An analysis of the evolution of science-technology linkage in biomedicine

Qing Ke*
Northeastern University, Boston, MA 02115, USA

Demonstrating the practical value of public research has been an important subject in science policy. Here we present a detailed study on the evolution of the citation linkage between life science related patents and biomedical research over a 37-year period. Our analysis relies on a newly-created dataset that systematically links millions of non-patent references to biomedical papers. We find a large disparity in the volume of science linkage among technology sectors, with biotechnology and drug patents dominating it. The linkage has been growing exponentially over a long period of time, doubling every 2.9 years. The U.S. has been the largest producer of cited science for years, receiving nearly half of the citations. More than half of citations goes to universities, and the cited papers are likely to be basic research. The U.S. National Institute of Health continues to be a major funder of cited science. For the majority of companies, more than half of citations in their patents are authored by public research. Taken together, these results indicate a continuous contribution of public science to private sector inventions.

Keywords: patent-to-paper citation; non-patent reference; science-technology interaction; biomedical research; public science

I. INTRODUCTION

There is a longstanding policy interest in unraveling how knowledge generated from public research is used in the private-sector. Studies towards this goal have heavily focused on patent data and considered citations between patents as evidence of knowledge flow. Despite some criticism [7], such notion has been widely accepted in the literature. Consequently, substantial attention has been paid to patents assigned to universities and other public organizations, examining how those patents are cited by other patents, especially by patents from companies [14, 16].

University patents, however, only account for a small portion of granted patents, and the main products of public research are scholarly papers rather than patents. Just as patents, papers can also be cited by patents, and indeed both the cited patents and cited papers are served as the "prior art" of a patent application, playing a significant role for patent examiner to determine the patentability of the application. While there has been a large literature on the patent-to-patent citation linkage, studies about patent-to-paper linkage have been relatively scarce, especially systematic studies, which is the goal of this paper.

Our primary interest in this work is in the life science sector. The last several decades have seen an unprecedented rapid progress of life science, both in basic scientific discoveries and clinical medicine. Recent studies have suggested that biotechnology and pharmaceutical patents have been the main driver for the overall growth of patents and exhibit a particularly prominent "science linkage" [8]. This has prompted us to ask: How has the patent-to-paper citation linkage of life science patents changed over time? In particular, we aim to answer the following lines of research questions:

1. How has the amount of science linkage changed over time? Does the change vary across different technology classes?

2. On the cited side of the linkage, which countries and types of institutions produce the cited papers? Whether basic or applied research are more likely to be cited?

3. On the citing side, to what extent company patents cite public science?

These questions are important due to their high relevance to the policy community. Although the study of science linkage of patents has a long history, initiated by Narin and his colleagues in the 1980s [6, 10, 11], an up-to-date "status report" of science linkage has been lacking in the literature, partially due to the daunting task of resolving non-patent references to corresponding scholarly papers. Even in Narin's landmark study [10], the analyzed patents were granted in two two-year periods (1987–1988 and 1993–1994). By contrast, our analysis covers patents from 1976 to 2012. Such a large-scale corpus allows us to probe how the science linkage has changed over time. By using a large sample over a 36-year period, we contribute to the literature a systematic accounting of linkage from technology to science.

II. DATA AND METHODS

Sample selection: The NBER patent database [2] has been one of the major sources for information about U.S. patents. However, it only covers patents granted until 2006, whereas we want to extend to later patents. We therefore used patent data directly from the USPTO and parsed the downloaded XML files (`https:`

* q.ke@northeastern.edu

TABLE I. Top 20 countries (or territories) with most patents.

Country	Patents	%	Country	Patents	%
US	605, 875	55.7	TW	9, 849	0.9
JP	156, 491	14.4	SE	9, 192	0.8
DE	87, 127	8.0	BE	7, 883	0.7
FR	35, 505	3.3	IL	6, 825	0.6
GB	33, 274	3.1	AU	6, 388	0.6
CA	20, 691	1.9	DT	5, 361	0.5
CH	17, 865	1.6	JA	4, 847	0.4
KR	14, 901	1.4	DK	4, 816	0.4
IT	14, 376	1.3	FI	4, 051	0.4
NL	10, 794	1.0	AT	3, 476	0.3

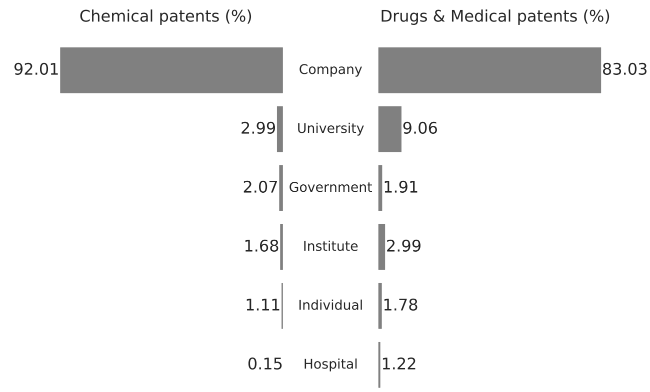

FIG. 1. Distribution of types of assignees.

//bulkdata.uspto.gov/) to obtain bibliographic information of patents. The NBER dataset instead is used as an auxiliary source when we infer various attributes of patents.

