A BANG FOR A BUCK: ECONOMIC IMPACTS OF NIH FUNDING OF RARE DISEASE RESEARCH

Kevin Chandra[1]

ABSTRACT

What is the magnitude of economic returns of publicly-funded research, and how large are the spillovers relative to the returns? Historically, estimating direct impacts of such research, as well as differentiating direct impacts from spillovers, has been fraught with empirical challenges due to limitations in access to available data, especially at the most granular level. With advances in micro-level data in recent years (e.g. patent-paper citations or NIH-to-paper-to-patent linkages), economists and policymakers are increasingly interested in examining the impacts of public research funding on the value of scientific progress and innovation, as well as their spillovers to other fields. Using a preliminary list of 700 rare diseases, I constructed a new dataset linking NIH grants, publications, and patents spanning nearly three decades (1995-2023). Based on disease names and equivalent medical terms obtained from the National Organization for Rare Disorders (NORD) In-Depth Report Database, I utilize these keywords to conduct search and collect relevant NIH projects, publications, patents citing these papers, as well as patents containing disease-specific keywords that may not necessarily cite these papers. The direct impact is proxied by the presence of patents containing rare disease keywords that cite same-disease papers acknowledging same-disease NIH grants. The spillovers are proxied by: 1) patents containing rare disease keywords that do not always cite same-disease papers, and; 2) any patents (regardless of their nature) that cite these publications. Some stylized facts are summarized as follows. First, based on the preliminary 700-disease sample, I found that NIH has been allocating, in nominal terms, over US$120 billion (out of the cumulative total of US$800 billion) within the three-decade period for 580 diseases, but more than 70% of the funding is concentrated in only 5% of the funded diseases. In contrast, the remaining 120 diseases have never received any NIH funding in the same period. Second, using patent value estimates from Kogan et al. (2017), I found that, on average, every US$1 in disease-specific funding is, under no-lag scenario, associated with US$5.3 in the value of same-disease innovations, and once spillovers are taken into account, the value goes up to about US$34.6. The estimates remain persistent and demonstrate an upward trajectory across growing time lags. Third, the distribution of returns across diseases is highly skewed, with high returns largely concentrated among diseases with above-median, above-average, and above-90th-percentile funding, as well as diseases with advocacy organizations and rare cancers.

---

[1] Department of Strategy and Innovation, Questrom School of Business, Boston University, email: kchandra@bu.edu

I would like to thank participants at the BU Questrom Strategy and Innovation (S&I) Brown Bag Seminar and UCSB HIWG Research Chat for their valuable feedback. I receive no funding sources in carrying out this research. All errors and omissions are my own.

A BANG FOR A BUCK: ECONOMIC IMPACTS OF NIH FUNDING OF RARE DISEASE RESEARCH

INTRODUCTION

What are the economic impacts of public research funding on science and innovation? And how large is the magnitude of spillovers from field-specific funding relative to direct impacts to the focal field itself? For decades, economists and policymakers have long been interested in trying to estimate the direct impacts of government support of R&D on growth, and the extent to which the resulting spillovers flow into other fields (e.g. Solow, 1957; Arrow, 1962; Romer, 1990; Jaffe, Trajtenberg, and Henderson, 1993). Nevertheless, empirical studies on these questions have long been hampered by methodological and identification challenges, particularly in: 1) delineating direct impacts from spillovers, and; 2) identifying the right proxies for private and social returns, which may differ from one field to another (Jones and Summers, 2022).

Let us consider the two following illustrations. On the one hand, some research—in a seemingly linear direction—does lead to intended outcomes as originally envisioned. This is demonstrated by the case of cystic fibrosis transmembrane conductance (CFTR) regulator, a vital protein that helps balancing the flow of water and salt in human bodies as encoded by the CFTR gene, in which any mutation could subsequently result in cystic fibrosis. Historically, a cystic fibrosis diagnosis meant that these patients most likely would have to live with artificial lung devices (colloquially known as “iron lungs”), often for the rest of their lives. A major shift occurred in the early 2000s with a paper published by a team of researchers based in the University of California—San Francisco (UCSF) (Ma et al., 2003), where they tested a series of

drug compounds to identify CFTR activators, the discovery by which would be cited in a series of patent filings by Vertex Pharmaceuticals in the late 2000s.[2] In turn, these patent filings constituted a major component of Ivacaftor, a major orphan drug subsequently authorized by the FDA to treat cystic fibrosis patients. This paper, in turn, acknowledged four NIH grants, one of which ("HL60288") was related to analyzing "airway surface liquid in cystic fibrosis".

