Appendix Materials: For Online Publication

Appendix A: A Primer on NIH Funding

The National Institutes of Health (NIH) is the primary organization within the United States government with responsibilities for health-related research. The NIH is the single largest funder of biomedical research, with an annual budget of approximately $30 billion. According to its own web site, NIH's mission is *"to seek fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability."*

NIH comprises 21 different Institutes (plus an assortment of centers that our analysis will ignore), each with a distinct, though sometimes overlapping, research agenda. For example, the National Institute for Mental Health, as the name suggests, focuses on mental health related research. It shares interests with the National Institute of Aging on issues related to dementia. All Institutes receive their funding directly from Congress, and manage their own budgets. Table A1 lists each of the agency's component institutes.

Figure A1(i) 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. Figure A1(ii) compiles 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 funding levels for particular Institutes are more highly correlated with disease burden than with scientific advances (Gillum et al. 2011).

Approximately 10% of the overall NIH budget is dedicated to the intramural research program, with almost all Institutes providing some support. The program directly supports about 6,000 scientists working within the federal laboratories on NIH Campuses. Unlike the intramural program, where allocation decisions are relatively opaque, the operations of the extramural program are quite transparent. More than 80% of the total budget supports extramural research through competitive grants that are awarded to universities, medical schools, and other research institutions, primarily in the United States. The largest and most established of these grant mechanisms is the R01, a project-based renewable research grant which constitutes half of all NIH grant spending and is the primary funding source for most academic biomedical labs in the United States. There are currently 27,000 outstanding awards, with 4,000 new projects approved each year. The average size of each award is 1.7 million dollars spread over 3 to 5 years and the application success rate is approximately 20 percent (Li 2016).

Requests for proposals identify priority areas, but investigators are also free to submit applications on unsolicited topics under the extramural research program. All applications are assigned to a review committee comprised of scientific peers, generally known as a study section (Table A2 lists the 173 study sections that currently exist). Reviewers are asked to ignore budgetary issues, limiting their attention to scientific and technical merit on the basis of five criteria: (1) Significance [does the project address an important issue?]; (2) Approach [is the methodology sound?]; (3) Innovation [is the research novel?]; (4) Investigator [are the skills of the research team well matched to the project?]; and (5) Environment [is the place in which the work will take place conducive to project success?]. Each reviewer assigns a two digit priority score ranging from 1.0 for the best application to 5.0 for the worst. At the study section meeting, three reviewers are typically asked to discuss an application and present their initial scores. This is followed by an open discussion by all reviewers and a brief period for everyone to revise their initial scoring based on the group deliberations before anonymously submitting their final scores. The overall priority score for the proposal is based on the average across all study section members. Those applications determined to be of the lowest quality by the study section do not receive priority scores. Scores are then normalized within review groups through the assignment of percentile scores to facilitate funding decisions.

Funding decisions are decoupled from the scientific review and determined by program areas within the Institutes. In essence, each decision making unit (e.g., Division, Program, Branch) within an Institute is allocated a fixed annual budget. Units then fund new projects in order of their priority score until their budget, net of encumbered funds for ongoing grants awarded in previous years, is exhausted. The highest percentile score that is funded is known as the payline. A grant's score is generally the sole determinant of the funding decision,[i] irrespective of proposal costs (assuming they are deemed reasonable). Researchers who do not receive funding are given the opportunity to respond to reviewer criticisms and submit an amended application.

Institutes considered in the econometric analysis. We exclude from our analytic sample observations corresponding to the National Library of Medicine (NLM), the National Institute of Nursing Research (NINR), and the National Institute on Minority Health and Health Disparities (NIMHD), which together represent less than 3% of NIH's total budget. We drop the NLM because it seldom supports extramural researchers. We drop NINR and NIMHD because we found no instances of the grants funded by these Institutes generating publications referenced in private-sector patents.

A cursory look at the names of the list of the 18 Institutes we do include in most of our analyses reveals that some of these Institutes may not be strictly disease-focused. This is certainly the case for NIGMS (which supports mostly untargeted laboratory research), for NHGRI (the Genome Institute), and NIBIB (which focuses on imaging technology). In a sensitivity test, we will explore whether our main results are robust to the exclusion of these three "science-focused" Institutes. Further, we will also investigate the effects of dropping NIA, NIDCD, NIEHS, and NICHD who traditionally support research on a broad spectrum of loosely related diseases.

Study sections. As mentioned above, the majority of grant evaluation occurs in approximately 200 standing review committees, known as "study sections." Each study section is organized around a scientific topic—for instance, "Cellular and Molecular Immunology"—and is responsible for evaluating the quality of applications in its area. Traditionally, the boundaries delineating study sections have changed only very slowly (too slowly for many NIH critics). Additions and deletions of study sections is relatively rare, and often controversial. In 2006, however, 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 frame and, throughout our analysis, we refer to the old system.

Allocation of Applications to Study Sections. Could applicants improve their odds of funding by sending their applications to study sections reputed to be "weaker"? Study section shopping of this type would be almost surely unproductive, given year-to-year fluctuations in funding and the vagaries of the reapplication process (most proposals are not funded at the first review).[ii] Formally, grant applicants do not choose the study section that will review their proposals. Rather, each application is assigned by staff within the Division of Receipt and Referral at the NIH to a study section based on the needed expertise to evaluate scientific and technical merit.[iii] While many investigators ask to be reviewed by a specific study section, the NIH grants such requests based on the scientific content of the proposal, a consideration of conflicts of interest, and the administrative viability of the request (Chacko 2014). More importantly, the typical advice received by new investigators is to petition to be reviewed in the study section that is most likely to have members on their roster whom are familiar with their narrowly-defined field, and then to stick to this initial

[i] Institute directors have the discretion to fund applications out of order if, for example, they are especially important to the Institute's mission. Since applications can only be submitted three times, Institutes may also choose to fund applications on their last evaluation cycle instead of newly submitted applications that can be reconsidered later. These exceptions appear rare (Jacob and Lefgren 2011).

[ii] Even grant administrators are usually unable to communicate to applicants how the score they received in committee is likely to translate into a final funding decision. It is implausible that grant applicants could be better informed than these knowledgeable insiders.

[iii] `http://public.csr.nih.gov/ApplicantResources/ReceiptReferal/Pages/Submission-and-Assignment-Process.aspx`, accessed August 30, 2014

choice. Consistent with this advice, an essential component of "grantsmanship" at NIH is to build a cordial relationship with the Scientific Review Officer, the staff person within NIH's Center for Scientific Review who administers the logistics of the review process. These informal practices would seem to run counter any temptation to "chase the money."

We see this in the data, where there is considerable inertia in scientist-study section pairings. In a typical five year-period, 88% of NIH grant recipients are evaluated by only one study section; eleven percent are evaluated by two study sections; and only one percent are evaluated by three study sections or more. Why would a given scientist's grant applications ever be reviewed by multiple study sections? One reason is that study sections are not immutable. Some are created; others are eliminated; yet others are merged. Intellectual breadth may also explain the anomalies: In a sample of 10,177 well-funded investigators for whom we have gathered a carefully curated list of publications (cf. Azoulay et al. 2012), intellectual breadth (as proxied by the diversity of MeSH keywords that tag the publications produced by these scientists in rolling five-year windows) is strongly correlated with the likelihood of having one's work reviewed by multiple study section (Table A3). This results holds even when controlling for the total level of funding received. This results hold even when controlling for the total level of funding received. This suggests that scientists have their work reviewed by two or more committees only to the extent that they are active in subfields that are sufficiently distant in intellectual space.

Disease/Science as a level of analysis. As highlighted in the introduction, the organization of the NIH into disease-based funding Institutes and science-based review committees plays an important role in our empirical work, since our independent and dependent variables will be computed at the level of the disease/science/year (DST, technically the IC/study section/year level). If applications evaluated by a study section were always funded by the same Institute, the distinction we emphasize between the disease/science level of analysis and disease-level variation over time would not be very meaningful. However, it is indeed the case that study sections cut across diseases, in the sense that the grant applications they pass favorable judgement on will go on to be funded by several different Institutes. Figure A2(i) shows that the majority, 75 percent, of study sections evaluated grants funded by at least two Institutes. Conversely, Figure A2(ii) shows that the typical Institute draws on applications stemming from more than 50 study sections, on average.

Not only is the DST level of analysis policy-relevant, it is tractable by using the structure of NIH grant review and mapping Institutes into disease areas, and study sections into science areas, respectively. And because of the "intellectual promiscuity" documented above, in practice, increases in funding for one disease can impact innovation in another by supporting research on the scientific foundations these two areas share.

Figure A3 plots residual variation in funding taking out, successively, fixed effects for calendar year, disease/science, disease/year, and science/year. These kernel density estimates make clear that there remains substantial unexplained variation in funding after controlling for all these fixed effects. It is this DST-level variation that we use to estimate the effect of funding on private-sector patenting.

Table A1: NIH Institutes and Centers (ICs)

Institute	Abbrev.	Established	Avg. Budget*
National Cancer Institute	NCI	1937	$4,019,793
National Heart, Lung, and Blood Institute	NHLBI	1948	$2,489,629
National Institute of Allergy and Infectious Diseases	NIAID	1948	$2,070,634
National Institute of Dental and Craniofacial Research	NIDCR	1948	$325,861
National Institute of Mental Health	NIMH	1949	$1,378,636
National Institute of Diabetes and Digestive and Kidney Diseases	NIDDK	1950	$1,491,613
National Institute of Neurological Disorders and Stroke	NINDS	1950	$1,244,241
National Eye Institute	NEI	1968	$562,126
National Institute on Alcohol Abuse and Alcoholism	NIAAA	1970	$423,341
National Institute on Drug Abuse	NIDA	1974	$960,637
National Institute of Arthritis and Musculoskeletal and Skin Diseases	NIAMS	1986	$458,273
National Institute of Child Health and Human Development	NICHD	1962	$1,043,447
National Institute of Environmental Health Sciences	NIEHS	1969	$557,645
National Institute on Aging	NIA	1974	$702,184
National Institute on Deafness and Other Communication Disorders	NIDCD	1988	$347,646
National Institute of General Medical Sciences	NIGMS	1962	$1,629,056
National Human Genome Research Institute	NHGRI	1989	$375,451
National Institute of Biomedical Imaging and Bioengineering	NIBIB	2000	$316,430
National Library of Medicine	NLM	1956	$229,442
National Institute of Nursing Research	NINR	1986	$106,880
National Institute on Minority Health and Health Disparities	NIMHD	1993	$228,287

*Over the 1980-2005 time period, In thousands of 2010 dollars (amounts deflated by the Biomedical R&D PPI)

Table A2: NIH Study Sections

Study Section	Description	Study Section	Description	Study Section	Description
ACE	AIDS Clinical Studies and Epidemiology	CPDD	Child Psychopathology and Developmental Disabilities	MSFA	Macromolecular Structure and Function A
ACTS	Arthritis, Connective Tissue and Skin	CRFS	Clinical Research and Field Studies of Infectious Diseases	MSFB	Macromolecular Structure and Function B
ADDT	AIDS Discovery and Development of Therapeutics	CSRS	Cellular Signaling and Regulatory Systems	MSFC	Macromolecular Structure and Function C
AICS	Atherosclerosis and Inflammation of the Cardiovascular System	DBD	Developmental Brain Disorders	MSFD	Macromolecular Structure and Function D
AIP	AIDS Immunology and Pathogenesis	DDNS	Drug Discovery for the Nervous System	MSFE	Macromolecular Structure and Function E
AMCB	AIDS Molecular and Cellular Biology	DDR	Drug Discovery and Mechanisms of Antimicrobial Resistance	MTE	Musculoskeletal Tissue Engineering
ANIE	Acute Neural Injury and Epilepsy	DEV1	Development - 1	NAED	NeuroAIDS and other End-Organ Diseases
AOIC	AIDS-associated Opportunistic Infections and Cancer	DEV2	Development - 2	NAL	Neurotoxicology and Alcohol
APDA	Adult Psychopathology and Disorders of Aging	DIRH	Dissemination and Implementation Research in Health	NAME	Neurological, Aging and Musculoskeletal Epidemiology
ASG	Aging Systems and Geriatrics	DMP	Drug Discovery and Molecular Pharmacology	NANO	Nanotechnology
AUD	Auditory System	DPVS	Diseases and Pathophysiology of the Visual System	NCF	Neurogenesis and Cell Fate
BACP	Bacterial Pathogenesis	DT	Developmental Therapeutics	NCSD	Nuclear and Cytoplasmic Structure/Function and Dynamics
BBM	Biochemistry and Biophysics of Membranes	EBIT	Enabling Bioanalytical and Imaging Technologies	NDPR	Neurodifferentiation, Plasticity, Regeneration and Rhythmicity
BCHI	Biomedical Computing and Health Informatics	EPIC	Epidemiology of Cancer	NMB	Neurobiology of Motivated Behavior
BDMA	Biodata Management and Analysis	ESTA	Electrical Signaling, Ion Transport, and Arrhythmias	NNRS	Neuroendocrinology, Neuroimmunology, Rhythms and Sleep
BGES	Behavioral Genetics and Epidemiology	GCAT	Genomics, Computational Biology and Technology	NOIT	Neuroscience and Ophthalmic Imaging Technologies
BINP	Brain Injury and Neurovascular Pathologies	GDD	Gene and Drug Delivery Systems	NOMD	Neural Oxidative Metabolism and Death
BMBI	Biomaterials and Biointerfaces	GHD	Genetics of Health and Disease	NPAS	Neural Basis of Psychopathology, Addictions and Sleep Disorders
BMCT	Basic Mechanisms of Cancer Therapeutics	GMPB	Gastrointestinal Mucosal Pathobiology	NRCS	Nursing and Related Clinical Sciences
BMIO	Behavioral Medicine, Interventions and Outcomes	GVE	Genetic Variation and Evolution	NTRC	Neurotransporters, Receptors, and Calcium Signaling
BMIT-A	Biomedical Imaging Technology A	HAI	Hypersensitivity, Autoimmune, and Immune-mediated Diseases	ODCS	Oral, Dental and Craniofacial Sciences
BMIT-B	Biomedical Imaging Technology B	HBPP	Hepatobiliary Pathophysiology	PBKD	Pathobiology of Kidney Disease
BMRD	Biostatistical Methods and Research Design	HDEP	Health Disparities and Equity Promotion	PCMB	Prokaryotic Cell and Molecular Biology
BNVT	Bioengineering of Neuroscience, Vision and Low Vision Technologies	HIBP	Host Interactions with Bacterial Pathogens	PDRP	Psychosocial Development, Risk and Prevention
BPNS	Biophysics of Neural Systems	HM	Hypertension and Microcirculation	PMDA	Pathophysiological Basis of Mental Disorders and Addictions
BRLE	Biobehavioral Regulation, Learning and Ethology	HSOD	Health Services Organization and Delivery	PN	Pregnancy and Neonatology
BSCH	Behavioral and Social Consequences of HIV/AIDS	HT	Hemostasis and Thrombosis	PRDP	Psychosocial Risk and Disease Prevention
BSPH	Behavioral and Social Science Approaches to Preventing HIV/AIDS	ICER	Integrative and Clinical Endocrinology and Reproduction	PTHE	Pathogenic Eukaryotes
BTSS	Bioengineering, Technology and Surgical Sciences	ICI	Intercellular Interactions	RIBT	Respiratory Integrative Biology and Translational Research
BVS	Biology of the Visual System	ICP1	International and Cooperative Projects - 1	RPIA	Risk, Prevention and Intervention for Addictions
CADO	Cellular Aspects of Diabetes and Obesity	IHD	Immunity and Host Defense	RTB	Radiation Therapeutics and Biology
CAMP	Cancer Molecular Pathobiology	III	Innate Immunity and Inflammation	SAT	Surgery, Anesthesiology and Trauma
CASE	Cardiovascular and Sleep Epidemiology	INMP	Integrative Nutrition and Metabolic Processes	SBCA	Synthetic and Biological Chemistry A
CBSS	Cancer Biomarkers	IPOD	Integrative Physiology of Obesity and Diabetes	SBCB	Synthetic and Biological Chemistry B
CCHF	Cardiac Contractility, Hypertrophy, and Failure	IRAP	Infectious Diseases, Reproductive Health, Asthma and Pulmonary Conditions	SBDD	Skeletal Biology Development and Disease
CDD	Cardiovascular Differentiation and Development	ISD	Instrumentation and Systems Development	SBSR	Skeletal Biology Structure and Regeneration
CDIN	Chronic Dysfunction and Integrative Neurodegeneration	KMBD	Kidney Molecular Biology and Genitourinary Organ Development	SCS	Somatosensory and Chemosensory Systems
CDP	Chemo/Dietary Prevention	KNOD	Kidney, Nutrition, Obesity and Diabetes	SEIR	Societal and Ethical Issues in Research
CE	Cancer Etiology	LAM	Neurobiology of Learning and Memory	SMEP	Skeletal Muscle and Exercise Physiology
CG	Cancer Genetics	LCMI	Lung Cellular, Molecular, and Immunobiology	SMI	Sensorimotor Integration
CICS	Clinical and Integrative Cardiovascular Sciences	LCOM	Language and Communication	SPC	Mechanisms of Sensory, Perceptual, and Cognitive Processes
CIDO	Clinical and Integrative Diabetes and Obesity	LIRR	Lung Injury, Repair, and Remodeling	SPIP	Social Psychology, Personality and Interpersonal Processes
CIHB	Community Influences on Health Behavior	MABS	Modeling and Analysis of Biological Systems	SSPA	Social Sciences and Population Studies A
CII	Cancer Immunopathology and Immunotherapy	MBPP	Membrane Biology and Protein Processing	SSPB	Social Sciences and Population Studies B
CIMG	Clinical, Integrative and Molecular Gastroenterology	MCE	Molecular and Cellular Endocrinology	SYN	Synapses, Cytoskeleton and Trafficking
CLHP	Community-Level Health Promotion	MCH	Molecular and Cellular Hematology	TAG	Therapeutic Approaches to Genetic Diseases
CMAD	Cellular Mechanisms in Aging and Development	MEDI	Medical Imaging	TCB	Tumor Cell Biology
CMBG	Cellular and Molecular Biology of Glia	MESH	Biobehavioral Mechanisms of Emotion, Stress and Health	TME	Tumor Microenvironment
CMIA	Cellular and Molecular Immunology - A	MFSR	Motor Function, Speech and Rehabilitation	TPM	Tumor Progression and Metastasis
CMIB	Cellular and Molecular Immunology - B	MGA	Molecular Genetics A	TTT	Transplantation, Tolerance, and Tumor Immunology
CMIP	Clinical Molecular Imaging and Probe Development	MGB	Molecular Genetics B	UGPP	Urologic and Genitourinary Physiology and Pathology
CMIR	Cellular, Molecular and Integrative Reproduction	MIM	Myocardial Ischemia and Metabolism	VACC	HIV/AIDS Vaccines
CMND	Cellular and Molecular Biology of Neurodegeneration	MIST	Molecular and Integrative Signal Transduction	VB	Vector Biology
CNBT	Clinical Neuroimmunology and Brain Tumors	MNG	Molecular Neurogenetics	VCMB	Vascular Cell and Molecular Biology
CNN	Clinical Neuroscience and Neurodegeneration	MNPS	Molecular Neuropharmacology and Signaling	VIRA	Virology - A
CNNT	Clinical Neuroplasticity and Neurotransmitters	MONC	Molecular Oncogenesis	VIRB	Virology - B
CONC	Clinical Oncology	MRS	Musculoskeletal Rehabilitation Sciences	VMD	Vaccines Against Microbial Diseases
CP	Cognition and Perception			XNDA	Xenobiotic and Nutrient Disposition and Action