As we are interested in science-technology linkage in the life science domain, we need to select life science patents. In doing this, we note that there is an inherent trade-off regarding sample coverage. On one hand, it may not be desirable to narrow our analysis to, for example, patents about drugs that treat certain diseases. On the other, selecting patents from other domains, such as the software industry, may bias our statistics about science linkage, since those patents seldom cite biomedical papers. Here we leverage the categorization developed by NBER [2], which segments patents into six categories. We define life science patents as those belonging to one of the two NBER technological categories, namely Chemical (Category 1) and Drugs & Medical (Category 3). Operationally, we selected not-withdrawn (`https://www.uspto.gov/patents-application-process/patent-search/withdrawn-patent-numbers`), utility patents granted between 1976 and 2012 whose primary, three-digit USPC (U.S. Patent Classification) technology codes are in the 92 codes corresponding to the two NBER categories (Appendix 1 in [2]). The final sample used in our study consists of 1, 088, 650 patents. Patents that are not included into our sample are from the following NBER categories: Computers & Communications (Category 2), Electrical & Electronic (Category 4), Mechanical (Category 5), and Others (Category 6).

Country origin of patent: To examine whether science linkage varies across patents from different countries, we need to identify the country origin of a patent. We do so by using the residence of the first inventor, a standard practice used in the literature [1, 2]. USPTO also employs this method to, for example, report patent summary statistics by country (`https://www.uspto.gov/web/offices/ac/ido/oeip/taf/stcasg/regions_stcorg.htm`). For a tiny portion (0.63%) of patents whose country of origin cannot be determined through this way, which is due to missing data of the first inventor's address, we use the NBER data to locate the country.

Table I lists the top 20 countries that have the largest share of patents, which in total contribute to 97.3% of all patents in our cohort. It is clear that the US patent system has granted life science patents from inventors originated from diverse countries, although US accounts for more than half of the patents.

Type of patent assignee: To study how the types of assignees may affect citations to scientific papers, we need to classify patent assignees. To do this, we again leverage the NBER patent dataset that has already classified assignees of patents in 1976–2006 into one of the following six types: corporation, university, institute, government, hospital, and individual. For later patents, we assign the type based on the exact match of assignee names. If this fails, we then classify by checking the role of the assignee provided by USPTO, whether the assignee name is the same as an applicant, and whether it contains certain keywords (e.g., "Ltd", "University"). There are 9.76% patents without any assignee listed.

Figure 1 gives a side-by-side comparison of the decomposition of the types of assignees for both chemical and drugs & medical (DM) patents. Not surprisingly, the overwhelmingly majority of patents are assigned to companies. A larger fraction of DM patents, however, come from universities. Many previous works have linked this to the Bayh–Dole Act that permits universities to own inventions that are funded by government [9].

Non-patent references in patent: Each and every NPR cited in the patents has been resolved previously to determine whether and which MEDLINE paper it refers to, with a high accuracy obtained [3]. MEDLINE is perhaps the most widely used database for the biomedical research literature, curated and maintained by the US National Library of Medicine (NLM). It is publicly available and provides a variety of meta data about papers indexed there, including common bibliographic information like authors, affiliations, journal, publication year, funding, etc. It also provides domain specific information like Medical Subject Headings (MeSH). Moreover, many additional resources that we rely on have been built on top of MEDLINE, and literature has been using MEDLINE for innovation study, such as operationalization of the triple-helix model based on MeSH terms [13].

Country and institution type of papers: To understand how public science contributes to knowledge cited in patents, we need to classify the types of institutions of papers. However, an important question before the classification is which author's (or authors') affiliation we should use, as modern science has become a collaboration endeavor [17]. Here we choose to look at only the first-author's affiliation for two reasons. First, as stated from the NLM, "until 2014, only the affiliation of the first author was included," (`https://www.nlm.nih.gov/bsd/mms/medlineelements.html#ad`) and the first author's affiliation was not included until 1988. This limitation is also reflected in the data: 87% of the 218,483 papers cited in patents and without author affiliations were published before 1988. Second, in biomedical research, it is generally accepted that the first and the last author get the most credit of a paper for performing and supervising the research, respectively, and the two authors share the same affiliations in most cases.

From the text of the first author's affiliation, we seek to extract the country and institution type information. This task, fortunately, has been fulfilled by an online tool called MapAffil (`http://abel.lis.illinois.edu/cgi-bin/mapaffil/search.py`; [15]). It returns geography information and institution type of the input MEDLINE paper and has a reported accuracy of 97.7%. MapAffil classifies institutions into eight categories, namely educational, hospital, educational hospital, organization, commercial, government, military, and unknown. For our study, we merge educational hospital into educational, since teaching hospitals still serve the education role for training medical students. Furthermore, we combine the organization, government, and military categories into a single one, called public research organization (PRO), because we primarily concern about whether or not cited research are performed by companies. Previous studies have also employed a similar procedure [1]. Therefore, there are five types of institutions of papers, namely educational (EDU), PRO, hospital (HOS), commercial (COM), and unknown (UNK).

