Linking NIH grants to drugs:
Bibliometric approaches and their limits

Bhaven N. Sampat (Columbia and NBER)

NIH SciSIP Workshop
8 April 2016
― “Discovery of new therapeutic agents and methods usually results from basic studies in medicine and the underlying sciences.” (Vannevar Bush, *STEF*, 1945)

― “Ultimately, the American taxpayer pays for the development of a lot of these drugs through NIH grants and other kinds of research grants” (Hillary Clinton, 2007)
Difficult to assess this role

- Unknown lags, serendipity makes this even harder
- Looking at patents from NIH and NIH researchers that are listed on the Orange Book one common approach (Sampat 2009, Sampat and Lichtenberg 2011; Stevens et al 2011)
- But a narrow lens:
  - Only 5 percent of grants generate patents (About 7 percent when corrected for RePORTER/Edison underreporting)
  - Qualitative, survey, and historical evidence suggests main impact of public sector research is more indirect, including through publications and other “open science” channels
The bibliometric approach

Background

The Bibliometric Approach

The Dataset

Examples

The Promise

Potential

Problems

The Road Ahead
An example: Tasigna (NILOTINIB HYDROCHLORIDE)

Non-patent references have long been used in bibliometric assessments of US science.

Patent-to-Literature Citations

Citations to the S&E literature on the cover pages of issued patents are one indicator of the contribution of research to the development of practical innovations. This indicator of how science links to invention increased sharply in the late 1980’s and early 1990’s (Narin, Hamilton, and Olivastro 1997), due in part to developments in U.S. policy, industry growth and maturation, and court interpretation. At the same time, patenting activity by academic institutions was increasing rapidly, as were patent citations to S&E literature produced across all sectors (NSB 2008, pp. 5-49 to 5-54).

Between 1998 and 2010, growth for this indicator was much slower. Of utility patents awarded to both U.S. and foreign assignees, 11% cited the S&E articles analyzed in this chapter in 2010 (appendix table 5-49). Concomitant with a growth in the percentage of U.S. utility patents awarded to foreign assignees, nearly 50% of the citations to the S&E literature in 2010 cited non-U.S. S&E articles.

In 2010, five broad S&E fields (biological sciences, medical sciences, chemistry, physics, and engineering) accounted for 96% of the citations to U.S. articles in USPTO patents of academic citations were notable in engineering (from 46% to 63%) and physics (from 43% to 66%).

Figure 5-34 shows, within the most cited S&E fields, the distribution by U.S. sector of citations to articles in U.S. patents in 2010. As noted above, academic articles dominate across all of the fields shown, from 62% in the biological sciences to 68% in chemistry. U.S. government-authored articles received 7% of the 2010 patent citations in both the biological and medical sciences. S&E articles from industry accounted for 27% of the engineering citations and about one-fifth of the articles cited in chemistry and physics. FFRDC-authored articles accounted for 6% of the physics citations.
Advantages over other approaches

- Patents directly resulting in the course of NIH grants
  - Patenting a minority activity
  - Publications a more common route through which firms benefit from public research (Cohen, Nelson, Walsh 2001)

- Patent-patent citations
  - A significant share (40-60 percent) from examiners (Alcacer, Gittleman, Sampat 2010)
  - But only 5 percent of patent-publication citations from examiners (Lemley and Sampat 2012)
  - Roach and Cohen (2012): Patent-publication citations provide a better signal of knowledge flows from public research than patent-patent citations

- But these data haven’t been widely used as other measures in evaluation of public funding
So what’s the problem?

Harold Varmus’s 1988 Science article Retroviruses is cited in many patents, but in numerous different formats, including:


(Note: EPO and WIPO moving towards better formatted journal references)
Solution

- Developed a matching algorithm to match all of PubMed against all citations in 3+ million USPTO patents
- Validated versus human matching
- Currently: ~1.5 million distinct PMID cited in ~500K distinct patents
- Link PubMed PMID to RePORTER grants, and patents to FDA Orange Book
Background
The Bibliometric Approach
The Dataset
Examples
The Promise
Potential Problems
The Road Ahead

Share of grants with a citing patent by IC

Figure 3
Grant-Patent Lags by Disease Area — Top 10 ICs

<table>
<thead>
<tr>
<th>Years after grant approval</th>
<th>% of grants with a patent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>10</td>
<td>40</td>
</tr>
<tr>
<td>15</td>
<td>60</td>
</tr>
</tbody>
</table>

- Cancer
- Heart, Lung, Blood
- General Med
- Allergy
- Diabetes
- Neurological
- Mental Health
- Drug Abuse
Changing lags between grant and citing patents

**Figure 4**
Grant-Patent Lags by Grant Cohort

![Graph showing the changing lags between grant cycle and patent issue over years after grant approval.

Years after grant approval

- 1980–1985
- 1986–1990
- 1996–2000

Years between grant cycle and patent issue.

