Public R&D Investments and Private Sector Patenting: Evidence from NIH Funding Rules *

Pierre Azoulay[†], Joshua Graff-Zivin[‡], Danielle Li[§], and Bhaven Sampat[¶]

PRELIMINARY & INCOMPLETE

Comments are very welcome, but please do not cite or distribute without permission

First draft: Sept 8, 2013

This version: Nov 1, 2013

Abstract

This paper measures the impact of public R&D investments on innovation by private sector firms. We quantify the returns to grant spending at the National Institutes of Health (NIH) in terms of the biomedical patents it generates. Our paper makes two contributions. First, we use newly constructed bibliometric data to develop a method for flexibly measuring the outcomes of basic science investments. Second, we take advantage of the institutional features of NIH peer review to address concerns about the endogeneity of grant funding. Our results show that NIH funding generates more private patents than it crowds out. A $10 million increase in NIH support generates 2.8 additional patents; given an average grant award of $1.34 million, this means that we expect one additional patent private sector patent to be produced for every three additional NIH grants. We document, moreover, substantial cross-disease spillovers in funding for biomedical research; approximately half of all patents generated by NIH funding for one disease area are primarily relevant for a different disease.

*We are grateful to Jason Abaluck, David Autor, Alex Frankel, Ben Jones, Kelly Shue, Heidi Williams, and numerous seminar participants at Columbia GSB, Georgia Institute of Technology, Harvard Business School, and Washington University in St. Louis for helpful comments and suggestions. The authors gratefully acknowledge the financial support of the National Science Foundation through its SciSIP Program [Award SBE-0738142]. All errors are our own.

†pazoulay@mit.edu

‡jgraffzivin@ucsd.edu

§danielle-li@kellogg.northwestern.edu

¶bns3@columbia.edu

1 Introduction

Governments of industrialized nations invest billions of dollars in scientific research with the intention of generating economic advantage. In the life sciences, the culmination of this process is the development of new diagnostics and treatments to improve health and productivity. Since these final products are generally brought to market by private firms, assessing whether public investments in research are delivering on their promise requires evaluating their impact on private innovation. While this has been a longstanding concern in the literature (e.g. Murphy and Topel, 2003), measurement and inference challenges have hampered the development of strong empirical evidence in this area.

In this paper, we attempt to overcome these challenges in order to analyze the impact of biomedical research funding by the National Institutes of Health (NIH) on patenting by private sector firms within the United States. Our paper makes two contributions. First, we take advantage of newly available bibliometric data to create novel measures of the linkages between public R&D investments and the innovations that they subsequently generate. Second, we exploit the unique features of the funding process at the National Institutes of Health (NIH) to develop causal estimates of this relationship.

Public investments in biomedical R&D are potentially very important for health. The drug company Novartis, for example, made use of decades of publicly funded research in the development of Gleevec, a pathbreaking drug that transformed an extremely lethal cancer, chronic myeloid leukemia (CML), into a manageable chronic condition. CML is caused by a single gene mutation that leads a common cell-signaling protein, tyrosine kinase, to become overactive, resulting in a proliferation of white blood cells. This scientific understanding, which emerged over decades (1960s through 1990s), pointed toward a strategy for treating CML: find a way to inhibit the production of tyrosine kinase. After initial laboratory tests funded by an NIH grant, work on this problem moved into the private sector, where Novartis scientists adapted research on the use of kinase inhibitors in a different setting—the treatment of vascular conditions associated with diabetes—to push Gleevec from the laboratory to the clinic (Pray, 2008). In the years since it received FDA approval in 2001, the success of Gleevec has inspired a wave of pharmaceutical research aimed at developing drugs that selectively target cancer cells (as opposed to all fast-growing cells, as in traditional chemotherapy).

While the story of Gleevec is frequently cited as evidence that public sector investments spur private innovation, it also illustrates the pitfalls that accompany attempts to test this claim empirically. The synthesis of imatinib mesylate, the chemical compound behind Gleevec, was the culmination of decades of both private and public investment in research not just into cancer, but into gene mutation, cellular signaling, and vascular disease as well. This complicated history means that attempts to assess the role of public funding in developing this—or any other—medical treatment must 1) track the unpredictable, and often convoluted, path between initial R&D investments and final commercial products; and 2) isolate variation in public investments that are uncorrelated with the factors that also drive private investments. We make progress on both these issues.

Our first contribution is to develop new ways of linking basic science investments with commercial outcomes. The most recent work in this area, Blume-Kohout (2012), Toole (2007), and Manton et al. (2009), examines the effects of funding for a disease area on outcomes that can be identified in that same area, with pre-specified lags. Instead our approach uses newly available data on grant acknowledgements in research articles and patents and adds data on patent-publication citations that we construct. This allows us to explicitly trace out the flow of knowledge between funding and patents. Because this strategy does not require us to make ex ante assumptions about when and where funding may have an effect, we are able to account for the products of public investments in R&D both over time and across disease areas. With this information, we create three outcome measures that allow us to assess whether and to what extent NIH funding 1) directly generates patents; 2) supports research that aids firms in developing patents; and 3) increases total private sector patenting.

Our second contribution is methodological. We use institutional features of the NIH to address concerns about the endogeneity of public funding. Both public and private R&D investments may respond, for instance, to changes in the disease burden associated with particular conditions (e.g. Acemoglu and Lin, 2004). In this case, we would observe a correlation between public funding and private patenting even if public investments were useless. The innovation in this paper is to recognize that scientists do not simply propose research "on cancer"; instead, they propose research on specific scientific questions as applied to cancer. This means that the total funding that the NIH allocates to a disease does not necessarily reflect the true amount of research funding that is relevant for any particular set of researchers. Funding for a cancer researcher using a mouse model to study the physiology of tumors is unlikely to be useful for a cancer researcher using high-throughput sequencing techniques to study gene expression. By recognizing that biomedical research has both a "science" component—such as cell signaling, as in the case of Gleevec—as well as a disease component, we are able to construct a finer measure of public investment in a research area.

This level of granularity helps our analysis in two ways. First, we are able to include detailed controls for omitted variables, such as unobserved, time-varying disease burden or specific scientific breakthroughs, that have hampered other work. Second, we are able to take advantage of idiosyncratic variation in NIH funding for research areas within a disease or science area in a given year. Specifically, NIH funding rules require that grant proposals be awarded on the basis of their ordinal ranking as opposed to cardinal measures of their quality (even though both measures are collected). We instrument for NIH funding in a given research area with the funding it receives by luck, e.g. because of differences in the rank assigned to its applications, holding constant their absolute quality. This isolates a source of funding variation generated by procedural rigidities rather than by conscious efforts to direct resources to areas with more unobserved potential (see Section 3.3 for more details).

We show that NIH funding increases total private sector patenting. On net, an additional $10 million in NIH funding for a research area generates 2.8 additional private sector patents.[1] Given that the average NIH grant award is $1.34 million, we expect three additional NIH grants to generate approximately one additional private sector patent.

Our results are also the first to quantitatively document the spillover effects that NIH-funding targeted toward one disease has on other disease areas. We show that fully half of the patents resulting from NIH funding are for disease applications different from the one that funded the initial research.

Our main estimates do not take into account the possibility that changes in NIH funding can lead firms to reallocate resources to or from other projects. Reallocation can affect the interpretation of our results in two ways: if increased funding in one area leads firms to divert resources away from

[1] If a patent is linked to more than one NIH research area, it counts as $1/N$th of a patent with respect to each NIH funding area.

other projects, then this would lead us to overestimate the overall effect of funding; if, on the other hand, it leads firms to divert their resources toward other areas, then we would underestimate the impact of funding. We show in Section 7 that firms which work in an area of increased NIH funding produce more patents in that area, but not at the cost of patenting in other areas of its portfolio. This suggests that NIH funding spurs private patenting by increasing total firm R&D expenditure.

We proceed as follows. In Section 2, we discuss the role of the NIH in biomedical innovation and explain the institutional details that will be relevant to our identification strategy. We describe our empirical strategy in Section 3 and Sections 4, 5, and 6 describe our data, results, and robustness checks. We discuss extensions in Section 7 and Section 8 concludes.

2 Background and Terminology

The NIH is the world's largest single funder of biomedical research. Its annual budget of $30 billion comprises almost one-third of all US spending on biomedical research and development, private and public sector combined.[2]

Our empirical strategy requires us to associate NIH-funded grants with specific disease and science topic areas. To do this, we rely on the fact that NIH requires all grant applications to specify both a disease focus and a scientific topic focus. Specifically, the NIH is composed of 27 semi-autonomous Institutes and Centers (ICs) that are typically organized around diseases (for example, the National Cancer Institute). These ICs receive separate Congressional appropriations and are responsible for funding relevant to their mission. In order to be considered, a grant proposal must first specify an area of disease application because this determines the Institute that will likely be responsible for funding the proposal in the event that it is approved.

Meanwhile, grant evaluation at the NIH is handled by a centralized peer review body. During the span of our data, 1980-2000, the majority of grant review occurred in approximately 200 standing review committees, known as "study sections." Each study section is organized around a scientific topic—for instance, "Cellular Signaling and Regulatory Systems"–and comprises 20 to 40 reviewers working on that topic.[3] These study sections are responsible for assessing the potential

[2] See CBO "Research and Development in the Pharmaceuticals Industry" (2006).

[3] In 2006, the NIH reorganized its standing study sections. This involved closing or consolidating some study sections, splitting others, and creating new study sections, for instance one on data analytics, to respond to new topics and tools. The overall review process stayed largely the same. This change happens outside of our sample

of the specific research that a grant proposes to conduct. In addition to indicating the disease area for which her work is relevant, a grant applicant must also specify its area of scientific focus, as this determines which study section will be responsible for evaluating the application.

This organizational structure allows us to clearly define groups of grants that share a similar disease interest and that benefit from an understanding of the same science. Using administrative records, we gather the disease and science area for all grants over the period 1980–2000. We then use this information to construct funding amounts for research at the disease *and* science topic level: that is, we can distinguish how much money the NIH spent on cancer research focused on cell signaling in 1996 from how much it spent on cancer-research focused on biomarkers in the same year. For ease of exposition, we use "disease area" as a general term to refer to research funded by a particular NIH Institute and "science area" to refer to research reviewed by a particular NIH study section. A "disease-science-time" area refers to all grants funded by a particular IC and evaluated by a particular study section in a particular year. We refer to this as a DST. In this paper, the central unit of observation will be an NIH research area, as represented by a DST.

The organization of NIH peer review will also matter for our empirical work because, having controlled for our full set of fixed effects, our remaining variation comes from the rankings of grant applications made by study sections. We will discuss the relevant details of NIH peer review when we discuss our identification strategy in Section 3.3.

3 Empirical Implementation

The goal of this paper is to quantify how much additional private sector patenting occurs as a result of basic science investments by the NIH. The size and sign of this impact is theoretically ambiguous. NIH funding in a research area (DST) may have the following effects on private patenting in the same area.

3.1 Possible Effects of NIH Funding

1. No impact: In this case, NIH funding has no impact on the behavior of private firms. This may occur if, for instance, publicly funded research is scientifically interesting but of little

frame and, throughout our analysis, we refer to the old system.

practical value.

2. Crowd-in: Increased NIH funding in an area leads to increased private investment in that same area. In many cases, firms may not invest in foundational research because of scientific uncertainty, the high fixed costs of R&D, or the inappropriability of basic scientific knowledge. NIH investments may increase the expected returns to private investment by generating knowledge that, for instance, sheds light on innovation opportunities or reduces the riskiness of R&D projects. NIH funding may also increase private investment by lowering a firm's costs, for instance by contributing to the fixed cost of R&D or by training new scientists who can serve as labor inputs to private firms. Under crowd-in, public investments in R&D in a particular research area are expected to have a positive effect on private innovation in that same research area.[4]

3. Crowd-out: Increased NIH funding in an area leads firms to decrease their own research efforts in this area. This could happen for a variety of reasons: public funds could crowd out private investments by subsidizing the cost of research investments that a firm would otherwise have been willing to make on its own; or public funding could crowd out private innovation if it makes capturing profit from those innovations more difficult—by, for example, lowering the costs of entry for other firms. Under crowd-out, public investments in R&D in a particular research area are expected to have a negative effect on private innovation in that same research area.

NIH funding for an area may also impact how firms allocate funds across their research portfolio. An analysis of the impact of NIH funding on *overall* private R&D output, then, should account for the potential changes in private sector patenting in other areas of a firm's R&D portfolio. In Section 7, we discuss these issues in more detail and provide some empirical tests of the extent and direction of firm-level reallocation in response to NIH funding.

What this paper does not attempt, however, is to address the welfare implications of public investment in R&D. This is because both crowd-in and crowd-out are difficult to map into social utility without more information: crowd-in, for example, may be socially inefficient if it encourages

[4] Crowd-in should have an unambiguously positive effect on R&D investments; whether this translates into more patents depends on the nature of the investment. If, for instance, NIH funding leads private firms to pursue more high risk research, this may actually translate into fewer patents.

firms to invest in R&D with the goal of simply stealing business from each other. Conversely, crowd-out from a research area may be socially efficient if it occurs because public investment has simply revealed that there is little innovative potential in that area.

3.2 Measuring Patent Outcomes

We assess the direct effect of NIH funding using several measures of private sector patenting outcomes. Linking public funding with subsequent innovations presents a serious challenge because it is difficult to predict where and when basic science may find a practical application. Advances in basic science, even if directed at a particular disease area, can impact other areas in ways that are difficult to anticipate. The time lag between basic and applied research, moreover, is also difficult to predict; basic science may take decades to percolate into applied work or its effects may be seen immediately, as in cases where publications and patents are released in tandem (Murray, 2005; David, Mowery, and Steinmueller 1992).

A key innovation in this paper is that we construct new data to track the output of each NIH funded grant using data on grant acknowledgments, patents, and patent citations. As such, we do not need to make ex ante assumptions about when or where NIH funded research may have an effect; we let the data speak for itself. This data and variable construction process is summarized in Figure 1. In particular, we construct the following measures of patenting outcomes associated with a DST.

<u>Patents directly resulting from NIH funding</u>. Our first outcome measure is designed to assess whether NIH funding directly produces patentable knowledge. This is the most immediate measure of NIH grant output and has become more common since the passage of the 1981 Bayh-Dole Act created incentives for academic and other public-sector researchers to patent and license their discoveries. To construct this measure, we use NIH administrative data to track all patents that directly acknowledge financial support from the NIH (Figure 1, first column). We then aggregate the number of patents acknowledging NIH grants to the DST level: we refer to this outcome measure as the set of patents that are produced by NIH-funded researchers.

<u>Patents citing NIH-funded research</u>. NIH funding may also impact patenting by producing research that private sector firms then build on. Both economic theory (e.g. Nelson 1984) and policymakers

(e.g. Bush 1945) characterize this channel as the dominant way in which public sector research affects private-sector innovation. We assess this claim directly by identifying the number of private-sector patents that explicitly cite NIH-funded research. This is done by first linking NIH grants to the publications they support using grant-acknowledgement data, and then by linking those publications to the patents that build on their findings (Figure 1, second column). To accomplish this second task, we construct a new dataset describing patent to publication citations. This represents an innovation over the patent-patent citation data used in previous work examining returns to public research.[5]

Our first two measures of patent outcomes allow us to identify innovations that build on NIH funding without having to make ex ante assumptions about when and how this may occur. These measures, however, are also subject to two important drawbacks. First, relying on grant-publication-patent citations limits the types of intellectual influence we can observe; we would not, for instance, credit NIH funding if it leads to patenting through more complicated citation patterns, informal interactions, or the hiring of NIH funded researchers and trainees. This omission may lead us to underestimate the impact of NIH funding. Second, measures of patent output based on citations of NIH-funded research may overstate the returns to public investments by conflating the creation of new patents with the reclassification of existing ones. To see this, suppose that a patent is produced with private support in the absence of NIH funding. In this case, that patent would exist but not be accounted for in our data because it does not cite NIH funding. If NIH funding crowded out private sector support for this patent, then the patent would enter in our data and our outcome measures would mistakenly credit the NIH with its creation. Failing to distinguish the true effect of funding from crowd-out would lead us to overstate the impact of public monies.

Patents in the intellectual vicinity of an NIH funding area. To address both these concerns, our final outcome measure identifies all patents in the same intellectual "area" as a DST, including those that do not cite NIH-funded work. If public funding merely crowds out private investment,

[5] This true for two reasons. First, publications rather than patents are the main output of scientific researchers and, second, the vast majority (over 90%) of patent-paper citations come from applicants rather than examiners and are thus more plausibly indicators of real knowledge flows than patent-patent citations, for which only 60% of citations are applicant generated. See Agrawal and Henderson 2001, Sampat 2010, Alcacer, Gittleman and Sampat 2008, and Alcacer and Gittleman 2006. Details of this matching process are discussed in Section 4 and Appendix A.

then an increase in NIH funding should not affect this measure of innovation because the number of NIH-linked patents would simply offset the number of non NIH-linked patents. Further, because this measure does not limit itself to simple citation links between funding and patents, it can capture a richer set of ways in which NIH funding may impact innovative output. This final measure allows us to assess the net impact of public funding for a research area on total private-sector innovation in that area.[6]

We construct this measure by beginning with the universe of biomedical patents (see Figure 1, third column). We then infer which DSTs are relevant for each patent by examining which DSTs supported publications *similar* to the ones cited by that patent. Specifically, we link a patent to the publications that it cites and then we use a similarity measure based on article keywords to identify a set of related publications. The final step in our matching process is to link this broader set of publications to NIH funding areas, again via grant acknowledgments. The main difference between this matching process and our previous approaches is that we assign *all* biomedical patents to at least one relevant DST[7]; it follows that a patent here need not actually cite research that is funded by a DST to be linked to a DST.