Table A3: Intellectual Breadth and Study Section Affiliations

	(1)	(2)	(3)	(4)
Two Study Sections	0.141^{**}	0.124^{**}	0.026^{**}	0.011^{**}
	(0.005)	(0.005)	(0.003)	(0.003)
Three Study Sections	0.249^{**}	0.222^{**}	0.042^{**}	0.018^{**}
	(0.011)	(0.012)	(0.006)	(0.007)
Four Study Sections	0.333^{**}	0.297^{**}	0.065^{**}	0.035^{*}
	(0.033)	(0.034)	(0.017)	(0.017)
Five Study Sections	0.354^{**}	0.313^{**}	0.037	0.003
	(0.084)	(0.084)	(0.055)	(0.055)
Ln(NIH Funding)		0.030^{**}		0.031^{**}
		(0.005)		(0.003)
Scientist Fixed Effects	Not Incl.	Not Incl.	Incl.	Incl.
Nb. of Scientists	10,177	10,177	10,177	10,177
Nb. of Observations	146,661	146,661	146,661	146,661
Adjusted R^2	0.226	0.227	0.711	0.712

The dependent variable is the log odds of intellectual diversity, computed as one minus the herfindahl of MeSH keywords in a sample of 10,177 "superstar scientists." The specifications in columns (1) and (2) include indicator variables for type of degree (MD, PhD, MD/PhD), year of highest degree, and gender. All specifications include a full suite of indicator variables for calendar year and for scientist age.

Standard errors in parentheses, clustered by scientist ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$)

Figure A1: Congressional Appropriations for NIH Institutes

(i) Example of Appropriations Language

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.

(ii) Word Frequency in Appropriations Documents

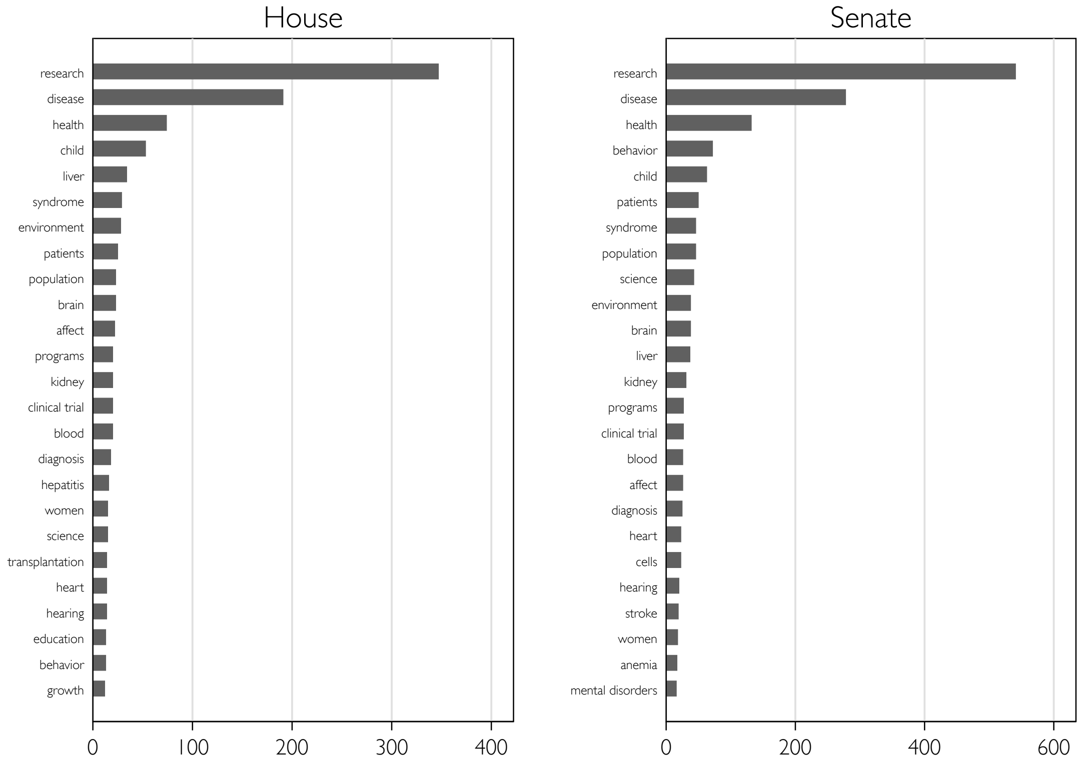
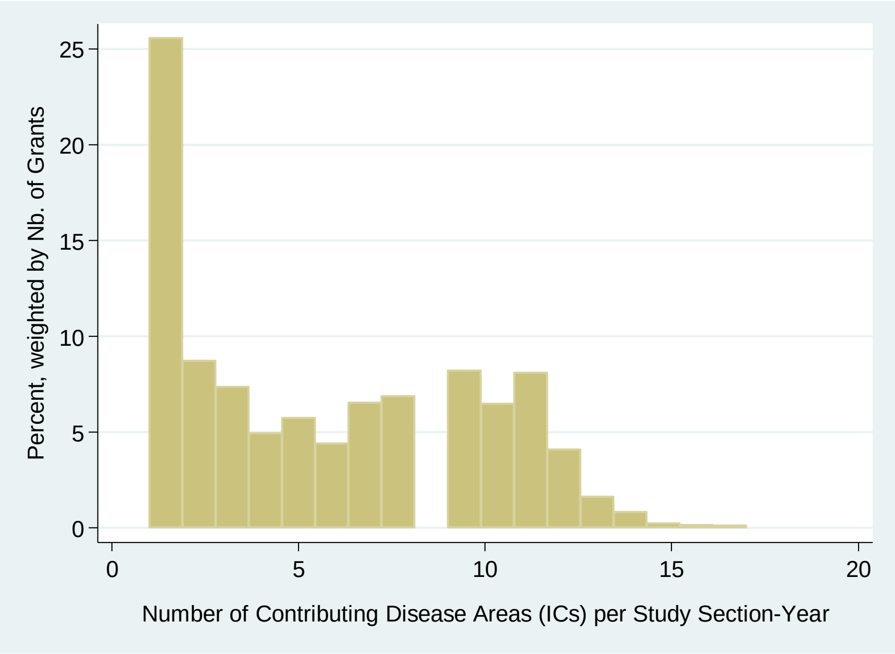


Figure A2: Institute and Study Section Overlap

(i) Number of Institutes per Study Section



(ii) Number of Study Sections per Institute

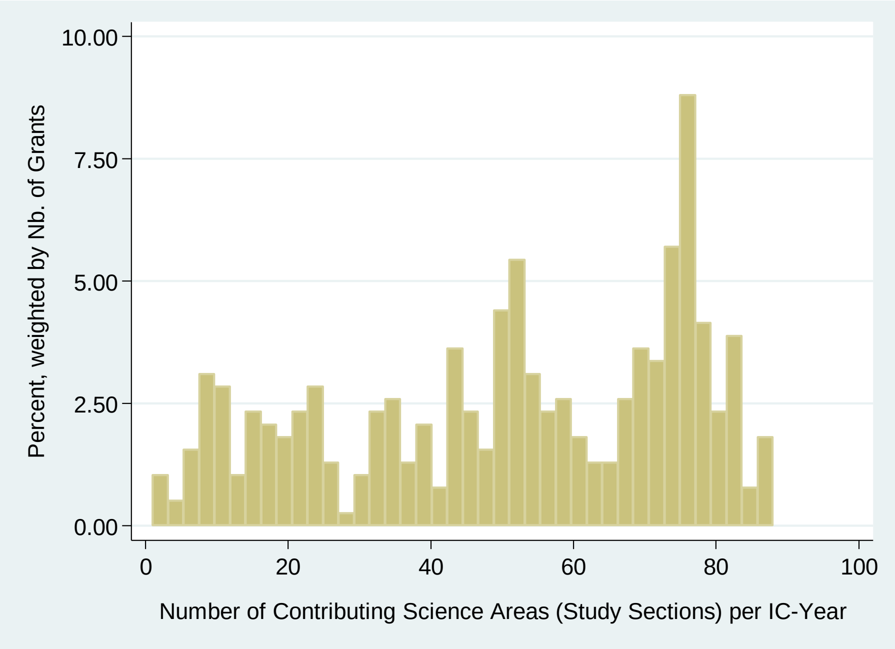
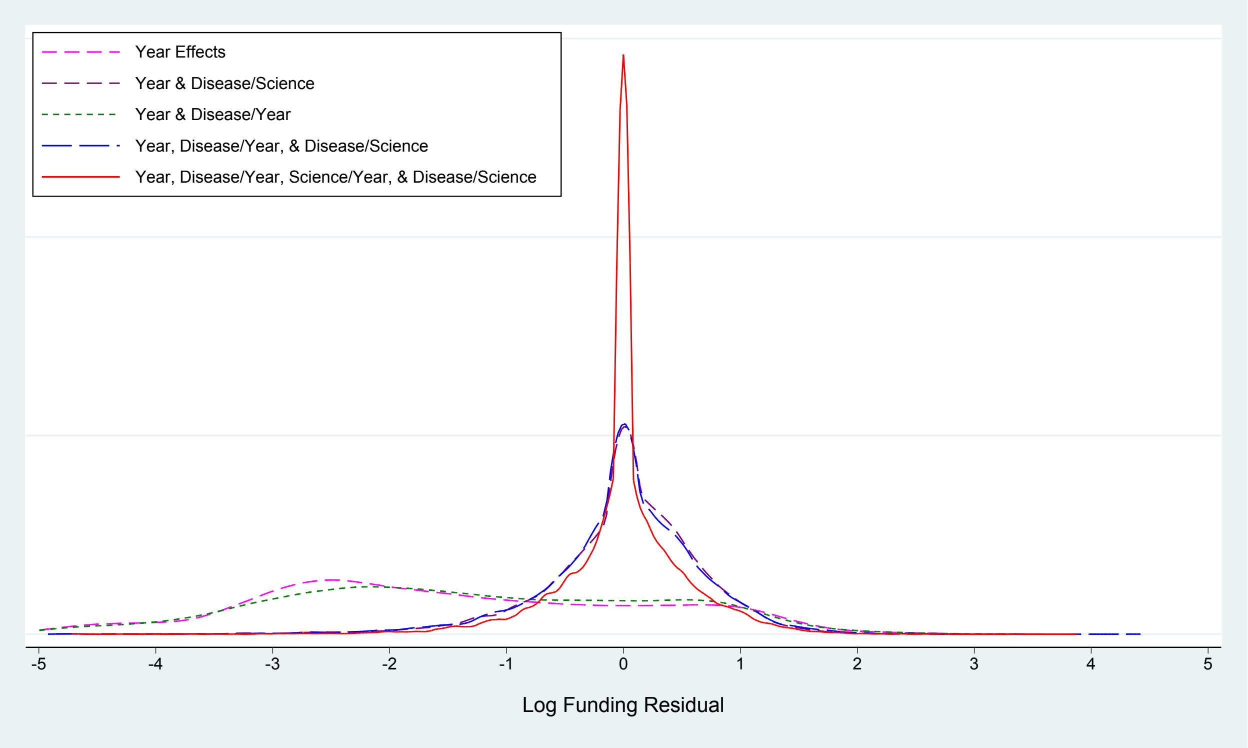


Figure A3: Residual Variation in DST Funding



Appendix B: “Life Science” Patents

To assess the impact of NIH funding, we need to define a universe of life science patents. While we do not want to impose strong restrictions on where NIH funding could have an effect (e.g., by looking in specific disease areas) focusing on a specific subset of the universe of issued patents is necessary for two reasons. From a substantive standpoint, it is important to assign most patents to one or more NIH research areas, and this would be infeasible were we to focus on all patents granted by the USPTO.[iv] From a pragmatic standpoint, linking NIH publications to patents requires probabilistic matching (see Appendix D2), and the rate of false positives is much lower if we restrict the set of potential matches.

To do so, we started with the 5,269,968 patents issued by the USPTO between 1980 and 2012. Then, using the NBER patent categorization described in Hall et al. (2001), we focused on patents in the classes belonging to NBER Categories 1 (Chemicals) and 3 (Drugs and Medical). This left 1,310,700 patents. Of these patents, 565,593 cite at least one non-patent reference. Using the algorithm described in Azoulay et al. (2012) and Sampat and Lichtenberg (2011) we determined that 312,903 patents cite an article indexed in PubMed. We refer to this set—patents in NBER Classes 1 and 3 that cite to at least one PubMed indexed article—as “life science patents.” Classes 1 and 3 cover a range of subcategories, listed in Table B1.

To provide a better sense of what this set includes, we took a random sample of 1,000 in the universe described above, and looked them up in the Thomson Reuters Innovation Database. This database includes information on the expert classification of each patent to one or more codes in the Derwent World Patents Index (Derwent World Patents Index 2012). Of the 1,000 patents, 656 had at least one DWPI “B” code, indicating they are in the “pharmaceuticals” category. According to DWPI 2012 (page 5) these pharmaceutical patents include:

- Compounds and proteins of pharmaceutical (or veterinary) interest;
- Compounds used as intermediates in the manufacture of pharmaceutical products;
- Compositions used for diagnosis and analysis in pharmaceuticals;
- Technologies dealing with production of tablets, pills, capsules, etc.
- Devices for dispensing pharmaceuticals.

Importantly, the “B” classes also include a range of biotechnology research tools and processes.

What about those without a “B” code, about one-third of the life science patents? The majority of these non-pharmaceutical patents are in five DWPI categories covering chemistry and medical devices: Class A (Polymers and Plastics), Class D (Food, Detergents, Water Treatment, and Associated Biotechnology), Class E (General Chemicals), Class S (Instrumentation, Measuring, and Testing), and Class P (General Human Necessities, including diagnosis/surgery).

Private sector vs. public sector patents. We are primarily interested in the effect of NIH funding on the rate of production of private-sector patents, excluding those assigned to public research entities such as universities, research institutes, academic medical centers, or government agencies (e.g., the intramural campus of NIH). This focus is justified by our desire to focus on disembodied knowledge flows. Since the Bayh-Dole act, life science academics have considerably increased their rate of patenting (Azoulay et al. 2007; 2009). Previous scholarship has documented the growing importance of patent-paper pairs (Murray and Stern 2007) where a given piece of academic knowledge gives rise to both an article and a patent listing the authors of the article as inventors and their employer (often a public institution) as assignee. Including these patents in our analyses would make the interpretation of our results (which emphasizes indirect spillovers of knowledge) difficult. To separate private-sector from public-sector patents, we adapted

[iv] e.g., class 150, “Purses, Wallets, and Protective Covers,” or Class 169, “Fire Extinguishers.”

Bronwyn Hall's patent assignee name matching algorithm to isolate private-sector assignees.[v] Using this method, we restrict the sample to 232,276 patents, or 74% of the life science patents (see Table 2 in the main body of the manuscript).

Patents on drug candidates and approved drugs. Though a substantial share of the life science patents are "pharmaceuticals" not all are therapeutic molecules or proteins. Even among those that are, there is substantial heterogeneity in value, since only a small share of drugs and biologics enter trials, and of these a small share receive marketing approval.

To examine heterogeneity of the effects of NIH funding, and to assess the effects on drug development, we isolated patents associated with important drugs and biologics. We began with all patents from current and archival versions of the FDA's Orange Book (officially named Approved Drug Product with Therapeutic Equivalence Evaluations). Since the 1984 Hatch-Waxman Act, branded firms are required to list on the Orange Book patent issued before drug approval with at least one claim covering a drug's active ingredient, formulation, or methods of use for approved indications. Though there is strong incentive to list patents issued after drug approval as well (Hemphill and Sampat 2012), strictly speaking this is not required. Moreover other drug patents (methods of manufacture, formulations not covering the marketed product, methods of use covering unapproved indications) are barred.

In parts of our analysis, we look at the effects of NIH funding on "important" life science patents associated with drugs that have been approved or entered late-stage clinical trials. For doing so, the Orange Book is restrictive, for several reasons. First, it does not list all patents on a drug, as already noted. Second, it does not list patents for all biologic drugs (since these drugs were historically covered by a division of the FDA exempt from Orange Book listing rules). Third, it does not include patents on drugs and biologics in late stage trials. Accordingly, we supplemented the patent list from the Orange Book with those from IMS Patent Focus, which includes patents on drugs and biologics in Phase III trials and above, and is less restrictive about the types of patents it includes than the Orange Book.[vi]

Together 4,718 of the 232,276 life science patents were listed in the Orange Book and/or IMS. We call this set of patents "Advanced Drug Candidates."

For welfare calculations, we multiply the effects of NIH patenting with measures of the value of new drugs. In order to do so, we need to isolate the patents associated with new molecular and biological entities (NMEs and NBEs), eliminating patents on drugs that associated with other drugs (e.g., line extensions) and unapproved drugs. This is not to say that drugs beyond NMEs and NBEs are unimportant. However, doing so is necessary since our measures of private and social value of drugs are based on data on new drugs that have been approved for marketing (as opposed to line extensions or unapproved drugs).

