × post_{t-1}					0.155	
					(0.132)	
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	12,070
Number of patents	1,367	1,366	1,366	1,366	1,366	1,366
R^2	.435	.443	.407	.404	.442	.082
Mean at t_0	20.207	20.207	19.954	19.974	20.207	.159

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 10-year impact weighted family citations for years $t-4$ to $t+5$ relative to the one-year lagged online date. Weighted citation counts are transformed to their natural logarithm, adding a small quantity (+1) to each count. 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).

Table 14. Differential effects for patents based on recency of grant, within AIDS DB

Dependent variable:	*Top 25%*	*Top 50%*	*Oldest 25%*	*Oldest 10%*
Number of patent citations	(1)	(2)	(3)	(4)
No reference × post_{t-1}	0.194***	0.167**	0.136***	0.151***
	(0.056)	(0.071)	(0.047)	(0.044)
Post_{t-1} × *recent*	0.300***	0.284***	-0.236***	-0.167*
	(0.095)	(0.081)	(0.078)	(0.091)
No reference × post_{t-1} × *recent*	-0.093	-0.038	0.027	-0.101
	(0.083)	(0.089)	(0.104)	(0.153)
Patent fixed effects	Yes	Yes	Yes	Yes
Time/ year fixed effects	Yes	Yes	Yes	Yes
Field time trends	Yes	Yes	Yes	Yes
Observations	12,192	12,183	12,192	12,183
Number of patents	1,366	1,366	1,366	1,366
R^2	.507	.508	.507	.507
Mean at t_0	1.210	1.210	1.210	1.210
SD at t_0	2.570	2.570	2.570	2.570

Notes: Each column reports parameter estimates of regression (1) split up as triple-differences for heterogeneity of effects on patents based on grant recency in the matched panel of U.S. AIDS DB patents without vs. with front-page reference to HIV/AIDS. "Top 25%" most recent patents are granted less than one year prior to online deposit. "Oldest 10%" patents are granted more than 4 years prior to online deposit. The $post_{t-1}$ parameter captures relative changes in citations to patents with front-page reference to HIV/AIDS in each split-sample category. 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. 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).

TABLE 15. SUMMARY STATISTICS, QUALITY AND REACH OF SPILLOVERS PRE-AIDS DB

Within citing patents comparison	Control	No reference		With reference		Diff
	Mean	Mean	SD	Mean	SD	p-val.
Re-occurrence of new words	.01	.00	.05	.01	.12	.00
Re-occurrence of novel SNPR	.02	.01	.10	.02	.15	.00
Share of international citations	.40	.29	.44	.38	.47	.00
Share of interregional citations	.90	.89	.28	.90	.28	.04
Share of intermetropolitan citations	.90	.90	.29	.90	.28	.48
Shortest path, author-inventor graph	3.58	3.47	1.48	3.49	1.49	.45
Shortest path, inventor graph	6.94	6.82	3.13	6.80	2.85	.71
Share shortest path > 1 (auth-invt)	.95	.94	.24	.95	.22	.01
Share shortest path > 2 (auth-invt)	.84	.84	.36	.83	.38	.13
Share shortest path > 3 (auth-invt)	.63	.60	.49	.60	.49	.78
Share shortest path $= \infty$ (auth-invt)	.24	.25	.43	.22	.42	.00
Number of citing patents	6,716	6,064		6,716		

Notes: The table reports group mean summary statistics for variable relating to the quality and reach of spillovers between between all pairwise links of citing and cited inventor locations for AIDS DB and control group patents in citing patents between 1991 and 1995. Inventor and applicant self-citations are excluded. Control group patents in column (2) are non-AIDS DB cited patents of similar timing within the same citing application. Columns (3) - (6) report means for AIDS DB patents without (3-4) and with (5-6) front-page reference to HIV/AIDS. Column (7) reports p-values from two-sample t-tests with unequal variances for differences in sample means. "New words" are unique keywords appearing for the first time on a patent in the universe of U.S. patents since 1976. "Novel SNPR" are non-patent references to scientific publications in PubMed which are in the top-5% of the distribution of new medical subject term combinations in a their respective year and field of publication, following Boudreau et al. (2016). Shares of international, -regional, and -metropolitan are the share of links in all pairwise links between inventor locations on a citing and cited patent that cross the respective geographic area for citing inventors. The bottom panel of the table reports group means for minimal social distance between all shortest paths of citing and cited inventors based on the network graph of all $>$ 5mio inventors on U.S. patents and all $>$ 16mio authors indexed in PubMed and their prior collaboration ties at the filing time of the citing patent. Citations to patents with all first-time inventors/authors are excluded. The data were collected by the author and combine web-scraped information from the *CNIDR* server archive with data from the *USPTO*, *PATSTAT* and *PubMed*. Patent key words are obtained from Arts, Cassiman and Gomez (2018). Scientific non-patent references come from Marx and Fuegi (2020). 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 16. Effects on the geographic reach of generated spillovers

Dependent variable:	*International*		*Interregional*		*Intermetropolitan*	
Probability of distant citation	(1)	(2)	(3)	(4)	(5)	(6)
No reference × post1994$_{t-1}$	0.020**	0.058***	-0.011***	0.021***	-0.020***	0.004
	(0.009)	(0.006)	(0.002)	(0.005)	(0.003)	(0.004)
With reference × post1994$_{t-1}$	-0.026***	-0.110***	0.040***	-0.015**	0.028***	-0.029***
	(0.009)	(0.022)	(0.012)	(0.006)	(0.006)	(0.009)
No reference × post1994$_{t-1}$		-0.049***		-0.047***		-0.036***
× private firm citing		(0.016)		(0.006)		(0.009)
With reference × post1994$_{t-1}$		0.119***		0.083***		0.083***
× private firm citing		(0.043)		(0.020)		(0.015)
Main category interactions		Incl		Incl		Incl
Citing patent × cited year FE	Yes	Yes	Yes	Yes	Yes	Yes
Citing region time trends	Yes	Yes	Yes	Yes	Yes	Yes
Observations	36,310	36,238	36,158	36,086	27,894	27,837
Number of citing clusters	8,524	8,505	8,509	8,490	6,986	6,972
R^2	.556	.556	.369	.369	.371	.372
Mean at t_0	.352	.352	.897	.897	.908	.908

Notes: Each column reports parameter estimates of regression (2) on citing patents-cited year pairs. The main category parameter is included. The dependent variables measure shares of international, interregional, and intermetropolitan 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. 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 country-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*, and *BvD Orbis*. Geo-coordinates of inventor locations are based on input data from Li et al. (2014); Morrison, Riccaboni and Pammolli (2017) (see Section III for details).

TABLE 17. EFFECTS ON THE SOCIAL DISTANCE (SD) REACH OF GENERATED SPILLOVERS, AUTHOR-INVENTOR GRAPH

Dependent variable:	$SD > 1$	$SD > 2$	$SD > 3$	$SD \infty$	*Log(SD)*
Shortest path of citation	(1)	(2)	(3)	(4)	(5)
No reference × post1994$_{t-1}$	0.004	0.030***	0.030***	0.007*	0.030***
	(0.003)	(0.004)	(0.005)	(0.004)	(0.005)
With reference × post1994$_{t-1}$	0.077***	0.060*	0.031*	-0.021**	0.079**
	(0.025)	(0.031)	(0.018)	(0.008)	(0.029)
Citing patent × cited year FE	Yes	Yes	Yes	Yes	Yes
Citing region time trends	Yes	Yes	Yes	Yes	Yes
Observations	36,933	36,933	36,933	36,933	31,055
Number of citing clusters	8,630	8,630	8,630	8,630	7,548
R^2	.441	.450	.595	.731	.584
Mean at t_0	.954	.870	.631	.201	1.250