Funding support of papers: An ongoing effort in the study of the patent-to-paper citation linkage is to understand to what extent cited papers are supported by public funding. We retrieve this information from the paper meta data provided in the MEDLINE database. First, we determine whether a paper is funded by the US government by looking at whether the "Publication Type" field has any of the following four terms: "Research Support, U.S. Gov't, Non-P.H.S.", "Research Support, U.S. Gov't, P.H.S.", "Research Support, N.I.H., Extramural", and "Research Support, N.I.H., Intramural". Second, we determine whether a paper is supported by the NIH by looking at the "Grant List" field and further record which NIH institutes support the paper.

"Basicness" of papers: A repeatedly occurring assumption in the literature about the role of public science is that public science institutions conduct basic science, while private firms perform applied research by utilizing findings from basic research. Yet, few studies have examined to some extent papers cited in patents are basic research, possibly because of the difficulty in the operationalization of the two notions to papers. One notable exception is [6] that used the four-level classification scheme developed by the CHI Research in the 1970s [12]. The scheme assigns journals to one of the four categories, which are, from the most basic to the most applied, "basic research", "clinical investigation", "clinical mix", and "clinical observation". In this study, however, we do not adopt this method for four reasons. First, we are not aware the scheme is publicly available. Second, it remains unclear whether a scheme developed in the 1970s is still applicable nowadays, with numerous journals established since then. Third, the scheme only considers journals indexed in the SCI, while many MEDLINE journals are not there. Lastly and most importantly, the scheme operates on journals rather than papers. One immediate implication of this is that papers published in all the journals belonging to the same category have the same "basicness." This is problematic, because many biomedical journals publish qualitatively different types of research, which can be basic or applied. As an example, *Circulation*, a prestige journal with a 2017 Impact Factor of 18.88, "publishes [...] related to cardiovascular health and disease, including observational studies, clinical trials, epidemiology, health services and outcomes studies, and advances in basic and translational research" (`https://www.ahajournals.org/circ/about`).

Here we use a method that was recently proposed to identify translational science in biomedicine [4]. Translation science is research that "translate" basic scientific discoveries (bench-side or basic research) to clinical applications (bed-side or applied research). The method quantifies the basicness of papers directly. It results in a paper-level indicator called level score (LS) ranging from -1 to 1, with LS closer to -1 meaning that the paper is, by construction, more basic and 1 more applied. The method learns similarities between MeSH terms based on their co-occurrences among papers, using modern representation learning techniques. It then identifies an axis that points from basic science terms to applied ones. The basicness of a MeSH term is its projected position onto the axis, expressed as the cosine similarity between the axis vector and the term vector. The LS of a paper is the average basicness of its MeSH terms. The method has been validated and is consistent with Narin's four-level classification and other existing methods.

TABLE II. Summary statistics of non-patent references (NPRs) cited by U.S. life science utility patents 1976–2012, grouped by their NBER categories. SNPR refers to an NPR that corresponds to a MEDLINE paper.

NBER Category 1: *Chemical*

		Patents		Total		Mean SNPRs by	
Sub-cat	Name	All	w/ SNPRs	NPRs	SNPRs	All	w/ SNPRs
11	Agriculture, food, textiles	22,166	1,019	54,183	2,853	0.129	2.800
12	Coating	58,326	1,873	127,440	9,402	0.161	5.020
13	Gas	20,196	319	32,179	1,350	0.067	4.232
14	Organic compounds	91,301	26,538	686,384	291,540	3.193	10.986
15	Resins	105,960	15,667	585,437	288,730	2.725	18.429
19	Miscellaneous	384,434	20,049	827,008	123,405	0.321	6.155
	Total:	682,383	65,465	2,312,631	717,280	1.051	10.957

NBER Category 3: *Drugs* & *Medical*

		Patents		Total		Mean SNPRs by	
Sub-cat	Name	All	w/ SNPRs	NPRs	SNPRs	All	w/ SNPRs
31	Drugs	158,665	89,008	2,225,049	1,395,016	8.792	15.673
32	Surgery & medical instruments	137,981	28,975	668,424	274,526	1.990	9.474
33	Biotechnology	79,148	63,625	1,618,241	1,141,578	14.423	17.942
39	Miscellaneous	30,473	5,748	123,833	46,377	1.522	8.068
	Total:	406,267	187,356	4,635,547	2,857,497	7.034	15.252

III. RESULTS

A. Summary statistics

Table II reports the overall statistics of NPRs cited in the 1,088,650 patents in our sample, grouped by their NBER subcategories. The first group of statistics in Table II concerns about the total number of patents. Chemical patents share 62.7%, and the rest are DM patents. Among chemical patents, resins and organic compounds are the two largest subcategories, whereas drug and surgery & medical instruments patents are most presented ones in the DM category. Overall only 252,821 (23.2%) patents have at least one NPR linked to a MEDLINE paper (hereafter, scientific NPR or SNPR). This fraction, however, varies significantly across the two categories: only 9.6% for chemical patents but 46.1% for DM ones. The variability also holds at the subcategory level. 29.1% of resins patents and 14.8% organic compounds patents have SNPRs; by contrast, 80% of biotechnology patents cite MEDLINE papers, while 56.1% of drugs patents and 21% of surgery & medical instruments patents do so.