Each of our outcome measures is designed to provide a different test of the effects of NIH funding on private innovation. One hypothesis is that NIH funding is useless to private firms. If this is the case, then an exogenous change in funding should have no effect on innovation as measured by any of our outcomes. If instead NIH funds research of use to private firms, then NIH funding should have a positive effect on explicitly linked patent output. Finally, a positive effect of NIH funding on total innovation in an area indicates that public funding leads to the net creation of private-sector patents that would not have otherwise been developed.

Finally, it is important to note that not all research outputs are patentable. While this is less of a concern in the life sciences than in other sectors (since the propensity to patent is higher in

[6]All of our outcome measures allow for funding from a particular NIH disease-science area to have impacts in other disease or science areas. Thus, when we say that NIH funding for a DST increases private sector patenting in its "area," we mean that funding for a DST leads to more patents in research areas for which that DST's funding is relevant. For example, if cancer-related research articles about cell signaling are similar to diabetes-related articles about cell signaling, then diabetes patents related to cell signaling would be considered "in the area" of funding for cancer-cell signaling. If a random shock to funding for cancer-cell-signaling leads to an increase in diabetes-related cell-signaling patents, our method would credit cancer funding with this increase.

[7]Our match rate is actually 82%. See Section 4 for a discussion of this.

drugs and devices than in other fields[8]), our analysis will still miss a range of economically and clinically important outputs of NIH research, e.g. epidemiological knowledge about health risks; clinical research on drugs, devices, or procedures, and practices that change clinical practice or program delivery.

3.3 Estimating Equations

Once we have constructed measures of private patenting associated with NIH funding areas, we estimate the following regression:

$$\text{Patents}_{dst} = a_0 + a_1 \text{Funding}_{dst} + X_{dst}\beta + \delta_{ds} + \gamma_{dt} + \nu_{s't} + \mathbf{X}_{dst}\mathbf{b} + \varepsilon_{dst} \quad (1)$$

Here Patents_{dst} describes the number of patents linked to funding from disease area d, science area s, at time (year) t using the process described in Section 3.2. The main explanatory variable, Funding_{dst}, is the amount of funding given to disease area d science area s at time t. The key to our identification strategy is that Funding_{dst} varies within disease-year and within science area–year. Given this, we proceed in two steps: first, we include a detailed set of fixed effects and other controls to account for factors that may affect both public and private investments. Second, we use an instrumental variables strategy to isolate the part of the remaining funding variation that is driven by randomness in NIH's peer review process.

Without any controls, the coefficient on funding, a_1, in Equation (1) may be biased in a number of ways. Some diseases, such as cancer for instance, may simply impact more people—and thus command both more public and private interest—than others. To address this, our regressions include fixed effects δ_{ds} to control for any time-invariant differences between research areas defined at the disease–science level. This not only controls for differences in disease burden, but also for more complex unobservables such as the possibility that some research areas may be more tractable than others, even within the same science area. Research into the genetic basis of CML, for instance, is easier than research into the genetic basis for schizophrenia because the former only involves one gene mutation.

Both the demand for innovations in an area and our understanding of its science, however,

[8] See Cohen, Nelson and Walsh (2000).

can change over time. We may be concerned, for instance, that a rise in obesity increases both the potential market for heart disease drugs and public funding devoted to developing them. To control for this and other disease specific changes in demand (e.g. disease burden, popular interest) or supply (e.g. changes in the feasibility of research), we also include disease–year fixed effects, γ_{dt}.

Another concern may be that NIH funding for a DST also changes in response to scientific advances. New DNA-sequencing technologies in the late 1990s, for instance, may have increased both public and private research funding for diseases with a genetic component. We also include science–year fixed effects, $\nu_{s't}$, to control for this type of variation.[9] Finally, we include additional controls for the the quality of grants in a DST and lagged measures of a DST's patent productivity.[10]

3.4 Identifying Variation

Figure 8 plots residual variation in funding taking out, successively, fixed effects for disease–science, disease–year, and science–year. At this point, our remaining variation in DST funding comes from within-disease-year or within-science-year changes. These types of funding changes present a problem for our identification strategy if they are also correlated with changes to the innovative or commercial potential of a research area. In the wake of Gleevec's success, for example, private firms invested more money into applications of cell-signaling research in cancer treatment. Equation (1) would not be able to identify the impact of any correlated changes in NIH-funding for cancer cell-signaling because we would expect changes in patenting for this area even in the absence of NIH-funds.

There are two potential sources of endogeneity in funding within the disease-year or science-year level. The first is what we refer to as "top-down" endogeneity; DST funding may be endogenous if Congress or an NIH Institute allocates more funding to DSTs on the basis of its innovative or commercial potential. In response to the success of Gleevec, for example, the National Cancer

[9] Our data contain 431 science areas as defined by study sections. Because it was difficult to estimate standard errors with study section by year fixed effects, we included Integrated Review Group (IRG) by year fixed effects instead. This is why we denote our fixed effects as $\nu_{s't}$, as opposed to as ν_{st}. IRGs, of which there are about 40 in our data, are the broader science areas by which study sections are grouped. A study section on "Cellular Signaling and Regulatory Systems," for instance, is included in an IRG on cell biology.

[10] Application quality controls include cubics in the average raw score received by awarded grants and cubics for the average score received by all applicants regardless of whether they are funded; ten annual lags of number of publications produced by the DST and number of applications to the DST; and dummies for number of DST applicants near the IC's funding threshold. We also include ten years of lagged funding controls with indicators if a DST is not observed in any year.

Institute may decide to devote a greater percent of its budget toward the study of cell-signaling or gene expression, scientific topics that are particularly relevant for targeted cancer therapies.

While it seems plausible—and, indeed, efficient—that the NIH might direct funding to DSTs on the basis of their evolving potential, in practice, NIH grant funding rules make this very difficult. In particular, the way that an NIH grant application is funded is as follows: when grant application is submitted, it is assigned, on the basis of its disease area, to an Institute (IC) for funding and, on the basis of its science area, to a study section for review. Study sections cut across ICs in the sense that they review applications in multiple disease areas as long as they share an interest in the same science. Study sections assign each grant application a score which is then normalized. After scoring, the NIH's funding rule is mechanical: an IC lines up all applications it is assigned and funds them in order of their normalized score until its budget has been exhausted (the worst score that is still funded is known as the payline).

As a result, ICs have little purview to adjust the number of grants they fund from each study section in response to their scientific relevance. Because scores are normalized within study sections before being compared, an IC effectively can only fund the second-ranked application from a study section after it has funded the first-ranked application from all other study sections. This means that if cell-signaling were particularly "hot" in one year, the NCI could not decide to fund the top 20 cancer-related cell-signaling applications without first funding the top 20 cancer-related applications in all other science areas; most likely, it would not have the budget to do so.[11] In fact, the rigidity of this system was cited in an NIH-commissioned report from 2000, urging reform:

> Researchers perceive that...applications describing some of the most productive, highest impact work may be assigned to too few study sections, causing too much of the "best science" to compete with itself; that the scope of some study sections is restricted to research with relatively low impact, resulting in undeserved "entitlements"....[12]

Even if Congress or the NIH is not able to allocate funding top down, another possibility is that NIH study sections may give higher scores to applications from disease areas with more

[11] The main way that ICs get around these rules is to either fund an application out of scoring order or to issue a request for proposals (RFP or RFA) on a specific topic. RFPs and RFAs also account for only a small portion of NIH grant spending. Grants responding to these are typically evaluated in one time "special" study sections, which we exclude from our sample because it is difficult to assign them to a scientific area. See Section 6 for a discussion of out of order grant funding.

[12] "Recommendations For Change At The NIH'S Center For Scientific Review" Final Phase 1 Report, Jan 14, 2000. Available online. http:public.csr.nih.govaboutcsrCSROrganizationDocumentspsbrfinalphase1reportjan2000.doc.

potential. Returning to the example of Gleevec, even if the National Cancer Institute were not able to earmark more funds for cancer cell-signaling research, the cell signaling study section may decide to give higher scores to cancer-related applications. Because higher raw scores generally translate into higher rank scores, study section decisions may induce endogeneity in DST funding from the bottom up, even after controlling for the fixed effects in Equation (1). We instrument for funding in order to address this specific concern.

3.5 Constructing an instrument for DST funding

Our instrument for DST funding is based on the fact that grant applications with the same score may have different funding outcomes because of differences in rankings.

Figure 9 illustrates an example of this variation. Here, the National Cancer Institute (NCI) and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) are responsible for funding grant applications to the cell-signaling and tumor physiology study sections. Panel 1 illustrates the raw scores that each study section assigns to the applications that they review. These raw scores translate into normalized scores, which can be thought of as within-study section ranks. Study sections, however, are not responsible for funding. Instead, Institutes are responsible for funding grant applications from multiple study sections. To make this funding determination, Institutes order applications from all study sections by their within study section rank. There are, then, three "scores" associated to an application: the raw score given by the study section, the within study section rank that this implies, and, finally, the within-Institute ranking that is obtained when study section ranks are combined. It is this final score, which we call a "priority rank," that determines whether an application is funded.

This is illustrated in Panel 2 of Figure 9, where we order our example applications by priority rank within their respective Institutes, using raw scores as tie breakers. The solid line is the pay line; applications with priority above the pay line are funded; ones with priority below are not. The highlighted cells in Panel two show that, within a disease-year, applications from two different science areas with the same score may have different founding outcomes. In this case, tumor physiology applications are less competitive than cell-signaling applications, meaning that—independent of quality—proposals to study tumor physiology in cancer are more likely to be funded than proposals to study cancer cell-signaling. Similarly, within a science-year, applications from different disease areas with the same score may also have different funding outcomes. In Panel 3, which recreates Panel 1 but with funding outcomes, cancer applications in tumor physiology have high rankings. These applications will thus take up more NCI funding, leaving less for cancer-cell-signaling, even if the latter are of higher absolute quality.

We construct an instrument that isolates variation in DST funding coming from differences in priority rank, controlling for raw scores. In Figure 9, the cancer-tumor physiology DST and the cancer-cell signaling DST both have applications with raw scores of 7.6, but only one is funded, the tumor physiology application. The additional funding that cancer-tumor physiology receives, then, can be thought of as "lucky" funding because it is not related to the merits of the application. This is illustrated in Figure 10.

To capture this more generally, we instrument for Funding_{dst} with funding for a DST's grants that are just above an Institute's funding threshold:

$$\text{Lucky\_Funding}_{dst} = \sum_{g \in \mathbf{G}_{dt}} F_{gdst}$$

where $\mathbf{G}_{dt}$ is the set of ten grant applications on either side of the funding threshold for disease area d in year t.[13] For the example given in Figure 9, $\mathbf{G}_{dt}$ would include all grant applications from either cell-signaling or tumor physiology within a five-grant window from the payline at a given disease-based Institute. For this set of applications near the threshold, we ask how many of a given DST's applications are funded and claim that, within the set $\mathbf{G}_{dt}$, whether an application is funded is uncorrelated with innovative potential in that application's DST.

In general, this may not be true. DSTs with higher quality applications overall may have more applications near the threshold, or conditional on being in near the threshold, some DSTs may have applications with higher absolute quality. Thus, we use our instrument in conjunction with a full set of indicator variables for the number of grant applications an given DST has near the threshold set $\mathbf{G}_{dt}$, as well as cubics in the average raw score of all grant applications to a DST, the average raw score of all funded applications in a DST, and the average raw score of all applications within the threshold set $\mathbf{G}_{dt}$ from a DST. Together, this isolates the portion of a DST's total funding that comes from grant applications near a funding threshold that are funded on the basis of their rank

[13] We can use different bandwidths; this changes the power of our instrument but not the magnitude of our results.

score, not their cardinal quality.

Finally, to deal with the possibility that ICs may fund grants out of order, we construct **G** using only grants that are funded in order.[14] Section 6 discusses specification checks associated with this IV strategy and presents results using two alternative instruments.

4 Data Construction and Descriptives

Our analysis combines data from several primary sources: 1) Administrative data on NIH funded grants from NIH RePORTER; 2) publication data from PubMed including information on grant acknowledgements; 3) patent data from the USPTO; and 4) information on FDA-approved drugs from the Orange book. In addition to these, we use data from iEdison to identify patents developed as a direct result of NIH funding and NIH's PMRA data linking all PubMed publications to related publications. Our final analytic sample captures linkages between the universe of NIH-funded grants from 1980-2000 at both the individual grant and DST levels, and the universe of all life science patents citing biomedical research from 1985-2005. Figure 1 provides an overview of our data construction.

4.1 Grant-level Patent Match

Our first set of outcome measures involves linking NIH grants to patents that they directly produce and to patents citing research that they support. To do this, we begin with data on all 123,478 NIH grants from 1980-2000 that are evaluated in chartered study sections and funded by biology-focused Institutes.[15] The characteristics of these grants are described in Table 1. These grants are funded by 16 ICs, representing disease areas and 443 study sections, representing science areas. In total, we have grant-level data on the activities of 11,110 DSTs.

[14]A grant is considered to be funded in order if there are no grants with higher scores that are unfunded¿

[15]For our primary analysis, we include grants funded by the National Cancer Institute, the National Eye Institute, the National Heart, Lung, and Blood Institute, the National Institute of Diabetes and Digestive and Kidney Diseases, the National Institute on Drug Abuse, the National Institute of Neurological Disorders and Stroke, National Institute of Mental Health, the National Institute of Allergy and Infectious Diseases, the National Institute of Arthritis and Musculoskeletal and Skin Diseases, the National Institute of Dental and Craniofacial Research, the National Institute on Alcohol Abuse and Alcoholism, the National Institute on Aging, the National Institute on Deafness and Other Communication Disorders, and the National Institute of Environmental Health Sciences, the National Institute for Human Genome Research, and the National Institute for Biomedical Imaging. This set excludes NIH ICs devoted to, for instance, nursing research or minority health disparities, which do not fit into our disease-science description of NIH research.

The average award size for grants in our sample is approximately \$1.34 million. The majority (68%) of grants are R01-equivalents—the R01 is a renewable, project-based grant that constitutes the majority of NIH's grant spending—and most (63%) are for new research projects (as opposed to renewals of existing projects).

NIH-funded scientists rarely patent their research: we find that only 1,283 out of 123,478 grants, or just over 1%, are directly acknowledged by private-sector patents. These grants tend to be much larger; the average award for a grant that financially supports at least one patent is \$3.08 million, reflecting the fact that translational and clinical research projects tend to be more expensive. The upper left hand side panel of Figure 2 plots the distribution of the number of patents by NIH-funded researchers.

Next, we consider how many patents build on NIH-funded research. This matching process is described in the second column of Figure 1: we use NIH administrative data to link grants with the publications that they produce and then we use our own algorithm to extract publication references from the texts of patent applications in the USPTO database. Figure 6 illustrates our grant to publication to patent match with an example. In its first three years of funding, the NIH grant CA-065823 was acknowledged by four publications, among which is the article published by Thiesing et al. in the leading hematology journal Blood. We observe this link because grant acknowledgements are reported for publications indexed in the National Library of Medicine's PubMed database. Next, the Thiesing et al. article is listed as prior art in patent number 7,125,875 issued in 2006 to the pharmaceutical firm Bristol Myers Squibb. To capture this link, we use natural language processing techniques to identify and standardize publication references in the text of patent applications at the USPTO (see Appendix A for more details).

This matching process links the majority of grants in our sample, 81%, to at least one publication and 34% to at least one patent (via a publication that that patent references). The cost of these grants, moreover, is not considerably different from the cost of the average grant overall: \$1.96 million vs. \$1.34 million. The middle left panel of Figure 2 plots the distribution of patents linked to grants in this manner.

Our final set of outcome measures considers the number of patents that are in the intellectual vicinity of NIH-funded research. These are privately held patents that need not be associated to the NIH in any way, but which may still be affected by changes to NIH funding. The third column of

Figure 1 illustrates this process. To construct this measure, we begin with all 173,631 science-based biomedical patents issued between 1985 and 2005. For each of the publications that a patent cites, we find the set of related publications published within five years of the original publication. Then, for this set of directly cited or related publication, we find the DSTs of any NIH funding that they acknowledge. The goal of this procedure is to infer what NIH funding areas are relevant for any given patent by finding the DSTs that funded research similar to the work cited by the patent.

Figure 7 illustrates this process. Patent number 6,894,051was issued to Novartis in May 2005, one of the five patents listed in the FDA Orange book as associated with the drug imatinib mesylate, better known by its brand name, Gleevec. Patent 6,894,051 does not cite any publications which are directly supported by the NIH so it would not be linked to an NIH DST under our citation-linkage measure of innovative output. It does, however, cite PubMed publication 8548747, published in *Cancer Research* in 1996. The PubMed Related Citation Algorithm [PMRA] indicates that this publication is closely related to PubMed article 9389713, which acknowledges funding from NIH grant CA-0011422. These indirect bibliometric linkages are valuable to us because they enable us to link the great majority of life science patents to an NIH disease-science area. In other words, most patents can be traced back to one or more NIH grants, because most patents cite publications that are similar to publications that acknowledge NIH funding.

We match 141,356 patents ($\sim$ 80 % of our universe) to at least one NIH DST using either the direct or PMRA citation approach. These patents, meanwhile, are not concentrated in a small subset of DSTs. Table 2 indicates that approximately 90% of DSTs produce research that is potentially relevant for patenting. Counting the number of patents in the vicinity of an DST leads to a significantly greater number of patent matches: the median DST is linked to 340 patents, the 75th percentile, to 1,649. Again, these grants do not require considerably more investment than the average grant: \$1.53 million vs. \$1.34 million. The bottom-left panel of Figure 2 graphs the distribution of the number of adjacent patents for both individual grants and DSTs.