Notes: Each column reports parameter estimates of regression (2) on citing patents-cited year pairs. The main category parameter is included. The dependent variables measures the probability that a citation originates from a research team at at different degrees of minimal social distance (shortest path) in the networks of direct and indirect prior collaborators of any cited inventor at the time of filing of the citing patent, based on the universe of all > 5mio inventors on U.S. patents and all > 16mio authors indexed in PubMed. The sample consists for AIDS DB and control group patents in citing patents between 1991 and 2000. 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 country-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*, and *BvD Orbis*. Inventor and 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 18. Effects on the social distance (SD) reach of generated spillovers, inventor graph

Dependent variable:	$SD > 1$	$SD > 2$	$SD > 3$	$SD \infty$	*Log(SD)*
Shortest path of citation	(1)	(2)	(3)	(4)	(5)
No reference × post1994$_{t-1}$	0.010***	0.026***	0.024***	0.020***	0.041***
	(0.002)	(0.004)	(0.008)	(0.006)	(0.013)
With reference × post1994$_{t-1}$	0.042*	0.053*	-0.000	-0.023**	0.111*
	(0.020)	(0.027)	(0.018)	(0.011)	(0.057)
Citing patent × cited year FE	Yes	Yes	Yes	Yes	Yes
Citing region time trends	Yes	Yes	Yes	Yes	Yes
Observations	36,925	36,925	36,925	36,925	24,621
Number of citing clusters	8,628	8,628	8,628	8,628	6,111
R^2	.410	.397	.389	.714	.507
Mean at t_0	.971	.951	.928	.439	1.794

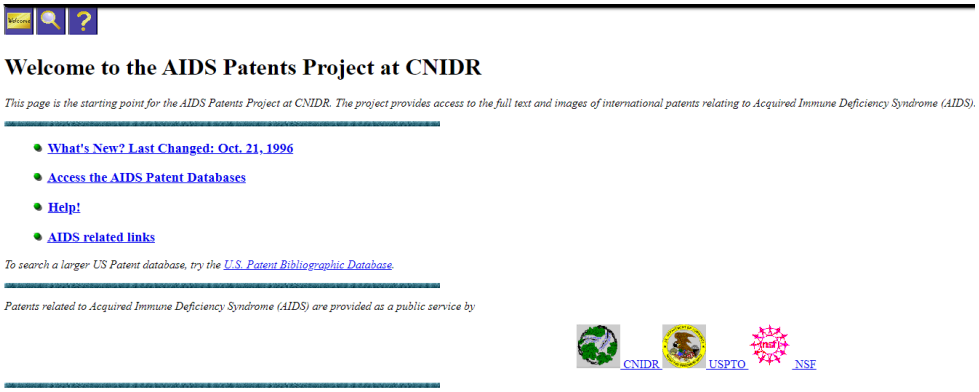
Notes: Each column reports parameter estimates of regression (2) on citing patents-cited year pairs. The main category parameter is included. The dependent variables measures the probability that a citation originates from a research team at at different degrees of minimal social distance (shortest path) in the networks of direct and indirect prior collaborators of any cited inventor at the time of filing of the citing patent, based on the universe of all > 5mio inventors on U.S. patents. The sample consists for AIDS DB and control group patents in citing patents between 1991 and 2000. 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 country-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*, and *BvD Orbis*. Inventor identities are disambiguated based on input data from Li et al. (2014); Morrison, Riccaboni and Pammolli (2017) (see Section III for details).

B Figures

FIGURE 6. AIDSDB LINK ON USPTO WEBSITE, FALL 1996

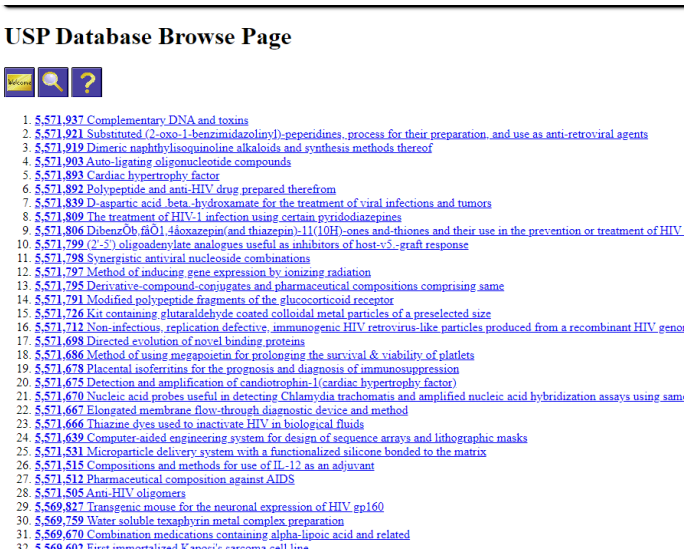
Notes: The figure shows a screenshot of the website of the USPTO in late 1996. The home page included a prominently positioned button linking to the access page of AIDS Patent Database hosted on the CNIDR server (see center-right). Wordwide access to both websites was possible with a dial-in modem and a telephone line.

FIGURE 7. AIDSDB ACCESS PAGE II, FALL 1996



Notes: The figure shows a screenshot to the access page to the AIDS Patent Database hosted on the CNIDR server in December 1996. 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. 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.

FIGURE 8. AIDSDB ACCESS PAGE III, FALL 1996

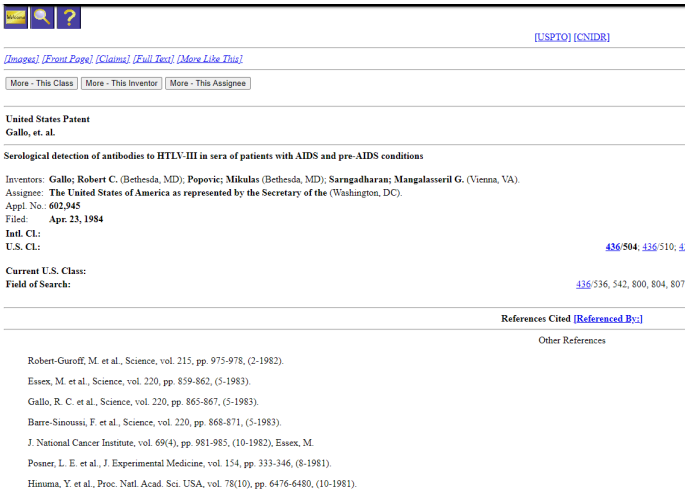


USP Database Browse Page

1. 5,571,937 Complementary DNA and toxins
2. 5,571,921 Substituted (2-oxo-1-benzimidazolinyl)-peperidines, process for their preparation, and use as anti-retroviral agents
3. 5,571,919 Dimeric naphthylisoquinoline alkaloids and synthesis methods thereof
4. 5,571,903 Auto-ligating oligonucleotide compounds
5. 5,571,893 Cardiac hypertrophy factor
6. 5,571,892 Polypeptide and anti-HIV drug prepared therefrom
7. 5,571,839 D-aspartic acid .beta.-hydroxamate for the treatment of viral infections and tumors
8. 5,571,809 The treatment of HIV-1 infection using certain pyridodiazepines
9. 5,571,806 DibenzÖb,fâÖ1,4âoxazepin(and thiazepin)-11(10H)-ones and-thiones and their use in the prevention or treatment of HIV
10. 5,571,799 (2'-5') oligoadenylate analogues useful as inhibitors of host-v5.-graft response
11. 5,571,798 Synergistic antiviral nucleoside combinations
12. 5,571,797 Method of inducing gene expression by ionizing radiation
13. 5,571,795 Derivative-compound-conjugates and pharmaceutical compositions comprising same
14. 5,571,791 Modified polypeptide fragments of the glucocorticoid receptor
15. 5,571,726 Kit containing glutaraldehyde coated colloidal metal particles of a preselected size
16. 5,571,712 Non-infectious, replication defective, immunogenic HIV retrovirus-like particles produced from a recombinant HIV genoi
17. 5,571,698 Directed evolution of novel binding proteins
18. 5,571,686 Method of using megapoietin for prolonging the survival & viability of platlets
19. 5,571,678 Placental isoferritins for the prognosis and diagnosis of immunosuppression
20. 5,571,675 Detection and amplification of candiotrophin-1(cardiac hypertrophy factor)
21. 5,571,670 Nucleic acid probes useful in detecting Chlamydia trachomatis and amplified nucleic acid hybridization assays using sam
22. 5,571,667 Elongated membrane flow-through diagnostic device and method
23. 5,571,666 Thiazine dyes used to inactivate HIV in biological fluids
24. 5,571,639 Computer-aided engineering system for design of sequence arrays and lithographic masks
25. 5,571,531 Microparticle delivery system with a functionalized silicone bonded to the matrix
26. 5,571,515 Compositions and methods for use of IL-12 as an adjuvant
27. 5,571,512 Pharmaceutical composition against AIDS
28. 5,571,505 Anti-HIV oligomers
29. 5,569,827 Transgenic mouse for the neuronal expression of HIV gp160
30. 5,569,759 Water soluble texaphyrin metal complex preparation
31. 5,569,670 Combination medications containing alpha-lipoic acid and related

Notes: The figure shows a screenshot to the browse view page to the AIDS Patent Database hosted on the CNIDR server in December 1996. Patents listed were sorted in descending order based on latest issue date. Links were clickable and allowed direct access to the patent view pages. 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.