The second group of statistics is the total number of NPRs and SNPRs. A total of 6,948,178 NPRs were emanated from our corpus of patents, among which 2,312,621 (33.3%) are from chemical ones. More than half (3,574,777; 51.4%) of the NPRs are SNPRs, which are dominated by DM patents (2,857,497; 79.9%). The rest (717,280; 20.1%) are from chemical patents. Therefore, although there is a larger portion of chemical patents, they generate less amount of NPRs and are less linked to science, when comparing to DM patents. As for the subcategories, 49.3% and 42.5% of NPRs in resins and organic compounds patents, respectively, refer to MEDLINE papers. Biotechnology and drug patents account for 88.8% of all the SNPRs in the DM category, and 70.5% and 62.7% of their NPRs are scientific.

The last group of statistics is the average number of SNPRs per patent. Here we average by both all patents and patents with at least one SNPR, since the majority of patents have no SNPRs. On average, a chemical patent cites one MEDLINE paper; a DM patent cites 7 papers. Such contrast, however, is much less evident if we use the second averaging procedure. For patents that have at least one SNPR, a chemical patent has 11 SNPRs while 15 for a DM patent. Delving into subcategories, there is a large variation of the extent of linkage to science. Organic compounds and resin patents on average cite 3.2 and 2.7 SNPRs respectively. A biotechnology patent has on average 14 SNPRs, larger than any other categories.

In summary, all the overall statistics suggest that there is a huge variation of the volume of science linkage, which is dominated by biotechnology and drug patents. This result is consistent with a previous small-scale study [6].

B. Overall characteristics over time

Next, we investigate how overall characteristics change over time. Figure 2A shows a steady increase of the total number of granted patents over the examined period, reaching from 21,151 in 1976 to 52,994 in 2012. Such in-

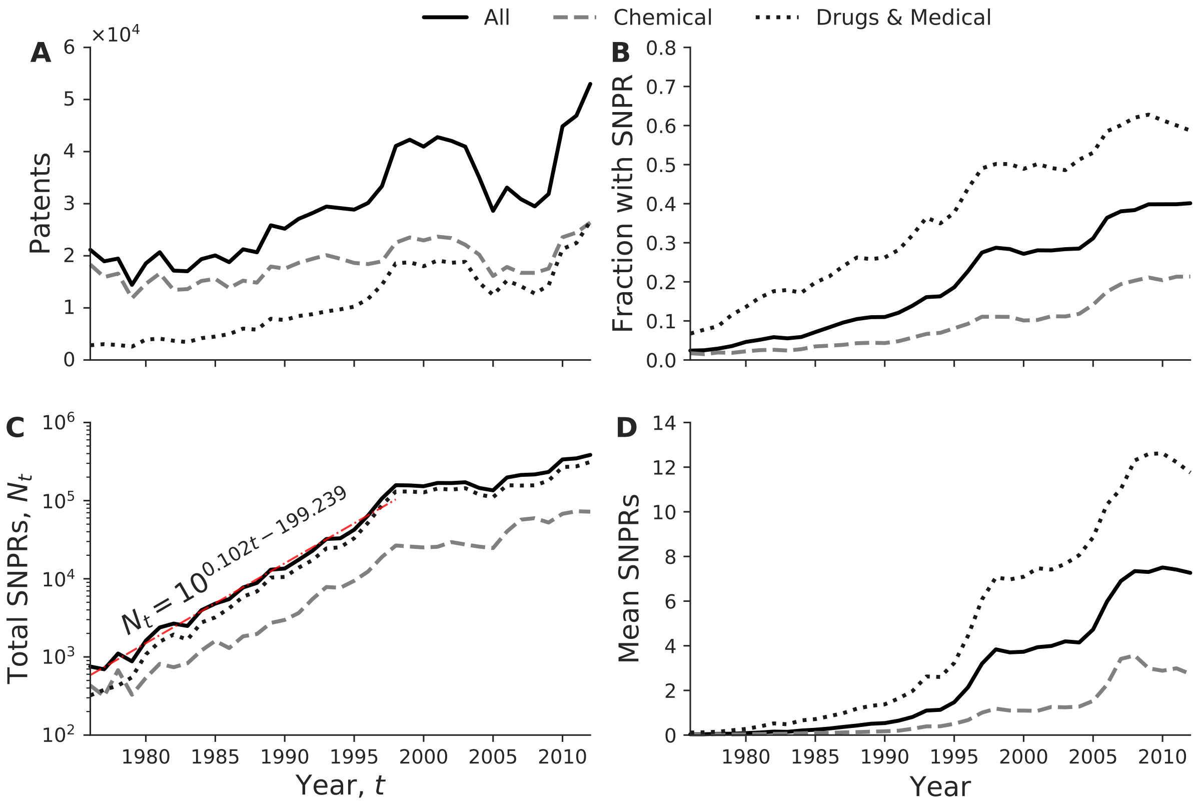

FIG. 2. Overall characteristics of the patent-to-paper citation linkage over time. (A) The number of patents. (B) The fraction of patents with at least one SNPR. (C) The total number of SNPRs. The red, dash-dotted line represent an exponential fit of the total number of SNPRs from 1976 to 1998, $N_t = 10^{0.102 \cdot t - 199.239}$. (D) The average number of SNPRs per patent.