On the other hand, not all research leads to outcomes originally envisaged by the focal scientist(s). Oftentimes, such research yields accidental, largely unexpected discoveries—and thus 'serendipities' (Ban, 2006; Yaqub, 2018; Busch, 2024)—or other forms of spillover. One example is the research on extremophile bacteria, which 'unwittingly' led to the development of polymerase chain reaction (PCR) technology. Another illustration is from Thibodeaux et al. (2009), better known for the "shrimp-on-a-treadmill" study that was subject to criticisms among some policymakers and media reports as an instance of 'wasteful spending'.[3] These critiques belie the fact that the scientists sought to examine how environmental changes, and especially climate change, affect the exposure of marine organisms to harmful bacteria, and thus their health. Findings from the study play an important role not only in marine biology, but also to aid scientists' efforts on climate change adaptation more broadly.

These cases illustrate the persistent challenges faced by economists and policymakers in seeking to address a question as innocuously framed as "what are the returns of investment for every dollar allocated to a certain field?". Economists, scientists, and policymakers alike are cognizant of the fact that every unit of resource invested in a field are inevitably bound to

---

[2] The seminal paper was referenced in the following title: Ma, T., Vetriviel, L., Yang, H., Pedemonte, N., Zegarra-Moran, O., Galietta, L. J. V., and Verkman, A. S. (2002). High-affinity activators of cystic fibrosis transmembrane conductance regulator (CFTR) chloride conductance identified by high-throughput screening. J Biol Chem. 2002 Oct 4; 277(40): 37235-41. doi: 10.1074/jbc.M205932200. Epub 2002 Aug 2.

[3] One of the authors, Louis E. Burnett, sought to address these issues with a blog post clarifying the purpose of the 'shrimp-on-a-treadmill' study. Link: https://burnettl.people.charleston.edu/research/treadmill.php

generate some degree of market failures (Arrow, 1962). In turn, these failures are manifest in various forms, such as complexity in the degree of appropriability of scientific discoveries (Teece, 1986), spillovers as proxied by social returns exceeding private returns to research (Jones and Summers, 2022), the long, right-skewed tail in the distribution of returns (Jones and Summers, 2022; Belenzon and Cioaca, 2025), as well as the dilemma on how much to prioritize moonshots vis-à-vis incremental research (Carnehl and Schneider, 2025).

Recent advances in archiving micro-level data have enabled studies to examine large-scale 'paper trails' of the influence of scientific research on innovation (Bryan and Williams, 2021; Jones and Summers, 2022; Gullo et al., 2025). Examples range from efforts to map out the linkages from NIH grants to acknowledging publications to subsequently patents citing these papers (Azoulay et al., 2019; Fleming et al., 2019), to the use of highly detailed and granular records of federal government procurement in mapping innovation activities (Belenzon and Cioaca, 2025), as well as the use of machine learning in matching topic-specific funding announcements to the subsequent publications and/or patents (e.g. Myers, 2020; Myers and Lanahan, 2022; Rezaei and Yao, 2025), and linking firm-level productivity to public R&D spillovers (Dyevre, 2023).

While these studies have provided a rich contribution to the literature on assessing the returns to R&D investments, it remains to be seen whether we can roughly estimate how much of the returns directly materializes to the focal field, and how much of the remainder is spillover. Being able to differentiate between these two types of impacts is of particular importance in areas such as health research, where cross-disease spillovers are more commonplace than previously assumed (Coburn et al., 2024). Based on an analysis of over 300,000 grant-publication pairs for four major diseases (Chagas disease, malaria, diabetes, and breast cancer),

the authors found that about 60-70% of grant-publication pairs are of different diseases (Coburn et al., 2024), and the results are robust across funding grants of different countries.