One important detail to note is that Figure 2 reports unweighted patent counts. That is, a patent which is linked to multiple DSTs continues to count as one for each DST. This means that the number of citation- and PMRA-linked patents are not directly comparable; under a PMRA-match, the same patent can be linked to many more DSTs through the inclusion of related publications so that the total patent count per DST is mechanically higher. In our regressions, we report estimates where patent counts are weighted in the following way: regardless of what outcome measure we use, if a patent is linked to N DSTs, it counts as $1/N$ of a patent in each DST. This means that patents are restricted to being counted once across all DSTs to which they are linked. The correct way to weight patents depends on the nature of the production function for patents with respect to NIH-funded research. Ideally the weight assigned to a patent at any particular DST should be the probability that support from that DST was pivotal for the production of the patent. We will discuss the implications of weighting for the interpretation of our results in Section 5.

Our data allow us to document, for the first time, new patterns in the dissemination of basic science investments over time and across disciplines. Figures 3, 4, and 5 describe how the time lag between NIH-funded research and patent output varies. Figure 3 documents substantial variation in the relevance of NIH funding for patenting across disease applications. Approximately 15 years after funding, almost 60% of grants funded by the National Institutes for Allergy and Infectious Diseases have produced research that has been cited by a patent. In contrast, this is true of only 20% of grants funded by the National Institutes of Mental Health. These differences likely reflect differences in the ease of biomedical innovation across disease areas, as opposed to differences in the efficacy of NIH funds. Figure 4, meanwhile, shows that time-to-patent has been decreasing over time. Only 20% of grants awarded between 1980 and 1985 produced research that is relevant for a patent in the ten years following. For grants awarded between 1991 and 1995, this figure was almost 40%. Finally, Figure 5 examines time-lags for different types of patents. We see that a grant is most likely to produce research relevant to patents in its own disease area. Perhaps more surprisingly, though, is that it is still quite likely to produce research relevant for patents in different disease areas. Grants are also just as likely to produce publications cited by private sector patents as public sector patents (though rates of citations differ, as there are more private sector patents). Finally, patents associated with successful drugs are still rare.

4.2 DST-level Patent Match

Our empirical variation, however, resides at the level of the DST, the set of grants in the same disease and science area in the same year. Table 2 describes our DSTs and how they are linked to patents. The average DST supports 11 grants totaling $14 million in funding.

Nine percent of DSTs produce research that is directly patentable, but the average number of patents associated with a DST through direct patent acknowledgements is small: even conditional on directly producing at least one patent, the median DST produces only half a private sector patent; the 75th percentile DST provides financial support for only one patent. If NIH funding has any impact on patenting behavior, it does not seem to be primarily through this route.

NIH funding plays a larger role in funding research that is cited by patents. Using this definition of a link, 65,745 patents or 38% of our universe of patents, can be matched to 7,133 DSTs, or 64%, of our DST sample. The median DST is linked to 7.5 private sector patents in this way; the 75th percentile to 19. Conditional on being linked to a patent, the median DST is linked to 9, the 75th percentile to 20. Our PMRA match increases the number of patents that we associate with DSTs: the median DST is linked to 19 weighted patents, the 75th percentile to 40. The right-hand-side panel of Figure 2 illustrates the distribution of patents linked at the DST-level.

5 Main Results

Tables 3 through 5 show our main results for each of our three sets of outcome measures. Table 3 describes the impact of NIH funding for a DST on the number of patents that it produces directly. Panel 1, Column 1 of Table 3 reports the raw correlation between the patents a DST produces and the amount of funding it receives, controlling only for year effects. The coefficient that we estimate, 0.035. This coefficient remains similar as we add increasingly detailed fixed effects. In our most detailed specification in Column 5, with fixed effects for disease–science area, disease–year, broad science area–year, and controls for application quality and lagged funding, our coefficient only drops to 0.033 and remains highly significant. A coefficient of 0.033 indicates that a $10 million increase in DST leads to the direct production of 0.033 additional patents. Given the fact that the average grant costs $1.34 million dollars, this says that we expect to produce one patent directly for every 226 NIH grants.

The results in Panel 1 report our estimates with patents weighted such that, if a patent is linked to N DSTs, each DST receives credit for $1/N$th of that patent. This assumes, essentially, that all DSTs are perfect substitutes for the production of patents. In general, the appropriate weighting will, of course, depend on the innovation production function and the degree to which any particular piece of knowledge is instrumental in generating the patent. If workarounds are straightforward, then the support from any given DST will have only a small effect on obtaining the patent, and thus should be down-weighted appropriately. If, instead, the contributions of various DSTs are complements, then a patent should count for more than $1/N$th; in the extreme, where all pieces are critical such that production is Leontief, every DST should receive full credit for the patents that acknowledge its support.

Panel 2 presents our results under this more generous assumption. This assumption is plausible in this case because our measure of DST-patent linkage is financial support; it is not hard to imagine that losing support from one source can endanger the entire project, especially in a setting where the public sector makes up such a large proportion of total R&D investment. Again, our estimates are very stable even as we move from only year fixed effects to our full set of fixed effects. The coefficient of 0.16 indicates that $10 million in funding leads to 0.16 patents or that, on average, we expect one patent for every 47 grants. The elasticity of NIH-funded patenting with respect to DST funding is similar under both weightings: approximately 0.8 to 1.0. Both weighting strategies indicate that while direct-patent production is relatively elastic with respect to the NIH funding, it accounts for only a small proportion of total patenting.

Table 4 expands the set of patents potentially affected by NIH funding to include those that cite publications supported by public funding. We find larger effects of NIH funding using this measure. Adding fixed effects for research areas (disease–science groupings) reduces this coefficient to 2.1. Interestingly, adding disease-year fixed effects does not change this coefficient much; this is consistent with the possibility that NIH funding does not respond to time-varying disease interests that are correlated with the success of private innovative efforts. With our full set of controls, we estimate that a $10 million increase in funding leads to 2.3 additional patents, or one patent for every three grants. Panel 2 reports these same results under the assumption that ever publication a patent cites is necessary for that patent's production and cannot be substituted with another non-NIH funded patent. The first part of this assumption is plausible; the median patent cites only four publications and 90% of those citations are added by the patent's inventors, as opposed to by a patent examiner (Sampat 2010, Alcacer, Gittleman and Sampat 2008). Those cited publications are likely to have contributed important insights in the development of the patent. Without weighting, we estimate that $10 million leads to 14.9 more patents, or about 2 patents for every NIH grant. NIH funding, then, appears to be much more effective in generating innovation through producing research that private firms can then later build on.

The estimates in Tables 3 and 4, however, may not reflect the true value of NIH funding if public support for science either crowds out private investment or if it spurs patenting in ways that cannot be captured by a direct grant-publication-patent link. Table 5 repeats our estimates using our third outcome measure, the number of patents in the same intellectual space as an NIH funding area. These specifications are meant to assess the net impact of NIH funding on total innovation in an area, accounting for both the possibility of crowd-out and the possibility that not all patents spurred by NIH funding can be linked via direct citations. Column 5 of Table 5 finds that, on net, a \$10 million increase in DST funding increases the number of private sector patents in its area by 2.8; or about one patent for every two to three NIH grants. The magnitude of the impact of NIH funding on total patenting in an area is slightly larger than its effect on patenting that can be directly linked to NIH funds. This indicates that even if NIH funding does crowd out private efforts, it generates more patents through non-direct citation means than it crowds out. This may occur if, for instance, NIH funding increases the productivity of private R&D investments by clarifying the scientific potential of various research areas. In this case, even if firms reduce their investments, total private patenting in an area may still increase.

Panel 2 reports these results with unweighted patent counts and estimates effects that are an order of magnitude larger. These results, however, are unlikely to reflect the true effect of NIH funding. Recall that this final outcome measure is designed to capture the influence that NIH funding may have on patenting that does not require a direct citation linkage between funding and patents. In this measure, patents are linked to study sections through shared intellectual foci, reflecting the notion that public funding in a particular area produces knowledge that enhances productivity of others working in that area. This approach is designed to assign at least one DST to *every* life science patent, meaning that, in practice, most patents are linked to many more, thus driving a large wedge between weighted and unweighted impacts. Unlike the direct approaches which connect patents to a small number of study sections, our indirect method often yields connections to hundreds of study sections in related intellectual realms. While all linkages may be important, it is harder to imagine that each unit of knowledge is instrumental and thus we favor the more conservative weighted approach in this case. Going forward, we will discuss estimates of the effect of funding on overall patent production using only our more conservative weighted counts. The unweighted results, however, are still reported in our tables.

Table 6 reports our estimates using our lucky funding instrument. Columns 1 and 2 report our first stage estimates of the relationship between total DST funding and DST funding coming just from grant applications that are funded because of their study section rankings (as opposed to their absolute quality). This relationship is statistically significant at the one percent level, but short of the conventional F-statistic of 10. Columns 4 and 6 present our preferred specifications with full controls for application quality and number of DST applications near the funding threshold for a given disease-year. Using the IV, we find a stronger effect of NIH funding on the number of directly cited patents (3.3 vs. 2.3) and a similar effect for the total number of patents related to an NIH research area. In sum, our estimates translate into an elasticity of about 0.7 to 1.0 for directly cited patents and 0.4 to 0.5 for net private sector patenting.

5.1 Heterogeneity

In addition to quantifying the impact of NIH funding on overall patenting, we also examine which types of patents rely most on NIH support. Holding constant its impact on total patenting, understanding the characteristics of the marginal patents that NIH funding supports is important for understanding the role that NIH funding plays in biomedical research.

Even though investments in R&D are generally targeted toward specific disease areas, our next set of results show that stories like Gleevec are common in the sense that drugs and patents often build on research originally intended for other diseases. To measure this, we assign a primary disease affiliation to each patent in our data by examining the set of publications that it cites. Using information on the DSTs that these publications are either directly supported by or in the same area as, we construct two different measures of disease affiliation. The first is a binary variable that identifies the disease area (IC) to which the majority of a patent's cited publications are linked. The second is a fractional measure: if one third of the publications that a patent cites are linked to, for instance, the National Cancer Institute, then we record that patent as having a cancer affiliation of one third.

We find that NIH funding directed toward one disease area is equality likely to translate into patenting in other disease areas as it is to translate into patenting in its own area. To see this, compare Columns 1 and 2 of Table 7. The coefficient reported in Column 1 indicates that a $10 million increase in funding for a DST generates 1.02 additional citation-linked patents with the same primary disease affiliation. This is likely the effect that Congress is interested in when allocating funds for particular diseases. Column 2, however, shows that this same funding also generates 1.3 additional patents in other disease areas. Similarly, Columns 3 and 4 show the similar results for net patenting: an additional $10 million dollars in NIH funding leads to 1.6 more patents in the same disease area and 1.15 additional patents in different disease areas.

This result highlights the importance of using a patent-linking strategy that does not assume that funding only impacts innovation in its intended area. Had we done this in our setting, we would have undervalued the returns to NIH R&D investments by almost 50%. Our results in Table 7, moreover, show that the elasticities of both same and other disease area patenting with respect to NIH funding appear identical. This means that, the marginal patent generated by NIH funding is also just as likely to be in another disease area as not. Appendix Table 11 reports these results for our second measure, which allows a patent to have multiple disease affiliations. Using this measure, we find similar effects.

We also examine the impact of NIH funding on the number of highly cited patents, patents associated with FDA approved drugs, and patenting by small or large firms. Columns 1-3 of Table 8 focus on highly cited patents, those which are among the top 5% most cited patents of their three-digit patent class–issue-year cohort. We estimate that $10 million in funding leads to net increase of 0.2 such patents in a DST area. This magnitude of this effect means that, at approximately $1.34 million per grant, one out of every 50 NIH grants can be expected to produce a highly cited patent. This translates into an elasticity of highly cited patents with respect to funding of 0.44, which is similar to the elasticity that we estimate for the production of private sector patents overall (0.48). We also report results for unweighted patent counts with, again, the caveat that this weighting scheme is unlikely to be appropriate for our broadest outcome measure, that of the number of high-value patents in the area of a DST.

Columns 4-6 repeat this exercise for patents associated with FDA-approved drugs. Column 6 indicates that a $10 million increase in funding leads to a net increase of 0.06 patents, meaning that one out of every 125 grants produces a patent that is linked with new pharmaceuticals. This translates into an elasticity of patenting with respect to NIH funding of 0.42, again, similar to the overall patent population. Table 9 asks whether small firms (those with fewer than 500 employees) or large firms are more responsive to NIH funding and finds no difference.

6 Alternative Explanations and Robustness Checks

The estimates in Section 5 identify the causal impact of NIH funding under the assumption that NIH funding for a DST does not respond to changes in the specific innovative potential of a disease–science area combination. There are several stories that would violate this assumption.

Congress may, for instance, allocate funding to NIH Institutes based on changes in the productivity of scientific methods relevant to their disease areas. Advances in genetics, for instance, may lead Congress to allocate relatively more funding to the National Cancer Institute than to the National Institute on Allergy and Infectious Disease. If this were the case, then differences in DST finding within a disease-year might reflect differences in scientific potential, thus violating our identifying assumption. Differences in the budgets for cancer-genetics vs. cancer-tumor-physiology, for instance, might reflect the fact that the NCI is responding to increased promise in genetics research.

Even though our IV strategy is designed to address this concern, there are several reasons to believe that this does not occur. The first is that, in practice, funding determinations for NIH Institutes tend to be justified on the basis of disease-level concerns: how great is the burden of disease associated with conditions that fall under an IC's purview and what are the trends associated with those conditions?[16] Appendix Figure A provides an example of language from an appropriations bill for the National Cancer Institute; here, Congress uses the disease burden associated with pancreatic cancer to underscore the need for more research in this field. Appendix Figure A also attempts to formalize this example by compiling a list of the mostly commonly used words in the Congressional appropriations documents for all NIH Institutes, for a sample year. The highest-frequency word in both House and Senate appropriations is, unsurprisingly, "research." The majority of the remaining list are medicine or disease focused: "disease," "health," "child," "behavior," "patients," "syndrome," etc. This reasoning is supported by research showing that NIH funding is more highly correlated with disease burden and public demand than with scientific advances (Gillum et al., 2011). These motivations, meanwhile, do not present a problem for our

[16] In practice, many critiques of the NIH are that ICs do not respond to even that–rather, IC funding may also be justified on the basis of politics or path dependencies.

identification because we include disease by year fixed effects.

Another way NIH may be able to direct funding toward areas with more potential is by funding grants out of the order in which they are scored. Approximately four to five percent of grants are funded as exceptions. While this usually occurs in response to the emergence of new data to strengthen the application, grants are also sometimes funded out of order if they were evaluated in an exceptionally strong committee and received a lower relative score than their absolute quality should indicate.[17]

We show that this possibility does not affect our results in two ways. First, if NIH Institutes do selectively fund grant applications from competitive, high-interest science areas out of order, then we would expect that the amount of funding for DSTs that share the same scientific interests should be correlated; that is, if the NCI (cancer) were allocating more money to genetics because of increased potential in that area, then we should weakly expect the NIDDK (diabetes) to do the same. Similarly, if Congress increased funding for all Institutes whose disease focus has a strong hereditary component, we would also expect cancer-genetics and heart disease-genetics funding to be positively correlated. Appendix Table B examines the correlation between own-disease funding for a science area, Funding_{dst}, and funding for that same science area from other diseases $\text{Funding}_{-d,st}$. Column 1, which includes only year fixed effects, shows a strong negative correlation between own and other funding. This, however, is likely due to the mechanical relationship between the size of one's own disease area in a given science area, and the size of other disease areas. Column 2 controls for this confounder by introducing disease by year fixed effects; we find no correlation between own and other disease funding. This is also true if we add disease by year fixed effects as we do in Column 3. Column 3 includes the same set of controls as we use in estimating our main results. Columns 4 through 6 repeat this exercise using the proportion of a disease area's funding devoted to a particular science area as the variable of interest. This asks: if the NCI begins spending a greater proportion of its budget on genetics, does it appear that other disease areas do the same? Again, we find that this does not appear to be the case.

Another way to address the possibility that out-of-order scoring matters is to instrument for DST funding using funding from grants that are not funded out of order. Ideally, we would add up requested funding amounts for the top ranked applications, regardless of whether they were

[17] Authors' conversation with Stefano Bertuzzi, NIH Center for Scientific Review.

actually funded, but we do not have data on funding requests for unfunded applications. Instead, we count funding amounts for the subset of DST grants that are funded in order. Appendix Table C presents our findings using this alternative strategy. Columns 1 and 2 indicate that we have a strong first stage and, using this instrument, we find that an additional $10 million in ordered funding increases net patenting by 3.7, compared with 2.8 in our main OLS specification and 2.9 in our preferred IV specification.[18] The implied elasticities of all these estimates are similar.

Our next test checks the plausibility of the exclusion restriction on our instrument. Specifically, Appendix Table D tests whether, after controlling for our primary set of regressors, our instrument for funding is correlated with any measures of lagged application quality or lagged patent output. Column 1 reports the F-test of the joint significance of 30 variables describing ten years of lagged patent output (and indicators if a lagged value is not available) and fails to reject a hypothesis of no effect. Column 2 does the same for 20 variables describing ten year lags of average raw scores of applicants to a DST. Again, we fail to reject a hypothesis of no effect.