FIGURE 9. AIDSDB ACCESS PAGE IV, PATENT VIEW, FALL 1996



[USPTO] [CNIDR]

[Images] [Front Page] [Claims] [Full Text] [More Like This]

More - This Class | More - This Inventor | More - This Assignee

United States Patent
Gallo, et. al.

Serological detection of antibodies to HTLV-III in sera of patients with AIDS and pre-AIDS conditions

Inventors: **Gallo; Robert C.** (Bethesda, MD); **Popovic; Mikulas** (Bethesda, MD); **Sarngadharan; Mangalasseril G.** (Vienna, VA).
Assignee: **The United States of America as represented by the Secretary of the** (Washington, DC).
Appl. No.: **602,945**
Filed: **Apr. 23, 1984**
Intl. Cl.:
U.S. Cl.: **436/504**; 436/510; 4
Current U.S. Class:
Field of Search: 436/536, 542, 800, 804, 807

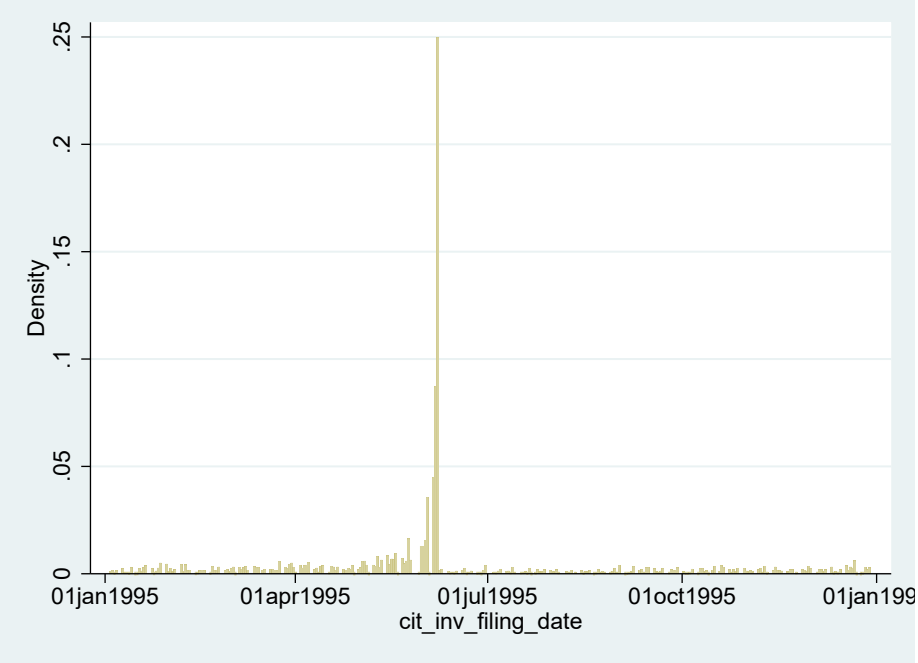
References Cited [Referenced By:]

Other References

Robert-Guroff, M. et al., Science, vol. 215, pp. 975-978, (2-1982).
Essex, M. et al., Science, vol. 220, pp. 859-862, (5-1983).
Gallo, R. C. et al., Science, vol. 220, pp. 865-867, (5-1983).
Barre-Sinoussi, F. et al., Science, vol. 220, pp. 868-871, (5-1983).
J. National Cancer Institute, vol. 69(4), pp. 981-985, (10-1982), Essex, M.
Posner, L. E. et al., J. Experimental Medicine, vol. 154, pp. 333-346, (8-1981).
Hinuma, Y. et al., Proc. Natl. Acad. Sci. USA, vol. 78(10), pp. 6476-6480, (10-1981).

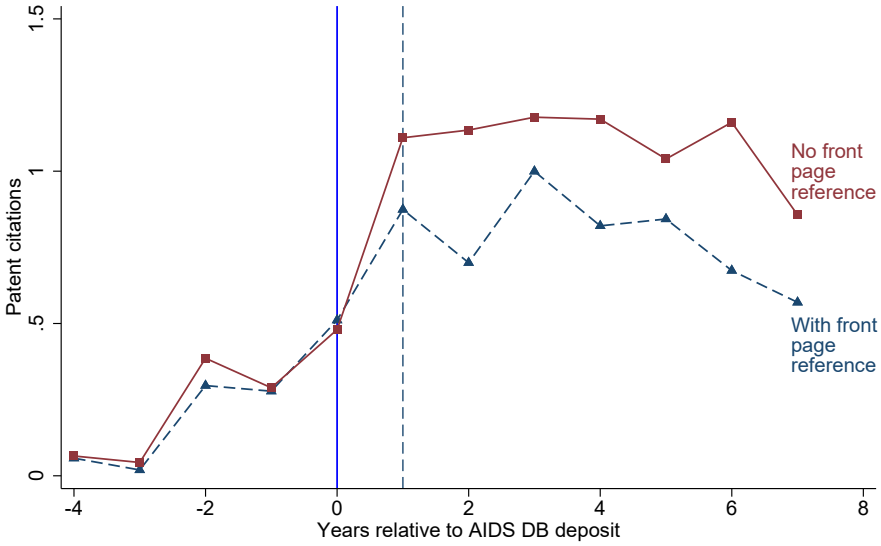
Notes: The figure shows a screenshot to the AIDS DB patent view page of patent # US 4,520,113, by Robert Gallo and co-inventors, granted in 1985, assigned to the U.S. Department of Health and Human Services in Washington, DC. The patent view page included further links to access full-text and high-resolution images and drawings, to this and all patents related to HIV/AIDS. Wordwide access to both websites was possible with a dial-in modem and a telephone line. Download pages were optimized for small (56k) bandwidths.

FIGURE 10. SPIKE IN DAILY U.S. PATENT FILINGS PRIOR TO TRIPS REFORMS ENACTMENT ON JUNE 8TH 1995



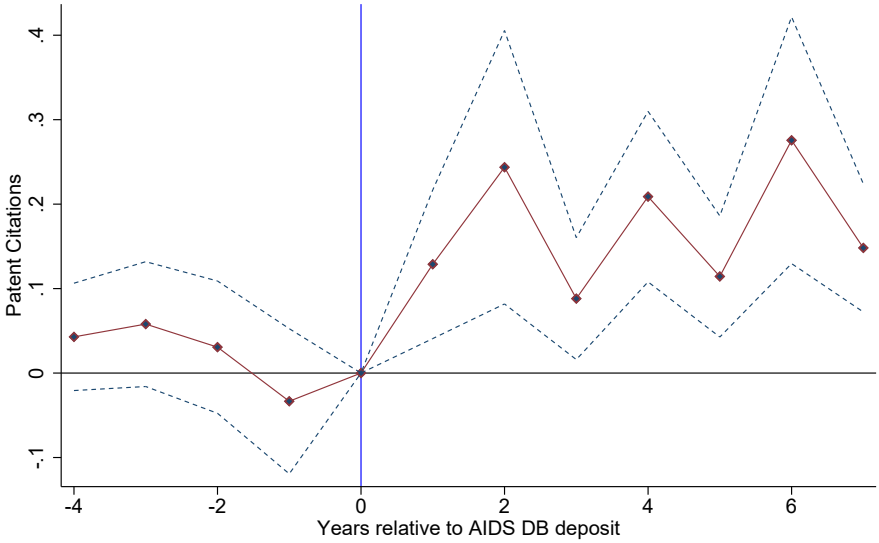
Notes: The figure shows spikes in the number of patent filings to the USPTO in the days and weeks leading up to June 8th of 1995, on which the U.S. patent system amended a significant part of its provision to be in line with the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS), which was negotiated at the end of the Uruguay Round of the General Agreement on Tariffs and Trade (GATT) between 1989 and 1990. The most important amendments included a change in patent term from 17 years after grant to 20 years after filing date of the application, the implementation of a domestic priority system in form of 'provisional applications', and the requirement to recognize foreign priority filings (Source: USPTO).