This paper aims to further the contribution made by existing works in two ways. The former is about following Azoulay et al. (2019) in mapping out the NIH grant-paper-patent linkages, but instead of using NIH study sections, I will center the analysis specifically at the disease level. My motivation to focus on the disease level relates to the latter contribution: research in one disease oftentimes leads to discoveries that may or may not be relevant to the focal disease (Coburn et al., 2024). In this study, I delineate the 'boundaries' in which research in a focal disease leads to scientific output and innovation in the same disease, and which ones end up as spillovers.

A major challenge in differentiating direct impacts from spillovers is that cross-disease spillovers could occur even at the grant level. Moreover, my interest for this study is to identify spillovers at the patent-to-paper, instead of grant-to-paper, linkages. To do so, I isolate the grant-paper linkages to only focus on focal-disease grants that end up with same-disease papers, and afterwards assess the types of innovations that emerge from the latter.

More importantly, with the availability of funding data for NIH projects, as well as estimates of private value of innovations (e.g. Kogan et al., 2017, henceforth KPSS), I am also able to estimate the subsequent amount of dollars that are associated with every single dollar of NIH funding allocated to one disease (both in nominal terms). In the meantime, I refer to a preliminary list of 700 rare diseases collected from the National Organization for Rare Disorders (NORD) In-Depth Report List, with a particular focus on grant-paper-patent linkages spanning a nearly three-decade period (1995-2023). For this study, innovation is proxied by USPTO patent applications that subsequently receive approvals (patent grants).

Several stylized facts are presented from this study. First, a vast majority, but not all, rare diseases have received NIH funding. A total of US$121.3 billion has been allocated by NIH to study 580 out of 700 diseases within the 1995-2023 period. Second, the share of NIH funding allocated to study rare diseases has been rising over time, from slightly above 6% in fiscal year 1995 to approximately 21% in fiscal year 2023. Third, there is a substantial degree of funding concentration among a small number of diseases: about 42 diseases account for about 75% of rare disease-specific NIH funding, and once the funding is limited to only those resulting in same-disease papers, the proportion remains at approximately 70%. Relatedly, nearly 200 rare diseases have received, on average, cumulative NIH funding of barely US$1 million throughout the three-decade period. It is also worth noting that over half of the most-funded rare diseases happen to be in the oncology-related space, e.g. Glioma, Glioblastoma, Hodgkin's Disease, Astrocytoma, Hepatocellular Carcinoma, and Acute Myeloid Leukemia. Fourth, rare disease research is positively associated with the significant value of resulting innovations. In aggregate, the cumulative value of patents that cite focal-disease research, contain disease-specific keywords, and/or acknowledge public funding (regardless of their disease linkages) is nearly US$3 trillion (or about US$9.1 trillion with disease double-counting). Once I accounted for only patents that cite same-disease patents acknowledging same-disease NIH grants, the aggregate value of these patents stands at nearly US$740 billion (approximately US$3.8 trillion with disease double-counting).

I also ran a series of OLS (ordinary least squares) regressions with different time lags in order to assess the extent to which a public research input is associated with both the direct impacts and the spillovers. On average, one NIH project is correlated with nearly 1.1 citing patents, of which about 10%—with the coefficient of ~0.1—is attributed to direct impacts (as in

these refer to a focal-disease patent citing a same-disease paper acknowledging a same-disease NIH project), and the remaining 90% to spillovers (basically these refer to any patents, regardless of their relevance to the focal-disease papers). Additionally, one NIH project is also associated with ~0.25 patents containing the same-disease keyword(s), regardless of their actual linkages to the project itself. These results remain persistently stable across different time lags. The size of these coefficients shrinks by about 80-90% once I only include patent grants with KPSS value, but the overall patterns, as well as the direct impact-to-spillover ratio, remain largely unchanged. In terms of economic value, every one million US$ of NIH funding is associated with about US$5.3 million in the value of same-disease innovations, and once spillovers are accounted for, the value rises to nearly US$34.6 million. The positive association persists across increasing time lags, with the value of the coefficients rising by 40-50% under a six-year time lag.