Another potential concern is that our patent-matching strategy may lead us to link more patents to DSTs when those DSTs receive more funding. This is because we associate patents to a DST if they cite publications related to those funded by DST grants. If DST funding increases the number of directly-cited publications, then this may make it more likely that a given patent is linked to that DST. If this is the case, any increase in the number of patents associated with that DST will be partially driven by our matching strategy, as opposed to a true effect of NIH funding. Using weighted counts of patents partially alleviates this concern by ensuring that a given patent can be counted for at most one weighted patent across all DSTs. If matching leads more patents to be linked to more DSTs, each patent will count for less at each DST.

Nonetheless, it is still possible that the number of citation-linked patents directly influences the number of patents associated with a DST's area. To check that our results are not driven by this effect, Appendix Table E examines the impact of NIH funding on the total number of patents in the intellectual vicinity of an NIH research area, controlling directly for the number of citation-linked patents. Holding constant the number of citation-linked patents, we still find a positive and significant effect of NIH funding: a $10 million increase in funding leads to 1.8 more weighted patents, for an estimated elasticity of 0.05. This is about a third smaller than our main estimates,

[18] Note that our original lucky funding instrument already purges funding dollars to out of order grants.

but our qualitative results continue to hold.

7 Assessing Reallocation

So far, our results have examined the impact of NIH funding on patenting within the same intellectual area. Yet in the cases of both crowd in and crowd out, the additional resources that a firm devotes to—or diverts from—a DST must come from somewhere else in its budget. One possibility is that these resources come from either an expansion in the firm's total R&D budget (in the case of crowd-in) or a contraction in the firm's R&D budget (in the case of crowd-out). In this case, the main effect of NIH funding in an area on investments in that same area is the effect on firm R&D.

Another possibility, however, is that firms respond to public investments by reallocating resources to and from other parts of their R&D portfolio. In this case, one needs to know the consequences of these investments in other areas in order to assess the full impact of NIH funding on private innovation. In terms of reallocation, there are two possibilities.

1. Reallocated crowd-in: Here, firms respond to increased public funding for an NIH-funded research area by reallocating funds from other parts of its research portfolio. In this case, the effect of NIH funding in a research area on private innovation is two-fold: the direct effect of NIH funding is to increase private innovation in the same area and the countervailing reallocation effect is to decrease private innovation in the areas that a firm diverts resources from.

2. Reallocated crowd-out: Similarly, firms may divert resources away from the NIH-funded area toward other research projects. Again, the overall effect of NIH funding on total private innovation will be twofold: the direct effect of public funding is to reduce private innovation within the same intellectual area, but the reallocation effect is to increase private innovation in areas to which funds are reallocated.

We attempt to directly measure the extent of firm reallocation in response to NIH funding. First, we note that our final outcome measure, that of the number of patents in the same area as a DST, is already likely to take into account some of the impact of reallocation. This is because our patent linking approach defines the area of a DST quite broadly. If the NIH increases spending on, for instance, cancer (D) cell signaling (S) research in 1990 (T), we measure net impact of this change on total innovation in *all* parts of the firm's R&D portfolio that are related to cancer and cell signaling research from 1990. This may include patents related to cell signaling in other disease areas, cancer patents unrelated to cell signaling, or any other set of projects similar to research that is supported by the DST. Firm relocation within this set of research would already be captured in our results from Table 5.

Firms, however, may also choose to reallocate funds to or from projects that are completely unrelated to a DST's research. If NIH funding in one DST leads firms to reallocate funds away from that DST, then we should observe an increase in non-DST patenting within that firm. If, instead, NIH investments in a DST lead firms to reallocate funding away from other projects toward the area of NIH investment, then we should observe a decrease in non-DST patenting within that firm.

To measure the extent of reallocation, we would ideally like to focus on the set of firms that actually faced a decision about whether to invest more or less in a DST as a result of NIH funding. In the absence of these data, we focus on firms that actively patent in a DST area and construct a measure of the number of non d, non s patents that they produce in the same year. We have two final variables of interest. Total Patents$_{-d,-s,t}$ measures the total number of non d, non s patents that are produced by firms that also produce a DST-linked patent in the same year. Average Patents$_{-d,-s,t}$ measures the average number of non d, non s patents a firm produces for every DST-linked patent it produces, averaged over all firms in that DST.

The advantage of this approach is that we restrict our analysis to firms that are definitely affected by changes in DST funding. If these firms spend more resources in another area, it is likely that these funds could have also been spent on DST research. The downside of this approach, however, is that it limits the kinds of reallocation we can study. If NIH funding pulls firm resources toward a DST and away from other areas, we will observe this to the extent that it leads to DST patenting. If, however, NIH funding pulls firm resources away from a DST, we will only observe this if that firm still produces at least one DST-linked patent. If NIH funding for a DST leads a firm to reallocate toward other areas entirely, then we will not be able to observe this. We think of our measures of reallocation, then, as an estimate of the extent of reallocation on the intensive margin, conditional on firms not switching away entirely.

Our results show that firms do reallocate resources in response to changes in NIH-funding. Table 10 shows that, in general, an increase in NIH-funding for one area of a firm's R&D portfolio increases the number of patents that those firms develop in other areas of its portfolio. This is consistent with crowd-out: more public investment in one area frees firms up to invest in other areas. Our estimate in Column 1 indicates that a $10 million increase in DST funding leads to the production of 10.6 additional patents in other areas. NIH funding for a DST also increases the average number of non-DST patents we identify, consistent again with reallocated crowd-out. Our results in Tables 4 and 5, however, tell us that overall patenting in the area of the funded DST also increases. In order for both these finding to be true, it must either be that NIH funding in a DST increases by more than the amount that firms divert to other areas or that NIH funding increases a firm's total R&D investment. Another interpretation for this finding is that there are more direct spillovers from NIH funding for a DST than we capture through our outcome measures. If, for instance, firms respond to increased NIH funding by expanding their scientific labor force, and these scientists work on a variety of projects, then an increase in NIH funding for one DST can impact other patenting areas in ways our outcome measures cannot observe.

The elasticities we estimate under all of these specifications are smaller than the ones we estimate for the direct effect of DST funding on patenting in the same area. This is smaller magnitude makes sense theoretically. In the case of reallocated crowd-in, the patents that are lost in the area from which the firm diverts funds should be fewer than the number that are gained, as long the firm is reallocating optimally. Similarly, in the case of reallocated crowd-out, the patents that are gained in the area to which firms divert funds should be fewer than the number that are lost in the original area, as long as firms had initially allocated their investments optimally.

We should note that while our results indicate NIH funding increases total private R&D, assessing the welfare implications of NIH funding is beyond the scope of this paper. In particular, this is because neither crowd-in nor crowd-out translate easily into welfare assessments. While crowd-in is generally thought of as welfare enhancing, at least from the perspective of increasing innovation it need not be: firms might divert funds from worthy non R&D purposes or crowd-in might encourage too many firms to enter a research area. The logic that the direct effect of NIH funding on same-area private research dominates reallocation effects applies only within one firm; if multiple firms divert funds from other, different R&D projects in order to pursue similar research in the area of NIH funding, this could result in business stealing and a decline in overall innovation. Similarly, to the extent that there is a shadow cost of public funds, crowd-out is generally thought of as welfare-reducing. This may not be true, however, if NIH funding improves the efficiency of private R&D allocation by showing that a particular area has little innovative potential.

8 Conclusion

This paper generates a causal estimate of the impact of public investments in biomedical research on subsequent private sector patenting. Our results show that NIH investments in an area increase subsequent private sector patenting in that area; a $10 million increase in funding for an area leads to 2.8 additional patents or, equivalently, we expect one private sector patent generated for every two to three NIH-funded grants. This positive effect, meanwhile, does not appear to be associated with lower private investments in other research areas. We also demonstrate that investments in basic research are difficult to target. In our sample, the returns to NIH funding are just as often felt outside of the disease area for which they were intended.

How large are the effects we find? Ideally, one would like an estimate of the return to public R&D investments in monetary terms. Coming up with this figure, however, requires making strong assumptions about the distribution of patent quality, the probability that patents translate into drugs, and the market or social value of those drugs. There is currently little agreement in the literature about those parameters. Instead, we prefer to interpret the returns to NIH funding in terms of a lottery. The distribution of patent quality is highly skewed and, in practice, the majority of life science patents have little value because they are often associated with molecular compounds that are later shown to be ineffective. The returns to NIH funding, then, might be more effectively thought of not in terms of the realized value of the patents it generates, but in terms of its ex ante chance of generating a blockbuster drug or treatment.

The probability that a patent will result in a breakthrough, of course, depends on the kind of research that firms choose to engage in. Our results focus, for the most part, on how NIH funding impacts the volume of private R&D investments and outcomes. Equally important is the impact of public funding on the types of projects that they pursue. Our finding that the marginal patent generated by NIH funding is less likely to be commercially successful, for instance, is consistent with the view that NIH funding lowers the risks associated with exploratory work. It would also, of course, be consistent with the view that the public sector funds lower-quality research. Distinguishing between these cases and, more generally, investigating how firms choose among potential R&D investments in light of public funding is an important area for future work.

References

[1] Acemoglu, Daron, and Joshua Linn. 2004. "Market Size in Innovation: Theory and Evidence from the Pharmaceutical Industry." Quarterly Journal of Economics 119(3): 1049-1090.

[2] Adams, James D. 1990. "Fundamental Stocks of Knowledge and Productivity Growth." Journal of Political Economy 98(4): 673-702.

[3] Alccer, Juan, and Michelle Gittelman. 2006. "Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations." Review of Economics and Statistics 88(4): 774-779.

[4] Alccer, Juan, Michelle Gittelman, and Bhaven Sampat. 2009. "Applicant and Examiner Citations in U.S. Patents: An Overview and Analysis." Research Policy 38(2): 415-427.

[5] Bachrach, C. A., and Thelma Charen. 1978. "Selection of MEDLINE Contents, the Development of its Thesaurus, and the Indexing Process." Medical Informatics (London) 3(3): 237-254.

[6] Blume-Kohout, Margaret E. 2012. "Does Targeted, Disease-Specific Public Research Funding Influence Pharmaceutical Innovation?" Journal of Policy Analysis and Management 31(3): 641-660.

[7] Cech, Thomas R. 2005. "Fostering Innovation and Discovery in Biomedical Research." Journal of the American Medical Association 294(11): 1390-1393.

[8] Congressional Budget Office. 2006. Research and Development in the Pharmaceuticals Industry (CBO Publication No. 2589). Washington, DC: U.S. Government Printing Office. Retrieved from http://www.cbo.gov/sites/default/files/cbofiles/ftpdocs/76xx/doc7615/10-02-drugrd.pdf

[9] Cockburn, Iain M., and Rebecca M. Henderson. 1998. "Absorptive Capacity, Coauthoring Behavior, and the Organization of Research in Drug Discovery." Journal of Industrial Economics 46(2): 157-182.

[10] Cockburn, Iain M., and Rebecca M. Henderson. 2000. "Publicly Funded Science and the Productivity of the Pharmaceutical Industry." Innovation Policy and the Economy 1: 1-34.

[11] 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)." NBER Working Paper #7552.

[12] Comroe, Julius H., and Robert D. Dripps. 1976. "Scientific Basis for the Support of Biomedical Science." Science 192(4235): 105-111.

[13] Connolly, Laura S. 1997. "Does External Funding of Academic Research Crowd Out Institutional Support?" Journal of Public Economics 64(3): 389-406.

[14] David, Paul, Mowery David, and W. Edward Steinmueller. 1992. "Assessing the Economic Payoff from Basic Research." Economics of Innovation and New Technology 2(1): 73-90.

[15] David, Paul A., Bronwyn H. Hall, and Andrew A. Toole. 2000. "Is Public R&D a Complement or Substitute for Private R&D? A Review of the Econometric Evidence." Research Policy 29(4-5): 497-529.

[16] Dorsey, E. Ray, Jason de Roulet, Joel P. Thompson, Jason I. Reminick, Ashley Thai, Zachary White-Stellato, Christopher A. Beck, Benjamin P. George, and Hamilton Moses III. 2013. "Funding of US Biomedical Research, 2003-2008." JAMA 303(2): 137-143.

[17] Druker, Brian J., Moshe Talpaz, Debra J. Resta, Bin Peng, Elisabeth Buchdunger, John M. Ford, Nicholas B. Lydon, Hagop Kantarjian, Renaud Capdeville, Sayuri Ohno-Jones, and Charles L. Sawyers. 2001. "Efficacy and Safety of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in Chronic Myeloid Leukemia." New England Journal of Medicine 344(14): 1031-1037.

[18] Druker, Brian J., Shu Tamura, Elisabeth Buchdunger, Sayuri Ohno, Gerald M. Segal, Shane Fanning, Jrg Zimmermann, and Nicholas B. Lydon. 1996. "Efficacy and Safety of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in Chronic Myeloid Leukemia." Nature Medicine 2(5): 561-566.

[19] Garber, Alan M., and Paul M. Romer. 1996. "Evaluating the Federal Role in Financing Health-Related Research." Proceedings of the National Academy of Sciences 93(23): 12717-12724.

[20] Gelijns, Annetine C., Nathan Rosenberg, and Alan J. Moskowitz. 1998. "Capturing the Unexpected Benefits of Medical Research." New England Journal of Medicine 339(10): 693-698.

[21] Gillum, Leslie A., Christopher Gouveia, E. Ray Dorsey, Mark Pletcher, Colin D. Mathers, Charles E. McCulloch, and S. Claiborne Johnston. 2011. "NIH Disease Funding Levels and Burden of Disease." PLoS ONE 6(2): e16837.

[22] Hall, Bronwyn H., Adam Jaffe, and Manuel Trajtenberg. 2005. "Market Value and Patent Citations." The RAND Journal of Economics 36(1): 16-38.

[23] Henderson, Rebecca. 1994. "The Evolution of Integrative Capability: Innovation in Cardiovascular Drug Discovery." Industrial & Coporate Change 3(3): 607-630.

[24] Henderson, Rebecca, Luigi Orsenigo, and Gary P. Pisano. 1999. "The Pharmaceutical Industry and the Revolution in Molecular Biology: Interactions Among Scientific, Institutional, and Organizational Change." In David C. Mowery, and Richard R. Nelson (Eds.), Sources of Industrial Leadership, pp. 267-311. New York: Cambridge University Press.

[25] Jaffe, Adam B. 1986. "Technological Opportunity and Spillovers from R&D: Evidence from Firms Patents, Profits, and Market Value." American Economic Review 76(5): 984-1001.

[26] Lemley, Mark A., and Bhaven Sampat. 2012. "Examiner Experience and Patent Office Outcomes." Review of Economics and Statistics 94(3): 817-827.

[27] Lichtenberg, Frank R. 2001. "The Allocation of Publicly Funded Biomedical Research." Chapter 15 in Ernst Berndt, and David Cutler (Eds.), Medical Care Output and Productivity, pp. 565-589. Washington, DC: University of Chicago Press.

[28] Manton, Kenneth G., Xi-Liang Gu, Gene Lowrimore, Arthur Ullian, and H. Dennis Tolley. 2009. "NIH Funding Trajectories and their Correlations with US Health Dynamics from 1950 to 2004." Proceedings of the National Academy of Sciences 106(27): 10981-10986.

[29] Mukherjee, Siddhartha. 2010. The Emperor of All Maladies: A Biography of Cancer. New York: Scribner.

[30] Murphy, Kevin M., and Robert H. Topel. 2003. "The Economic Value of Medical Research." In Kevin M. Murphy, and Robert H. Topel (Eds.), Measuring the Gains from Medical Research: An Economic Approach, pp. 41-73. Chicago: University of Chicago.

[31] Murray, Fiona. 2010. "The Oncomouse that Roared: Hybrid Exchange Strategies as a Source of Productive Tension at the Boundary of Overlapping Institutions." American Journal of Sociology 116(2): 341-388.

[32] Narin, Francis, and Dominic Olivastro. 1992. "Status Report: Linkage between Technology and Science." Research Policy 21(3): 237-249.

[33] National Institute for Health Care Management Research and Educational Foundation. 2006. Changing Patterns of Pharmaceutical Innovation. Washington, DC: NIHCM Foundation. Retrieved from http://nihcm.org/pdf/innovations.pdf Nvol, Aurlie, Rezarta Islamaj Dogan, and Zhiyong Lu. 2010. "Author Keywords in Biomedical Journal Articles." AMIA Symposium Proceedings 537-541.

[34] Nowell, Peter C., and David A. Hungerford. 1960. "A Minute Chromosome in Human Chronic Granulocytic Leukemia." Science 132(3438): 1497.

[35] Pray, Leslie A. 2008. "Gleevec: the Breakthrough in Cancer Treatment." Nature Education 1(1).

[36] Roach, Michael, and Wesley M. Cohen. 2013. "Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research." Management Science 59(2): 504-525.

[37] Sampat, Bhaven N. 2012. "Mission-oriented Biomedical Research at the NIH." Research Policy 41(10): 1729-1741.

[38] Sampat, Bhaven N., Kristen Buterbaugh, and Marcel Perl. 2013. "New Evidence on the Allocation of NIH Funds across Diseases." Milbank Quarterly 91(1): 163-185.

[39] Sampat, Bhaven N., and Frank R. Lichtenberg. 2011. "What Are the Respective Roles of the Public and Private Sectors in Pharmaceutical Innovation?" Health Affairs 30(2): 332-339.

[40] Stevens, Ashley J., Jonathan J. Jensen, Katrine Wyller, Patrick C. Kilgore, Sabarni Chatterjee, and Mark L. Rohrbaugh. 2011. "The Role of Public-Sector Research in the Discovery of Drugs and Vaccines." New England Journal of Medicine 364(6): 535-541.

[41] Szalinski, Christina. 2013. "Why Pharma Needs the NIH." ACSB Newsletter 36(7): 13.