FIGURE 11. GROUP MEANS *within AIDS DB* COMPARISON YEARLY PATENT CITATIONS, MATCHED ON PRE-TRENDS



Notes: The figure plots trends in group means across AIDS DB patents without vs. with front-page reference to HIV/AIDS, additionally matched on pre-period yearly citation levels. 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).

FIGURE 12. YEARLY EFFECT FOR PATENTS WITHOUT HIV/AIDS FRONT PAGE REFERENCE, MATCHED ON PRE-TRENDS



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, additionally matched on pre-period yearly citation levels. 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).

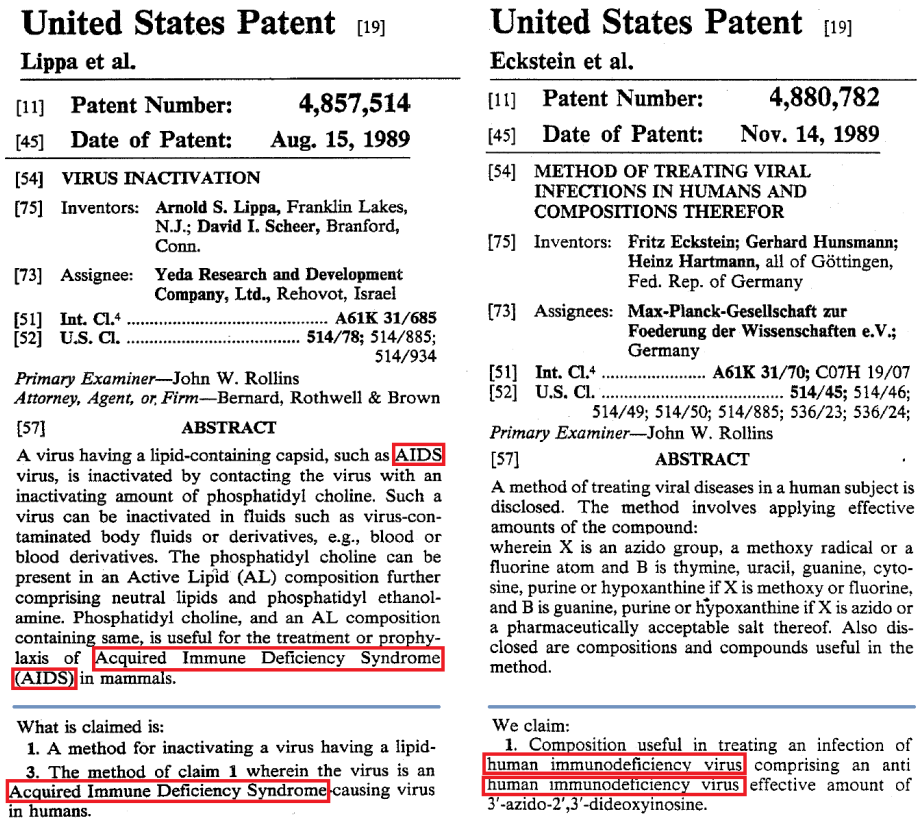
C Examples of AIDS DB patents with vs. without front-page reference to HIV/AIDS

In this appendix sections, I present several examples of cases of AIDS DB patents, that are paired in the sample by the matching approach outlined in Section IV.B, that differ in the extent to which the disease-link to HIV/AIDS applications becomes obvious from the front-page information. [64] This evidence is, of course, anecdotal and non-systematic. It is important to highlight that my design neither claims nor requires these patent pairs to be (almost) identical in technical content nor that a specialized inventor would not recognize the link to HIV/AIDS of those patents without obvious front-page references upon more careful inspection. Merely, my assumption is that the front-page reference of specific terms relating to the disease increases the likelihood that an inventor, searching for prior art art inputs, would retrieve a patent as relevant out of a bibliographic index containing numerous, diverse patents.

The first example, shown in Figure 13, is the case of two patents granted in the second half of 1989 on methods, and related compounds, inhibiting viral replication after HIV-infection, targeted specifically at the prodromal (pre-AIDS) stage of disease. The patent on the left-hand side (patent #: US 4,857,514) has been invented by Arnold Lippa and David Sheer (both based in the U.S.) and has been first filed in 1985. It is assigned to the "Yeda Research and Development Company Ltd.", which is a technology transfer organization of the Weizmann Institute of Science, in Rehovot, Israel, one of the leading public research institutions in natural and exact sciences. The patent on the right in Figure 13 (patent #: US 4,880,782) is an invention of Fritz Eckstein, Gerhard Hunsmann and Heinz Hartmann (all based in Germany), filed first in 1986, and is assigned to the Max-Planck Society and the German Primate Research Centre, both renown non-profit

[64] To identify this in the paper, I query the front-pages of AIDS DB patents for explicit textual references to terms relating to HIV/AIDS, as defined by the National Library of Medicine (see: https://meshb.nlm.nih.gov/).Queried keywords for HIV/AIDS-related terms are: *AIDS; Acquired Immune Deficiency Syndrome; Acquired Immuno-Deficiency Syndrome; Immunodeficiency Syndrome, Acquired; AIDS Arteritis, Central Nervous System; AIDS Dementia Complex;AIDS Serodiagnosis; HIV Seropositivity; HIV Seroprevalence; Lymphoma, AIDS-Related; HTLV-III; Human Immunodeficiency Virus; Human T Cell Lymphotropic Virus Type III; Human T Lymphotropic Virus Type III; Human T-Cell Leukemia Virus Type III; Human T-Cell Lymphotropic Virus Type III; Human T-Lymphotropic Virus Type III; Immunodeficiency Virus, Human; Immunodeficiency Viruses, Human; LAV-HTLV-III; Lymphadenopathy-Associated Virus; Virus, Human Immunodeficiency; Human T-Cell Leukemia Virus*

FIGURE 13. EXAMPLE 1- LIPPA ET AL. VS. ECKSTEIN ET AL. (1989)



United States Patent [19]

Lippa et al.

[11] **Patent Number: 4,857,514**

[45] **Date of Patent: Aug. 15, 1989**

[54] **VIRUS INACTIVATION**

[75] Inventors: **Arnold S. Lippa**, Franklin Lakes, N.J.; **David I. Scheer**, Branford, Conn.

[73] Assignee: **Yeda Research and Development Company, Ltd.**, Rehovot, Israel

[51] **Int. Cl.**[4] **A61K 31/685**

[52] **U.S. Cl.** **514/78**; 514/885; 514/934

Primary Examiner—John W. Rollins

Attorney, Agent, or Firm—Bernard, Rothwell & Brown

[57] **ABSTRACT**

A virus having a lipid-containing capsid, such as AIDS virus, is inactivated by contacting the virus with an inactivating amount of phosphatidyl choline. Such a virus can be inactivated in fluids such as virus-contaminated body fluids or derivatives, e.g., blood or blood derivatives. The phosphatidyl choline can be present in an Active Lipid (AL) composition further comprising neutral lipids and phosphatidyl ethanolamine. Phosphatidyl choline, and an AL composition containing same, is useful for the treatment or prophylaxis of Acquired Immune Deficiency Syndrome (AIDS) in mammals.