Nevertheless, the results mask substantial heterogeneities in the distribution of resources across diseases, as well as the potential returns on investment (ROI). A small number of diseases account for a large increase in 'multipliers' (proxied by the total value of disease-specific innovations divided by disease-specific research funding). Moreover, additional heterogeneity analyses provide evidence of considerable discrepancies in the levels of ROI. Diseases with above-median, above-average, and/or above-90th-percentile cumulative NIH funding record coefficients that are almost similar or above the baseline results, and almost no significance is detected for diseases below these thresholds. The same pattern is also observed for diseases with advocacy organizations, as well as rare cancers.

RELATED LITERATURE

This study offers insights into a few streams of literature. To begin with, the analysis contributes to the economics of science literature, especially on the questions surrounding the impacts of public R&D on incentivizing innovation (Goolsbee, 1998; Diamond Jr., 1999; David, Hall, and Toole, 2000; Toole, 2007). Historically, existing works in this stream have been largely ambivalent on this question (David et al., 2000), especially on the crowding-out concerns (Goolsbee, 1998), but evidence from more sector-specific analyses—and especially those related to federal research funding on biomedical sciences—demonstrate that publicly funded research leads to a crowding-in effect of private R&D (Diamond Jr., 1999; Toole, 2007). In other words, and at least in biomedical sciences, government-funded research does incentivize innovation.

While these studies have shed light on the importance of government support, further efforts to explore the mechanisms at the micro level, as well as address the spillover questions, have—until recent years—been hindered by data limitations. With increasing advances in digitizing archival data (Garber and Romer, 1996), a growing number of researchers are now able to establish more precise linkages, especially in linking administrative data to science and/or innovation-related indicators, e.g. linking publication data to R&D at publicly traded firms (Arora, Belenzon, and Patacconi, 2018), NIH-to-paper-to-patent linkages (Azoulay et al., 2019), patents citing government-funded publications (Fleming et al., 2019), patent-to-paper citations and whether they were added in-front or within the text (Marx and Fuegi, 2020; Bryan, Ozcan, and Sampat, 2020), and identifying patents that are directly supported by federal research (Gross and Sampat, 2025). This study further contributes to the existing discussions by better refining the boundaries between direct impacts (where government funding of research in one field leads to outputs that are germane to the focal field) and spillovers (in which the outputs are not immediately relevant to the focal field).

Another dimension in which the study offers important insights relates to the broader debate in political economy surrounding resource allocation, especially in areas where different stakeholders compete for policymakers' attention (Cohen et al., 1972; Hilgartner and Bosk, 1988; Baumgartner et al., 2009). As policymakers have limited attention spans due to a plethora of issues and short electoral cycles, it is inevitable that they will prioritize issues—thus the 'garbage can model'—that either hold considerable significance for their political careers or affect the largest number of constituents (Cohen et al., 1972; Baumgartner et al., 2009). This is particularly the case for disease advocacy, in which advocacy groups have played an important role in raising awareness and/or redirecting policymakers' attention to their grievances (Best, 2012; Hegde and Sampat, 2015). While this study does not directly examine the role of advocacy groups, it provides a snapshot of the current landscape of disease-specific funding by NIH, and sheds light on the degree to which funding resources have been concentrated among a small number of diseases. Rare diseases represent a salient context, given the large number of diseases identified under this setting (Richter et al., 2015; Haendel et al., 2020; Nguengang-Wakap et al., 2020). Relatedly, scholars representing a broad range of fields, from economists to philosophers, have been raising questions about the optimal mode of resource allocation and parameters to consider in the allocation process, such as whether to consider disease burden or severity (Varmus, 1999; Millum, 2023), potential future market revenues (Garber and Romer, 1996), and/or patient population size (Acemoglu and Linn, 2004).[4]

[4] Scholarly debate remains ongoing with regard to which parameters would be considered most appropriate for policymakers and regulators in deciding the best mode of resource allocation, especially in disease-specific research. Millum (2023) suggested that health research funding can be made in proportion to the degree of severity of a focal disease, but Varmus (1999) cautioned that prioritizing disease burden as the main parameter—and he also raised concerns about pressures raised by advocacy groups—could raise questions about the impartiality of NIH, highlighting that the responsibility of federal agencies should be on thorough examination of all relevant factors, including direct and indirect costs to the society, in addition to disease burden. To the best of my knowledge, I have yet to identify a paper that seeks to formalize the resource allocation process in a multi-disease setting.