[42] Toole, Andrew. 2007. "Does Public Scientific Research Complement Private Investment in R&D in the Pharmaceutical Industry?" Journal of Law & Economics 50(1): 81-104.

[43] Toole, Andrew A. 2012. "The Impact of Public Basic Research on Industrial Innovation: Evidence from the Pharmaceutical Industry." Research Policy 41(1): 1-12.

[44] Trajtenberg, Manuel, Rebecca M. Henderson, and Adam B. Jaffe. 1997. "University vs. Corporate Patents: A Window on the Basicness of Innovations." Economics of Innovation and New Technology 5(1): 19-50.

[45] Varmus, Harold. 2009. The Art and Politics of Science: W. W. Norton & Company.

[46] Wurtman, Robert J., and Robert L. Bettiker. 1995. "The Slowing of Treatment Discovery, 1965-1995." Nature Medicine 1(11): 1122-1125.

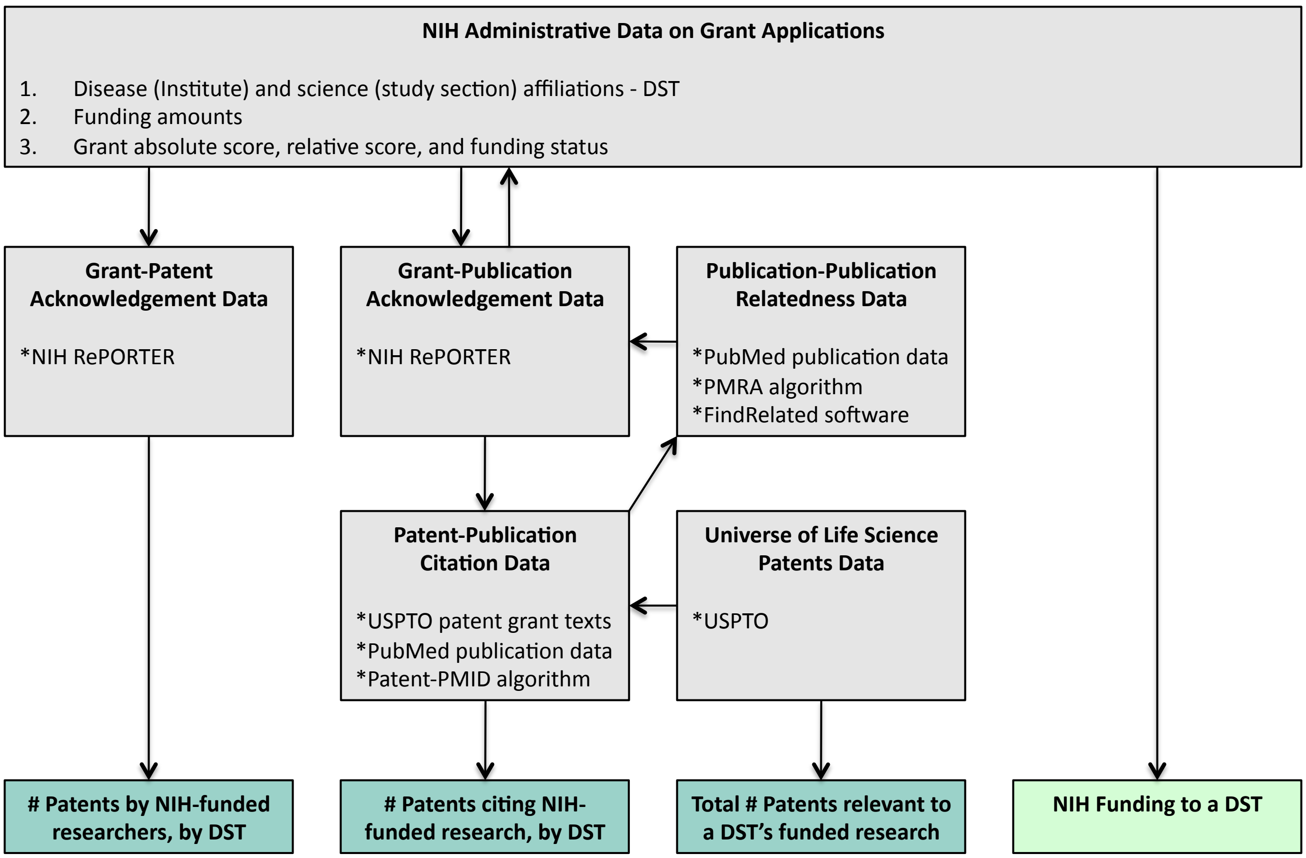


Figure 1: Overview of Data and Construction of Patent Outcome Measures

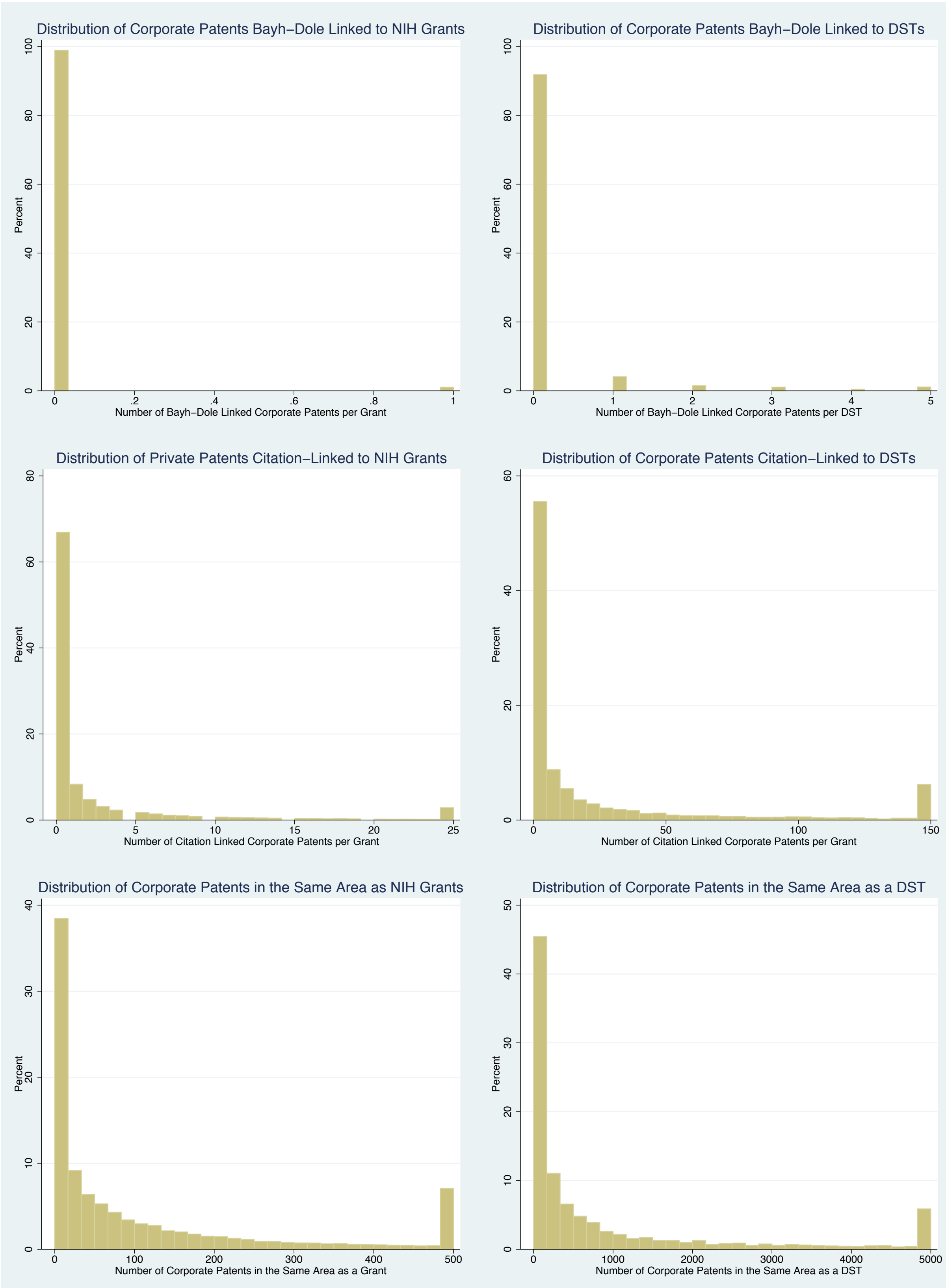


Figure 2: Outcome Measures by Grant and DST (Patent counts are unweighted)