What is claimed is:

1. A method for inactivating a virus having a lipid-

3. The method of claim **1** wherein the virus is an Acquired Immune Deficiency Syndrome-causing virus in humans.

United States Patent [19]

Eckstein et al.

[11] **Patent Number: 4,880,782**

[45] **Date of Patent: Nov. 14, 1989**

[54] **METHOD OF TREATING VIRAL INFECTIONS IN HUMANS AND COMPOSITIONS THEREFOR**

[75] Inventors: **Fritz Eckstein; Gerhard Hunsmann; Heinz Hartmann**, all of Göttingen, Fed. Rep. of Germany

[73] Assignees: **Max-Planck-Gesellschaft zur Foederung der Wissenschaften e.V.**; Germany

[51] **Int. Cl.**[4] **A61K 31/70**; C07H 19/07

[52] **U.S. Cl.** **514/45**; 514/46; 514/49; 514/50; 514/885; 536/23; 536/24;

Primary Examiner—John W. Rollins

[57] **ABSTRACT**

A method of treating viral diseases in a human subject is disclosed. The method involves applying effective amounts of the compound:

wherein X is an azido group, a methoxy radical or a fluorine atom and B is thymine, uracil, guanine, cytosine, purine or hypoxanthine if X is methoxy or fluorine, and B is guanine, purine or hypoxanthine if X is azido or a pharmaceutically acceptable salt thereof. Also disclosed are compositions and compounds useful in the method.

We claim:

1. Composition useful in treating an infection of human immunodeficiency virus comprising an anti human immunodeficiency virus effective amount of 3'-azido-2',3'-dideoxyinosine.

Notes: The upper part of the figure shows (shortened) front-page information of the original patent documents. The left-hand side patent contains explicit textual references to terms relating to HIV/AIDS, as defined by the National Library of Medicine (see: https://meshb.nlm.nih.gov/). The patent on the right does not contain such front-page references. The lower part of the figure (below the blue separating line) contains relevant excerpts from the full-patent text, which was not retrievable by inventors from standard bibliographic search.

public research organizations located in Göttingen, Germany. The technological similarity of the inventions can been seen from the assigned patent classes in Figure 13: Both share the same 6-digit IPC class ("A61K 31"), as well as the same USPC 3-digit class ("514"). This similarity is further evidenced in Figure 13 by the fact that both patents have been reviewed by the same primary USPTO patent examiner (John W. Rollins). Each of the two patents reports several scientific publications as prior art references. Figure 13 further shows that, while the Lippa et al. patent makes two explicit references to "*AIDS*" and the usefulness of the invention for "*the treatment and prophylaxis of Acquired Immune*

Deficiency Syndrome (AIDS)", the front page of the patent by the German inventors does not include any such reference regarding the applicability to HIV/AIDS, but remains very technical and focused primarily on the chemical compounds related to the invention.

The bottom part of Figure 13, below the inserted blue line, shows information extracted from the full text body of the two patents, which has not been visible to inventors in standard bibliographic prior art search. It can be clearly seen that both inventions apply to the treatment of the AIDS-causing, human immunodeficiency virus. Most noticeably, this becomes visible from the inspection of the patent claims - which represent the core of the invention for which the patent is claimed; In the case of the Eckstein et al. patent, which included no front page reference to HIV/AIDS, already the first (main) claim made reflects the usefulness of the invention for "*treating an infection of human immunodeficiency virus*". Analogously, though only in the 3rd claim, this can be seen for the Lippa et al. patent.

The second example, shown in Figure, is a patent pair from 1990 of two U.S. based single-inventors; Patent # US 4,944,920 of Alan Rubinstein, assigned to the University of Southern California in Los Angeles, CA, and patent # US 4,880,602 of Habib Al-Sioufi, a pathologist at the University of Massachusetts Hospital in Boston, MA. Both inventions relate to methods for disinfecting viral contaminants from HIV in human blood and other body fluids, without affecting the integrity of the blood specimen for transfusion or further clinical evaluation. In each of the two cases, the usefulness of the invention is targeted at the avoidance of contamination from HIV, but not limited to this specific viral agent. Similar to the previous case, the strong technological relatedness between the two patents is shown in Figure 14, namely by sharing the same IPC-class ("A16L 2/18") and USPC (sub-)classes ("422", "514"), as well as the same assistant USPTO patent examiner (Jill Johnston). While the Rubinstein patent makes a clear reference to the "*risk of transmission of AIDS*" in the abstract on the front page, no such reference is found for the Al-Sioufi patent cover page, which remains relatively unspecific with regard to usefulness and areas of application (see Figure 14). However, when inspecting the remainder of the patent

FIGURE 14. EXAMPLE 2- RUBINSTEIN VS. AL-SIOUFI (1990)

United States Patent [19]

Rubinstein

[11] **Patent Number: 4,944,920**

[45] **Date of Patent: Jul. 31, 1990**

[54] **NOVEL METHOD TO TREAT BLOOD**

[75] Inventor: **Alan I. Rubinstein**, Los Angeles, Calif.

[73] Assignee: **University of Southern California**, Los Angeles, Calif.

[51] **Int. Cl.**[5] **A61L 2/18**; C07K 15/06

[52] **U.S. Cl.** **422/37**; 422/28; 424/533; 435/2; 435/288; 514/833; 530/385

Primary Examiner—Robert J. Warden

Assistant Examiner—Jill Johnston

[57] **ABSTRACT**

It is well known that the transfusion of human blood and blood components carriers with it a substantial risk of transmission of AIDS and many other diseases. This disclosure describes a method of disinfecting red blood cells to make them safer for human transfusion, while maintaining their biologic activity. A sterilizing solution is prepared from, e.g., a commercially available disinfectant (LD TM Alcide Corporation) containing primarily lactic acid and sodium chlorite. Normal saline solution is used as diluent instead of distilled water. The red cells are exposed to the disinfectant for a time sufficient to inactivate or reduce the infectivity of disease agents. The normal-saline environment prevents or deters hemolysis. The blood cells are then washed with normal saline solution until the disinfectant concentration is insignificant. The blood is then safe for human transfusion.

SUMMARY OF THE DISCLOSURE

My invention is a method for treating a blood substance comprising red blood cells, to inactivate or greatly reduce the activity therein of certain harmful contaminants. The blood substance may itself be red blood cells.

The harmful contaminants of concern may include HTLV-III or any other AIDS-causing agents; or cyto-

United States Patent [19]

Al-Sioufi

[11] **Patent Number: 4,880,602**

[45] **Date of Patent: * Nov. 14, 1989**

[54] **METHOD AND DEVICE FOR DISINFECTING BIOLOGICAL FLUIDS AND CONTAINER FOR SAME**

[76] Inventor: **Habib Al-Sioufi**, P.O. Box 654, Brookline, Mass. 02146

[51] **Int. Cl.**[4] **A61L 2/18**

[52] **U.S. Cl.** **422/28**; 422/36; 422/37; 514/635; 514/642; 514/693; 514/694; 514/731

Primary Examiner—Garry S. Richman

Assistant Examiner—Jill Johnston

[57] **ABSTRACT**

A technique and receptacle for disinfecting biological fluids such as whole blood is described in which the disinfectant is prepositioned in a receptacle for biological fluids utilized for clinical evaluation in an amount which is sufficient to disinfect the fluid without interfering with subsequent clinical evaluation. The invention is specifically directed to disinfecting viral contaminants in blood by providing a closed container for the blood specimen which contains an amount of a disinfectant sufficient to destroy without otherwise affecting the integrity of the specimen for future evaluation. The amount of disinfectant positioned in the container is adjusted to provide an ultimate concentration in the blood specimen of aldehyde of about 0.001 to 5.0 weight percent and is buffered to a pH of about 7.2 to 8.5 percent preferably about 7.4. To increase the stability and shelf life of the sample container and disinfectant, activation or buffering to the indicated pH range does not take place until or just prior to introduction of the specimen into the container. In a particularly preferred embodiment of the invention, the close sample container is evacuated and provided with an elastomeric stopper adapted to receive the hollow needle of a syringe whereby the blood specimen is introduced into the container directly from the donor. The aldehyde based disinfectant used in accordance with the invention have also been found to facilitate separation of the fluid components of the blood by causing gelling of cellular blood components.