To construct this set, we used information on all NMEs and NBEs approved by the FDA between 1984 and 2012. Specifically, we collected information on all new molecular entities and biological license applications approved by the FDA. We searched for patents on each of these in the Orange Book using application numbers, and supplemented with searches in IMS patent focus using drug names. About 30 percent of these patents were listed both in the Orange Book and IMS, 67 percent in IMS only, and 3 percent in the Orange Book only. On average, there were 7.6 patents per drug in the dataset (7.3 for NME and 9.6 for biologics). After limiting to private sector patents (see above), we were left with a set of 1,999 private sector life science patents associated with new molecules and biologics.

[v] http://eml.berkeley.edu/~bhhall/pat/namematch.html

[vi] http://www.imshealth.com/deployedfiles/imshealth/Global/Content/Technology/Syndicated%20Analytics/Lifecycle%20and%20Portfolio%20Management/IMS_LifeCycle_Patent_Focus_Global_Brochure.pdf

Table B1: Relevant Patent Classes

Cat. Code	Category Name	Sub-Cat. Code	Sub-Category Name	Patent Classes
1	Chemical	11	Agriculture, Food, Textiles	8, 19, 71, 127, 442, 504
		12	Coating	106,118, 401, 427
		13	Gas	48, 55, 95, 96
		14	Organic Compounds	534, 536, 540, 544, 546, 548, 549, 552, 554, 556, 558, 560, 562, 564, 568, 570
		15	Resins	520, 521, 522, 523, 524, 525, 526, 527, 528, 530
		19	Miscellaneous	23, 34, 44, 102, 117, 149, 156, 159, 162, 196, 201, 202, 203, 204, 205, 208, 210, 216, 222, 252, 260, 261, 349, 366, 416, 422, 423, 430, 436, 494, 501, 502, 510, 512, 516, 518, 585, 588
3	Drugs & Medical	31	Drugs	424, 514
		32	Surgery & Medical Instruments	128, 600, 601, 602, 604, 606, 607
		33	Biotechnology	435, 800
		39	Miscellaneous	351, 433, 623

Appendix C: Why use DSTs as our Unit of Analysis?

Our conceptual model motivates our approach of tracing the patenting impact of research investments in each of r "research areas." In theory, a research area can be defined in many ways: narrowly at the level of an individual grant or broadly at the level of a disease. We choose to define research areas at the disease-science-time (DST) level for two reasons. First, DSTs represent coherent research areas and therefore capture a unit of funding variation that is policy-relevant. A more disaggregated level of analysis, such as the individual grant, has a different interpretation. To see this, consider an analogous regression at the grant level:

$$Patents_{\widetilde{g}} = \alpha_0 + \alpha_1 Funding_g + Controls_g + \varepsilon_g \tag{c1}$$

In Equation (c1), α_1 captures the impact of changes in funding for grant g on patenting outputs related to g (the comparison is implicitly to a grant g' that receives less funding). Since we typically only observe outcomes for funded grants, α_1 captures the intensive margin effect of budget increases for already funded grants, but would not incorporate any extensive margin impacts of funding additional grants.[vii]

To capture the impact of NIH funding at the extensive margin, one would need to examine patenting outcomes related to all grant applications, both funded and unfunded. This is challenging because unfunded applications do not generate acknowledgement data, making it difficult to track downstream outcomes using bibliometric linkages. Jacob and Lefgren (2011) circumvent this issue by studying the impact of NIH funding on the publication output of individual scientists. By focusing on the individual, they are able to link publications to scientists using authorship information rather than grant acknowledgements.

In our setting, however, estimating the impact of funding on individual scientists is of less policy interest. Fundamentally, policy makers care about overall innovation in a research area, not about whether a given applicant is funded. If an individual applicant is able to produce more research as a result of being funded, it does not necessarily generate more innovation in a research area because funding for one applicant may simply come at the expense of funding for other applicants with similar ideas: α_1 may therefore misstate the impact of NIH funding on overall innovation in a research area.

By aggregating to the level of a research area, we eliminate the concern that we simply identify the advantage that funded individuals have over unfunded ones. While it is still the case that funding for one DST could come at the expense of funding for other DSTs, this variation is more likely to impact the substantive content of innovation, relative to funding variation at the investigator level. This is because different D-S combinations correspond to different intellectual areas and are therefore less likely to support overlapping research ideas.[viii]

Policy makers are perhaps more interested in the impact of funding at the disease level, rather than the disease/science level. Our second reason for examining DSTs is that it is important for our identification strategy. Funding for a DST is a byproduct of funding decisions for diseases—made at the Congressional level—and scientific evaluations for individual grant applications—made by peer reviewers. Because no one explicitly allocates funding to a DST, we are able to exploit funding rules that generate incidental variation in funding across research areas. This is described in more detail in Section 3.4.

[vii] This is problematic because the NIH has a stated policy of funding the anticipated cost of an accepted research proposal, regardless of its peer review score. As as result, there is relatively less scope for increases in a grant's budget, conditional on being funded, to affect its innovative potential. More likely, when the NIH provides more funding for a research area, this funding is used to support additional grant applications that would not have been funded otherwise. These grants go on to produce publications that, in turn, later inspire commercial applications.

[viii] This does not address the concern that public funds may crowd out private investment. We discussed this form of crowd out in Section 2.1. Section 3.3 discusses how we address this issue empirically.

Appendix D1: Linking NIH Grants to Publications that Acknowledge NIH Support

The NIH asks of its grantees to include acknowledgements to agency support in any publications resulting from the grant, and to do so in a very specific format.[ix] Since the early 1980s, PubMed has recorded these acknowledgements in a separate field, and we use this data to link every grant in the NIH Compound Grant Applicant File (CGAF) with the publications that result. The process used to systematically map publication-to-grant linkages is relatively straightforward, but may be prone to measurement error. We discuss three potential issues below, and investigate the bias they might create for the reported results.

Dynamic linking inconsistency. In the vast majority of the cases, a grant acknowledgement provides a grant mechanism, a funding institute, and a grant serial number (as in `R01GM987654`), but typically no reference to a particular grant cycle. This limitation is potentially serious, since we need to be able to assign each publication to a particular DST, and not simply to a particular DS. Our final dataset relies on 987,799 unique publications that acknowledge a grant funded by NIH. 100% of these acknowledgements occur in a window of ten years before the year in which the article appeared in print. 93% of these publications are linked to the same grant within seven years, 83% within five years, and 47% within two years. To find the relevant grant cycle for each publication acknowledging a grant, we adopted the following procedure: (i) look up the year of publication t_{pub} for the acknowledging publication; (ii) create a five year "catchment window" $[t_{pub} - 5; t_{pub}]$; (iii) identify the most recent fiscal year t_{grant} in that window during which the grant was funded either as a new grant or as a competitive renewal; and (iv) link the publication to the funding institute identified in the grant acknowledgement, the study section that evaluated this grant according to NIH records, in the year t_{grant}.

While we cannot directly observe whether a publication was funded by a different grant cycle, we have verified that our benchmark results are robust to alternative choices for the length of the catchment window: $[t_{pub} - 2; t_{pub}]$, $[t_{pub} - 7; t_{pub}]$, $[t_{pub} - 10; t_{pub}]$.

Overclaiming of publications. NIH grant renewal is dependent on the research and publications stemming from that stream of funding. To our knowledge, NIH does not audit the acknowledgement trail systematically—this is left to the discretion of scientific review officers (the federal employees who manage the flow of information between reviewers in a particular study section and the NIH funding apparatus). Therefore, grantees may have an incentive to "over-attribute" publications—e.g., to credit some publications to the support of a grant, even if they were in fact enabled by other streams of funding. This raises the concern that increases in DST funding, even if exogenous, can lead us to identify more related patents, but only through the spurious channel of false attributions.

We believe that our results are unlikely to be driven by this behavior for two reasons. First, the vast majority of public biomedical research funding in the US comes from NIH, meaning that most scientists do not have meaningful amounts of funding from other sources to support their research.[x] While scientists often use grant funding to subsidize research projects that are not directly related to the topic of their grant, these projects should still be counted as a product of grant funding.

Second, if misattribution were driving our results, we would expect to see that boosts in NIH funding increase the number of patents directly linked to NIH funding (our "citation-linked" measure of patenting, see Table 6), but it would not increase the total number of patents in a DST's intellectual area (our "PMRA" measure of patenting, see Table 7). Our PMRA measure is designed to capture, through related publications, patents building on research related to a DST, regardless of whether that research is NIH-funded. If increases in

[ix] `http://grants.nih.gov/grants/acknow.htm`

[x] NIH accounted for 70% of the research budget of academic medical centers in 1997 (Commonwealth Fund Task Force on Academic Health Centers 1999); within Graduate Schools of Arts and Sciences, who cannot rely on clinical income to support the research mission, one would expect the NIH share to be greater still.

DST funding merely induce scientists to acknowledge these grants, we would not see the overall increase in innovation that we document in Tables 7 and 8.

Underclaiming of publications. Given the incentives created by the funding renewal decision, it seems unlikely that researchers would err by failing to credit their grant upon publication when they legitimately could. However, the number of NIH grant acknowledgements in PubMed jumps from 25,466 for articles appearing in 1980 to 56,308 for articles appearing in 1981 before stabilizing on a slow upward trend that correlates with the growth in funding thereafter. This is likely because the National Library of Medicine only gradually moved to a regime where grant acknowledgement data was systematically captured. Although the grants acknowledged in these early publications likely predate the start of our observation period (1980), this is an additional source of measurement error to which we must attend. In contrast to the second issue, however, there is no reason to suspect that erroneous capture of these data is related to the size of a DST. Year effects, included in all our specifications, should deal adequately with any secular change in NLM's propensity to accurately capture information related to grant acknowledgment.

Example. We illustrate the procedure with the case of particular publication, *Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions*, by Bowie et al., which appeared in the journal *Science* on March 16th, 1990 (see the left side of Figure D1-1). The publication credits grant support from NIH, specifically grant `AI-15706`. Despite the fact that this acknowledgement appears at the very end of the paper as the ultimate reference in the bibliography (reference #46 on page 1310), PubMed captures this data accurately (see the right side of Figure D1-1). Note that the acknowledgement omits the grant mechanism, as well as the leading zero in the grant serial number. These issues, which are typical in the PubMed grant acknowledgement data, turn out to be unimportant. In particular, the National Institute of Allergy and Infectious Diseases (NIAID, code-named `AI`) has only one grant with serial number `015706`: A project R01 grant first awarded to Robert T. Sauer, an investigator in the biology department at MIT, in 1979, and competitively renewed in 1982, 1987, 1992, 1997, and 2002. The grant was evaluated by the `BBCA` (Molecular and Cellular Biophysics) study section; its title is *Sequence Determinants of Protein Structure & Stability*, with a budget of $1,211,685 for the cycle that began in 1987, three years before the date of the publication above (whose last author is also Robert Sauer). As a result, the publication is linked to the DST corresponding to the combination `AI` (Institute)/`BBCA` (study section)/`1987` (year).

Distribution of Grant Acknowledgement Lag. In order for NIH funding to have an impact on total innovation, it must be that NIH funding enables the production of new knowledge. Yet, a common critique of NIH funding is that it is based on nearly completed research. Under this view, NIH funding essentially functions as "a prize for work well done" (Lazear 1997), rather than an input into future research effort. Below, we show that this is not entirely the case. While some publications may have been more or less complete at the time of grant application, there are many more publications that are based on research that was likely enabled by receiving the grant. Figure D1-2 shows that, in fact, only 3.27% of publications occur in the year of grant receipt (year 0), with an additional 13.81% occurring in year 1. Thus, the timing of publications seems to suggest that while some research is very advanced when a grant is funded, the funding is largely being used to generate new research.

Topic Drift. We can also examine the claim that NIH funding is generating new research by examining how closely the publications relate to the specific aims of the research proposal they acknowledge. To do so, we measure the extent to which the MeSH keywords used in publications that acknowledge the grant deviate, or at least drift away, from the MeSH keywords that characterize the research proposed in the initial grant application, based on its abstract. The Medical Text Indexer (MTI) developed by a team of researchers at the National Library of Medicine is a natural language processing tool that enables researchers to map full text paragraphs onto the MeSH controlled thesaurus.[xi] We batch process each grant abstract with the MTI tool to identify the core concepts and themes within each proposal. On average, MTI maps a grant to 13 MeSH terms (the median is also 13; the number of mapped terms ranges from one to 101).

[xi] `https://ii.nlm.nih.gov/MTI/`. MeSH is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity. There are 27,455 descriptors in the 2015 MeSH edition used in this manuscript. See also Appendix E.

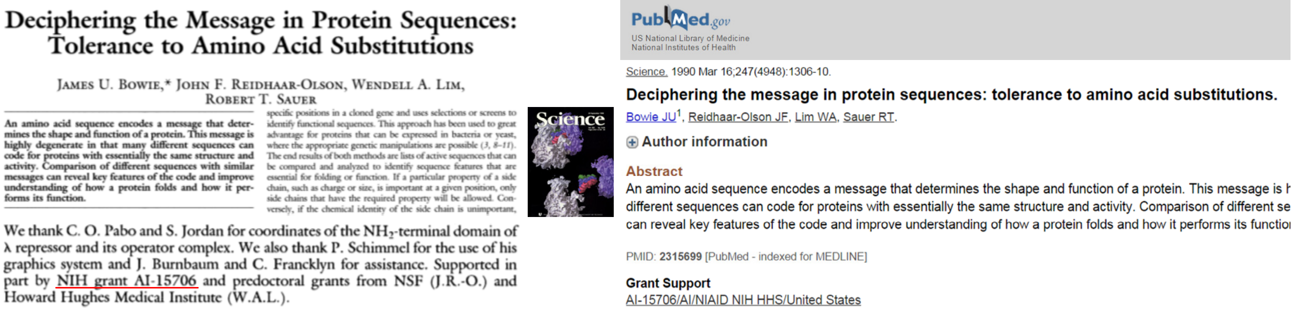
We can then compute, for each grant/publication pair, the number of MeSH terms that overlap between the grant and the publication, giving us a fine-grained measure of the similarity in content between the grant proposal and subsequent publications that acknowledge that grant funding. To analyze that data, we run count model specifications of the following type (which we estimate by Quasi Maximum Likelihood):

$$E\left[\#OVRLP\_KWRDS_{ij}|X_{ij}\right] = \#TTL\_KWRDS_i \times exp\Big[\gamma_{t(i)} + \delta_{IC(i)ss(i)} + \sum_{k=1}^{5} \beta_k 1_{t(j)-t(i)=k}\Big] \qquad (d1)$$

where $\#OVRLP\_KWRDS_{ij}$ is the number of MeSH keywords that are common between grant i and publication j (where j acknowledges i), $\#TTL\_KWRDS_i$ is the total number of MeSH keywords for grant i (so that the outcome variable is effectively the proportion of keywords that overlap between i and j), $\gamma_{t(i)}$ is a series of indicator variables for the fiscal years $t(i)$ in which grant i is funded, and $\delta_{IC(i)ss(i)}$ is a series of indicator variables for the institute $IC(i)$ that funded grant i and the study section $ss(i)$ that evaluated it. The coefficients of interest in this regression are the β_k's: they pin down the keyword drift for publications that appear k years after $t(i)$.

Table D1-3 shows the results. Figure D1-2 also provides the estimates of the β_k's in graphical form (corresponding to Column 3 of Table D1-1). It shows that the work published in the first year of the grant is most closely tied to the intellectual content of the grant proposal, and that the topic of publications in later years increasingly deviate from the ideas laid out in the investigator's original proposal. This is further evidence that much of the output attributed to a grant represents new research that was not (nearly) complete at the time of grant submission.

Figure D1-1: Example of Grant Acknowledgement



Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions

James U. Bowie,* John F. Reidhaar-Olson, Wendell A. Lim, Robert T. Sauer

An amino acid sequence encodes a message that determines the shape and function of a protein. This message is highly degenerate in that many different sequences can code for proteins with essentially the same structure and activity. Comparison of different sequences with similar messages can reveal key features of the code and improve understanding of how a protein folds and how it performs its function.

specific positions in a cloned gene and uses selections or screens to identify functional sequences. This approach has been used to great advantage for proteins that can be expressed in bacteria or yeast, where the appropriate genetic manipulations are possible (*3, 8–11*). The end results of both methods are lists of active sequences that can be compared and analyzed to identify sequence features that are essential for folding or function. If a particular property of a side chain, such as charge or size, is important at a given position, only side chains that have the required property will be allowed. Conversely, if the chemical identity of the side chain is unimportant,

We thank C. O. Pabo and S. Jordan for coordinates of the NH_2-terminal domain of λ repressor and its operator complex. We also thank P. Schimmel for the use of his graphics system and J. Burnbaum and C. Francklyn for assistance. Supported in part by NIH grant AI-15706 and predoctoral grants from NSF (J.R.-O.) and Howard Hughes Medical Institute (W.A.L.).

PubMed.gov
US National Library of Medicine
National Institutes of Health

Science. 1990 Mar 16;247(4948):1306-10.

Deciphering the message in protein sequences: tolerance to amino acid substitutions.

Bowie JU[1], Reidhaar-Olson JF, Lim WA, Sauer RT.