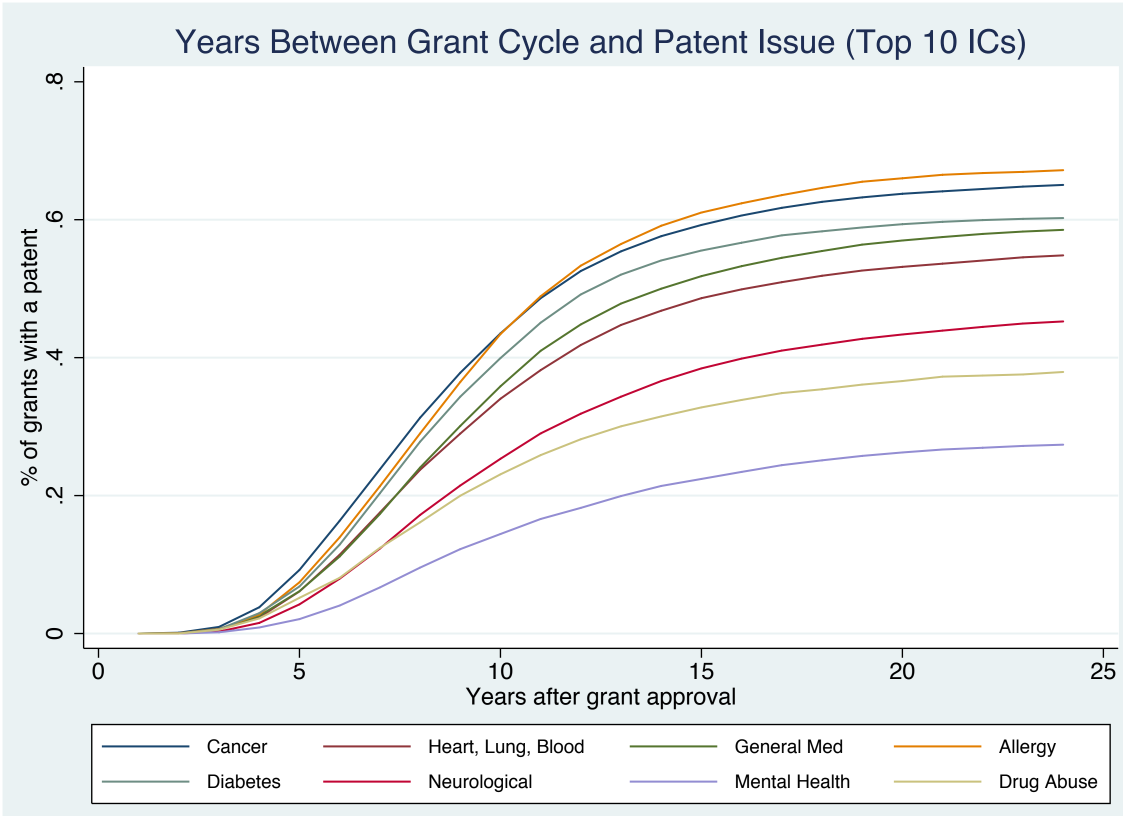


Figure 3: Grant-Patent Lags by Disease Area

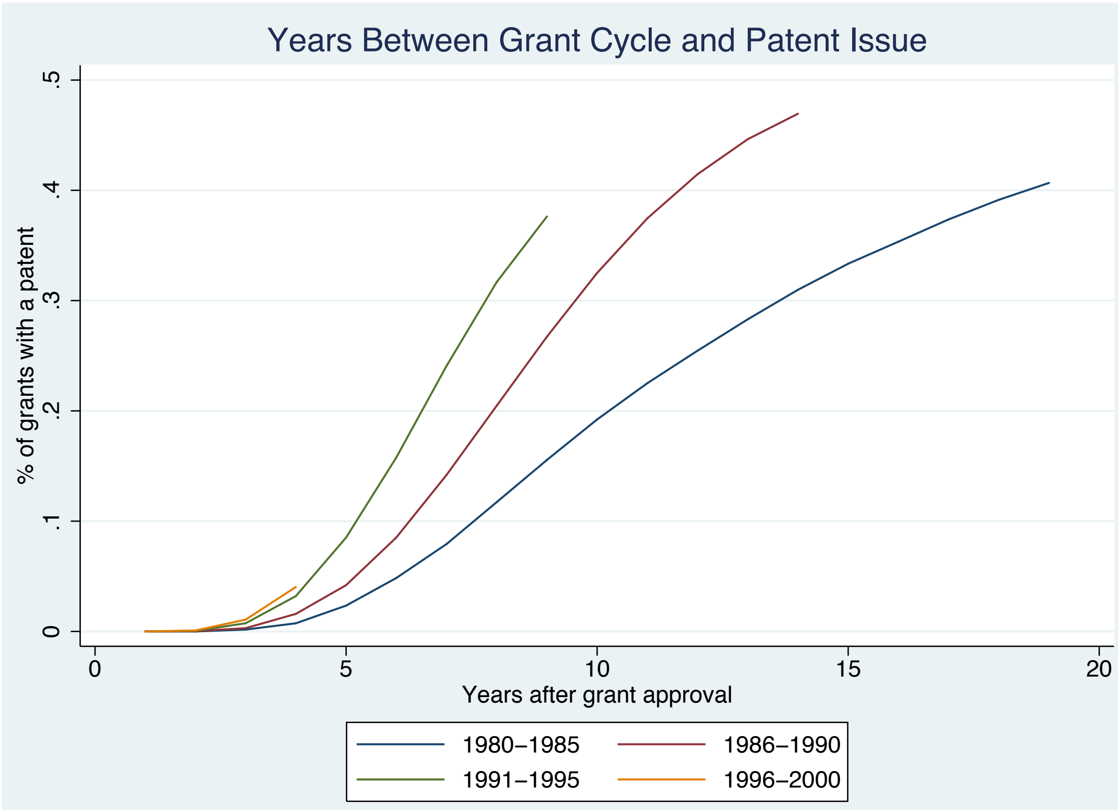


Figure 4: Grant-Patent Lags by Cohort

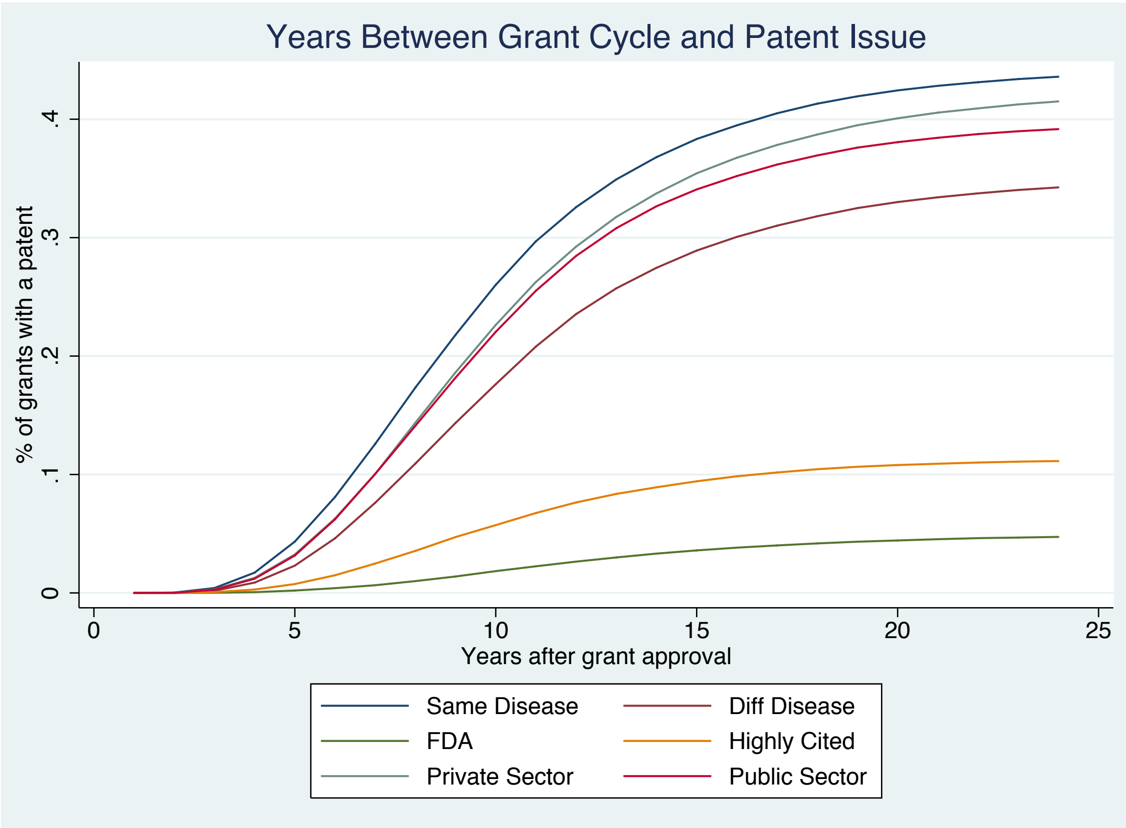
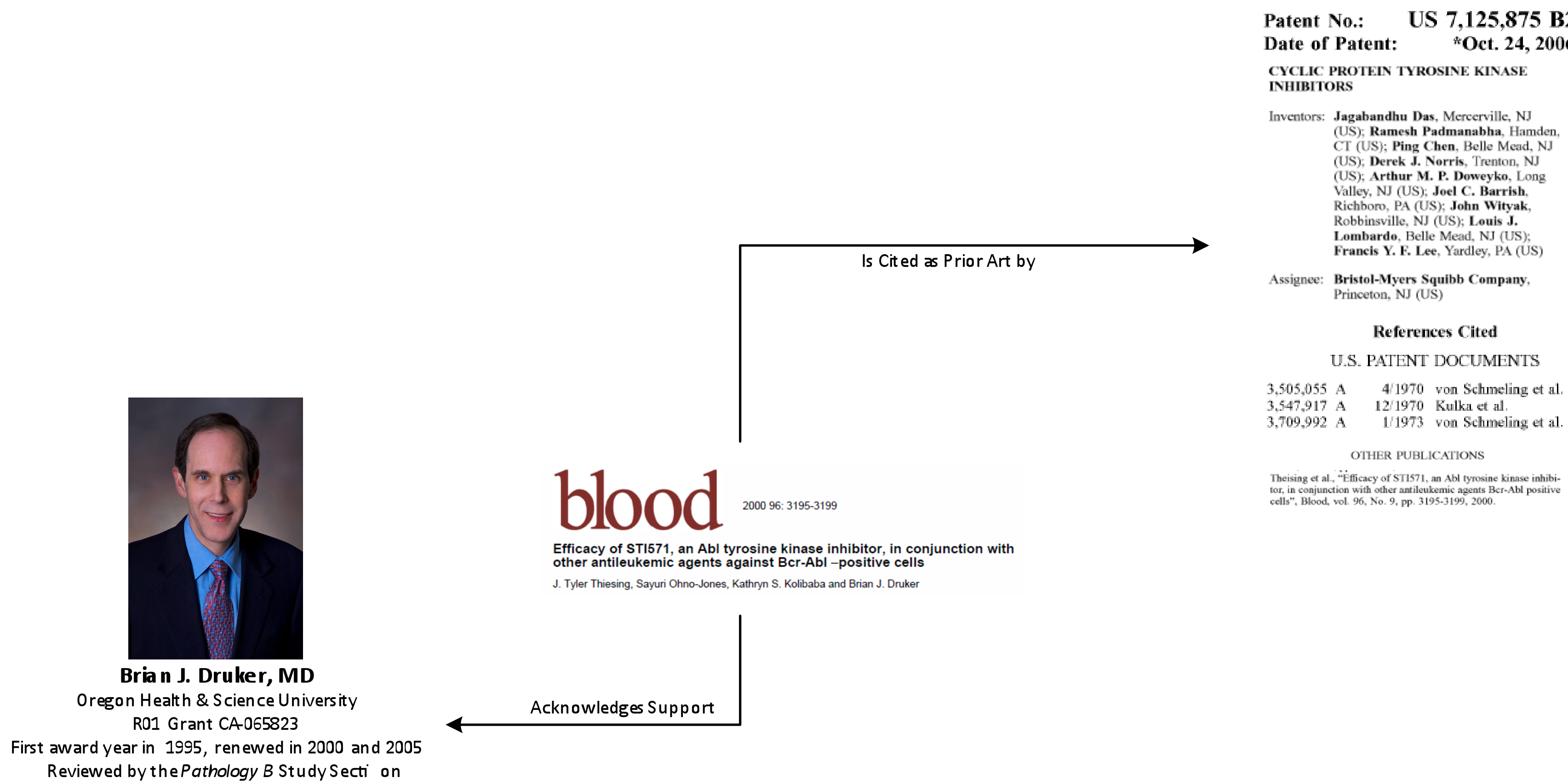


Figure 5: Grant-Patent Lags by Patent Type



Note: The grant CA-065823 in its first cycle acknowledges 4 publications indexed in PubMed, among which is the article published by Thiesing et al. in the leading Hematology journal *Blood*. In turn, this article is listed as prior art in the 7,125,875 patent issued in 2006 to the pharmaceutical firm Bristol Myers Squibb. In this fiscal year, the Pathology B study section evaluated 66 proposals that were eventually funded, 63 of them by the National Cancer Institute (the same institute that funded Druker). Two of the remaining three proposals were funded by the National Institute of Aging (NIA), and the last was funded by the National Eye Institute. These three grants are acknowledged by 15 publications in PubMed, which are themselves cited by 11 distinct patents in the USPTO database.

Figure 6: Example of Citation-Linked Patent Matching

Note: The grant CA-001422 is acknowledged by 10 publications, among which is the article by Carroll et al. in the journal *Blood*. In turn, this article is listed as prior art in the patent 7,232,842 issued in 2007 to Stanford University. In addition to this direct bibliometric linkage (cf. Figure 4A), we focus on indirect linkages by matching the Carroll et al. publication with its intellectual neighbors through the use of the PubMed Related Citation Algorithm [PMRA]. As can be seen above, the fifth most related publication was published in the journal *Cancer Research* in 1996. We focus on this publication because it is cited as prior art by the patent 6,894,051 issued to Novartis in May 2005. This patent is valuable indeed, since it is listed in the FDA Orange Book as one of the five patents associated with the registration of Imatinib Mesylate, better known by its brand name, *Gleevec*. These indirect bibliometric linkages are valuable to us because they enable us to link the great majority of patents in biopharmaceutical classes to a study section × institute × year strata. In other words, most patents can be traced back to one (or more) NIH grant, because most patents cite publications as prior art that are related in ideas space to another publication which acknowledges NIH funding.

FIGURE 7: EXAMPLE OF SAME INTELLECTUAL AREA PATENT MATCHING (USING PMRA)

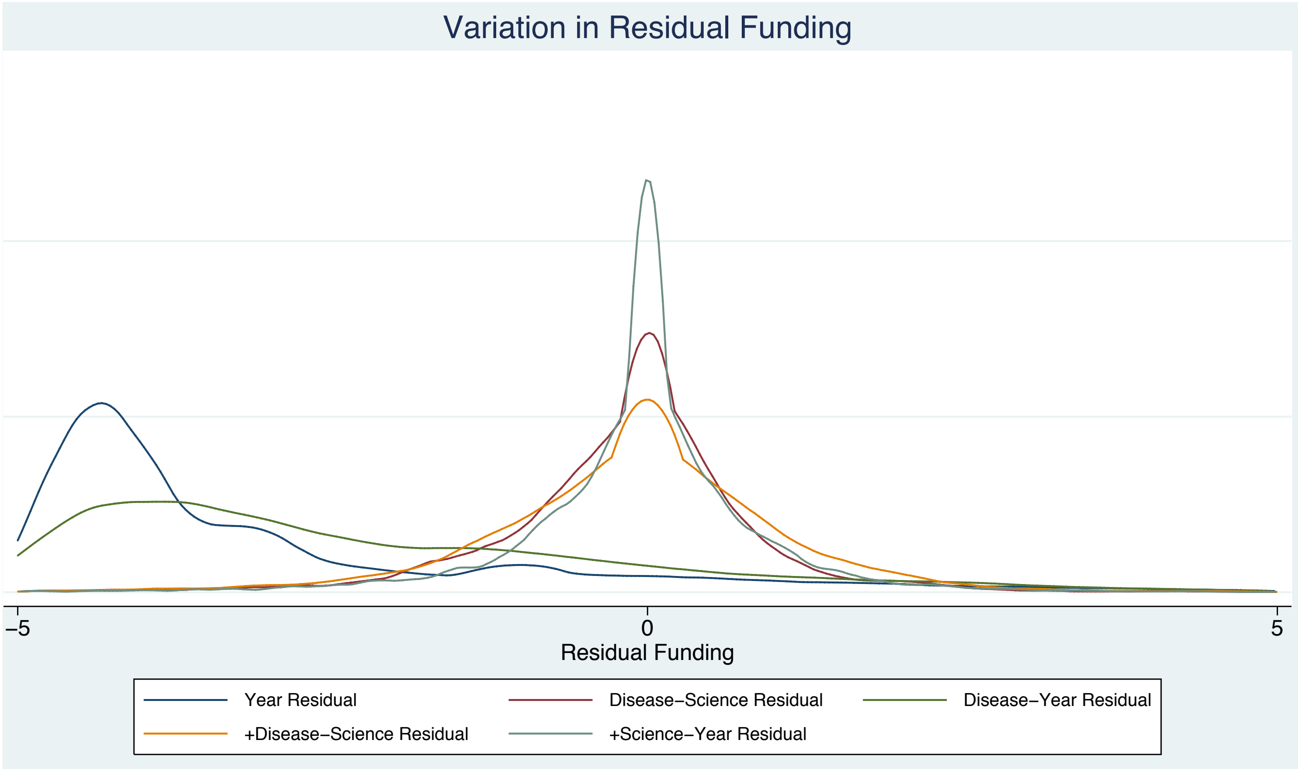


Figure 8: Identifying Variation in DST Funding

Panel 1: Initial Study Section Scores and Rankings

Cell Signaling Study Section				Tumor Physiology Study Section			
Priority	**Rank**	**Disease**	**Raw Score**	**Priority**	**Rank**	**Disease**	**Raw Score**
1	1	Cancer	10	2	1	Cancer	8.2
1	2	Diabetes	9.8	3	2	Cancer	8.1
4	3	Cancer	9.2	5	3	Cancer	7.6
6	4	Cancer	9.1	7	4	Cancer	6.4
8	5	Cancer	8.3	9	5	Cancer	5.4
2	6	Diabetes	7.6	3	6	Diabetes	5.2
10	7	Cancer	7.6	4	7	Diabetes	4.8
5	8	Diabetes	7.5	6	8	Diabetes	4.4

Panel 2: Institute Funding Outcomes

Cancer Institute (NCI)				Diabetes Institute (NIDDK)			
Priority	**Rank**	**Study Section**	**Raw Score**	**Priority**	**Rank**	**Study Section**	**Raw Score**
1	1	Cell	10	1	2	Cell	9.8
2	1	Tumor	8.2	2	6	Cell	7.6
3	2	Tumor	8.1	3	6	Tumor	5.2
4	3	Cell	9.2	4	7	Tumor	4.8
5	3	Tumor	7.6	~~5~~	~~8~~	~~Cell~~	~~7.5~~
~~6~~	~~4~~	~~Cell~~	~~9.1~~	~~6~~	~~8~~	~~Tumor~~	~~4.4~~
~~7~~	~~4~~	~~Tumor~~	~~6.4~~				
~~8~~	~~5~~	~~Cell~~	~~8.3~~				
~~9~~	~~5~~	~~Tumor~~	~~5.4~~				
~~10~~	~~7~~	~~Cell~~	~~7.6~~				

Panel 3: Study Section Funding Outcomes

Cell Signaling Study Section				Tumor Physiology Study Section			
Priority	**Rank**	**Disease**	**Raw Score**	**Priority**	**Rank**	**Disease**	**Raw Score**
1	1	Cancer	10	2	1	Cancer	8.2
1	2	Diabetes	9.8	3	2	Cancer	8.1
4	3	Cancer	9.2	5	3	Cancer	7.6
~~6~~	~~4~~	~~Cancer~~	~~9.1~~	~~7~~	~~4~~	~~Cancer~~	~~6.4~~
~~8~~	~~5~~	~~Cancer~~	~~8.3~~	~~9~~	~~5~~	~~Cancer~~	~~5.4~~
2	6	Diabetes	7.6	3	6	Diabetes	5.2
~~10~~	~~7~~	~~Cancer~~	~~7.6~~	4	7	Diabetes	4.8
~~5~~	~~8~~	~~Diabetes~~	~~7.5~~	~~6~~	~~8~~	~~Diabetes~~	~~4.4~~

Figure 9: Example of variation in funding unrelated to quality

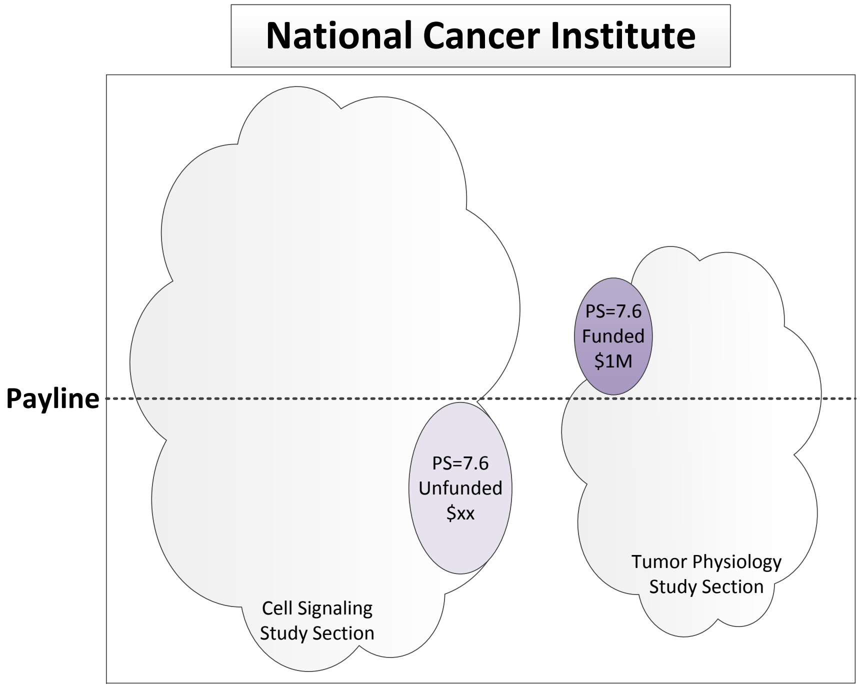


Figure 10: Analogy to a regression discontinuity

Table 1: Grant Characteristics, 1980-2000

		Grants Linked to Private Sector Patents		
	Full Sample	**NIH-funded**	**Citation**	**PMRA**
Sample Coverage				
# Grants (Type 1 and 2 only)	123,478	1,283	41,369	96,557
# Disease Areas (Institutes/Centers)	16	16	16	16
# Science Areas (Study Sections)	443	179	371	416
# DSTs	11,110	984	7,133	10,172
Grant Characteristics				
% R01 equivalent	68.09	79.50	77.70	75.42
% Center grants	3.26	11.46	5.64	3.44
% Teaching grants	15.58	5.07	10.37	12.26
% New (Type 1)	63.40	30.79	48.30	57.12
Funding amount (Total project allocation, 2010 dollars)	\$1,344,661	\$3,078,722	\$1,958,900	\$1,527,456
	(1,972,101)	*(4,617,934)*	*(2,881,154)*	*(2,132,951)*
Publication Match				
% with at least one matched publication	81.49	100.00	100.00	100.00
# of matched publications	6.90	20.47	13.18	8.75
	(12.70)	*(35.26)*	*(19.33)*	*(13.80)*
Patent Match				
# of private sector patents by NIH-funded researchers (weighted sum)	0.01	0.55	0.01	0.01
	(0.10)	*(0.86)*	*(0.17)*	*(118.00)*
# of citation-linked private sector patents (weighted sum)	0.43	2.53	1.27	0.55
	(2.19)	*(5.32)*	*(3.64)*	*(2.46)*
# of PRMA linked private sector patents (weighted sum)	0.82	2.95	-1.80	1.04
	(2.00)	*(4.93)*	*(3.05)*	*(2.20)*

Notes: Sample is the set of all NIH-funded grants from 1980-2000, with the restriction that these grants be funded by disease or body systems focused Institutes (see text for a full list) and evaluated by chartered study sections. The sample is restricted to new and competitive renewal grants so that there is one observation per successful grant application cycle. A grant is defined to be matched with a publication if it acknowledges the project number of the grant and is published within 5 years of the grant's funding year. A patent is considered to be by an NIH-funded researcher if it directly acknowledges funding from that grant. A patent is citation-linked to a grant if it cites a publication that is linked to a grant. A patent is PRMA linked if it cites a publication that is similar (as defined by the PubMed Relatedness Matching Algorithm--see Appendix) to a publication that is linked to a patent. In this paper, we require that similar publications be published within 5 years of each other. Patent counts are weighted in the sense that if a patent is matched to N distinct grants, it counts as 1/Nth of a patent for each grant. A patent is defined as private sector if its assignee is a US or foreign firm.

Table 2: Characteristics of NIH Research Areas (DSTs), 1980-2000

		DSTs Linked to Patents		
	Full Sample	**NIH-Funded**	**Citation**	**PMRA**
Average # of Grants	11.11	31.19	15.77	12.01
	(16.88)	*(23.82)*	*(19.25)*	*(17.37)*
Output Characteristics (weighted by DST size)				
Funding Amount (DST)	$46,733,370	$74,337,140	$50,293,620	$47,226,450
	(49,248,420)	*(68,411,190)*	*(49,969,720)*	*(49,290,640)*
# of Private Sector Patents by NIH Funded Researchers (weighted counts)	0.19	0.76	0.21	0.19
	(0.61)	*(1.03)*	*(0.64)*	*(0.61)*
Unweighted	0.71	2.84	0.77	0.71
	(2.01)	*(3.18)*	*(2.09)*	*(2.02)*
# Citation-Linked Private Sector Patents (weighted counts)	14.43	26.29	15.83	14.59
	(18.86)	*(22.97)*	*(19.19)*	*(18.91)*
Unweighted	104.00	200.00	115.00	106.00
	(141.00)	*(186.00)*	*(144.00)*	*(142.00)*
# PMRA Linked Private Sector Patents (weighted counts)	27.24	45.15	29.76	27.54
	(27.36)	*(33.27)*	*(27.38)*	*(27.36)*
Unweighted	3248.00	5611.00	3556.00	3283.00
	(3151.00)	*(3771.00)*	*(3135.00)*	*(3149.00)*
N	11,110	984	7,133	10,172

Notes: Sample is same as from Table 1, except aggregated to the NIH Disease-Science-Time level. Please see the notes to Table 1 for additional definitions. The weighting on patent counts is modified from the grant-level weights so that if a patent is matched to N distinct DSTs, it counts as 1/Nth of a patent for each DST. Funding amounts are in 2010 dollars.