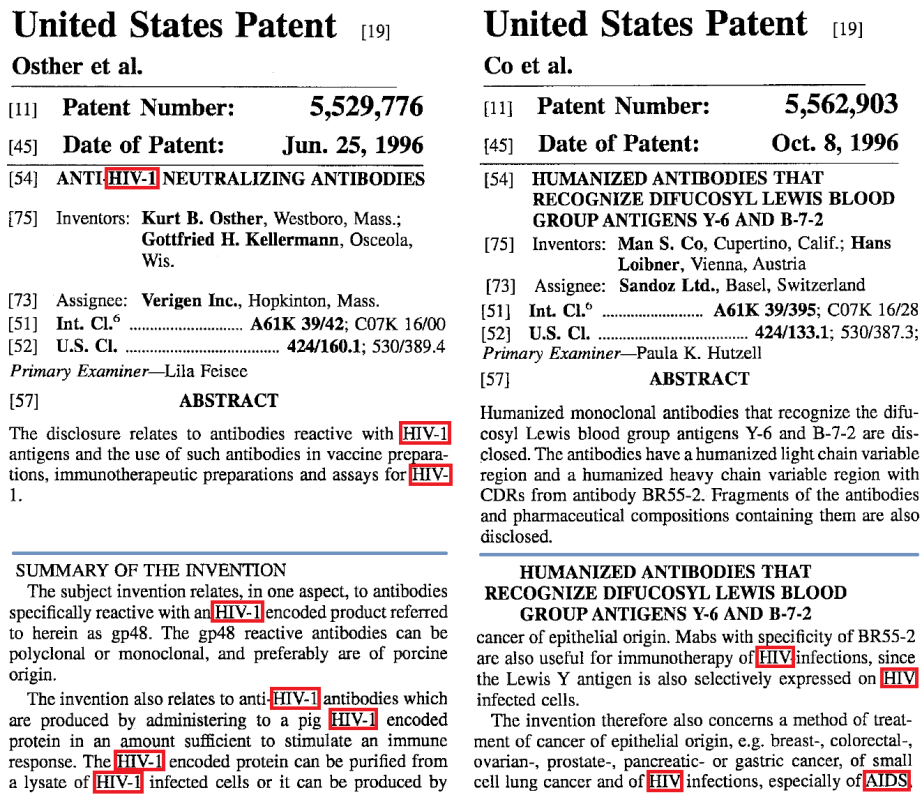
SUMMARY OF THE INVENTION

specifically, the present invention is particularly concerned with disinfecting viral contamination in biological specimens to avoid infecting those coming in contact either with the specimen itself or the receptacles and equipment used to contain and evaluate the specimen. Of particular concern in the present invention is the avoidance of contamination by HTLV-III Virus responsible for Acquired Immune Deficiency Syndrome and Hepatitis Virus which may be present in

Notes: The upper part of the figure shows (shortened) front-page information of the original patent documents. The left-hand side patent contains explicit textual references to terms relating to HIV/AIDS, as defined by the National Library of Medicine (see: https://meshb.nlm.nih.gov/). The patent on the right does not contain such front-page references. The lower part of the figure (below the blue separating line) contains relevant excerpts from the full-patent text, which was not retrievable by inventors from standard bibliographic search.

document, when outlying the content of the invention, also the latter patent points out that "*of particular concern in the present invention is the avoidance of contamination by HTLV-III virus responsible for Acquired Immune Deficiency Syndrome(..)*"[65], as reported in Figure 14 below the blue separation line. On the left hand side at the bottom of Figure 14 it can be seen that the usefulness for decontamination from HIV-agents is outlined in the Rubinstein patent full text with almost identical wording.

FIGURE 15. EXAMPLE 3- OSTHER ET AL. VS. CO ET AL. (1996)



United States Patent [19]

Osther et al.

[11] **Patent Number: 5,529,776**

[45] **Date of Patent: Jun. 25, 1996**

[54] **ANTI-HIV-1 NEUTRALIZING ANTIBODIES**

[75] Inventors: **Kurt B. Osther**, Westboro, Mass.; **Gottfried H. Kellermann**, Osceola, Wis.

[73] Assignee: **Verigen Inc.**, Hopkinton, Mass.

[51] **Int. Cl.**[6] **A61K 39/42**; C07K 16/00

[52] **U.S. Cl.** **424/160.1**; 530/389.4

Primary Examiner—Lila Feisee

[57] **ABSTRACT**

The disclosure relates to antibodies reactive with HIV-1 antigens and the use of such antibodies in vaccine preparations, immunotherapeutic preparations and assays for HIV-1.

United States Patent [19]

Co et al.

[11] **Patent Number: 5,562,903**

[45] **Date of Patent: Oct. 8, 1996**

[54] **HUMANIZED ANTIBODIES THAT RECOGNIZE DIFUCOSYL LEWIS BLOOD GROUP ANTIGENS Y-6 AND B-7-2**

[75] Inventors: **Man S. Co**, Cupertino, Calif.; **Hans Loibner**, Vienna, Austria

[73] Assignee: **Sandoz Ltd.**, Basel, Switzerland

[51] **Int. Cl.**[6] **A61K 39/395**; C07K 16/28

[52] **U.S. Cl.** **424/133.1**; 530/387.3;

Primary Examiner—Paula K. Hutzell

[57] **ABSTRACT**

Humanized monoclonal antibodies that recognize the difucosyl Lewis blood group antigens Y-6 and B-7-2 are disclosed. The antibodies have a humanized light chain variable region and a humanized heavy chain variable region with CDRs from antibody BR55-2. Fragments of the antibodies and pharmaceutical compositions containing them are also disclosed.

SUMMARY OF THE INVENTION

The subject invention relates, in one aspect, to antibodies specifically reactive with an HIV-1 encoded product referred to herein as gp48. The gp48 reactive antibodies can be polyclonal or monoclonal, and preferably are of porcine origin.

The invention also relates to anti-HIV-1 antibodies which are produced by administering to a pig HIV-1 encoded protein in an amount sufficient to stimulate an immune response. The HIV-1 encoded protein can be purified from a lysate of HIV-1 infected cells or it can be produced by

HUMANIZED ANTIBODIES THAT RECOGNIZE DIFUCOSYL LEWIS BLOOD GROUP ANTIGENS Y-6 AND B-7-2

cancer of epithelial origin. Mabs with specificity of BR55-2 are also useful for immunotherapy of HIV infections, since the Lewis Y antigen is also selectively expressed on HIV infected cells.

The invention therefore also concerns a method of treatment of cancer of epithelial origin, e.g. breast-, colorectal-, ovarian-, prostate-, pancreatic- or gastric cancer, of small cell lung cancer and of HIV infections, especially of AIDS.

Notes: The upper part of the figure shows (shortened) front-page information of the original patent documents. The left-hand side patent contains explicit textual references to terms relating to HIV/AIDS, as defined by the National Library of Medicine (see: https://meshb.nlm.nih.gov/). The patent on the right does not contain such front-page references. The lower part of the figure (below the blue separating line) contains relevant excerpts from the full-patent text, which was not retrievable by inventors from standard bibliographic search.

[65] HTLV-III is the abbreviation for "Human T-cell Lymphotropic Virus Type-III", which is the term used to denote the causative agent of AIDS in the original paper by Gallo et al. *Science* 220:865-867 (1983), later adopted to be referred to as Human Immunodeficiency Virus (HIV) by the scientific community.