crease is largely driven by the remarkable growth of DM patents: a nearly ten-fold increase from only $2,827$ in 1976 to $26,616$ in 2012. The number of chemical patents, on the other hand, has increased relatively slowly—44%. We notice that there is a flatten period followed by a decreasing period from 1998 to 2005, for both chemical and DM patents. Accompanying the increase of the raw number of patents is an increasing fraction of patents that cite MEDLINE-indexed papers, as presented in Figure 2B. In 1976, only 1.7% chemical and 6.8% DM patents had SNPRs, and in 2012 the number reached to 21.3% and 58.7%, respectively. Figure 2C plots the total number of patent-to-paper citations for patents granted in each year, demonstrating a remarkable increase of science linkage. We fit the growth from 1976 to 1998, obtaining $N_t = 10^{0.102 \cdot t - 199.239}$, where t is the calendar year and N_t is total citations at t. This means that there is an exponential growth of the total number of SNPRs, which doubled every $\log_{10} 2/0.102 = 2.94$ years. DM patents, again, drive the increase, and generate more SNPRs than chemical patents across years. Finally, the increase of the total number of SNPRs is not due to the increase of the number of patents, but rooted at patents themselves, as confirmed in Figure 2D which shows that the average number of SNPRs per patent also increases. Yet, DM patents have a faster increase than chemical patents.

We then add the country dimension to the analysis of patent-to-paper citations. Figure 3A shows that the average number of SNPRs per patent has been increasing for patents originated from the top 6 countries with the largest number of patents. The extent, however, varies by countries. For patents from Canada, the U.S., and the U.K., the average increases faster than the overall case, while for patents from France, Germany, and Japan, it increases slower than the overall case. What is noteworthy is that, starting from around 1996, Canada has surpassed US in generating more SNPRs per patent. Figures 3B–G further look at chemical and DM patents separately for each of the top 6 countries. From these figures, we can conclude that (1) across the top countries and categories, there is an increasing citation linkage from life science patents to biomedical research; and (2) DM patents exhibit a faster increase than chemical patents across years and countries.

C. Cited science

In this section, we explore the characteristics of papers that are cited by patents. We do so at the reference level; that is, a paper that is cited by multiple patents is counted multiple times, since the number of citations a

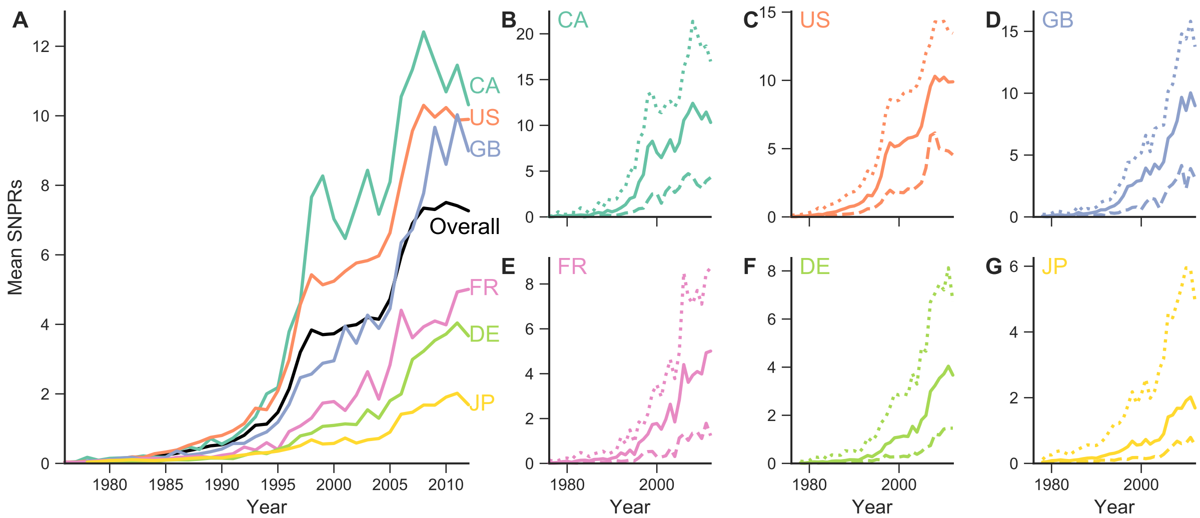

FIG. 3. The country dimension of the patent-to-paper citation linkage. (A) The average number of SNPRs per patent over all patents and over patents from the six most-patented countries. (B–G) The average number of SNPRs in patents originated from (B) Canada, (C) the U.S., (D) the United Kingdom, (E) France, (F) Germany, and (G) Japan. Solid lines in (B–D) represent all patents in the country, dashed lines chemical patents, and dotted lines drugs and medical patents.

paper receives from patents displays a heavy-tailed distribution, similar to the case of citations from papers [3].

First, we study the distributions of countries where cited papers are produced. Figure 4A plots the fractions of SNPRs authored by different countries over time. Here we display the results separately for the six individual countries that have the largest shares at 2012 and combine the shares of the rest countries together. The distribution at a particular year is derived as follows. We first get all the patents granted in that year, and then count the number of SNPRs produced by a given country and normalize it by the total number of SNPRs cited by all the patents in that year.

Figure 4A shows that the U.S. has been consistently the largest producer of cited science, accounting for almost half (49%) of the SNPRs cited by patents in 2012. Other top countries contribute to significantly smaller fractions: 6.8% for the UK and 5.5% for Japan. Note that here one may refrain to conclude that US science has been increasingly cited by patents over time, because the apparent increase of the fraction of US science could simply due to an increasing portion of cited papers with affiliation information available. This is corroborated by the observation that the share of US science has been stable since around 2000.