% of grants with a patent
Citations by different types of patents

**Figure 5**
Grant-Patent Lags by Patent Quality

![Graph showing the percentage of grants with a patent over years after grant approval for different categories: Advanced Drug Candidates, FDA Approved, All Private Sector.](image)
Other examples of analyses using these data

- Sampat and Lichtenberg (2011): A third of all marketed drugs, and over half of priority review drugs, link back to NIH grants through citation links.
- Azoulay, Graff-Zivin, Sampat (2012): How does the geography of private sector patent citations to academic articles change when the academic inventor moves?
- Sampat and Yaqub (2015): How commonly is an NIH grant in one disease area cited by a patent associated with a drug in another disease area?
The promise

- What was the impact of NIH doubling on publications? On patents citing publications? On drugs associated with these patents?
- Do open access mandates affect innovation (patent citations to the affected articles)?
- Reproducability: Are publications based on contaminated cell lines less likely to be cited by patents? In important patents? What is the cost of improper validation?
- Returns to different kinds of grants ("Basic" vs "applied"), "triangle of biomedicine", and different grant mechanisms, grant sizes
Problems: Conceptual

- Endogeneity of funding choices
- Too linear!
Limits: False negatives

- Misses other channels of knowledge flows to drug development (conferences, labor markets)
- Misses impact of NIH funding on outcomes that are not embodied in final products
- Harder to do for medical products other than small molecule drugs (biologics; devices)
Limits: False positives

- Not all citations represent real knowledge flows.
- Legal reasons for citation in patents (whether by applicants or examiners) may not map to intellectual influence.
- If we see many citations, are all essential? Or substitutes?
- In general, heterogeneity in what contribution each citation represents.
In text

Heterocyclic is especially a five, six or seven-membered heterocyclic system with one or two heteroatoms selected from the group comprising nitrogen, oxygen, and sulfur, which may be unsaturated or wholly or partly saturated, and is unsubstituted or substituted especially by lower alkyl, such as methyl, phenyl-lower alkyl, such as benzyl, o xo, or heteraryl, such as 2-piperazinyl; heterocyclic is especially 2- or 3-pyrrolidinyl, 2-oxo-5-pyrrolidinyl, piperidinyl, N-benzyl-4-piperidinyl, N-lower alkyl-4-piperidinyl, N-lower alkyl-piperazinyl, morpholinyl, e.g. 2- or 3-morpholyl, 2-oxo-1H-sazepin-3-yl, 2-tetrahydrofuranyl, or 2-methyl-1,3-dioxol-2-yl.

Salts are especially the pharmaceutically acceptable salts of compounds of formula I.

Such salts are formed, for example, as acid addition salts, preferably with organic or inorganic acids, from compounds of formula I with a basic nitrogen atom, especially the pharmaceutically acceptable salts. Suitable inorganic acids are, for example, halogen acids, such as hydrochloric acid, tors of c-Abl, Bcr-Abl, and VEGF-receptor tyrosine kinase activity can be demonstrated as follows:

Test for activity against c-Abl protein tyrosine kinase. The test is conducted as a filter binding assay as follows: The His-tagged kinase domain of c-Abl is cloned and expressed in the baculovirus/Sf9 system as described by Bhat et al., J. Biol. Chem. 272, 16170-5 (1997). A protein of 37 kD (c-Abl kinase) is purified by a two-step procedure over a Cobalt metal chelate column followed by an anion exchange column with a yield of 1-2 mg/L of Sf9 cells. The purity of the c-Abl kinase is >90% as judged by SDS-PAGE after Coo massie blue staining. The assay contains: c-Abl kinase (50 ng), 20 mM Tris-HCl pH 7.5, 10 mM MgCl₂, 10 μM Na₂VO₃, 1 mM DTT and 0.06 μCi assay [γ³²P] ATP (5 μM ATP) using 30 μg/mL poly-Ala,Glu,Lys,Tyr-6:2:5:1 (Poly-AEKY, Sigma P1152) in the presence of 1% DMSO, total volume of 30 μL. Reactions are terminated by adding 10 μL of 250 mM EDTA, and 30 μL of the reaction mixture is transferred onto Immobilon-PVDF membrane (Millipore, Bedford, Mass., USA) previously soaked for 5 min with
Sampat and Pincus (2015)

Examined citations in 52 drugs (103 patents) citing public sector research. A small number of the patents (15) actually discussed the cited article in the full-text. Among these, several types of relationships:

1. The article provided knowledge directly related to the therapy or its evaluation (including testing in animals/humans, providing process or synthesis technology).

2. The article was about a specific target or class of compounds that guided research efforts.

3. The article provided general background on the disease or intervention.
What is to be done?

1. Build on Jaffe, Fogarty, Banks (1994), survey cited researchers to assess extent and type of contribution of their articles. (Citing researcher as well?)

2. One hypothesis: Articles discussed in the text of the patent more crucial than those just cited; possible to extract these “in text” citations using fuzzy matching approaches to test directly

3. Moreover, could use natural language processing techniques and machine learning to examine features of citations (including semantic context, for in-text citations) associated with crucial knowledge inputs (cf Lacetera et al 2015)

4. In depth qualitative study of what linkages the bibliometric approaches are picking up (and what missed) in specific disease areas

5. Combine with data (e.g. STAR Metrics) capturing other channels through which NIH research linked to patents on marketed drugs
Thanks

- Any takers?
- bns3@columbia.edu