A final example of AIDS DB patent pairs in the sample with vs. without front-page reference to HIV/AIDS is provided in Figure 15 comparing the 1996 patents of Osther et al. (patent # US 5,529,776) and Co et al. (patent # US 5,562,903). The are two private firm patents, the former assigned to Verigen Inc., of Hopkinton, MA, in the U.S., and the latter to Sandoz Ltd., of Basel in Switzerland.

Both patented inventions relate to the production of specific antibodies, of porcine or murine chimeric origin, useful as immunotherapeutics to humans infected by HIV. These can be further used to isolate HIV antigens, particularly useful for vaccine development. Each of the two documents constitutes a continuation of previous applications, with filing year 1993. As Figure 15 shows, they are assigned to the same IPC-6 classes ("A61K 39") and USPC 3-digit classes ("424", "530"). They were examined by the same art unit, but with different examiners. While the Osther et al. patent makes several explicit front-page references to HIV, in both title and abstract, the Co et al. patent is entirely missing such an indication (see Figure 15). When inspecting the full text body of the latter patent, however, the link to HIV/AIDS becomes clearly evident starting from the first paragraph of patent, containing the summary of the invention (excerpts of which can be seen in the bottom-right part of Figure 15, below the blue separating line). The bottom-left part of Figure 14 shows the corresponding text-excerpt from the invention summary of the Osther et al. patent, with several references regarding the applicability to HIV infections of the produced antibodies, in line with the explicit indication provided on the patent front page.

D External matched control group results

Complementing the main findings of the paper on internal search costs, in this appendix section, I attempt to establish a baseline of the marginal impact of *online access* on cumulative citations by comparing AIDS DB patents to a control group of external patents with same timing, which were not included in the repository. To account for the selection at the researcher level and the particular institutional environments of HIV-research, I sample the control group exclusively from the population of non-AIDS DB patents of the same inventors whose patents were deposited in the database. To avoid confounding influences stemming from productivity changes or switching research focus over time, I only consider those patents by HIV-inventors that were filed at the same time or since the filing of their first included AIDS DB patent.[66] Further, I only consider patent family members filed at the USPTO. To avoid comparing patents on different types of technologies within broader technological fields, of different institutional context, scientific background and timing, I apply the same selection criteria to this control group as in the main analysis and apply the weights of Iacus, King and Porro (2012) to ensure balance in the estimation (see also Section IV.A).

Table 19 summarizes the main invention-level characteristics of patents in the matched sample. My strict selection criteria yield to at least one control patent for 1,979 AIDS DB patents.[67] A total of 17 technology fields and 57 art units are represented in the sample. Previous to deposit, AIDS DB patents in the sample receive on average 1.3 citations per year, which is a difference in citation levels of about .2 compared to the control group (significant at 1%). The sample groups are highly comparable on a range of ex-ante characteristics, however, some small, significant, differences persist, namely in the number of scientific references and the inventor team composition. Average lags between deposit and patent publication (18 months) and application (55 months) indicate that I observe a large

[66] Unfortunately, it is not possible to compare patents within exactly the same inventor, given that my design requires matching timing of invention, and only a very small group of inventors files more than one patent within the same narrow time window.

[67] By this, my sample covers 65.5% of all originally deposited patents in the AIDS DB. Inference is limited to this subset.

Table 19. Summary statistics, until year of AIDS DB deposit

Matched sample	AIDSDB patents		Control group		Diff
	Mean	SD	Mean	SD	p-val.
Yearly patent family citations until AIDSDB deposit	1.348	2.909	1.147	2.599	.009
Number of patent references	9.145	10.729	8.768	10.028	.196
Share with scientific reference	.798		.799		.930
Number of scientific references	11.636	21.403	9.271	15.110	.000
Share of novel technologies	.247		.227		.096
Share introducing new words	.311		.304		.592
Number of inventors	2.996	2.110	3.373	2.004	.000
Share with author-inventors	.916		.951		.000
Number of author-inventors	2.469	1.826	2.872	1.798	.000
Patent family size	6.300	6.898	6.271	6.885	.882
Share private firm patents	.667		.666		.940
Assignee prior patent families	2.989k	10.436k	3.554k	9.123k	.056
DB-to-publication lag (m)	18.533	20.997	18.010	21.921	.393
DB-to-application lag (m)	54.746	27.297	54.639	27.714	.891
Number of patents	1,979		3,361		
Number of technology fields	14		17		
Number of examining art units	57		57		

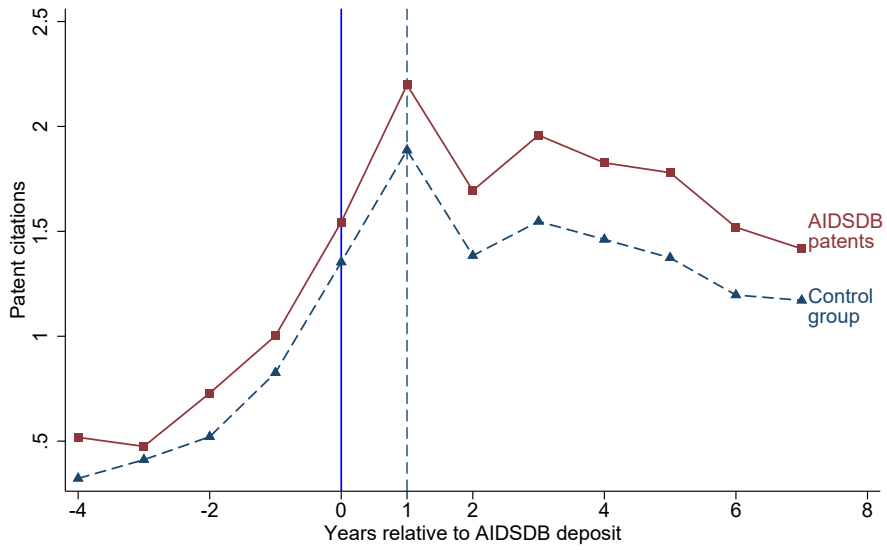
Notes: Row (1) reports the group mean and standard deviation for yearly patent family citations to deposited and control patents between year $t-4$ and t relative to AIDS DB deposit. Inventor and applicant self-citations are removed from the counts. The following rows report ex-ante time-invariant characteristics. Control group patents consist of not-online deposited patents of AIDS DB inventors, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired AIDS DB 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).

share of the sample patents for a significant time span before database inclusion.[68]

In order to evaluate whether the significant initial difference in levels of yearly patent citations between the two groups (as reported in Table 19) is stable in the pre-period, I inspect trends in groups means over time for the years preceding the database deposit in

[68]Online deposit of AIDS DB patents is assumed to take place on average one month after patent publication, compare Section III.A

FIGURE 16. GROUP MEANS COMPARISON YEARLY PATENT CITATIONS



Notes: The figure plots trends in group means across AIDS DB deposited and control patents. Control group patents consist of *not* online-deposited patents of AIDS DB inventors, filed and granted around the same time, with similar institutional and scientific background, and examined in the same USPTO art unit as paired AIDS DB patents. The y-axis scale reports levels of yearly patent family citations to patents in sample. Inventor and applicant 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).

Figure 16. The graph shows that, despite the constant small difference in levels, trends of group means between AIDS DB and control patents remain basically parallel over the entire pre-period from $t-4$ to t (with the minor exception from $t-4$ to $t-3$), which is a necessary condition for inference of an average treatment effect on the treated, and suggests that the control group is well selected. Figure 16 also reveals that starting from $t+1$ differences in group means can be seen to substantially increase, while trends still follow largely parallel patterns.