⊕ **Author information**

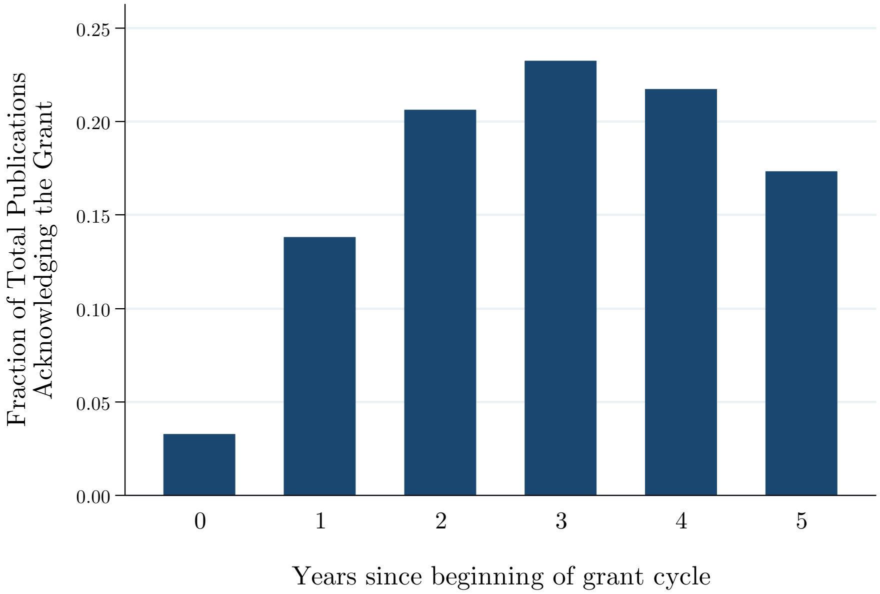
Abstract

An amino acid sequence encodes a message that determines the shape and function of a protein. This message is h
different sequences can code for proteins with essentially the same structure and activity. Comparison of different se
can reveal key features of the code and improve understanding of how a protein folds and how it performs its functio

PMID: **2315699** [PubMed - indexed for MEDLINE]

Grant Support
AI-15706/AI/NIAID NIH HHS/United States

Figure D1-2: Publication-Grant Acknowledgements Over Time



Note: We exclude from the sample of new grants those that were renewed, to avoid any confounding with the publications that accrue to competing continuations

Figure D1-3: Coefficient Estimates from Eqn. (d1)

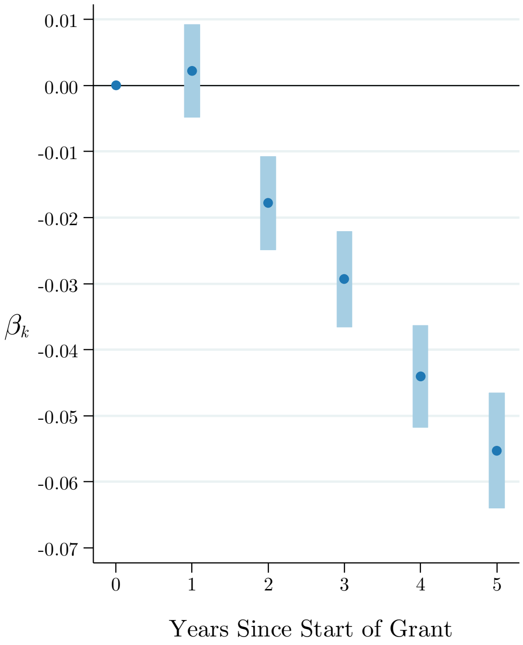


TABLE D1-1: KEYWORD DRIFT REGRESSIONS

	All Grants	All Grants	All Grants	New Grants Only	Competing Continuation Only	New Grants Only, never renewed
	(1)	(2)	(3)	(4)	(5)	(6)
One year after grant	0.004	0.003	0.003	-0.033**	-0.005	-0.039**
	(0.004)	(0.004)	(0.004)	(0.007)	(0.004)	(0.008)
Two years after grant	-0.017**	-0.017**	-0.017**	-0.059**	-0.024**	-0.067**
	(0.004)	(0.004)	(0.004)	(0.007)	(0.004)	(0.008)
Three years after grant	-0.027**	-0.028**	-0.029**	-0.075**	-0.030**	-0.079**
	(0.004)	(0.004)	(0.004)	(0.007)	(0.005)	(0.008)
Four years after grant	-0.040**	-0.041**	-0.043**	-0.084**	-0.057**	-0.085**
	(0.004)	(0.004)	(0.004)	(0.007)	(0.005)	(0.008)
Five years after grant	-0.050**	-0.052**	-0.054**	-0.098**	-0.064**	-0.098**
	0.004	0.003	0.003	-0.033**	(0.006)	(0.009)
Institute (IC) Fixed Effects	Incl.					
Study Section Fixed Effects		Incl.				
IC×Study Section Fixed Effects			Incl.	Incl.	Incl.	Incl.
Year Fixed Effects	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.
Nb. of ICs	17					
Nb. of Study Sections		465				
Nb. of IC×Study Sections Combos			2,049	1,913	1,187	1,815
Nb. of Grants	59,224	59,224	59,224	40,803	18,421	30,497
Nb. of Grant/Article Pairs	459,041	459,041	459,041	259,580	199,461	198,063
Log Likelihood	-840,770	-834,898	-830,791	-471,415	-356,626	-359,475

Estimates stem from QML Poisson specifications. An observation is a grant/application pair. The dependent variable is the number of MeSH keywords that overlap between the abstract of the grant at the application stage, and the published article acknowledging the grant. An offset for the total number of MeSH keywords for the grant application is included on the right hand side, so that the outcome variable is in effect the *fraction* of overlapping keywords between the grant and the publication. "n years after the grant" is an indicator variable that turns to one if the publication appeared n years after the grant was funded. The indicator variable corresponding to the year of the grant is omitted. Standard errors in parentheses, clustered at the level of the grant application. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$

Appendix D2: Linking PubMed References to USPTO Patents

We use patent-publication citation information to identify patents that build on NIH-funded research. Patent applicants are required to disclose any previous patents 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). Recent work (Sampat 2010, Alcácer, Gittleman and Sampat 2009), 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 (Alcácer and Gittleman 2006).

Building on the idea that citations in journal articles can be used to track knowledge flows, the pioneering work of Francis Narin and colleagues at CHI research in the 1970s used references on the front page of patents to scientific articles (part of the "non-patent references" cited in the patent), to examine the "science dependence" of technology (Carpenter and Narin 1983) and linkages between science and technology (Narin and Olivastro 1992, 1998). This research also found that life-science patents cite non-patent references more intensively than do firms from other fields. In the economics literature, the count of non-patent references (or the share of non-patent references in all citations) has been used a proxy for the extent to which patents are science-based (e.g., Trajtenberg et al. 1997).

For our purposes, leveraging patent-to-publication citation information is appealing for two reasons. First, publications, rather than patents, are the main output of academic researchers (Agrawal and Henderson 2002); second, the vast majority of patent-to-paper citations, over 90 percent, come from applicants rather than examiners, and are thus more plausibly indicators of real knowledge flows than patent-to-patent citations (Lemley and Sampat 2012; Roach and Cohen 2013).[xii] Our paper builds on and extends this approach, by linking life-science patents back to the articles that cite them, and the specific NIH grants funding these articles.

Determining whether patents cite publications is more difficult than tracing patent citations: while the cited patents are unique seven-digit numbers, cited publications are free-form text (Callaert et al. 2006). 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.

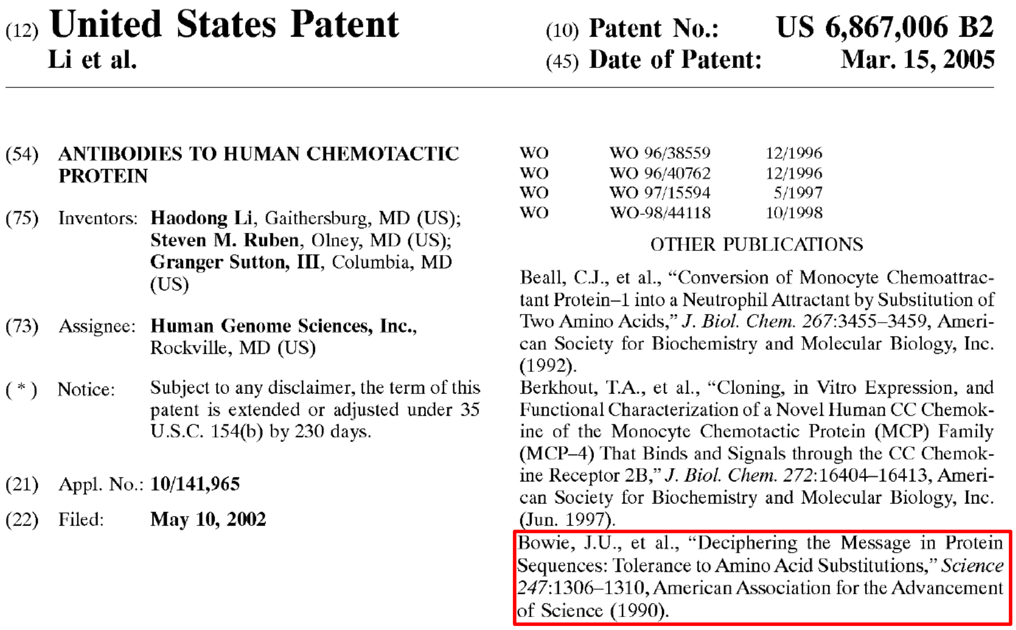
To address these difficulties, we developed a matching algorithm that compared each of several PubMed fields — first author, page numbers, volume, and the beginning of the title, publication year, or journal name — to all references in all biomedical and chemical patents issued by the USPTO since 1976. Biomedical patents are identified by technology class, using the patent class-field concordance developed by the National Bureau of Economic Research (Hall, Jaffe, and Trajtenberg 2001). We considered a dyad to be a match if four of the fields from PubMed were listed in a USPTO reference.

Overall, the algorithm returned 1,058,893 distinct PMIDs cited in distinct 322,385 patents. Azoulay, Graff Zivin and Sampat (2012) discuss the performance of this algorithm against manual searching, and tradeoffs involved in calibrating the algorithm.

[xii] Ozcan and Bryan (2017) stress the distinction between "in-text" as opposed to "front-page" citations. In their view, in-text citations play a role similar to that of citations in academic papers and tend to come from inventors, whereas the front-page citations we leverage could reflect the thinking of anyone involved in the invention or preparation of the patent document, including examiners and attorneys. However, at the time we began this project, the extraction and parsing of in-text citations at scale represented a difficult technical challenge.

Example. We illustrate the procedure with the case of particular patent, #6,687,006, issued on March 15, 2005 and assigned to the biopharmaceutical firm Human Genome Sciences, Inc. In the section of the patent entitled OTHER PUBLICATIONS, we can find a citation to "Bowie, J.U., et al., Deciphering the Message in Protein Sequences. . . ," precisely the publication we took as an example in Appendix D1. Our text-parsing algorithm identifies this reference and associates it with PUBMED article identifier `2315699`. As a result, this patent will participate in the patent count corresponding to the DST `AI/BBCA/1987` (see Appendix D1).

Figure D2: Example of Patent-to-Publication Citation



(12) **United States Patent**
Li et al.

(10) **Patent No.: US 6,867,006 B2**
(45) **Date of Patent: Mar. 15, 2005**

(54) **ANTIBODIES TO HUMAN CHEMOTACTIC PROTEIN**

(75) Inventors: **Haodong Li**, Gaithersburg, MD (US); **Steven M. Ruben**, Olney, MD (US); **Granger Sutton, III**, Columbia, MD (US)

(73) Assignee: **Human Genome Sciences, Inc.**, Rockville, MD (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 230 days.

(21) Appl. No.: **10/141,965**

(22) Filed: **May 10, 2002**

WO	WO 96/38559	12/1996
WO	WO 96/40762	12/1996
WO	WO 97/15594	5/1997
WO	WO-98/44118	10/1998

OTHER PUBLICATIONS

Beall, C.J., et al., "Conversion of Monocyte Chemoattractant Protein–1 into a Neutrophil Attractant by Substitution of Two Amino Acids," *J. Biol. Chem. 267*:3455–3459, American Society for Biochemistry and Molecular Biology, Inc. (1992).

Berkhout, T.A., et al., "Cloning, in Vitro Expression, and Functional Characterization of a Novel Human CC Chemokine of the Monocyte Chemotactic Protein (MCP) Family (MCP–4) That Binds and Signals through the CC Chemokine Receptor 2B," *J. Biol. Chem. 272*:16404–16413, American Society for Biochemistry and Molecular Biology, Inc. (Jun. 1997).

Bowie, J.U., et al., "Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions," *Science 247*:1306–1310, American Association for the Advancement of Science (1990).

Appendix E: PubMed Related Citations Algorithm [PMRA]

One of our outcome measures (described in more detail in Appendix G) captures all patents in the intellectual vicinity of an NIH funding area. A crucial input in the construction of this measure is the National Library of Medicine's PubMed Related Citations Algorithm (PMRA), which provides a way of determining the degree of intellectual similarity between any two publications. The following paragraphs were extracted from a brief description of PMRA:[xiii]

> *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.*
>
> *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.*

In Table E1, 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

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

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`.

Table E1: 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	

Appendix F: Structure of the Disease/Science Panel Dataset

As explained in Section 3.1, the level of analysis chosen for the econometric exercise is the disease/science/year level. With 17 NIH institutes (the "D" in DST), 624 standing study sections (the "S"), and 25 years (the "T"), one might expect our analytical sample to 265,200 DST observations (and 10,608 distinct DS research areas), but a quick perusal of the tables reveal only 14,085 DSTs, or 5.31% of the total number of potential DSTs (respectively 2,942 actual DS, or 27.73% of the total number of potential DS). Why such a seemingly high number of missing DSTs? This appendix (i) clarifies that there are different types of "missing DSTs"; (ii) explains why most of these missing DSTs are missing for benign reasons; and (iii) investigates the robustness of our results to the concern that some DSTs are missing for substantive reasons. Figure F1 provides a graphical representation of the structure of our panel dataset. For example, the purple line corresponds to the combination of the National Institute of Allergy and Infectious Diseases [NIAID] and the Molecular and Cellular Biophysics [BBCA] study section. In every year between 1980 and 2005, NIAID awarded at least three grants that were reviewed by the BBCA study sections. Therefore, in this case, all the 26 potential DSTs are accounted for.

Missing DSTs: A Taxonomy. A full 191,650 DSTs (72.27%) are missing from our data because the corresponding DS combinations are never observed. One can think of these instances as cases where the pairing of a disease with a science area would be intellectually incongruous. Consider, for instance, the pairing of the National Institute of Mental Health (NIMH) and the Tropical Medicine and Parasitology [TMP] study section. Not only are there no grants awarded by NIMH that were reviewed by the TMP study section, there is also no evidence of any *unfunded* grant application reviewed by TMP whose author designated NIMH as the funding institute. This case is represented by the orange dotted line in Figure F1.

We are left with 2,942 disease/science research areas that awarded at least one grant in at least one year during the observation period, or $2,942 \times 25 = 73,550$ potential DSTs. 55,058 of these 73,550 DSTs are missing because many study sections are not in continuous existence between 1980 and 2005: our sample is unbalanced. At regular intervals in the history of NIH, study sections have been added, dropped, split, or merged to accommodate changes in the structure of scientific disciplines as well as shifting patterns of momentum for some research areas, relative to others. DSTs that are missing because of the natural life cycle of study sections need not concern us, as long as we make the reasonable assumption that every grant application, at a given point time, has a study section that is fit to assess its scientific merits.

Figure F1 displays three examples that fall into this category. Consider first the red line, corresponding to the combination of the National Heart, Lung, and Blood Institute [NHLBI] and the Physiology [PHY] study section. The Physiology study section ceased to exist in 1998, so the NHLBI/PHY combination "misses" seven DSTs. What happened to the applications received in 2000 that would have been reviewed by the PHY study section had they been received in 1998? The answer is that newly created study sections, such as Integrative Physiology of Obesity and Diabetes [IPOD] or Skeletal Muscle Biology and Exercise Physiology [SMEP] almost certainly reviewed them. Similarly, the combination of NIDDK and the Biochemistry study section (which was born in 1991) is "missing" observations between 1980 and 1990, while the combination between NIA and the Neurology B-2 study section is missing observations between in 1980, 1981, 1982, and observations from 1998 to 2005. Notice that in all three of these cases, DSTs are not missing "in the middle," but only at the extremities.

Potentially more problematic for our analysis is the case of DS combinations that display intermediate sequences of starts and stops. Consider for example the blue line in Figure F1, which corresponds to the combination of the National Cancer Institute [NCI] and the Reproductive Biology [REB] study section. Ten of the potential 22 observations for this combination are missing between 1980 and 2001 (the REB study section ceased to exist after 2001). The story is similar for the combination of the National Eye Institute [NEI] and the Epidemiology and Disease Control 1 [EDC-1] study section. All together, out of the 2,942 DS combinations in our dataset, 2,101 (71.41%) are contiguous, and 841 are "hole-y" (for a total of 4,407 missing DSTs). We are concerned about these cases because it is possible that research was proposed in these areas, and that at least some of it got done (maybe thanks to alternative sources of funding), leading to patents downstream which we have no way of linking back to publicly-funded research efforts. One piece of evidence that allays these concerns is that in the great majority of cases (80%), we do not observe any application in the corresponding DSTs—if no funds were awarded, it is because no research was in fact proposed to NIH for funding consideration. In light of this fact, it seems harder to imagine that patents could be linked to these areas via some alternative method which does not rely on bibliometric linkages.

Robustness check: Contiguous DSTs. In addition, we probe the robustness of our results by replicating the main specifications while restricting the sample to the set of 2,101 intact, contiguous DS areas, for a total of 7,966 DSTs (57 percent of our original dataset). In Table F1, we report the results of specifications modeled after those used to generate the estimates in Table 6, our benchmark set of results. Using this approach, we obtain coefficients that are numerically very similar to those presented in Table 6, and estimated very precisely.

In summary, the great majority of the DSTs that appear to be missing from our data are not really missing, but rather, not in existence. And the small minority of DSTs that could genuinely said to be "missing" cannot be expected to change our conclusions, since limiting the analysis to the set of intact DS areas yields identical results.