Figure 4B presents the fractions of cited references that are produced by different types of institutions over time, derived using the same procedure described above. Universities have been consistently the largest producer; 57.7% of references that are cited by patents granted in 2012 are written by them. PRO, which includes institutes and government, are the second major player, contributing to 9.8%. Public science, therefore, share 67.5% of cited science in patents. Companies account for only 10%.

We also examine what are the funding agency that supported the science cited by patents. Figure 4C shows the portion of references that are supported by U.S. government and by NIH specifically. Since 1990, more than 30% of cited science are supported by U.S. government and 20% by NIH. Table III further shows the top NIH institutes by the amount of citations they receive.

The last effort to characterize the cited science is to examine to what extent they are basic or applied research. We use the LS indicator described in Section II to measure the basicness of each paper. First, we plot in Figure 4D the histogram of LS for all the $14,916,511$ MEDLINE papers published between 1980 and 2012, illustrating how the entire biomedical literature is distributed along the basic-applied spectrum. This serves as the baseline set, against which we compare with the set of papers cited by patents. We observe a bimodal distribution. Figure 4E shows the same plot, but for the references that are cited by patents. We clearly observe that the vast majority of these SNPRs situate at the basic end. As a simple calibration, the bimodality in Figure 4D allows us to empirically find a threshold th to separate the two modes, which is 0.16. For 42.7% of all papers in Figure 4D, their score is smaller than th; By contrast, 85.2% of SNPRS in Figure 4E fall into this category. This result is robust if we instead look at the paper level. We further make additional measurements to ascertain that the observation is not driven by patents with many SNPRs. For each patent, we calculate (1) the average value of LS of its cited papers; and (2) the fraction of papers with LS smaller than th. The results confirm that for the vast majority of patents, most of their references are papers from the basic side.

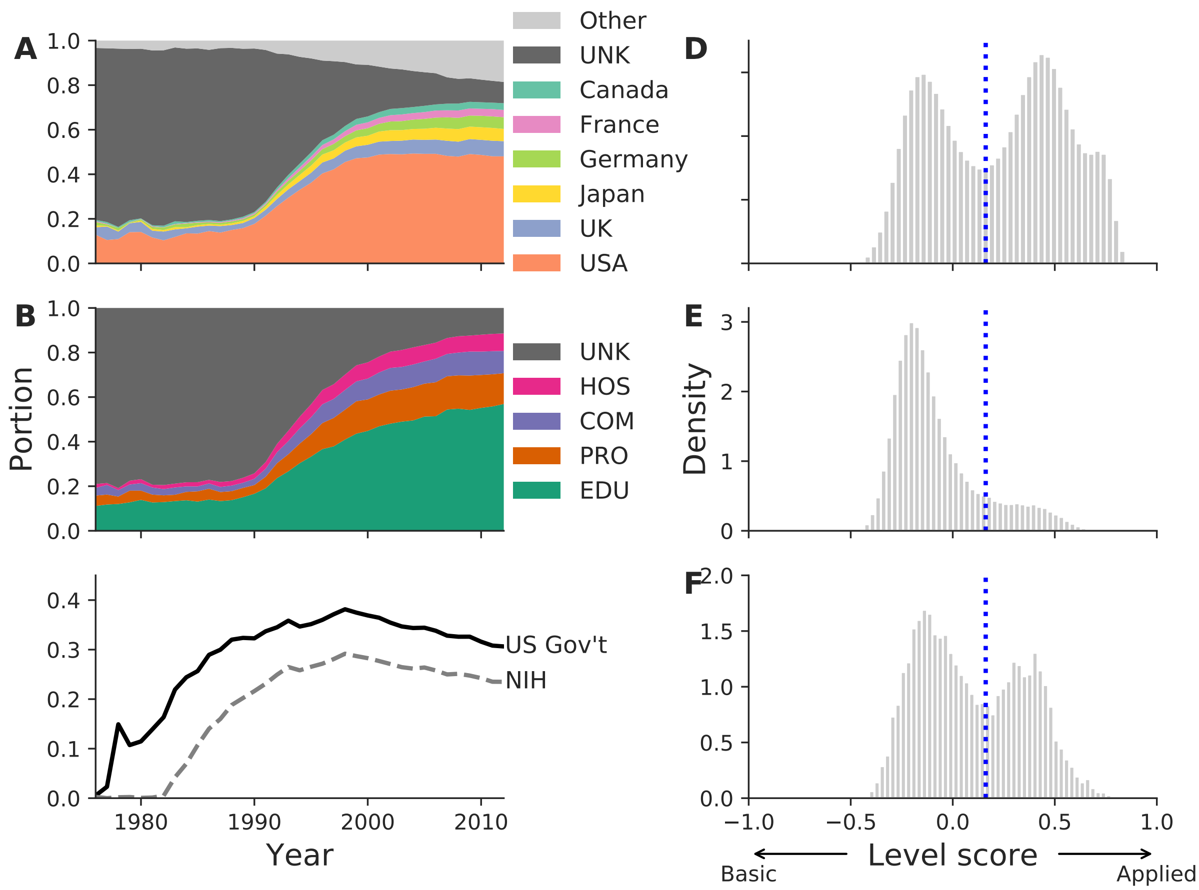

FIG. 4. Characteristics of cited science. (A–C) Fraction of SNPRs produced (A) by countries, (B) by institution types, and (C) supported by the U.S. government and the NIH in particular. (D) Histogram of level score for all MEDLINE-indexed papers published between 1980 and 2012. The blue dotted line indicates the score (0.16) corresponding to the local minimum of density. (E) Histogram of level score of references cited in USPTO-issued patents. (F) The same as (E) but based on patents associated with FDA-approved drugs. For 42.7% of papers (D) and 85.2% (E) and 59.6% (F) of references, the score is smaller than 0.16.