Table 3: Effect of NIH Investments on Private Sector Patenting by NIH-Funded Researchers

	(1)	(2)	(3)	(4)	(5)
Weighted Patent Counts					
DST Funding ($10 mill)	0.035***	0.031***	0.033***	0.030***	0.033***
	(0.004)	*(0.008)*	*(0.008)*	*(0.006)*	*(0.007)*
Elasticity	0.868	0.762	0.816	0.728	0.811
R-squared	0.089	0.307	0.371	0.766	0.772
Unweighted Patent Counts					
DST Funding ($10 mill)	0.149***	0.117***	0.117***	0.129***	0.157***
	(0.016)	*(0.030)*	*(0.027)*	*(0.020)*	*(0.028)*
Elasticity	0.989	0.772	0.778	0.854	1.033
R-squared	0.139	0.421	0.461	0.795	0.802
Observations	11,110	11,110	11,110	11,110	11,110
Year FEs	X	X	X	X	X
Disease X Science FEs		X	X	X	X
Disease X Year FEs			X	X	X
Science X Year FEs				X	X
Application Quality and Lagged Funding Controls					X

Notes: Each observation is Disease-Science Area-Time (DST) combination. A patent is considered to be by an NIH-funded researcher if it directly acknowledges funding from an NIH grant. For more details on this sample, please see the notes to Tables 1 and 2. Funding is defined by the sum of project-cycle allocations for all associated new and competing renewal grants. A patent is defined as private sector if its assignee is a US or foreign firm. Elasticities are evaluated at sample means. Application quality controls include cubics in the average raw score received by awarded grants and cubics for the average; 10 year lags of number of publications produced by the DST and number of applications to the DST; and dummies for number of DST applicants near the IC's funding threshold. We also include 10 years of lagged funding controls with indicators if a DST is not observed in any year.

Table 4: Effect of NIH Investments on Citation-Linked Patenting

	(1)	(2)	(3)	(4)	(5)
Weighted Patent Counts					
DST Funding ($10 mill)	2.352*** *(0.228)*	2.055*** *(0.376)*	2.045*** *(0.318)*	1.911*** *(0.202)*	2.306*** *(0.242)*
Elasticity	0.762	0.666	0.662	0.619	0.746
R-squared	0.414	0.650	0.695	0.877	0.890
Unweighted Patent Counts					
DST Funding ($10 mill)	17.985*** *(1.696)*	13.803*** *(3.217)*	14.209*** *(2.740)*	11.409*** *(1.553)*	14.857*** *(1.571)*
Elasticity	0.802	0.616	0.634	0.509	0.667
R-squared	0.426	0.706	0.746	0.905	0.919
Observations	11,110	11,110	11,110	11,110	11,110
Year FEs	X	X	X	X	X
Disease X Science FEs		X	X	X	X
Disease X Year FEs			X	X	X
Science X Year FEs				X	X
Application Quality and Lagged Funding Controls					X

Notes: Each observation is Disease-Science Area-Time (DST) combination. A patent is citation-linked to a DST if it cites research that acknowledges funding from that DST. For more details on this sample, please see the notes to Tables 1 and 2. Funding is defined by the sum of project-cycle allocations for all associated new and competing renewal grants. A patent is defined as private sector if its assignee is a US or foreign firm. Elasticities are evaluated at sample means. Application quality controls include cubics in the average raw score received by awarded grants and cubics for the average; 10 year lags of number of publications produced by the DST and number of applications to the DST; and dummies for number of DST applicants near the IC's funding threshold. We also include 10 years of lagged funding controls with indicators if a DST is not observed in any year.

Table 5: Effect of NIH Investments on PMRA-Linked Patenting

	(1)	(2)	(3)	(4)	(5)
Weighted Patent Counts					
DST Funding ($10 mill)	3.921***	2.748***	2.839***	2.255***	2.786***
	(0.415)	*(0.741)*	*(0.605)*	*(0.330)*	*(0.238)*
Elasticity	0.673	0.471	0.487	0.387	0.478
R-squared	0.520	0.807	0.829	0.954	0.965
Unweighted Patent Counts					
DST Funding ($10 mill)	438.041***	284.866***	292.464***	243.370***	307.134***
	(44.197)	*(73.635)*	*(62.360)*	*(37.219)*	*(30.074)*
Elasticity	0.63	0.41	0.421	0.350	0.442
R-squared	0.494	0.857	0.876	0.959	0.967
Observations	11,110	11,110	11,110	11,110	11,110
Year FEs	X	X	X	X	X
Disease X Science FEs		X	X	X	X
Disease X Year FEs			X	X	X
Science X Year FEs				X	X
Application Quality and Lagged Funding Controls					X

Notes: Each observation is Disease-Science Area-Time (DST) combination. A patent is PMRA linked to a DST if it cites publications that are related to publications supported by a DST. For more details on this sample, please see the notes to Tables 1 and 2. Funding is defined by the sum of project-cycle allocations for all associated new and competing renewal grants. A patent is defined as private sector if its assignee is a US or foreign firm. Elasticities are evaluated at sample means. Application quality controls include cubics in the average raw score received by awarded grants and cubics for the average score received by all applicants regardless of whether they are funded 10 year lags of number of publications produced by the DST and number of applications to the DST; and dummies for number of DST applicants near the IC's funding threshold. We also include 10 years of lagged funding controls with indicators if a DST is not observed in any year.

Table 6: Instrumental Variables Estimate, Lucky DST Funding

	(1)	(2)		(3)	(4)	(5)	(6)
	First Stage			**IV Estimates**			
	DST Funding ($10 mill)			Citation-Linked		PMRA-Linked	
DST Funding, just awarded grants ($10 mill)	2.772*** *(0.732)*	1.718*** *(0.544)*	DST Funding ($10 mill)	3.140*** *(1.193)*	3.319** *(1.666)*	3.885*** *(1.433)*	2.890* *(1.579)*
			Elasticity	1.016	1.074	0.666	0.495
R-squared	0.414	0.414		0.349	0.417	0.453	0.564
Observations	11,110	11,110		10,536	10,536	10,536	10,536
Year FEs	X	X		X	X	X	X
Disease X Science FEs	X	X		X	X	X	X
Disease X Year FEs	X	X		X	X	X	X
Science X Year Linear Trends	X	X		X	X	X	X
Application Quality and Lagged Funding Controls		X			X		X

Notes: The instrument is the total amount of funding for awarded DST grants within 5 grants of the award cutoff. Columns 2, 4, and 6 control for cardinal application quality and number of grants near the threshold. Each observation is Disease-Science Area-Time (DST) combination. A patent is citation-linked to a DST if it cites research that acknowledges funding from that DST. A patent is PMRA linked to a DST if it cites publications that are related to publications supported by a DST. For more details on this sample, please see the notes to Tables 1 and 2. Funding is defined by the sum of project-cycle allocations for all associated new and competing renewal grants. A patent is defined as private sector if its assignee is a US or foreign firm. Elasticities are evaluated at sample means. Application quality controls include cubics in the average raw score received by awarded grants and cubics for the average; 10 year lags of number of publications produced by the DST and number of applications to the DST; and dummies for number of DST applicants near the IC's funding threshold. We also include 10 years of lagged funding controls with indicators if a DST is not observed in any year.

Table 7: Effect of NIH Investments on Own vs. Other Disease Area Patenting

	(1)	(2)	(3)	(4)
	Citation-Linked Patents		**PMRA-Linked Patents**	
	Same Disease	Different Disease	Same Disease	Different Disease
Weighted Patent Counts				
DST Funding ($10 mill)	1.015***	1.292***	1.633***	1.152***
	(0.153)	*(0.139)*	*(0.145)*	*(0.116)*
Elasticity	0.777	0.724	0.476	0.479
R-squared	0.835	0.903	0.956	0.964
Unweighted Patent Counts				
DST Funding ($10 mill)	1.633***	13.225***	51.116***	256.018***
	(0.228)	*(1.453)*	*(4.154)*	*(26.487)*
Elasticity	0.654	0.663	0.496	0.432
R-squared	0.876	0.917	0.968	0.965
Observations	11,110	11,110	11,110	11,110
Year FEs	X	X	X	X
Disease X Science FEs	X	X	X	X
Disease X Year FEs	X	X	X	X
Science X Year FEs	X	X	X	X
Application Quality and Lagged Funding Controls	X	X	X	X

Notes: Each observation is Disease-Science Area-Time (DST) combination. A patent is citation-linked to a DST if it cites research that acknowledges funding from that DST. A patent is PMRA linked to a DST if it cites publications that are related to publications supported by a DST. For more details on this sample, please see the notes to Tables 1 and 2. Funding is defined by the sum of project-cycle allocations for all associated new and competing renewal grants. A patent is defined as private sector if its assignee is a US or foreign firm. Elasticities are evaluated at sample means. Application quality controls include cubics in the average raw score received by awarded grants and cubics for the average; 10 year lags of number of publications produced by the DST and number of applications to the DST; and dummies for number of DST applicants near the IC's funding threshold. We also include 10 years of lagged funding controls with indicators if a DST is not observed in any year. A patent is affiliated to the disease area to which it is most often associated. If this disease area is not the same as the DST, then this patent is considered to be in a different disease area.

Table 8: Effect of NIH Investments on Private Sector High-Value Patenting

	(1)	(2)	(3)	(4)	(5)	(6)
	Highly Cited Patents			**FDA Approved Drugs**		
	NIH-Funded	Citation	PMRA	NIH-Funded	Citation	PMRA
Weighted Patent Counts						
DST Funding ($10 mill)	0.002 *(0.001)*	0.090*** *(0.022)*	0.148*** *(0.018)*	0.001 *(0.001)*	0.062*** *(0.016)*	0.066*** *(0.012)*
Elasticity	0.753	0.512	0.435	0.778	0.669	0.416
R-squared	0.747	0.752	0.927	0.879	0.783	0.916
Unweighted Patent Counts						
DST Funding ($10 mill)	0.007* *(0.004)*	0.896*** *(0.152)*	16.538*** *(1.738)*	0.003 *(0.002)*	0.393*** *(0.102)*	8.537*** *(0.982)*
Elasticity	0.646	0.594	0.397	0.796	0.465	0.384
R-squared	0.819	0.835	0.960	0.824	0.465	0.384
Observations	11,110	11,110	11,110	11,110	11,110	11,110
Year FEs	X	X	X	X	X	X
Disease X Science FEs	X	X	X	X	X	X
Disease X Year FEs	X	X	X	X	X	X
Science X Year FEs	X	X	X	X	X	X
Application Quality and Lagged Funding Controls	X	X	X	X	X	X

Notes: Each observation is Disease-Science Area-Time (DST) combination. A patent is considered to be by an NIH-funded researcher if it directly acknowledges funding from an NIH grant. A patent is citation-linked to a DST if it cites research that acknowledges funding from that DST. A patent is PMRA linked to a DST if it cites publications that are related to publications supported by a DST. For more details on this sample, please see the notes to Tables 1 and 2. Funding is defined by the sum of project-cycle allocations for all associated new and competing renewal grants. A patent is defined as private sector if its assignee is a US or foreign firm. Elasticities are evaluated at sample means. Application quality controls include cubics in the average raw score received by awarded grants and cubics for the average; 10 year lags of number of publications produced by the DST and number of applications to the DST; and dummies for number of DST applicants near the IC's funding threshold. We also include 10 years of lagged funding controls with indicators if a DST is not observed in any year. A patent is highly cited if it is in the top 5 percentile for citations among patents in the same three digit class, issued in the same year. A patent is associated with an FDA approved drug if it is listed in the Orange Book.

Table 9: Effect of NIH Investments By Firm Size

	(1)	(2)	(3)	(4)	(5)	(6)
		Small			**Large**	
	NIH-Funded	Citation	PMRA	NIH-Funded	Citation	PMRA
Weighted Patent Counts						
DST Funding ($10 mill)	0.011***	0.510***	0.446***	0.022***	1.796***	2.340***
	(0.004)	*(0.056)*	*(0.044)*	*(0.006)*	*(0.214)*	*(0.199)*
Elasticity	0.566	0.811	0.450	1.032	0.730	0.483
R-squared	0.785	0.878	0.958	0.741	0.877	0.963
Unweighted Patent Counts						
DST Funding ($10 mill)	0.070***	3.429***	61.954***	0.088***	11.429***	245.180***
	(0.014)	*(0.315)*	*(6.116)*	*(0.021)*	*(1.309)*	*(24.074)*
Elasticity	1.004	0.678	0.437	1.079	0.657	0.443
R-squared	0.766	0.921	0.969	0.790	0.909	0.965
Observations	11,110	11,110	11,110	11,110	11,110	11,110
Year FEs	X	X	X	X	X	X
Disease X Science FEs	X	X	X	X	X	X
Disease X Year FEs	X	X	X	X	X	X
Science X Year FEs	X	X	X	X	X	X
Application Quality and Lagged Funding Controls	X	X	X	X	X	X

Notes: Each observation is Disease-Science Area-Time (DST) combination. A patent is considered to be by an NIH-funded researcher if it directly acknowledges funding from an NIH grant. A patent is citation-linked to a DST if it cites research that acknowledges funding from that DST. A patent is PMRA linked to a DST if it cites publications that are related to publications supported by a DST. For more details on this sample, please see the notes to Tables 1 and 2. Funding is defined by the sum of project-cycle allocations for all associated new and competing renewal grants. A patent is defined as private sector if its assignee is a US or foreign firm. Elasticities are evaluated at sample means. Application quality controls include cubics in the average raw score received by awarded grants and cubics for the average; 10 year lags of number of publications produced by the DST and number of applications to the DST; and dummies for number of DST applicants near the IC's funding threshold. We also include 10 years of lagged funding controls with indicators if a DST is not observed in any year. A firm is considered small if it has fewer than 500 employees.

Table 10: Effect of NIH Investments on Firm Reallocation of R&D Investments

	(1)	(2)	(3)	(4)
	Total non-DST patents		**Average non-DST patents, per DST-linked patent**	
	Citation	PMRA	Citation	PMRA
Weighted Patent Counts				
Annual Research Area (DST) Funding ($10 mill)	0.812	4.890***	1.062***	0.935***
	(5.859)	*(1.003)*	*(0.114)*	*(0.130)*
Elasticity	0.027	0.113***	0.165	0.075***
	(0.196)	*(0.023)*	*(0.148)*	*(0.026)*
R-squared	0.272	0.321	0.483	0.866
Unweighted Patent Counts				
Annual Research Area (DST) Funding ($10 mill)	28.430	971.446***	6.044	8.416***
	(61.775)	*(339.737)*	*(6.055)*	*(3.102)*
Elasticity	0.060	0.104***	0.124	0.057***
	(0.131)	*(0.036)*	*(0.124)*	*(0.021)*
R-squared	0.318	0.450	0.788	0.935
Observations	11110	11110	11110	11110
Year FEs	X	X	X	X
Disease X Science FEs	X	X	X	X
Disease X Year FEs	X	X	X	X
Science X Year FEs	X	X	X	X

A DST-Patent Matching Details

A.1 Patents directly resulting from NIH research

The first is measure a count of patents directly resulting from NIH-funded grants from an DST. The 1981 Bayh-Dole Act created incentives for these researchers and their institutions to patent their discoveries, so that they could be licensed to private firms. The Act also required that patents resulting from public funding acknowledge this fact and list specific grants in their "Government Interest" statements. We obtained this information from the NIH's iEdison database and count the number of unique patents that acknowledge support from each NIH DST.

A.2 Patents citing NIH-funded research

Our next measure of innovative output uses patent-publication citation information to identify patents that build on NIH-funded research. Patent applicants are required to disclose any previous patents and publications that are related to their research. Failure to do so can result in strong penalties for the applicant and attorney, and invalidation of the patent (Sampat 2009). There is a long history of using citation data as measures of intellectual influence or knowledge flows between public and private sector research (Jaffe and Trajtenberg 2005; Narin and Olivastro 1992). Recent work (Sampat 2010, Alcacer, Gittleman and Sampat 2008), however, shows that patent examiners rather than applicants insert many of these citations, casting doubt on their utility as measures of knowledge flows or spillovers (Alcacer and Gittleman 2006).

We will instead use information on patent citations to published scientific articles. This is appealing both because publications rather than patents are the main output of scientific researchers (Agrawal and Henderson 2001), but also because the vast majority of patent-paper citations, over 90 percent, come from applicants rather than examiners, and are thus more plausibly indicators of real knowledge flows than patent-patent citations (Lemley and Sampat 2010). Roach and Cohen (2012) provide empirical evidence on this point.

In previous work, systematic analyses of these non-patent references has been limited, since they are free-form text and difficult to link to other data. Our work relies on a novel match between non-patent references and biomedical articles indexed in PubMed. Developing this match was more difficult than for patent-patent citations: while the cited patents are unique seven-digit numbers, cited publications are free-form text. Moreover, the USPTO does not require that applicants submit references to literature in a standard format. For example, Harold Varmus's 1988 Science article "Retroviruses" is cited in 29 distinct patents, but in numerous different formats, including Varmus. "Retroviruses" Science 240:1427-1435 (1988) (in patent 6794141) and Varmus et al., 1988, Science 240:1427-1439 (in patent 6805882). As this example illustrates, there can be errors in author lists and page numbers. Even more problematic, in some cases certain fields (e.g. author name) are included, in others they are not. Journal names may be abbreviated in some patents, but not in others. We use a fuzzy-matching algorithm to overcome these difficulties, and thus have matched all non-patent references in over 3 million biomedical patents (issued from 1976 onwards) to articles in PubMed (Sampat and Lichthenberg 2010). We also link the full set of NIH funded grants from 1972-2002 to the set of scientific articles that it supports using grant acknowledgement data from PubMed. We combine these datasets to create links from grants funded by an DST to articles to citing patents.