Figure F1: A Taxonomy of DSTs

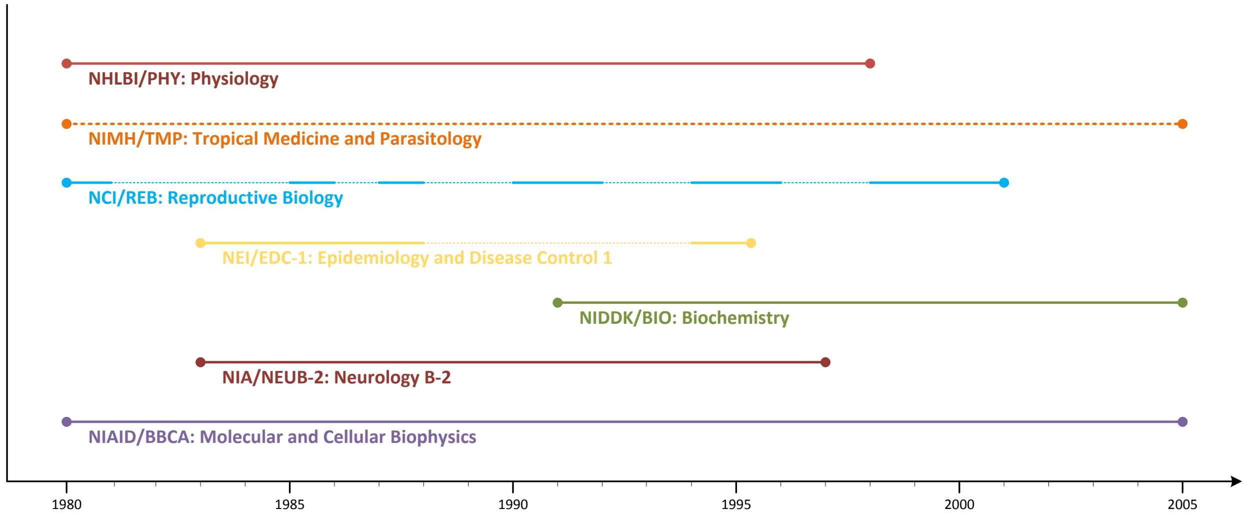


Table F1: Contiguous Disease-Science Categories Only

	First Stage		**Citation Linked**		**Total Related**	
	DST Funding (×$10 mln.)		Mean=14.2; SD=19.89		Mean=27.2; SD=28.5	
			OLS	IV	OLS	IV
	(1)		(2)	(3)	(4)	(5)
Windfall Funding (×$10 mln.)	1.031*** *(0.195)*	DST Funding ($10 mln.) Mean=4.49; SD=4.44	2.458*** *(0.799)*	2.138 *(1.368)*	3.671*** *(1.237)*	2.251 *(1.608)*
		Elasticity	0.796	0.649	0.604	0.349
R^2	0.918		0.751	0.550	0.861	0.631
Observations	7,966		7,966	7,966	7,966	7,966
Year FEs	Incl.		Incl.	Incl.	Incl.	Incl.
Disease × Science FEs	Incl.		Incl.	Incl.	Incl.	Incl.
Disease × Year FEs	Incl.		Incl.	Incl.	Incl.	Incl.
Science × Year Linear Trends	Incl.		Incl.	Incl.	Incl.	Incl.
Application Controls	Incl.		Incl.	Incl.	Incl.	Incl.

Note: See notes to Tables 6, 7, and 8 for details about the sample and IV. Application controls include (i) FEs for the number of applications that a DST receives; (ii) FEs for the number of applications associated with a DST that are also in a 50-grant window around the relevant IC payline, as well as (iii) cubics in the average raw and rank scores of applications associated with a DST that are also in a 50-grant window around the payline. Only contiguous disease-science areas, as defined in the text, are included.

Standard errors in parentheses, two-way clustered at the disease and science level ($^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$).

Appendix G: Linking NIH Research Areas (DSTs) to Patents

We begin by linking the universe of funded NIH grants between 1980 and 2005 to the set of articles that it supports using grant acknowledgement data from PubMed. We then link these publications to private-sector patents using two alternative procedures; in turn, the outcome measures that build on these procedures are designed to answer slightly different questions about the impact of NIH funding. The first measure asks whether private firms build on NIH-funded research in their patented inventions. The second measure asks whether NIH funding leads to the net creation of private-sector patents that would not have otherwise been developed. We describe the two procedures below; the overall data and variable construction process is summarized in Figure 1 in the main body of the manuscript.

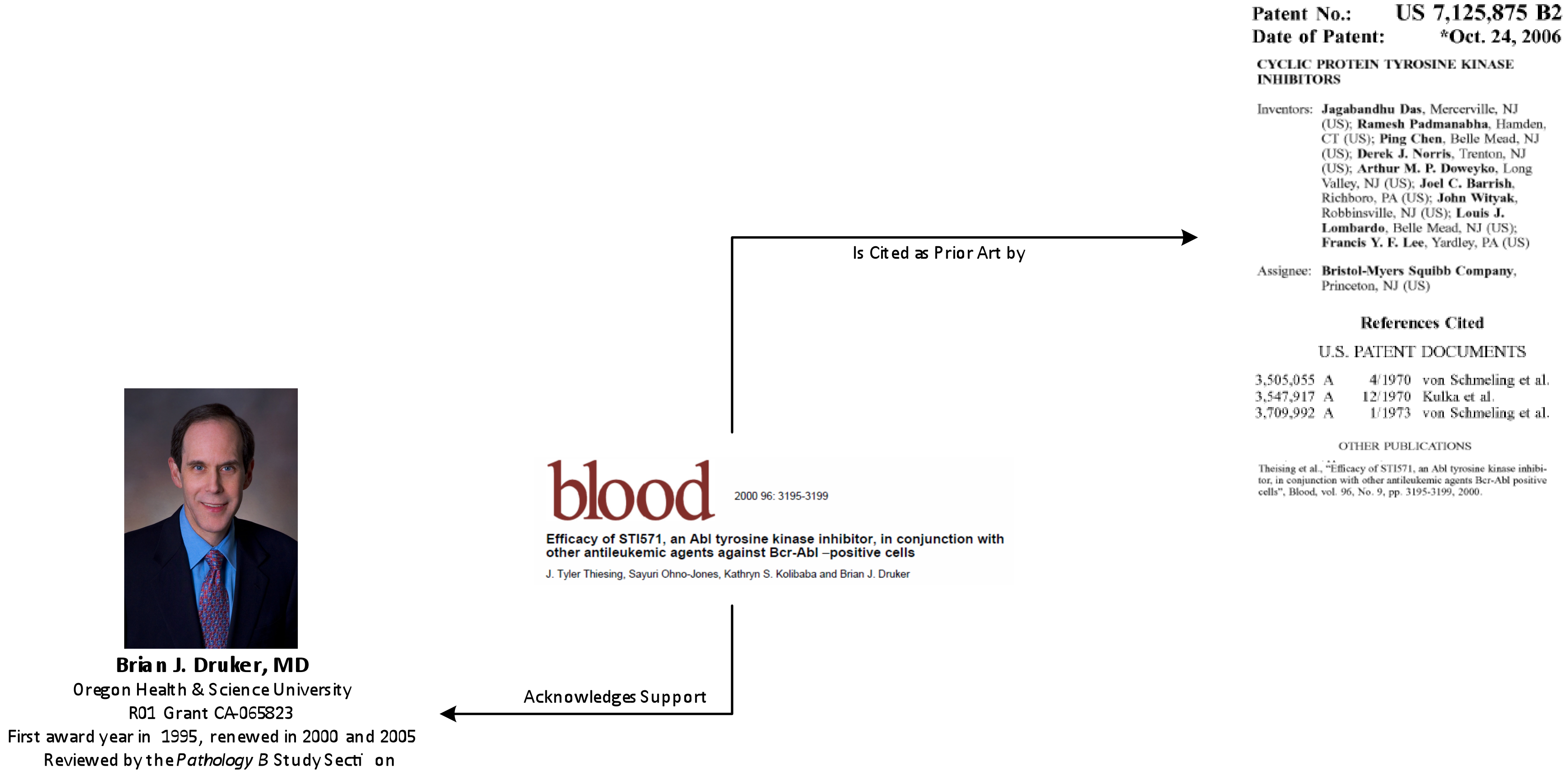
Patents building on NIH-funded research: Direct linkages. We consider how many patents explicitly build on NIH-funded research. Figure G1 illustrates the procedure with an example. In its first three years of funding, the NIH grant CA-065823 was acknowledged by four publications, among which is the article 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.

Patents building on NIH-funded research: Indirect linkages. The second procedure links a patent to a grant if this patent refers to a publication that is "intellectually similar" to a publication that does acknowledge NIH funding. In other words, these linkages are indirect: from a grant, to a publication that acknowledges it, to the publications that are proximate in intellectual space, to the patents that in turn cite these related publications. The grant linked to patents in this way delineates the pool of research expenditures that is intellectually relevant for the creation of these patents, even in the absence of a direct linkage between the patent and the grant. Figure G2 illustrates this process. Patent number 6,894,051 was 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, see Appendix E] indicates that this publication is closely related to PubMed article 9389713, which acknowledges funding from NIH grant CA-0011422. Using these second procedure, we can link the vast majority of life science patents to an NIH disease-science area. In other words, most patents cite publications that are similar to publications that acknowledge NIH funding.

Under the indirect procedure, the same patent can be linked to many distinct grants through the inclusion of related publications. In our regressions, we adjust for this by weighting patents in the following way: regardless of what outcome measure we use, if a patent is linked to N grants, it counts as $1/N$ of a patent in each NIH research area. This means that a patent is restricted to being counted once across all NIH research areas to which it is linked.

Aggregation from the individual grant-patent linkage up to the NIH research area level [DST]. The procedures outlined above describe how to link patents to specific NIH grants. However, we do not perform the econometric analysis at the grant level. Rather, we aggregate grants up to the disease/science/time (DST) level, as explained in Section 3. Understanding the impact of NIH funding at the DST level offers conceptual advantages apart from its econometric ones. Because DSTs are defined to be intellectually coherent units in which knowledge generated by one projects is likely to benefit other projects, our estimate of the impact of NIH funding on DST-level outcomes, then, captures the benefits of potential complementarities between research in the same area. This would not be true of an analysis of grant-level funding on grant-level patenting.

Figure G1: Grant-Patent Match, Direct Linkages



Brian J. Druker, MD
Oregon Health & Science University
R01 Grant CA-065823
First award year in 1995, renewed in 2000 and 2005
Reviewed by the *Pathology B* Study Section

Acknowledges Support

blood 2000 96: 3195-3199

Efficacy of STI571, an Abl tyrosine kinase inhibitor, in conjunction with other antileukemic agents against Bcr-Abl –positive cells

J. Tyler Thiesing, Sayuri Ohno-Jones, Kathryn S. Kolibaba and Brian J. Druker

Is Cited as Prior Art by

Patent No.: US 7,125,875 B2
Date of Patent: *Oct. 24, 2006

CYCLIC PROTEIN TYROSINE KINASE INHIBITORS

Inventors: **Jagabandhu Das**, Mercerville, NJ (US); **Ramesh Padmanabha**, Hamden, CT (US); **Ping Chen**, Belle Mead, NJ (US); **Derek J. Norris**, Trenton, NJ (US); **Arthur M. P. Doweyko**, Long Valley, NJ (US); **Joel C. Barrish**, Richboro, PA (US); **John Wityak**, Robbinsville, NJ (US); **Louis J. Lombardo**, Belle Mead, NJ (US); **Francis Y. F. Lee**, Yardley, PA (US)

Assignee: **Bristol-Myers Squibb Company**, Princeton, NJ (US)

References Cited

U.S. PATENT DOCUMENTS

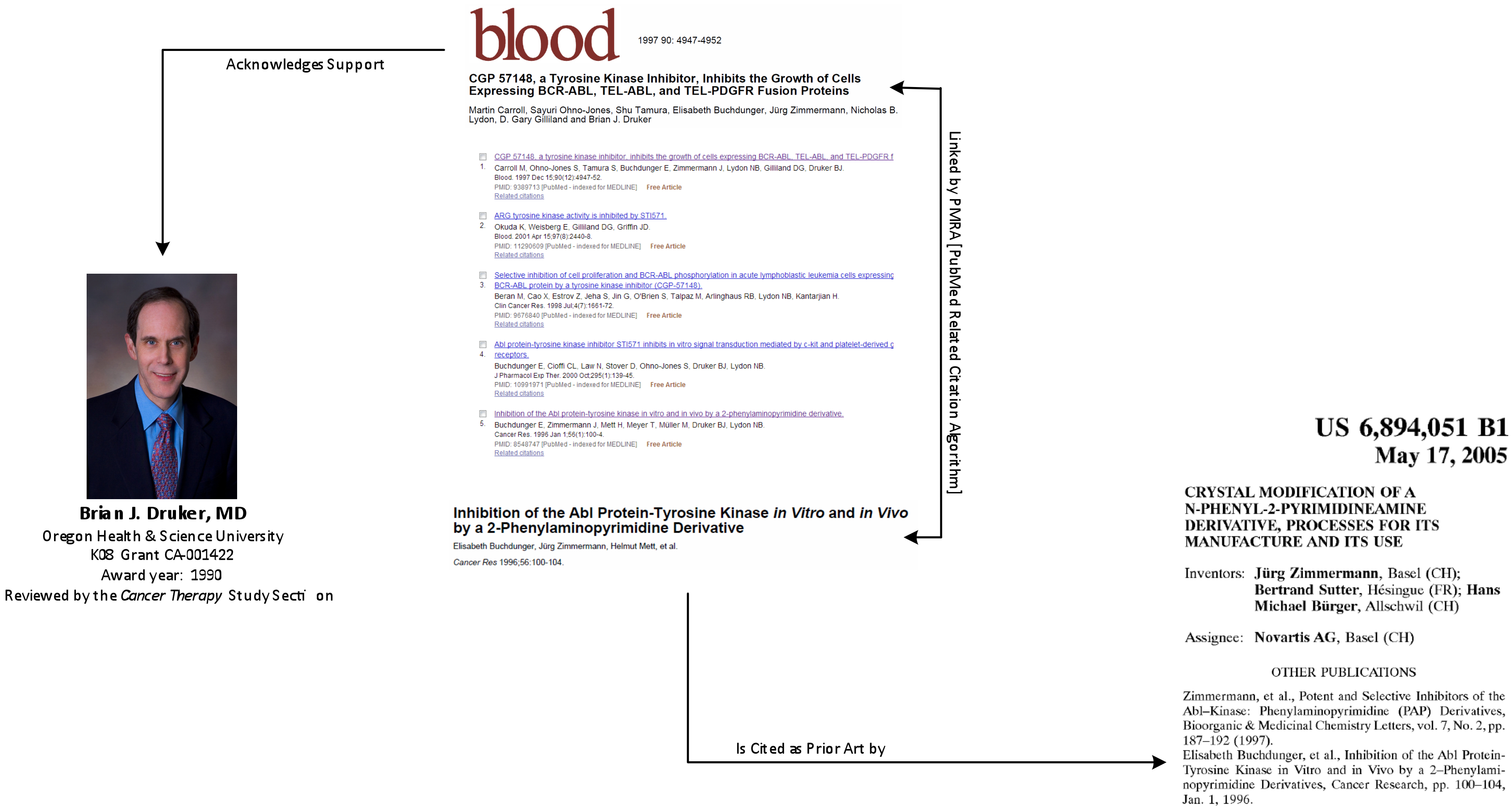
3,505,055	A	4/1970	von Schmeling et al.
3,547,917	A	12/1970	Kulka et al.
3,709,992	A	1/1973	von Schmeling et al.

OTHER PUBLICATIONS

Theising et al., "Efficacy of STI571, an Abl tyrosine kinase inhibitor, in conjunction with other antileukemic agents Bcr-Abl positive cells", Blood, vol. 96, No. 9, pp. 3195-3199, 2000.

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 G2: Grant-Patent Match, Indirect Linkages



Acknowledges Support
Linked by PMRA [PubMed Related Citation Algorithm]
Is Cited as Prior Art by
Brian J. Druker, MD
Oregon Health & Science University
K08 Grant CA-001422
Award year: 1990
Reviewed by the *Cancer Therapy* Study Section

blood 1997 90: 4947-4952

CGP 57148, a Tyrosine Kinase Inhibitor, Inhibits the Growth of Cells Expressing BCR-ABL, TEL-ABL, and TEL-PDGFR Fusion Proteins

Martin Carroll, Sayuri Ohno-Jones, Shu Tamura, Elisabeth Buchdunger, Jürg Zimmermann, Nicholas B. Lydon, D. Gary Gilliland and Brian J. Druker

□ 1. CGP 57148, a tyrosine kinase inhibitor, inhibits the growth of cells expressing BCR-ABL, TEL-ABL, and TEL-PDGFR f
Carroll M, Ohno-Jones S, Tamura S, Buchdunger E, Zimmermann J, Lydon NB, Gilliland DG, Druker BJ.
Blood. 1997 Dec 15;90(12):4947-52.
PMID: 9389713 [PubMed - indexed for MEDLINE] Free Article
Related citations

□ 2. ARG tyrosine kinase activity is inhibited by STI571.
Okuda K, Weisberg E, Gilliland DG, Griffin JD.
Blood. 2001 Apr 15;97(8):2440-8.
PMID: 11290609 [PubMed - indexed for MEDLINE] Free Article
Related citations

□ 3. Selective inhibition of cell proliferation and BCR-ABL phosphorylation in acute lymphoblastic leukemia cells expressing BCR-ABL protein by a tyrosine kinase inhibitor (CGP-57148).
Beran M, Cao X, Estrov Z, Jeha S, Jin G, O'Brien S, Talpaz M, Arlinghaus RB, Lydon NB, Kantarjian H.
Clin Cancer Res. 1998 Jul;4(7):1661-72.
PMID: 9676840 [PubMed - indexed for MEDLINE] Free Article
Related citations

□ 4. Abl protein-tyrosine kinase inhibitor STI571 inhibits in vitro signal transduction mediated by c-kit and platelet-derived g receptors.
Buchdunger E, Cioffi CL, Law N, Stover D, Ohno-Jones S, Druker BJ, Lydon NB.
J Pharmacol Exp Ther. 2000 Oct;295(1):139-45.
PMID: 10991971 [PubMed - indexed for MEDLINE] Free Article
Related citations

□ 5. Inhibition of the Abl protein-tyrosine kinase in vitro and in vivo by a 2-phenylaminopyrimidine derivative.
Buchdunger E, Zimmermann J, Mett H, Meyer T, Müller M, Druker BJ, Lydon NB.
Cancer Res. 1996 Jan 1;56(1):100-4.
PMID: 8548747 [PubMed - indexed for MEDLINE] Free Article
Related citations

Inhibition of the Abl Protein-Tyrosine Kinase *in Vitro* and *in Vivo* by a 2-Phenylaminopyrimidine Derivative

Elisabeth Buchdunger, Jürg Zimmermann, Helmut Mett, et al.

Cancer Res 1996;56:100-104.