TABLE III. Number of citations for top NIH IC.

IC	Citations	%
National Cancer Institute	416, 642	22.3
National Institute of General Medical Sciences	251, 171	13.5
National Institute of Allergy and Infectious Diseases	248, 842	13.3
National Heart, Lung, and Blood Institute	239, 139	12.8
National Institute of Diabetes and Digestive and Kidney Diseases	128, 801	6.9
National Institute of Neurological Disorders and Stroke	89, 904	4.8
National Institute of Child Health and Human Development	55, 184	3.0
National Center for Research Resources	54, 456	2.9
National Institute on Aging	49, 538	2.7
National Institute of Diabetes and Digestive and Kidney Diseases	40, 886	2.2

As a separate case study, we examine SNPRs from patents that are associated with drugs approved by the U.S. Food and Drug Administration (FDA). The Hatch–Waxman Act mandates that drug innovators to provide FDA with the list of patents that covers the drug, and FDA included these patents in the *Approved Drug Products With Therapeutic Equivalence Evaluations* (also known as the Orange Book), although it is not FDA's task to actually evaluate the coverage. Such patents may possess economic value for their owner to

TABLE IV. Percentage of SNPRs originated from companies to different types of institutions.

	All		Chemical		DM	
	All	US	All	US	All	US
EDU	48.1	47.6	47.7	47.7	48.2	47.6
PRO	13.4	13.0	14.8	14.5	13.1	12.6
COM	11.5	11.6	13.1	13.2	11.1	11.2
HOS	7.1	7.4	5.4	5.6	7.5	7.9
UNK	19.9	20.4	19.0	19.0	20.1	20.8

surpass the cost of the development of drugs, and at the same time have the cure value for patients. We get this list of patents from `https://www.fda.gov/drugs/informationondrugs/ucm129689.htm`. We find a much smaller number (4,380) of such patents, which cite 28,512 SNPRs in our sample of papers. Figure 4F shows that, although most (59.6%) of these SNPRs are on the basic side, substantial amount are on the applied side, yielding a bimodal distribution that is not present in the overall case in Figure 4E. This may be related to the underlying process of drug development where pharmaceutical companies need to test the safety and effectiveness of drugs on human—which is applied research by definition.

D. Private-sector patents

In this section, we analyze science linkage of patents assigned to companies. Table IV presents the overall percentages of citations originated from company patents to papers authored by different types of affiliations. We observe that about 48% of citations form company patents go to university papers, and this varies little if we focus on chemical or DM patents separately or US patents only. Other public research organization ranks the second, contributing to 13–15%. Companies share only 11–13% of the science base of their patents.

We further look at the science linkage of individual companies. By way of example, Medtronic, a global medical device company, owns the largest number (3,565) of DM patents in our sample. We find 11,242 SNPRs in those patents, among which 5,824 are from universities, 579 from PRO, and only 297 from companies. The fraction of SNPRs authored by the public science section (universities and PRO), therefore, is 0.57. Table V extends this calculation to the top 10 companies that have the largest number of chemical and DM patents, indicating a significant linkage to public science. We make one more step and repeat this calculation to all the companies whose patents have at least one SNPR. Figure 5 shows the cumulative distributions of fraction of public science SNPRs for all those companies, across the chemical and DM categories. For more than 60% of companies, more than half of SNPRs cited in their patents are from public science.

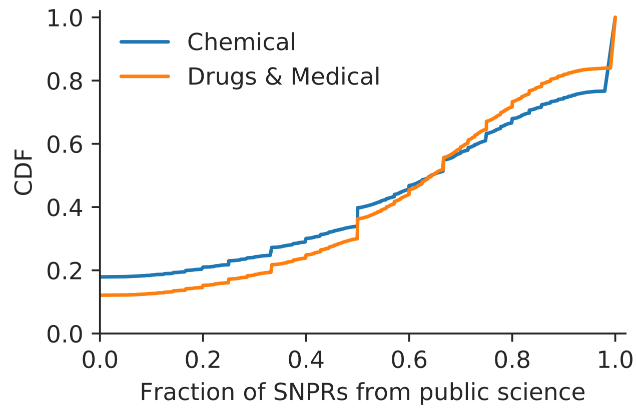

FIG. 5. Cumulative distributions of fraction of public science SNPRs cited in company patents.

IV. DISCUSSION

We have uncovered several empirical findings regarding how science linkage of US life science patents has changed over a 37-year period. First, the overall growth of life science patents are largely driven by the increase of drug and medical patents. The volume of science linkage are increasing exponentially, doubling every 2.9 years. The increase happens in both chemical and drugs & medical patents, as well as patents originated from different countries. Second, almost half of the SNPRs are produced in the US; the majority of them are from the public science sector. The overwhelming of them are basic research; yet, the nuance is for patents associated with drugs, with a non-negligible portion of them are applied research. US government and NIH in particular continue to be found as funders of research cited in patents. Third, for the majority of companies, most of their patents cite public science. Our results are important because they suggest a continuous linkage of public science to private sector inventions.