A.3 Patents in the intellectual vicinity of NIH research

Our final outcome measure captures all patents in the intellectual vicinity of an NIH funding area. To do this, we rely on the National Library of Medicine's PubMed Related Citations Algorithm (PMRA) to determined which publications are similar to each other. The following paragraphs were extracted from a brief description of PMRA:[19]

> *The neighbors of a document are those documents in the database that are the most similar to it. The similarity between documents is measured by the words they have in common, with some adjustment for document lengths. To carry out such a program, one must first define what a word is. For us, a word is basically an unbroken string of letters and numerals with at least one letter of the alphabet in it. Words end at hyphens, spaces, new lines, and punctuation. A list of 310 common, but uninformative, words (also known as stopwords) are eliminated from processing at this stage. Next, a limited amount of stemming of words is done, but no thesaurus is used in processing. Words from the abstract of a document are classified as text words. Words from titles are also classified as text words, but words from titles are added in a second time to give them a small advantage in the local weighting scheme. MeSH terms are placed in a third category, and a MeSH term with a subheading qualifier is entered twice, once without the qualifier and once with it. If a MeSH term is starred (indicating a major concept in a document), the star is ignored. These three categories of words (or phrases in the case of MeSH) comprise the representation of a document. No other fields, such as Author or Journal, enter into the calculations.*

[19] Available at `http://ii.nlm.nih.gov/MTI/related.shtml`

Having obtained the set of terms that represent each document, the next step is to recognize that not all words are of equal value. Each time a word is used, it is assigned a numerical weight. This numerical weight is based on information that the computer can obtain by automatic processing. Automatic processing is important because the number of different terms that have to be assigned weights is close to two million for this system. The weight or value of a term is dependent on three types of information: 1) the number of different documents in the database that contain the term; 2) the number of times the term occurs in a particular document; and 3) the number of term occurrences in the document. The first of these pieces of information is used to produce a number called the global weight of the term. The global weight is used in weighting the term throughout the database. The second and third pieces of information pertain only to a particular document and are used to produce a number called the local weight of the term in that specific document. When a word occurs in two documents, its weight is computed as the product of the global weight times the two local weights (one pertaining to each of the documents).

The global weight of a term is greater for the less frequent terms. This is reasonable because the presence of a term that occurred in most of the documents would really tell one very little about a document. On the other hand, a term that occurred in only 100 documents of one million would be very helpful in limiting the set of documents of interest. A word that occurred in only 10 documents is likely to be even more informative and will receive an even higher weight.

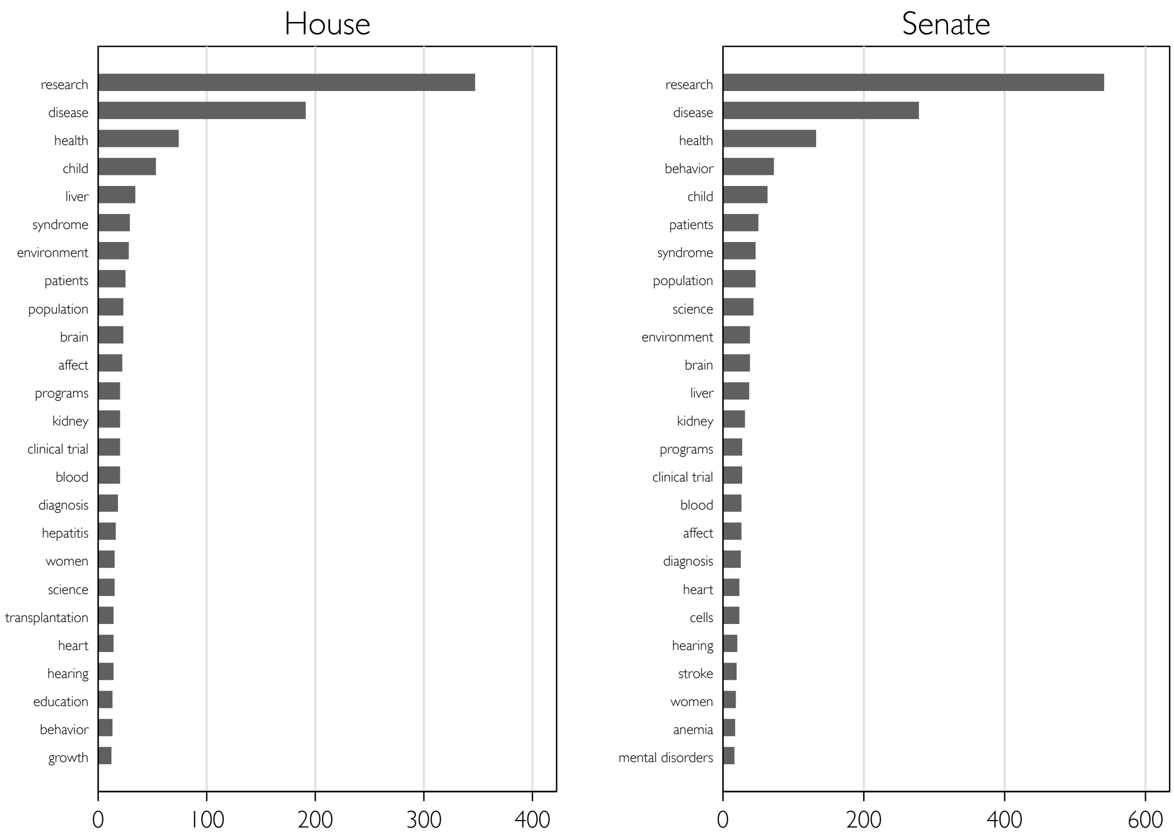
The local weight of a term is the measure of its importance in a particular document. Generally, the more frequent a term is within a document, the more important it is in representing the content of that document. However, this relationship is saturating, i.e., as the frequency continues to go up, the importance of the word increases less rapidly and finally comes to a finite limit. In addition, we do not want a longer document to be considered more important just because it is longer; therefore, a length correction is applied.

The similarity between two documents is computed by adding up the weights of all of the terms the two documents have in common. Once the similarity score of a document in relation to each of the other documents in the database has been computed, that document's neighbors are identified as the most similar (highest scoring) documents found. These closely related documents are pre-computed for each document in PubMed so that when one selects Related Articles, the system has only to retrieve this list. This enables a fast response time for such queries.

We illustrate the use of PMRA with an example taken from our sample. Brian Druker is a faculty member at the University of Oregon whose NIH grant CA-001422 (first awarded in 1990) yielded 9 publications. *"CGP 57148, a tyrosine kinase inhibitor, inhibits the growth of cells expressing BCR-ABL, TEL-ABL, and TEL-PDGFR fusion proteins"* (PubMed ID #9389713) appeared in the December 1997 issue of the journal *Blood* and lists 16 MeSH terms. PubMed ID #8548747 is its fifth-most related paper according to the PMRA algorithm; it appeared in *Cancer Research* in January 1996 and has 13 MeSH terms, 6 of which overlap with the Druker article. These terms include common terms such as Mice and Pyrimidines as well as more specific keywords including Oncogene Proteins v-abl and Receptors, Platelet-Derived Growth Factor.

PMRA and MeSH Terms Overlap — An Example

Source Article	**PMRA-Linked Article**
Carroll et al., "CGP 57148, a tyrosine kinase inhibitor, inhibits the growth of cells expressing BCR-ABL, TEL-ABL, and TEL-PDGFR fusion proteins." *Blood*, 1997.	Buchdunger et al. "Inhibition of the Abl protein-tyrosine kinase in vitro and in vivo by a 2-phenylaminopyrimidine derivative." *Cancer Research*, 1996.
PMID #9389713	**PMID #8548747**
MeSH Terms	**MeSH Terms**
Animals	3T3 Cells
Antineoplastic Agents	Animals
Cell Division	Cell Line, Transformed
Cell Line	Growth Substances
DNA-Binding Proteins*	Mice
Enzyme Inhibitors*	Mice, Inbred BALB C
Fusion Proteins, bcr-abl*	Oncogene Proteins v-abl*
Mice	Piperazines*
Oncogene Proteins v-abl*	Piperidines*
Piperazines*	Proto-Oncogene Proteins c-fos
Protein-Tyrosine Kinases*	Pyrimidines*
Proto-Oncogene Proteins c-ets	Receptors, Platelet-Derived Growth Factor*
Pyrimidines*	Tumor Cells, Cultured
Receptors, Platelet-Derived Growth Factor*	
Repressor Proteins*	
Transcription Factors*	
Substances	**Substances**
Antineoplastic Agents	Growth Substances
DNA-Binding Proteins	Oncogene Proteins v-abl
ETS translocation variant 6 protein	Piperazines
Enzyme Inhibitors	Piperidines
Fusion Proteins, bcr-abl	Proto-Oncogene Proteins c-fos
Oncogene Proteins v-abl	Pyrimidines
Piperazines	imatinib
Proto-Oncogene Proteins c-ets	Receptors, Platelet-Derived Growth Factor
Pyrimidines	
Repressor Proteins	
Transcription Factors	
imatinib	
Protein-Tyrosine Kinases	
Receptors, Platelet-Derived Growth Factor	



Pancreatic cancer.—Pancreatic cancer is the country's fourth leading cause of cancer death. Most patients present with advanced disease at diagnosis and the median overall survival rate for people diagnosed with metastatic disease is only about six months. The Committee is concerned that there are too few scientists researching pancreatic cancer and compliments the NCI's past efforts for increasing the research field through its program of a 50 percent formalized extended payline for grants that were 100 percent relevant to pancreatic cancer. The Committee considers this an important method for attracting both young and experienced investigators to develop careers in pancreatic cancer. In 2004, the NCI established a new policy for awarding additional grants in pancreatic cancer research and extended this initiative to research that is 50 percent relevant to pancreatic cancer. The Committee requests NCI to report in February, 2006 on how the two changes in policy have affected the pancreatic cancer portfolio, including the percentage relevancy of each grant to pancreatic cancer, and urges NCI to continue its commitment to fertilize the pancreatic cancer field.

Appendix Figure A: Language in NIH Congressional Appropriations

Appendix Table A: Effect of NIH Investments on Own vs. Other Disease Area Patenting; Disease affiliations defined as fractions

	(1)	(2)	(3)	(4)
	Citation-Linked Patents		**PMRA-Linked Patents**	
	Same Disease	Different Disease	Same Disease	Different Disease
Weighted Patent Counts				
DST Funding ($10 mill)	1.656***	0.651***	1.028***	1.758***
	(0.194)	*(0.069)*	*(0.098)*	*(0.168)*
Elasticity	0.774	0.685	0.424	0.515
R-squared	0.875	0.894	0.956	0.964
Unweighted Patent Counts				
DST Funding ($10 mill)	7.188***	7.669***	62.507***	244.626***
	(0.838)	*(0.833)*	*(6.064)*	*(26.277)*
Elasticity	0.672	0.653	0.389	0.457
R-squared	0.920	0.904	0.976	0.963
Observations	11,110	11,110	11,110	11,110
Year FEs	X	X	X	X
Disease X Science FEs	X	X	X	X
Disease X Year FEs	X	X	X	X
Science X Year FEs	X	X	X	X
Application Quality and Lagged Funding Controls	X	X	X	X

Notes: Each observation is Disease-Science Area-Time (DST) combination. A patent is citation-linked to a DST if it cites research that acknowledges funding from that DST. A patent is PMRA linked to a DST if it cites publications that are related to publications supported by a DST. For more details on this sample, please see the notes to Tables 1 and 2. Funding is defined by the sum of project-cycle allocations for all associated new and competing renewal grants. A patent is defined as private sector if its assignee is a US or foreign firm. Elasticities are evaluated at sample means. Application quality controls include cubics in the average raw score received by awarded grants and cubics for the average; 10 year lags of number of publications produced by the DST and number of applications to the DST; and dummies for number of DST applicants near the IC's funding threshold. We also include 10 years of lagged funding controls with indicators if a DST is not observed in any year. If X percent of a patent's publications are linked to a disease area, then that patent is counted as 0.X toward that disease area.

Appendix Table B: Relationship Between Own DST Funding and Funding by Other Ics for the same Science Area

	(1)	(2)	(3)	(4)	(5)	(6)
	DST Funding (10 mill) Mean: 4.67, SD: 4.92			**DST Funding/DT Funding** Mean: 0.079, SD: 0.105		
D′ST Funding, Other Diseases, Same Science (10 mill)	-0.467*** *(0.025)*	0.017 *(0.046)*	0.021 *(0.051)*			
D′ST Funding/D′T Funding, Other Diseases, Same Science				0.064 *(0.072)*	0.060 *(0.070)*	-0.052 *(0.035)*
Observations	11,110	11,110	11,110	11,110	11,110	11,110
R-squared	0.103	0.800	0.830	0.873	0.876	0.909
Year FEs	X	X	X	X	X	X
Disease X Science FEs		X	X		X	X
Disease X Year FEs			X			X

Notes: Each cell is a study section - IC - year. Funding is defined by the sum of project-cycle allocations for all Type I and II grants reviewed by that study section.

Appendix Table C: Instrumental Variables Estimate, Grants Funded in Order Only

	(1)	(2)		(3)	(4)	(5)	(6)
	First Stage			**IV Estimates**			
	DST Funding ($10 mill)			Citation-Linked		PMRA-Linked	
DST Funding, grants funded in order only	0.989*** *(0.122)*	0.910*** *(0.098)*	DST Funding ($10 mill)	2.341*** *(0.253)*	2.573*** *(0.214)*	3.185*** *(0.582)*	3.658*** *(0.378)*
			Elasticity	0.758	0.833	0.546	0.627
R-squared	0.945	0.953		0.376	0.433	0.473	0.562
Observations	11,110	11,110		10,536	10,536	10,536	10,536
Year FEs	X	X		X	X	X	X
Disease X Science FEs	X	X		X	X	X	X
Disease X Year FEs	X	X		X	X	X	X
Science X Year Linear Trends	X	X		X	X	X	X
Application Quality and Lagged Funding Controls		X			X		X

Notes: The instrument is the total amount of funding for awarded DST grants within 5 grants of the award cutoff, divided by the total number of applications within 5 grants of the cutoff on either side. Columns 2, 4, and 6 control for cardinal application quality and number of grants near the threshold. Each observation is Disease-Science Area-Time (DST) combination. A patent is citation-linked to a DST if it cites research that acknowledges funding from that DST. A patent is PMRA linked to a DST if it cites publications that are related to publications supported by a DST. For more details on this sample, please see the notes to Tables 1 and 2. Funding is defined by the sum of project-cycle allocations for all associated new and competing renewal grants. A patent is defined as private sector if its assignee is a US or foreign firm. Elasticities are evaluated at sample means. Application quality controls include cubics in the average raw score received by awarded grants and cubics for the average; 10 year lags of number of publications produced by the DST and number of applications to the DST; and dummies for number of DST applicants near the IC′s funding threshold. We also include 10 years of lagged funding controls with indicators if a DST is not observed in any year.

Appendix Table D: Correlation of Instrument with Measures of DST Quality

	(1)	(2)
	Lucky Funding	
RHS includes 10 Years of Lags for:	Past Patent Output	Raw Application Scores
F-test for Joint Signficance	1.508 *(1.109)*	0.039 *(0.049)*
Year FEs	X	X
Disease X Science FEs	X	X
Disease X Year FEs	X	X
Science X Year FEs	X	X
Application Quality and Lagged Funding Controls	X	X

Notes: Each observation is Disease-Science Area-Time (DST) combination. Each column reports a regression of our lucky funding instrument on measures of DST input and output quality. Column 1 reports an F-test for the joint significance of one to ten year lags of past DST patent production: NIH-funded, citation-linked, and total (30 variables). Column 2 reports the same but for one to ten year lags of average raw scores among funded grants and all applicants to a DST (20 variables).

Appendix Table E: Effect of NIH Investments on PMRA-Linked Patenting (with controls for # of citation-linked patents)

	(1) Private Sector	(2) Same disease	(3) Different disease	(4) FDA Approved	(5) Highly Cited	(6) Small	(7) Large
DST Funding ($10 mill)	1.717***	1.297***	0.662***	0.071***	0.181***	0.636***	1.850***
	(0.387)	*(0.252)*	*(0.138)*	*(0.017)*	*(0.035)*	*(0.138)*	*(0.361)*
Elasticity	0.294	0.378	0.275	0.448	0.532	0.642	0.382
Observations	11,110	11,110	11,110	11,110	11,110	11,110	11,110
R-squared	0.902	0.856	0.925	0.803	0.806	0.915	0.903
# of Citation-Linked Patents	X	X	X	X	X	X	X
Year FEs	X	X	X	X	X	X	X
Disease X Science FEs	X	X	X	X	X	X	X
Disease X Year FEs	X	X	X	X	X	X	X
Science X Year Linear Trends	X	X	X	X	X	X	X
Application Quality and Lagged Funding Controls	X	X	X	X	X	X	X

Notes: Each observation is Disease-Science Area-Time (DST) combination. A patent is PRMA-linked to a DST if it cites a publication is published within five years of a similar publication that acknowledges funding from a grant associated with that DST. For more details on this sample, please see the notes to Table 5. This table includes controls for the number of citation-linked patents associated with the same DST. See text for details.