US 6,894,051 B1
May 17, 2005

CRYSTAL MODIFICATION OF A N-PHENYL-2-PYRIMIDINEAMINE DERIVATIVE, PROCESSES FOR ITS MANUFACTURE AND ITS USE

Inventors: **Jürg Zimmermann**, Basel (CH); **Bertrand Sutter**, Hésingue (FR); **Hans Michael Bürger**, Allschwil (CH)

Assignee: **Novartis AG**, Basel (CH)

OTHER PUBLICATIONS

Zimmermann, et al., Potent and Selective Inhibitors of the Abl–Kinase: Phenylaminopyrimidine (PAP) Derivatives, Bioorganic & Medicinal Chemistry Letters, vol. 7, No. 2, pp. 187–192 (1997).
Elisabeth Buchdunger, et al., Inhibition of the Abl Protein-Tyrosine Kinase in Vitro and in Vivo by a 2–Phenylaminopyrimidine Derivatives, Cancer Research, pp. 100–104, Jan. 1, 1996.

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.

Appendix H: Conceptual Framework

We would like to identify how private-sector, patented innovations follow from public investments in fundamental knowledge. In this appendix, we present a stylized framework that motivates our empirical strategy. Let the space of ideas $\Re$ consist of R distinct fields indexed by the letter r. Our starting point is an innovation production function in which patenting output in a research area ν at time τ is determined by knowledge inputs from a variety of research areas r, at potentially different times t.[xiv] This can be summarized in matrix notation as:

$$P = \Omega K \tag{$h1$}$$

where P is a vector with elements $p_{\nu\tau}$, K is a vector of knowledge inputs k_{rt}, and Ω is a matrix with elements $\omega_{\nu\tau,rt}$ describing how knowledge inputs in research area r at time t impact innovation in area ν at time τ. The number of patents in area ν at time τ can be expressed as a function of the relative importance of the knowledge inputs k_{rt}:

$$p_{\nu\tau} = \sum_{r,t\leq\tau} \omega_{\nu\tau,rt} k_{rt} \tag{$h2$}$$

While Equation $(h2)$ has the familiar look of a typical knowledge production function in log-linearized form, it departs from it in one essential respect. The key inputs, investments in science, are public goods. Their non-rivalrous nature means that each input can be "consumed" by multiple production processes. Indeed, one insight can lead to patents in multiple areas. Their non-excludability, which obviates the need to "purchase" inputs, makes it particularly difficult to ascertain which knowledge inputs are employed in the production of any given innovation.

To overcome these challenges, the literature has traditionally made several restrictions on the structure of the matrix Ω. First, innovation in area ν is assumed to draw on knowledge stocks related to the same area only, ignoring potential spillovers. This means that the elements of the production matrix $\omega_{\nu\tau,rt} = 0$ for all $\nu \neq r$. Second, a fixed lag structure typically governs the relationship between the stream of expenditures $k_{rt}, k_{r,t+1}, ..., k_{r\tau}$ and $p_{r\tau}$. Together, these assumptions entail that public investments may only impact private innovation in the same area, within a well-defined time horizon.[xv] A generic concern with this type of approach is that it will fail to capture any benefits that may accrue to seemingly unrelated research areas or with unexpected time lags. In the case of basic R&D, where the intent is to enhance the understanding of building block relationships with often unanticipated, and potentially far-reaching implications, these assumptions may be particularly limiting. For example, much of the research underlying the development of anti-retrovirals used in the treatment of HIV infection in the 1990s was originally funded by the National Cancer Institute in the 1950s and 1960s, at a time when research on the causes of cancer centered on viruses.[xvi]

In this paper, we address these concerns by relaxing the traditionally restrictive assumptions about the matrix Ω. Instead of focusing on all the research areas r that contribute to patenting in a particular area ν, as described by Equation $(h1)$, we trace the impact of a single knowledge input, k_{rt} on patenting in a range of areas $\widetilde{r}$ and time periods $\widetilde{t}$. This can be thought of as the "dual" problem relative to the "primal problem" described in Equation $(h2)$:

$$P_{\widetilde{rt}} = \alpha_{rt} k_{rt} \tag{$h3$}$$

where $P_{\widetilde{rt}} = \sum_{p\in S_{rt}} p_{rt}$. S_{rt} consists of all patents, regardless of area, that draw on research done in area r at time t. The coefficient α_{rt} describes the impact of a unit increase in research input on aggregate innovation.

[xiv] This approach is standard in the literature. See, *inter alia*, Pakes and Griliches (1980) and Hall et al. (1986).

[xv] Toole (2012), for instance, regresses patenting in a given disease-year on 12 years of lagged funding for that same disease.

[xvi] Gleevec provides another example: Varmus (2009) recounts that that Ciba-Geigy was working with scientists of the Dana Farber Cancer Institute to find drugs that would block the action of a tyrosine kinase that contributes to atherosclerosis in blood vessels, a disorder that is very different from CML. The development of Gleevec also relied heavily on knowledge about the genetic causes of CML that was established in the 1960s and 70s (e.g., Nowell and Hungerford 1960). In this case, the availability of treatment lagged behind basic research by over forty years. In other settings, basic research percolates almost immediately into applications work, such as when publications and patents are released in tandem (Murray 2002).

We are interested in estimating the average of these α_{rt} terms across all research areas and time periods. This represents the average return to public investments in biomedical research, taking into account potentially unanticipated spillovers across areas and over time.

The key to estimating Equation $(h3)$ is defining the set of patents S_{rt} that draw on k_{rt} as an input. Instead of assuming a simple structure for S_{rt}, we implement a flexible procedure relying on bibliometric linkages to uncover the relevant connections. In Appendix I, we compare estimates using our approach with a more traditional production function estimation approach.

Appendix I: Impact of NIH Funding, Traditional Fixed Lag Approach

Our approach differs from traditional estimates of the impact of public R&D funding in that, instead of making *ex ante* assumptions about where and when to look for its effects, the structure of the bibliometric linkages naturally reveals, *ex post*, where and with what kind of lags the effects are being felt.

Relative to the traditional approach, one might worry that our estimates reflect in part idiosyncrasies of the linking process, rather than the effect of funding. For example, if scientists over-attribute publications to their grants in order to appear productive, then DSTs with more grants will exhibit a higher number of bibliometric linkages to patents, regardless of whether the funding in these DSTs actually contributed to the development of those patents. This will artificially inflate our estimates of the impact of NIH funding on citation-linked patents in Table 6 (though it should not increase the total number of patents in a research area, as estimated in Table 7).

In this appendix, we repeat our empirical exercise using the traditional method of examining the relationship between funding in a year and patenting in subsequent years, assuming a series of fixed lags between funding and innovation. The results are broadly similar in magnitude to those obtained in the benchmark specification using our preferred "ex post" methodology, with some important caveats that we detail below. We continue to favor the *ex post* approach because bibliometric linkages offer a variety of benefits, including the ability to track innovations across disease areas.

In order to follow the traditional approach, we must find a way to identify the research area(s) that is/are likely to be responsible for a particular patented innovation. Toole (2012), for instance, assumes that funding in a given disease area impacts drug development in the same disease area, and then goes on to examine the impact of funding on new drug approvals using a distributed lag structure. Here we replicate the spirit of his work, but with two important twists: (i) our outcome variable is patents, not drug approvals, and patents are more challenging to associate *ex ante* with disease areas; (ii) we perform the exercise both using a more aggregated disease level to partition funding into research areas (the unit of analysis used in Toole (2012) and most of the literature to date), and also using a finer-grained disease/science level, which parallels the level of analysis used throughout the main body of the manuscript.

Patent mapping. We create an *ex ante* mapping of patents to research areas by exploiting the fact that NIH grants sometimes directly generate patented innovations. The 1980 Bayh-Dole Act created incentives for researchers and their institutions to patent the discoveries derived from federal funding. 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. In total, 1,799 NIH grants generated 1,010 distinct patents.[xvii] We examine the three digit main patent class in each of these 1,010 patents to create a probabilistic mapping of each patent class to research areas, where a research area is defined as a funding institute (roughly isomorphic to a broad disease area, see Appendix A). For each funding institute/patent class combination, we construct the fraction of that class' patents that are supported by funding for the institute associated with that disease:

$$F_{cd} = \frac{\#\text{ of class } c \text{ patents acknowledging funding from NIH Institute } d}{\#\text{ class } c \text{ patents}}$$

So for instance, if a patent is part of a class that includes 100 patents, 10 of which are supported by the National Cancer Institute (NCI) and 15 of which are supported by the National Heart Lung and Blood Institute (NHLBI), then it will count as 0.10 of a patent to the NCI and 0.15 to the NHLBI. Note that this mapping only relies on the empirical distribution of Bayh-Dole patents across funding institutes. Within our universe of 315,982 life science patents, 269,839 (85%) have a main patent class that is represented in the

[xvii] While these patents are also issued between 1980 and 2012, they do not overlap with those in our main analyses because they are overwhelmingly assigned to universities or to the NIH intramural campus, as opposed to private-sector firms.

much smaller set of Bayh-Dole patents. We use our class-to-research area mapping to allocate each of these 269,385 patents in one or more funding institute using the weights described above.

We proceed in a similar fashion to create a mapping between disease/science areas and patent classes:

$$F_{cds} = \frac{\#\text{ of class } c \text{ patents acknowledging funding from NIH Institute } d \text{ and reviewed by study section } s}{\#\text{ class } c \text{ patents}}$$

The next step is to construct the number of patents in a research area issued in a particular year t. In the case of research areas defined at the disease level:

$$Patents_{dt} = \sum_{c} F_{cd} \cdot \#\text{ of patents in class } c \text{ issued in year } t$$

In the case of research areas defined at the disease/science level:

$$Patents_{dst} = \sum_{c} F_{cds} \cdot \#\text{ of patents in class } c \text{ issued in year } t$$

i.e., the number of patents issued in a particular year t as the proportion of class c's patents that can be mapped to the NIH research area defined by disease d and science area s. Since the weights F_{cd} and F_{cds} are time-invariant, the allocation of patents to research areas is not influenced by changes in funding and other potentially endogenous factors.

Estimation. Using these outcome variables, we estimate the following regressions:

$$\text{Patents}_{d,t+k} = \alpha_0 + \alpha_{1k}\text{Funding}_{dt} + \delta_d + \gamma_t + \varepsilon_{dt} \text{ for } k = 1, \ldots, 20 \tag{1}$$

at the disease level, and

$$\text{Patents}_{ds,t+k} = \beta_0 + \beta_{1k}\text{Funding}_{dst} + \delta_{ds} + \mu_{dt} + \nu_{st} + \varepsilon_{dt} \text{ for } k = 1, \ldots, 20 \tag{2}$$

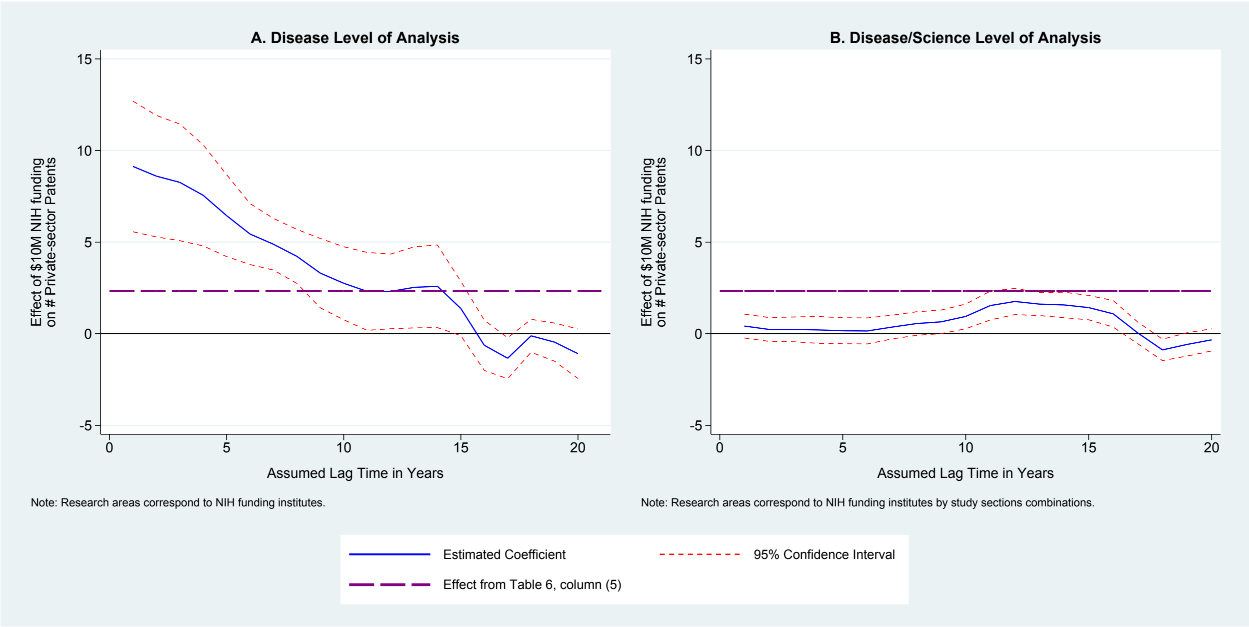
at the disease/science level. The coefficients of interests are α_{1k} and β_{1k} for $k = 1, \ldots, 20$, and we display them graphically in Panels A and B of Figures I1, together with their 95% confidence intervals. For comparison, we represent our benchmark result—from Table 6, column (5)—as an horizontal line (since this estimate does not depend on pre-specified lags).

Results. Figure I1, Panel A shows that, averaged over all the possible lags, the *ex ante* approach using the disease level of analysis yields effects whose magnitudes are quite comparable to our main *ex post* benchmark (2.33 patents for a \$10 million boost in funding), and in fact surprisingly similar to it for lags of 11 to 14 years. Interestingly, however, the *ex ante* approach appears to "overshoot" in the short run, and "undershoot" in the long run. For instance, we estimate that a \$10 million boost in funding to an institute would increase private-sector patenting by about 10 patents in the next year. Given the time needed both to perform the research and to complete the patent prosecution process, a near-term return to public funding of this magnitude seems highly implausible. This highlights some of the concerns with the fixed-lag approach; by assuming different lag structures, one could get very different estimates of the impact of funding, not all of which appear plausible. For this reason, we prefer the *ex post* approach.

Figure I1, Panel B, repeats the fixed lag approach using the DST as unit of analysis, paralleling our primary specifications. Here, the *ex ante* approach yields smaller estimates relative to the *ex post* benchmark (though the differences are not statistically significant for lags 11 to 14). The lack of congruence between the results in Panel A and Panel B makes sense in light of the different levels of analysis used to generate these figures. In Panel B, we do not capture in the outcome variable any patent that can be mapped *ex ante* to the same disease area unless it can also be mapped to the same science area. This is obviously very restrictive. Panel B therefore highlights another benefit of the *ex post* approach: it allows one to track innovation across research areas where *ex ante* mappings would simply assume the lack of any relation between funding and downstream innovation.

To explore the hypothesis that our disease/science level regressions yield smaller coefficients because they restrict associated patents to be ones in a narrow disease/science area, we reproduce Figure I1 using a slightly broader measure of "science area." Study sections are organized into slightly broader categories known as integrated review groups (IRGs). In our data, there are 624 study sections, and 327 IRGs. Figure I2 plots coefficients from a version of Equation (2), with patents matched to the relevant IC-IRG. Here, we find larger estimates, within range of our *ex post* results for at least some of the lags.

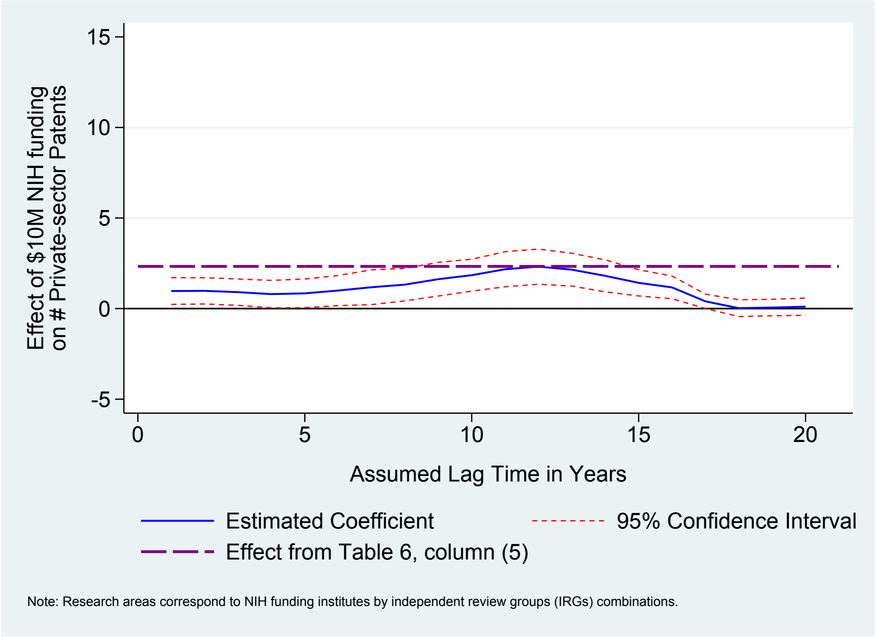
Figure I1: Effect of NIH Funding on Private-Sector Patenting *ex ante* Approach with Fixed Lags



Note: Research areas correspond to NIH funding institutes.

Note: Research areas correspond to NIH funding institutes by study sections combinations.

Figure I2: Reprise of Figure I1, Panel B but with broader, IRG-based level measure of science area



Note: Research areas correspond to NIH funding institutes by independent review groups (IRGs) combinations.

Appendix J: Identification Robustness Checks

The fixed effect estimation strategy outlined in Section 3 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. In the main body of the paper, we showed that funding for a given DS does not appear correlated with funding for the same science area in different diseases. We also showed that windfall funding does not appear to be correlated with past or future windfalls, nor with non-windfall funding. In this Section, we present several further tests of our identifying assumptions.

First, in Figure J1, we provide descriptive evidence that there is wide variation in windfall funding across DSTs: 28% of the 14,085 DSTs in our sample receive windfall funding. Table J1 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 10 year lags in the number of patents that acknowledge NIH funding from a disease/science area, as well as the number of patents that cite publications supported by that area or which cite publications related to those funded by that area. We also examine whether windfall funding is correlated with lagged applicant scores or lagged windfall funding. Again, we fail to reject the null hypothesis in all these cases.

Next, Table J2 presents the IV estimates and the corresponding reduced-form estimates side-by-side. We find that the reduced-form coefficient estimates for windfall funding (Columns 1 and 3) are quite similar in magnitude with the IV coefficient estimates for actual funding in a DST, instrumented by windfall funding (Columns 2 and 4).[xviii]

One potential concern is that the NIH occasionally funds grant applications out of the order in which they are scored. As discussed in Section 3.2 and Appendix A, peer review rules at the NIH make it difficult for NIH's component Institutes to direct resources to DSTs. ICs, however, do have the discretion to fund grant applications as exceptions to the standard scoring rules; approximately four to five percent of grants are funded in this way. 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.[xix] This practice has the potential of generating a correlation between DST funding and its unobserved potential.