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, similar to the one in the main analysis (cf. 1):

$$Y_{it} = \beta_1 * AIDSDB_i \times post_{t-1} + patentFE_i + yearFE_t + \phi_y + \theta_{fy} + \epsilon_{it}, \tag{3}$$

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. Control patents have the function to provide a reference level of citations that would have been received by a matched AIDS DB patent in the absence of the database treatment. Therefore, it is necessary to assign each control patent to a 'fictitious' deposit date, relative to which the treatment effect is observed. I assign to each group of two or more matched patents a unique database deposit date, based on the most frequent occurring actual deposit date in the group.[69] The coefficient β_1 measures changes in citation rates to AIDS DB patents, after deposit, relative to the control group patents, which are the excluded reference category. The interacted $post_{t-1}$ indicator denotes the one-year lagged post-deposit status. The regression model includes a full set of patent fixed effects and fixed effects for relative years to the AIDS DB deposit date. 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 ϕ_y). Finally, I include linear field-year trends (θ_{fy}) to control for idiosyncratic variation in productivity of specific technology fields. I estimate regression (3) on a symmetric sample window of five years preceding and five years following the switching of the $post_{t-1}$ indicator, i.e., for example, ranging from October 1991 to October 2000 for patents deposited in the initial launch of the AIDS DB database on October 26$^{\text{th}}$ 1994. Finally, I cluster all standard errors at the patent-level. For a more extensive discussion of the econometric framework, see Section IV.A.

[69] In case of multiple, I assign the earliest date. By construction determined deposit dates within matched groups are very close to each other.

Table 20 includes the main econometric results from the estimation of regression (3) on the matched sample. Column (1) of Table 20 shows that, starting from one year after deposit, AIDS DB patents received on average .04 standard deviations in cumulative citations more relative to control group patents (significant at the 10% level). Compared to the pre-deposit mean of citations (see Table 19), this implies a relative increase of .13 citations per year (roughly +10%) for the average AIDS DB patent. When including technology field specific time trends (Table 20, column (2)), the results become slightly larger and more significant (+.05 std. deviation, p-value $<$ .05). This constitutes my preferred specification and provides the first main finding of the paper.

Table 20. Main effect on Patent Citations, Matched Sample

Dependent variable:	OLS				Poisson
Number of patent citations	(1)	(2)	(3)	(4)	(5)
AIDSDB $\times$ post$_{t-1}$	0.044*	0.049**	0.050**	-0.004	0.052**
	(0.024)	(0.024)	(0.024)	(0.025)	(0.022)
Patent fixed effects	Yes	Yes	Yes	Yes	Yes
Time/ year fixed effects	Yes	Yes	Yes	Yes	Yes
Field time trends		Yes	Yes	Yes	Yes
Excl. firm self-cites			Yes		
Prosecution cites only				Yes	
Observations	45,818	45,818	45,731	44,997	39,862
Number of patents	5,339	5,339	5,339	5,329	4,618
Mean dependent					1.395

Notes: Each column reports parameter estimates of regression (3) for the matched panel of U.S. AIDS DB and control patents. 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 self-citations are removed from the counts. Control group patents are the reference category and consist of *not* online-deposited patents of AIDS DB inventors, filed and granted at the same time, with similar institutional and scientific background, and examined by the same art unit as paired AIDS DB patents. In columns (1)-(4), the number of citations is standardized to mean zero and standard deviation one within technology fields (fields based on Schmoch (2008)). Column (5) estimates Poisson pseudo-maximum likelihood regressions (correct for over-dispersion by construction) to address the count nature of citations. 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).

I check the robustness of this finding across several alternative models: In column (3) of Table 20, I re-estimate regression (3) excluding all firm-level self-citations from the citation counts. The point estimate in column (3) shows that this further increases the result slightly (significant at the 5% level), suggesting the effect to be largely due to an increase in external spillovers. In column (4) of Table 20, I re-estimate regression (3) *only* counting citations given from patents that were already under prosecution at the time of online deposit of the cited AIDS DB patent (or its linked control patent), i.e. filed before and granted after the AIDS DB deposit date.[70] These citations are very likely to be given by examiners rather than by the applicants.[71] 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. Results in column (4) show no significant difference in citations added during prosecution after database deposit across AIDS DB and control patents. The point estimate for β_1 is even slightly negative, but close to zero and highly insignificant. This suggests that the launch of the AIDS DB had no influence on citation practices of patent examiners. Finally, in Table 20 column (5), I estimate robustness of the preferred specification with Poisson pseudo-maximum likelihood, in order to address the count nature of citations. The Poisson estimates fully confirm the main result, indicating a relative increase in citations to AIDS DB patents of 5.3% (+.074 citations for the average patent) compared to control group patents (significant at the 5% level), starting from one year after online deposit, which is slightly smaller in magnitude compared to the marginal impact estimated with OLS.[72]

In order to investigate the timing of these effects, I re-estimate regression (3) with yearly coefficients, by interacting the *AIDSDB* 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

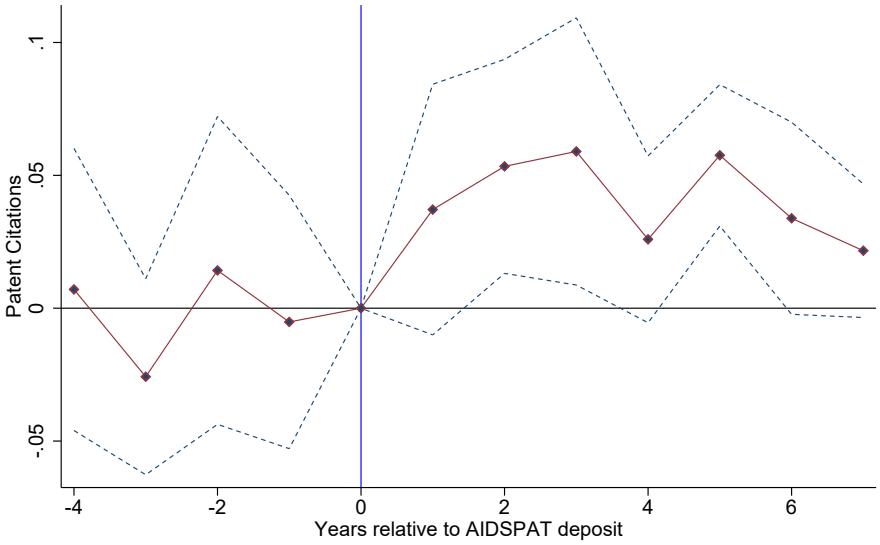
[70] In this case, deviant from my standard approach, I consider as citation date the grant date of a citing patent, which is arguably closest to the examination moment.

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

[72] The percentage increase associated with the point estimate of β_1 in the Poisson pseudo-maximum likelihood model is given by e^{β_1}

year). Figure 17 plots the corresponding point estimates within 95% confidence intervals. There are no significant differences estimated between citation trends of AIDS DB and control 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 AIDS DB patents in the years following their online availability, setting in highly significantly after one year, and reaching a peak around the third year post-deposit. Figure 17 also indicates a non-monotonic effect over time, as esti-

FIGURE 17. YEARLY EFFECT ON PATENT CITATIONS



Notes: The figure plots parameter estimates from regression (3) with yearly coefficients for $t-4$ to $t+7$ relative to online deposit for the matched panel of U.S. AIDS DB and control patents. The dependent variable measures the yearly number of family citations (inventor self-citations excluded). The year of deposit is omitted from the regression. Control group patents consist of *not* online-deposited patents of AIDS DB inventors, filed and granted at the same time, with similar institutional and scientific background, and examined by the same art unit as paired AIDS DB 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).

mated differences, first, decline sharply for one period after four years, and, subsequently, seem to gradually level out starting from the fifth year post-AIDS DB. For most patents in the sample (those of the initial cohort of patents uploaded in 1994), these periods coincide

with the launch, first, of the comprehensive bibliographic online database of the USPTO (in 1997) and, second, the full-text and images online catalogue including all U.S. patents (in 1998) and EPO Espacenet (1999). This provides further support for the believe that the observed effects are, indeed, caused by enhanced-access to AIDS DB patents by means of the online repository. Table **??** in Appendix .A reports results for heterogeneous effects on the subset of recent patents. In line with predictions of faster access to patents through full-document online availability, estimates show that differences were strongly significant and magnified for patents granted less then two years prior to AIDS DB deposit. While here, however, I also observe a significant increase in the rate of (likely) examiner-added citations, for citing patents under prosecution, the relative effect size on these appears much smaller.