That the cited papers are more likely to be basic research resonates with earlier results [6, 10], but appears to be in a sharp contrast with a recent finding that declares no relationship about whether basic or applied research are more likely to be cited by patents [5]. The inconsistency may be due to the difference in entities analyzed. While we focused on papers, they focused on grants and made the basic/applied dichotomy based on grant abstracts. Furthermore, it remains to be seen to what extent one short grant abstract can represent the actual research performed and how different the level scores are for papers produced under the same grant.

Future work is needed to model patent-to-paper knowledge flow among different types of institutions and compare how it is different from patent-to-patent knowledge flow. That the science linkage is dominated by biotechnology and drug patents may suggest a finer level categorization of these patents that goes beyond existing

TABLE V. The top 10 companies that own the largest number of chemical (left) and DM (right) patents. The Fraction columns refer to the fraction of SNPRs that are authored by the public science section (universities and PRO).

Chemical			DM		
Company	Patents	Fraction	Company	Patents	Fraction
BASF AG	8,523	0.57	Medtronic Inc.	3,565	0.57
Bayer AG	8,450	0.42	Merck & Co., Inc.	3,000	0.47
E. I. du Pont de Nemours and Company	7,462	0.58	The Procter & Gamble Company	2,349	0.52
General Electric Company	7,276	0.73	Eli Lilly and Company	2,314	0.42
Eastman Kodak Company	7,028	0.46	Bayer AG	2,258	0.54
Fuji Photo Film Co., Ltd.	6,463	0.56	Pioneer Hi-Bred International, Inc.	2,064	0.81
The Dow Chemical Company	5,545	0.47	Cardiac Pacemakers, Inc.	1,955	0.61
Ciba-Geigy Corporation	5,059	0.31	Pfizer Inc.	1,816	0.55
Hoechst AG	4,468	0.36	Abbott Laboratories	1,736	0.58
Shell Oil Company	4,076	0.73	Monsanto Technology LLC	1,696	0.80

schemes is needed. Future work can base the linkage to science to cluster these patents and compare how the data-driven derived clusters align with traditional schemes.

[1] E. Bacchiocchi and F. Montobbio. Knowledge diffusion from university and public research. a comparison between us, japan and europe using patent citations. *The Journal of Technology Transfer*, 34(2):169–181, 2009.

[2] B. H. Hall, A. B. Jaffe, and M. Trajtenberg. The nber patent citation data file: Lessons, insights and methodological tools. Working Paper 8498, National Bureau of Economic Research, 2001.

[3] Q. Ke. Comparing scientific and technological impact of biomedical research. *Journal of Informetrics*, 12(3):706–717, 2018.

[4] Q. Ke. Identifying translational science through embeddings of controlled vocabularies. *Journal of the American Medical Informatics Association*, 2019.

[5] D. Li, P. Azoulay, and B. N. Sampat. The applied value of public investments in biomedical research. *Science*, 356(6333):78–81, 2017.

[6] G. S. McMillan, F. Narin, and D. L. Deeds. An analysis of the critical role of public science in innovation: the case of biotechnology. *Research Policy*, 29(1):1–8, 2000.

[7] M. Meyer. Does science push technology? patents citing scientific literature. *Research Policy*, 29(3):409–434, 2000.

[8] D. C. Mowery, R. R. Nelson, B. N. Sampat, and A. A. Ziedonis. *Ivory Tower and Industrial Innovation: University-Industry Technology Transfer Before and After the Bayh-Dole Act*. Stanford University Press, 2004.

[9] D. C. Mowery and B. N. Sampat. The bayh-dole act of 1980 and university–industry technology transfer: A model for other oecd governments? *The Journal of Technology Transfer*, 30(1):115–127, 2004.

[10] F. Narin, K. S. Hamilton, and D. Olivastro. The increasing linkage between U.S. technology and public science. *Research Policy*, 26(3):317–330, 1997.

[11] F. Narin and D. Olivastro. Status report: Linkage between technology and science. *Research Policy*, 21(3):237–249, 1992.

[12] F. Narin, G. Pinski, and H. H. Gee. Structure of the biomedical literature. *Journal of the American Society for Information Science*, 27(1):25–45, 1976.

[13] A. M. Petersen, D. Rotolo, and L. Leydesdorff. A triple helix model of medical innovation: Supply, demand, and technological capabilities in terms of medical subject headings. *Research Policy*, 45(3):666–681, 2016.

[14] C. Rosell and A. Agrawal. Have university knowledge flows narrowed?: Evidence from patent data. *Research Policy*, 38(1):1–13, 2009.

[15] V. I. Torvik. Mapaffil: A bibliographic tool for mapping author affiliation strings to cities and their geocodes worldwide. *D-Lib Magazine*, 21(11/12), 2015.

[16] M. Trajtenberg, R. Henderson, and A. Jaffe. University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and New Technology*, 5(1):19–50, 1997.

[17] S. Wuchty, B. F. Jones, and B. Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.