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. To do this, we modify how we construct our surprise windfall instrument and compute deviations from expected funding using actual funding amounts coming only from grants that are funded in order. Table J3 presents our findings using this alternative strategy. Using this instrument, we find that an additional $10 million in ordered funding increases net patenting by 3.8, compared with 3.6 in our main OLS specification and 2.7 in our preferred IV specification. The implied elasticities of all these estimates are similar.

[xviii] We note that our IV estimates are more precise than our reduced form, which is somewhat unusual. Intuitively, this can happen when the first stage and reduced form estimates are correlated; in this case, taking the ratio of the two estimates can reduce noise that is separately present in both the first stage and the reduced form, making the IV relatively more precise.

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

Figure J1: Distribution of Windfall Funding

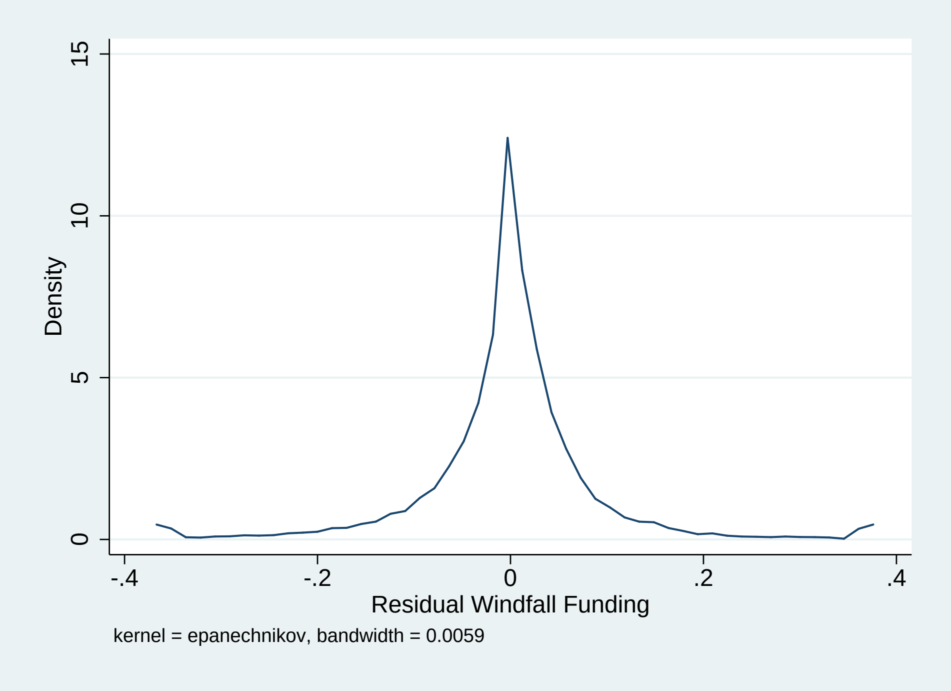


Table J1: Correlation Between Windfall Funding and Measures of DST Quality

RHS includes 10 Years of Lags for:	***F*-stat of Joint Significance**
# of Patents Citing Research Acknowledging NIH Funding	1.07
Raw and Rank Scores	1.110
All of the Above	1.090

Note: Each observation is a disease/science/time (DST) combination. Each column reports a regression of our windfall funding instrument on measures of DST input and output quality. We controls for the same set of variables as in our most detailed specification in Tables 6 and 7.

Table J2: Reduced Form and IV Estimates

	Citation Linked		**Total Related**	
	Mean=12.82; SD=19.17		Mean=24.8; SD=28.0	
	Reduced Form	IV	Reduced Form	IV
	(1)	(2)	(3)	(4)
Windfall Funding (\$10 mln.) Mean=0.20; SD 0.52	2.498 *(2.284)*		2.931 *(2.403)*	
DST Funding (\$10 mln.) Mean=4.06; SD 4.87		2.274* *(1.228)*		2.668* *(1.368)*
R^2	0.711	0.511	0.836	0.624
Observations	14,085	14,085	14,085	14,085
Year FEs	Incl.	Incl.	Incl.	Incl.
Disease × Science FEs	Incl.	Incl.	Incl.	Incl.
Disease × Year FEs	Incl.	Incl.	Incl.	Incl.
Science × Year Linear Trends	Incl.	Incl.	Incl.	Incl.
Application Controls	Incl.	Incl.	Incl.	Incl.

Note: See notes to Table 6 for details about the sample, and to Table 8 for details about the instrument. Application controls include (i) FEs for the number of applications that a DST receives; (ii) FEs for the number of applications associated with a DST that are also in a 50-grant window around the relevant IC payline, as well as (iii) cubics in the average raw and rank scores of applications associated with a DST that are also in a 50-grant window around the payline.

Standard errors in parentheses, two-way clustered at the disease and science level (*p $<$ 0.10, **p $<$ 0.05, ***p $<$ 0.01).

Table J3: Instrumenting DST Funding with Funding for Grants Funded in Order Only

	First Stage		**Citation Linked**		**Total Related**	
	DST Funding		Mean=12.82; SD=19.17		Mean=24.8; SD=28.0	
			OLS	IV	OLS	IV
	(1)		(2)	(3)	(4)	(5)
DST Funding, Grants in Order Only (×$10 mln.)	0.634*** *(0.080)*	DST Funding ($10 mln.) Mean=4.06; SD=4.36	2.408*** *(0.649)*	3.480*** *(0.943)*	3.625*** *(0.807)*	3.762*** *(0.968)*
		Elasticity	0.763	1.102	0.593	0.616
R^2	0.953		0.735	0.502	0.862	0.629
Observations	14,085		14,085	14,085	14,085	14,085
Year FEs	Incl.		Incl.	Incl.	Incl.	Incl.
Disease × Science FEs	Incl.		Incl.	Incl.	Incl.	Incl.
Disease × Year FEs	Incl.		Incl.	Incl.	Incl.	Incl.
Science × Year Linear Trends	Incl.		Incl.	Incl.	Incl.	Incl.
Application Controls	Incl.		Incl.	Incl.	Incl.	Incl.

Note: The outcome variables are fractional patent counts. The instrument is the total amount of funding for awarded DST grants that are funded in order of score (i.e., which are not exceptions) within a 50-grant window, minus the expected amount of funding. Expected amount of funding takes the number of grants within that window (excluding applications that were funded out of order) and multiplies this by 1/2 times the average amount awarded to each funded grant in that disease-year. For more details on this sample, see the notes to Tables 6. Application controls include (i) FEs for the number of applications that a DST receives; (ii) FEs for the number of applications associated with a DST that are also in a 50-grant window around the relevant IC payline, as well as (iii) cubics in the average raw and rank scores of applications associated with a DST that are also in a 50-grant window around the payline.

Standard errors in parentheses, two-way clustered at the disease and science level (*p < 0.10, **p < 0.05, ***p < 0.01).

Appendix K: Alternative Specifications and Samples

Another set of robustness checks explores the implications of using alternative specifications and/or samples. All of the results in the body of the manuscript rely on sample weights, where each observation is weighted by the yearly average of awarded grants for a disease-by-science combination. Weighting is justified by our desire to prevent small DSTs from influencing the results too strongly, relative to large DSTs. Table K1 replicates the benchmark results of Table 8, but without weighting the sample. The difference in results between the weighted and unweighted version are minor. Though we believe that weighting by average DST size (measured by yearly number of grants in a DS) is appropriate, this choice does not affect our substantive conclusions.

Our main results rely on linear fixed effects and IV models; this may be problematic because patenting outcomes tend to be very skewed. Table K2 shows that our results hold in logs as well. Columns 1 and 2 rerun our main results for our first outcome measure, the number of patents that cite research funded by that DST; Column 1 uses the same set of controls as our main fixed effects estimates from Table 6 and Column 2 uses our IV controls. On the subsample of DSTs with nonzero patenting under this measure (63% of our main DST sample), we show that a one percent increase in DST funding increases patenting by between 0.8 and 0.9 percent. This is similar, though slightly higher, to the elasticities we find in our main results. Columns 3 and 4 repeat this exercise using our second outcome measure, the total number of related patents. Again, we find elasticities between 0.8 and 0.9, which are slightly higher than in our main results.

A shortcoming of the log-log parametrization is that it entails dropping 1,062 DST observations that are not linked to any private-sector patent. Many researchers have dealt with the problem of excess zeros through the use of *ad hoc* transformations of the dependent variable, such as $log(1 + y)$. Because of Jensen's inequality, the estimates corresponding to the transformed outcome are difficult to compare numerically to the estimates when the dependent variable is left untransformed. A better approach in our view is to estimate our specifications using Quasi-Maximum Likelihood Poisson, which is consistent under very mild regularity conditions and allows us to deal with the skewness of the outcome variable as well as with its mass point at zero (Wooldridge 1997; Santos Silva and Tenreyro 2006). Table K3 estimates our benchmark specifications using the QML-Poisson approach, with one important caveat. The likelihood function fails to converge when we fully saturate the model with disease-by-science fixed effects, disease-by-year fixed effects, and science-by-year fixed effects. We are able to achieve convergence and to generate QML estimates when including disease-by-year fixed effects (columns 1 and 3), and when we combine disease-by-year and disease-by-science fixed effects (columns 2 and 4). While these specifications are not strictly analogous to the most saturated models presented in Tables 6 and 7, they remain very close to them in spirit. The magnitudes obtained with the Poisson parametrization, and the elasticities they imply, are numerically similar to the elasticities computed in Tables 6 and 7.

Next, we restrict our sample to only a subset of NIH's component institutes (ICs). In our paper, we refer to Institutes as representing diseases or body systems. In practice, however, not all ICs are organized in this way. The National Institute on Aging, for instance, does not focus on diseases in the same way as the National Cancer Institute. Other Institutes are even more difficult to think of as representing a disease or body system. For instance, the National Human Genome Research Institute (NHGRI) focuses on particular scientific techniques rather than on a set of related diseases. The fact that ICs do not always correspond to diseases does not impact the validity of our instrument, which relies only on the fact that ICs span study sections and vice versa.

It does, however, raise the concern that the IC by year fixed effects in our specifications may not, for some grants, be capturing changes in the innovative or commercial potential of their actual disease areas. For example, if the NHGRI funds research on cancer genetics, the IC by year FE associated with this grant will control for time varying potential in genetics, but not in cancer more generally. In Table K4, we restrict our sample to ICs that are more closely affiliated with disease and body system areas. Columns 1 and 2 reproduce our main results; Columns 3 and 4 exclude three science-focused ICs (general medicine, genome research, and biomedical imaging), and Columns 5 and 6 keep only ICs clearly associated with a disease or body system.

We also replicate our design using public-sector patents—rather than private-sector patents—as the outcome variable. Public-sector patents are patents assigned to universities, non-profit foundations and research institutes, government entities (including the intramural research campus of the NIH), and academic medical centers. There are fewer such patents: only 47,461 can be linked "directly" through a publication they cite to a DST, compared with 91,462 private-sector patents. Our analysis focuses on the private sector because the meaning of citations to publications contained in patents is likely different for biopharmaceutical firms, and corresponds more closely to the idea of a knowledge spillover. Life science academics sometimes patent, and yet other times found biopharmaceutical firms, typically with a license to a patent assigned to the researcher's academic employer. In other words, the same individuals might obtain NIH funding, publish results from research made possible by this funding, and choose to apply for a patent whose claims will cover these very same results. We might still be interested in assessing the magnitude of the patent-to-funding elasticity in this case. Although the question of crowd-out arises in the case of public-sector patents as well, it is probably capturing a different dynamic.

These objections notwithstanding, Table K5 replicates our benchmark results with public-sector patents as the outcome. Though the coefficient estimates differ from those displayed in Table 6, the elasticities are quite similar.

A final set of robustness analyses separates linkages that rely on the circulation of human capital at the interface between academia and industry, from those where the mechanism for knowledge transfer is the mere availability of research results in the scientific literature. We do so by examining the overlap between (i) the names of the PIs for each grant (typically a single individual); (ii) the names of the authors on publications that acknowledge the grant (this will pick up the names of trainees whom we would not expect to be PIs but could be the carriers of the knowledge produced by the grant); and (iii) the roster of inventor names on the patent. We call a linkage "disembodied" if there is no overlap between the inventor names on the patent and either the PI of the grant or any author of a publication that acknowledges that grant.

Overlap between the name of the PI on the grant and the list of inventor on the patent is vanishingly rare (less that 0.2% of linkages); however, overlap between the authors of papers that acknowledge the grant and names of inventors of the patent is less rare (about 7.5% of linkages with private-sector patents). Table K6 breaks down our benchmark set of results according to a name overlap split. Excluding the cases of linkages that involve author/grantee/inventor name overlap produces elasticities very close to those we obtain when ignoring the distinction between embodied and disembodied linkages. When we focus exclusively on the set of "embodied" linkages, the magnitudes of the OLS estimates are much smaller, but this simply reflects that the pool of patents with name overlap eligible for linking is also smaller. In contrast, the elasticities are quite similar (column 2a vs column 1a; column 4a vs. column 3a). The IV estimates for patents with embodied linkages are smaller and imprecisely estimated.

Table K7 provides a version of our benchmark results excluding from the universe of patents eligible to be linked any of the following: (i) Bayh-Dole patents (patents with a government interest statement); (ii) patents with author/grantee/inventor name overlap; and (iii) "hybrid" patents, i.e., patents assigned to a private sector firm as well as a public sector/non-profit/academic organization, or even a private individual (so-called "unassigned" patents). The results are once again quite close numerically to those presented in Table 6 and 7.[xx]

[xx] The remaining patents should be immune to Thursby et al.'s (2009) observation that up to a third of "academic" patents (in the sense that the team of inventors are academics) are not assigned to a university, but rather unassigned, or assigned to a private-sector firm, maybe in contravention to the formal rules adopted by most academic institutions.

Table K1: Benchmark Results with No Weights

	First Stage		**Citation Linked**		**Total Related**	
	DST Funding (× $10 mln.)		Mean=4.72; SD=12.56		Mean=9.25; SD=18.68	
			OLS	IV	OLS	IV
	(1)		(2)	(3)	(4)	(5)
Windfall Funding (×$10 mln.)	1.047*** *(0.275)*	DST Funding (×$10 mln.) Mean=1.52; SD=2.91	2.094*** *(0.454)*	3.029*** *(0.576)*	3.367*** *(0.718)*	3.136*** *(0.874)*
		Elasticity	0.674	0.975	0.553	0.515
R^2	0.905		0.639	0.290	0.853	0.476
Observations	14,085		14,085	14,085	14,085	14,085
Year FEs	Incl.		Incl.	Incl.	Incl.	Incl.
Disease × Science FEs	Incl.		Incl.	Incl.	Incl.	Incl.
Disease × Year FEs	Incl.		Incl.	Incl.	Incl.	Incl.
Science × Year Linear Trends	Incl.		Incl.	Incl.	Incl.	Incl.
Application Controls	Incl.		Incl.	Incl.	Incl.	Incl.

Note: See notes to Tables 6 and 7 for details about the sample, and Table 8 for details about the instrument. Application controls include (i) FEs for the number of applications that a DST receives; (ii) FEs for the number of applications associated with a DST that are also in a 50-grant window around the relevant IC payline, as well as (iii) cubics in the average raw and rank scores of applications associated with a DST that are also in a 50-grant window around the payline. Elasticities are evaluated at the sample means.

Standard errors in parentheses, two-way clustered at the disease and science level (*p < 0.10, **p < 0.05, ***p < 0.01).

Table K2: Log Patents-Log Funding Parametrization

	Log(# Citation Linked Patents)		**Log(# Related Patents)**	
	(1)	(2)	(3)	(4)
Log(DST Funding)	0.790^{***} *(0.129)*	0.874^{***} *(0.093)*	0.899^{***} *(0.034)*	0.899^{***} *(0.029)*
R^2	0.937	0.837	0.954	0.909
Observations	8,880	8,880	13,013	13,013
Full OLS Controls	Incl.		Incl.	
Full IV Controls		Incl.		Incl.

Note: The dependent variable in Columns 1 and 2 is the log of citation-linked fractional patents, with zeros treated as missing. There are 14,085-8,880=5,205 DSTs that do not produce research ever cited by a patent. Full OLS controls are the controls used in the most saturated specification of Tables 6 and 7 (see notes to those tables). Full IV controls are those used in Table 8. Log(#Related Patents) is the log of the number of fractional patents related by our second outcome measure, using PMRA. There are 14,085-13,023=1,062 DSTs that do not produce resarch that is related to a patent in our sample.

Standard errors in parentheses, two-way clustered at the disease and science level ($^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$).

Table K3: Poisson Specification

	# Citation-Linked Patents		# Related Patents	
	(1)	(2)	(3)	(4)
DST Funding (×$10 mln.) Mean=4.06; SD=4.36	0.091*** *(0.007)*	0.084*** *(0.013)*	0.088*** *(0.007)*	0.074*** *(0.009)*
% Change in Dep. Var. for additional $10 mln. in DST Funding	*9.53%*	*8.76%*	*9.20%*	*7.68%*
Pseudo-R^2	0.776	0.537	0.886	0.630
Observations	14,085	14,085	14,085	14,085
IC × Year FEs	Incl.	Incl.	Incl.	Incl.
IC × Study Section FEs		Incl.		Incl.

Note: See notes to Tables 6 and 7 for details about the sample.

Standard errors in parentheses, two-way clustered at the disease and science level ($^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$).

Table K4: Disease- or Body System-Specific ICs Only

	All ICs		**Excluding Science-based ICs**		**Core Disease/Body System ICs**	
	Mean=24.8; SD=28.0		Mean=24.10; SD=27.82		Mean=23.81; SD=26.80	
	OLS	IV	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)	(5)	(6)
DST Funding (×$10 mln.)	3.625***	2.668*	3.360***	3.525***	3.302***	2.845
Mean=4.06; SD=4.36	*(0.807)*	*(1.368)*	*(0.304)*	*(1.233)*	*(0.732)*	*(2.978)*
Elasticity	0.593	0.437	0.566	0.594	0.563	0.485
R^2	0.862	0.624	0.898	0.676	0.897	0.680
Observations	14,085	14,085	12,432	12,432	10,382	10,382

Note: Columns 1 and 2 reproduce the results from our primary sample. Columns 3 and 4 remove three IC based on methods or scientific topics. These are the National Institute of General Medical Sciences (NIGMS), the National Human Genome Research Institute (NHGRI), and the National Institute of Biomedical Imaging and Bioengineering (NIBIB). Columns 5 and 6 further restrict to a core set of ICs focused on diseases or body systems. See Appendix A for a list of these ICs. The outcome variables are fractional patent counts.

Standard errors in parentheses, two-way clustered at the disease and science level ($^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$).

Table K5: Effect on Public-Sector Patenting

	Citation Linked		**Total Related**	
	Mean=6.75; SD=10.01		Mean=9.97; SD=11.05	
	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
DST Funding ($10 mln.)	1.160***	1.043	1.368***	0.969*
Mean=4.06; SD=4.36	*(0.289)*	*(0.701)*	*(0.292)*	*(0.536)*
Elasticity	0.671	0.627	0.557	0.395
R^2	0.789	0.557	0.895	0.684
Observations	14,085	13,043	14,085	13,043
Year FEs	Incl.	Incl.	Incl.	Incl.
Disease × Science FEs	Incl.	Incl.	Incl.	Incl.
Disease × Year FEs	Incl.	Incl.	Incl.	Incl.
Science × Year Linear Trends	Incl.	Incl.	Incl.	Incl.
Application Controls	Incl.	Incl.	Incl.	Incl.

Note: See notes to Table 6 for details about the sample, and Table 8 for notes about the instrument. The outcome variables are fractional patent counts. Application controls include (i) FEs for the number of applications that a DST receives; (ii) FEs for the number of applications associated with a DST that are also in a 50-grant window around the relevant IC payline, as well as (iii) cubics in the average raw and rank scores of applications associated with a DST that are also in a 50-grant window around the payline. Public sector patents are defined as those assigned to government, non-profit foundations, academic, or hospital entities.

Standard errors in parentheses, two-way clustered at the disease and science level ($^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$).

Table K6: "Embodied" vs. "Disembodied" Linkages

	Citation Linked				Total Related			
	No Overlapping Names		Only Overlapping Names		No Overlapping Names		Only Overlapping Names	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
DST Funding ($10 mln.) Mean=4.06: SD=4.36	2.327*** *(0.619)*	2.201* *(1.147)*	0.695*** *(0.176)*	0.595 *(0.477)*	3.447*** *(0.762)*	2.603** *(1.282)*	0.518*** *(0.120)*	0.206 *(0.245)*
Elasticity	0.865	0.709	0.794	0.677	0.598	0.394	0.562	0.227
R^2	0.730	0.509	0.808	0.560	0.859	0.622	0.878	0.616
Observations	14085	13043	14085	13043	14085	13043	14085	13043
Year FEs	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.
Disease × Science FEs	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.
Disease × Year FEs	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.
Science × Year Linear Trends	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.
Application Controls	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.

Note: See notes to Table 6 for details about the sample and Table 8 for notes about the instrument. The outcome variables are fractional patent counts. Application controls include (i) FEs for the number of applications that a DST receives; (ii) FEs for the number of applications associated with a DST that are also in a 50-grant window around the relevant IC payline, as well as (iii) cubics in the average raw and rank scores of applications associated with a DST that are also in a 50-grant window around the payline.

Standard errors in parentheses, two-way clustered at the disease and science level ($^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$).

Table K7: Benchmark Results with Minimal Set of Patents

	Citation Linked		**Total Related**	
	Mean=11.54; SD=17.54		Mean=22.71; SD=25.75	
	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
DST Funding ($10 mln.)	2.230***	2.168**	3.357***	2.527**
Mean=4.06; SD=4.36	*(0.604)*	*(1.098)*	*(0.745)*	*(1.258)*
Elasticity	0.869	0.819	0.600	0.452
R^2	0.725	0.505	0.857	0.620
Observations	14085	13043	14085	13043
Year FEs	Incl.	Incl.	Incl.	Incl.
Disease × Science FEs	Incl.	Incl.	Incl.	Incl.
Disease × Year FEs	Incl.	Incl.	Incl.	Incl.
Science × Year Linear Trends	Incl.	Incl.	Incl.	Incl.
Application Controls	Incl.	Incl.	Incl.	Incl.

Note: See notes to Table 6 and 7 for details about the sample. The outcome variables are fractional patent counts. The instrument is the total amount of funding (2010 dollars) for the subset of grants funded by a DST whose rank of rank scores were marginal, i.e., were within 25 applications of the award cutoff for their specific disease area (Institute). Application controls include (i) FEs for the number of applications that a DST receives; (ii) FEs for the number of applications associated with a DST that are also in a 50-grant window around the relevant IC payline, as well as (iii) cubics in the average raw and rank scores of applications associated with a DST that are also in a 50-grant window around the payline.

Standard errors in parentheses, two-way clustered at the disease and science level ($^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$).

Appendix L: "Stable" Keywords Indirect Linking Strategy

Recall that our preferred outcome measure identifies *all* patents related to an NIH funding area, whether or not these patents actually cite NIH-funded research. This allows us to account for a richer set of channels through which NIH funding may impact private-sector patenting. "Related" patents may include patents linked to NIH funding via a longer citation chain or patents by NIH-trained scientists who end up in the private sector. Crucially, these related patents may also be the result of private sector investments in related research areas; they need not be financially dependent on the NIH at all. Capturing the total number of private sector patents in an intellectual area is important because it allows us to take into account the possibility that NIH funding crowds out private investments. If this were the case, then we would not expect NIH funds to increase the total number of patents in a given research area: it would simply change the funding source for those patents. The impact of NIH funding on total innovation in a research area captures the net effect of potential crowd-in and crowd-out.

A potential drawback with this approach is that our definition of a DST's "intellectual area" can vary over time. If funding allows a disease/science area to expand the set of topics that it supports, then we may associate increased funding with more patents simply because higher levels of grant expenditures leads us to credit DSTs with patents over a wider slice of technological space.

To ensure that our results are not driven by this phenomenon, it is important that the breadth of the space over which we attempt to link patents with grants in a DST is exogenous to the amount of funding a DST receives. One way to ensure this is true is to verify that this space is stable over time, within each disease/science (DS) area.

To do this, we categorize all MeSH keywords associated with a publication funded by a DS combination into one of two types: "stable" MeSH keywords are ones that appear in publications funded by that DS across all years in the observation window, whereas "peripheral" keywords appear only in a subset of years in the data. We then restrict our set of related publications to those that match to a DS on stable keywords only. This fixes the boundaries of an intellectual area over time and therefore breaks any mechanical relationship that might exist between funding and the number of indirectly linked patents.

Concretely, for each DS, across all years in the observation window, we list all the MeSH keywords tagging the publications that *directly* acknowledge the grants in the DS. We then compute the frequency distribution of keywords within each DS. To fix ideas, in the DS corresponding to the National Institute of General Medical Sciences (NIGMS) and the Microbial Physiology II study section (MBC-2), the MeSH keyword `DNA-Binding proteins` sits above the 80^{th} percentile of the frequency distribution; `E coli` sits above the 95^{th} percentile; `Structure-Activity Relationship` sits above the 50^{th} percentile; and `Glucosephosphates` lies below the fifth percentile.

In the next step, we once again link each acknowledged article to the related articles identified by PMRA. However, we can now track whether these related articles are themselves tagged by keywords that our previous analysis has identified as "stable" within the DS—those keywords that are at the median or above of the DS-specific MeSH keyword frequency distribution.[xxi] The last step is to identify the patents that cite these indirectly linked articles, but we now restrict the citations to exist between patents and only the subset of "stable" related articles.

We experimented with several alternative ways to characterize "stable" indirectly linked articles. We report the results of specifications modeled after those used to generate the estimates in columns 4 and 5 of Table 8, our benchmark set of results. We manipulate two characteristics of keywords to generate the four variations of the strategy presented in the table below. First, for each article indexed by PubMed, some keywords are designated as main keywords, in the sense that they pertain to the article's central theme(s). We generate the keyword frequency distributions using all keywords and only main keywords, separately.

[xxi] In unreported results, we also experimented with a top quartile threshold, with little change to the results.

Second, MeSH keywords are arrayed in a hierarchical tree with 13 levels, with keywords for each article potentially sitting at any of these levels. Eighty percent of keywords that occur in PubMed belong to the third level of the hierarchy or below. For each keyword below the third level, we climb up the MeSH hierarchy to the third level to find its third-level ancestor (in the case of keywords that belong to multiple branches in the tree, we pick the ancestor at random). We recompute the keyword frequency distribution at this coarser, but more homogeneous level. Combining these two characteristics (main vs. all keywords; any levels vs. third level of the MeSH tree) provides us with four distinct keyword frequency distributions to identify the set of stable, indirectly-linked articles. Each of these in turn correspond to a column in Table L1.

Two features of the results in this table deserve mention. First, the magnitudes of the coefficients are slightly smaller than those observed in Table 6. This is to be expected, since our "stable" linking strategy shrinks the number of opportunities to associate patents with DSTs. The IV estimates are more imprecisely estimated (statistically significant at the 10% level for three out of four specifications). Second, the elasticities are comparable in magnitude to those computed in Table 8 (columns 4 and 5).

In conclusion, the results corresponding to these alternative linking strategies bolster our claim that the indirect linking strategy presented in the main body of the manuscript allows us to identify total private-sector innovation in a DST in a way that is not mechanically related to the amount of funding this DST receives.

Table L1: Effect of NIH Investments on Total Related Private-Sector Patenting, Stable Research Area Keywords Only

	Main Keywords		**All Keywords**	
	Level Adjusted	Raw	Level Adjusted	Raw
	Mean=14.8; *SD=17.0*	*Mean=12.5;* *SD=14.9*	*Mean=23.1;* *SD=25.8*	*Mean=22.5;* *SD=25.2*
	(1)	(2)	(3)	(4)
OLS				
DST Funding (×\$10 mln.)	2.234***	2.023***	3.380***	3.305***
Mean=4.06; SD=4.36	*(0.435)*	*(0.381)*	*(0.696)*	*(0.686)*
Elasticity	0.613	0.657	0.594	0.596
IV				
DST Funding (×\$10 mln.)	1.712*	1.336	2.549**	2.531*
Mean=4.06; SD=4.36	*(0.942)*	*(0.858)*	*(1.293)*	*(1.297)*
Elasticity	0.470	0.434	0.448	0.457
Observations	14,085	14,085	14,085	14,085

Note: The dependent variable is the number of fractional patents in the same area as a given DST, but using a more restrictive definition of relatedness than in our benchmark specification. If a patent cites a publication that directly acknowledges an NIH grant, but which does not contain any keywords that have commonly been used in that D-S, then the linked patent is not counted under this approach. See Appendix L for more details regarding this matching method. Columns 1 and 2 apply this method counting only keywords that are designated as main keywords; Columns 3 and 4 do this for all keywords. Columns 1 and 3 match two different keywords if they share the same level 3 parent keyword in the National Library of Medicine's semantic keyword tree. Columns 2 and 4 do not.

Standard errors in parentheses, two-way clustered at the disease and science level ($^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$).

Appendix M: Assessing Firm Reallocation of R&D Expenditures

The results in the main body of the manuscript examine the impact of NIH funding on firm patenting in related research areas. 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 impact of NIH expenditures estimated in Tables 7 and 8 is the same as its impact on overall 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 NIH investments on firm investments in other areas in order to assess its full impact on private innovation.

If firms respond to increased NIH funding for a DST by adjusting their portfolio of investments, then the effect of NIH funding for a DST would be two-fold: the direct effect on private innovation in the area of that same DST, and the countervailing reallocation effect on private innovation in the other research areas that a firm reallocates to or from. If firms divert funds from other areas in order to invest in the DST with increased NIH funding, we think of this as "reallocated crowd-in." Conversely, firms may divert resources away from a DST with increased NIH funding toward other research areas; we refer to this as "reallocated crowd-out."

We attempt to directly measure the extent of firm reallocation in response to NIH funding. First, we note that our second outcome measure—the total number of patents that draw on research related to 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/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. Reallocation within this set would already be captured in the results displayed in Table 7.

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. $TotalPatents_{-d,-s,t}$ 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. $AveragePatents_{-d,-s,t}$ 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 indeed affected by changes in funding for a particular DST. 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 DST funding leads a firm to reallocate toward other areas entirely, then we would no longer be able to associate it to the original DST. Our results, then, document the impact of DST funding on the reallocation of firm investments on the intensive margin, conditional on firms not switching away entirely.

Table M1 shows that, in general, an increase in NIH funding for one area of a firm's R&D portfolio does not decrease the number of patents that those firms develop in other areas. Our estimates in Columns 1 and 2 indicate that a $10 million increase in DST funding leads to an additional four to five patents, although these estimates are noisy. NIH funding does not appear to increase the average number of non-DST patents assigned to firms.

These findings, when combined with our previous results, indicate that overall firm patenting appears to increase in response to NIH funding. This finding suggests that NIH investments lead firms to weakly increase their overall patenting. Another interpretation for this finding is that there is a larger direct impact of NIH funding for a DST than we capture through our main 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 main outcome measures cannot capture; some of those effects may be reflected in Table M1.

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. These smaller magnitudes are to be expected. 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 as 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.

Table M1: Effect of NIH Investments on Firm Reallocation of R&D Investments

	Total non-DST patents		**Average non-DST patents, per DST-linked patent**	
	Citation	Related	Citation	Related

	Mean=122.6;
	SD=289.1
		Mean=178.1;
		SD=197.7
			Mean=2.57
			SD=3.20
				Mean=21.05;
				SD=66.9

	(1)	(2)	(3)	(4)
DST Funding (×$10 mln.)	5.537 *(3.736)*	6.141*** *(1.991)*	0.035 *(0.457)*	-0.004 *(0.025)*
Elasticity	0.183	0.140	0.055	-0.001
R^2	0.898	0.983	0.825	0.908
Observations	14,085	14,085	14,085	14,085

Note: Each observation is Disease-Science Area-Time (DST) combination. The outcome variables are fractional patent counts. Total non-DST patents are calculated by first identifying all assignees that produce a patent linked to a DST (either through citations or through PMRA relatedness). We then find all non-D, non-S patents issued to that restricted set of assignees in the same year. This is our "Total non-DST" patent count. "Average non-DST" patents normalizes this by the number of DST-linked patents. A patent is assigned to the disease area to which it is most often associated. All regressions include disease-science FEs, disease-year FEs, science-year FEs, and FEs for the number of applications to the DST, and cubics in the number of DST-linked patents that are matched.

Standard errors in parentheses, two-way clustered at the disease and science level ($^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$).

Appendix N: Linking NIH Grants to Patents Directly [Bayh-Dole Linkage]

Recipients of NIH grants and contracts are allowed to seek patent protection on project results. This practice emerged in the 1970s under Institutional Patent Arrangements between individual grantees (and contractors) and the Department of Health, Education, and Welfare, and intensified after the implementation of the Bayh-Dole Act in 1981.

One Bayh-Dole requirement is for recipients of federal research funds to report to the funding agency any patent application they file. This information is stored in the Interagency Edison (iEdison) database. Another requirement is to acknowledge on patent documents the existence of federal funding and the fact that the government retains certain rights, in so-called "government interest" statements.

The iEdison database has typically not been public, and grants are acknowledged on government interest statements in a format that is not standardized. Recently iEdison data has been made available on the web,[xxii] although there is in all likelihood undercompliance in the early part of our sample (Rai and Sampat 2012). Accordingly, we complement iEdison data with information from government interest statements in granted patents. Grant numbers contained in government interest statements within patents are reported haphazardly. We extract them through the use of regular expression matching, looking for any mention of an NIH institute code followed by a grant number, possibly with punctuation (e.g., a dash) in between.[xxiii]

We find that 9,821 of these grants (6.4 percent of the total) generate patents directly, leading to 12,485 U.S. patents that are assigned primarily to universities and hospitals. These raw statistics are informative, since this represent only one fourth of the number of private-sector patents that can be linked through publications. Clearly, an assessment of patenting outcomes based on "Bayh-Dole" acknowledgments would miss a very large part of the impact we document in the main body of the manuscript. Just as in the case of our direct citation measure, we can assign each and every one of the "Bayh-Dole patents" to a DST, and run regression specifications analogous to those displayed in Table 6. The results are presented in Table N1 below. The OLS estimates (columns 1 and 3) imply an elasticity only approximately half as large as that yielded by the citation and PMRA-linking methods.

Our 2SLS estimates are negative and noisy. This is likely due to the fact that there are a relatively small number of grants that generate patents directly. If too few of these grants fall in the narrow window around an IC's payline, then our IV strategy is unlikely to be able to identify an effect.

[xxii] `http://www.iedison.gov`

[xxiii] See Sampat (2016) for more detail.

Table N1: Effect of NIH Investments on Downstream Patenting by Grantees

	Fractional Counts		**Unit Counts**	
	Mean=2.00; SD=3.12		Mean=16.7; SD=25.8	
	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
DST Funding (\$10 mln.) Mean=4.06; SD=4.36	0.117** *(0.049)*	-0.112 *(0.136)*	1.112** *(0.436)*	-0.488 *(1.036)*
Elasticity	0.255	-0.244	0.270	-0.119
R^2	0.877	0.594	0.894	0.641
Observations	14085	13043	14085	13043
Year FEs	Incl.	Incl.	Incl.	Incl.
Disease × Science FEs	Incl.	Incl.	Incl.	Incl.
Disease × Year FEs	Incl.	Incl.	Incl.	Incl.
Science × Year Linear Trends	Incl.	Incl.	Incl.	Incl.
Application Controls	Incl.	Incl.	Incl.	Incl.

Note: See notes to Table 6 for details about the sample, and to Table 8 for details about the instrument. Application controls include (i) FEs for the number of applications that a DST receives; (ii) FEs for the number of applications associated with a DST that are also in a 50-grant window around the relevant IC payline, as well as (iii) cubics in the average raw and rank scores of applications associated with a DST that are also in a 50-grant window around the payline. Public sector patents are defined as those assigned to government, non-profit foundations, academic, or hospital entities.

Standard errors in parentheses, two-way clustered at the disease and science level ($^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p <$ 0.01).