Another no less important dimension that the study contributes relates to the estimation of returns to R&D, and especially on publicly-funded research. Since the landmark study by Griliches (1958) on the returns of research on hybrid corn, numerous studies have sought to provide estimates of 'multipliers' that could be generated from R&D, but a vast majority of these studies are typically centered on the private sector, either at the industry-level or firm-level analyses (i.e. Hall, 1996; Hall, Mairesse, and Mohnen, 2010). Detailed analyses of the impacts of government-funded research, in contrast, are relatively nascent, as shown by Azoulay et al. (2019) on NIH-paper-patent linkages or Myers and Lanahan (2022) on the impacts of SBIR grants on energy-related innovations. More recently, with increasing concerns over potential budget cuts to federally funded research and the consequent welfare costs due to 'lost innovations' (Babina et al., 2023; Azoulay et al., 2025), there is a growing urgency among scholars to more accurately estimate the returns to publicly supported research.

BRIEF INTRODUCTION ON RARE DISEASES

Rare diseases represent a particularly unique context for a few reasons. First, in spite of the lack of consensus among the medical community (Richter et al., 2015; Haendel et al., 2020; Nguengang-Wakap et al., 2020), the number of such diseases is more common than previously assumed. At a global level, the number of people living with at least one rare disease is estimated at a range of 300-450 million people (Nguengang-Wakap et al., 2020). In the United States, the number is thought to be close to 30 million people, or about nearly one-tenth of the US population (National Organization for Rare Disorders, 2025).

In aggregate, the population of rare disease patients is highly prevalent, but these diseases, historically, have been largely understudied among the broader scientific community and the biotech/pharma industry due to their limited diagnosis and market size fragmentation

(Acemoglu and Linn, 2004).[5] Relatedly, there has also been—until recent years—limited attention on rare diseases in the economics and management literature, and social sciences more broadly. Among the studies that explore rare diseases, the emphasis tends to be on the role of advocacy groups (e.g. Best, 2012; Hegde and Sampat, 2015; Kucukkeles, Ben-Menahem, and von Krogh, 2019) or policy incentives (e.g. Yin, 2008; Chandra et al., 2019).

Overall, the challenge faced by rare disease populations is the large number of such diseases and the smaller average population relative to those of more prevalent diseases, which complicates the diagnostic process for the affected patients (Institute of Medicine of the National Academies of Science, 2010). One consequence of such phenomenon is the proliferation of disease-specific advocacy groups representing the disparate collective interests of these patients (Best, 2012; Wood et al., 2013). A major channel in which these groups have been actively involved pertains to congressional lobbying, particularly for additional resources, e.g. congressional earmarks, in supporting disease-specific research at federal agencies such as NIH (Hegde and Sampat, 2015). Existing literature on lobbying demonstrates that the relationship among advocacy groups tends to be highly competitive, and oftentimes bordering on adversarial (Hilgartner and Bosk, 1988; Best, 2012; Hegde and Sampat, 2015), due to a combination of resource constraints, policymakers' limited attention spans, as well as the garbage-can model of decision-making as a consequence of these limitations (Cohen, March, and Olsen, 1972; Hilgartner and Bosk, 1988; Baumgartner et al., 2009).

The considerable unmet need for treatments among rare disease populations also translates into substantial economic costs—whether direct or indirect—for both patients and their

[5] Existing works on market size and incentives for firms to innovate have been largely centered on highly prevalent (or common) diseases, such as cardiovascular conditions (e.g. Cockburn and Henderson, 1998; Cockburn, Henderson, and Stern, 2000; Acemoglu and Linn, 2004), or broad categories of diseases, e.g. seven classes of diseases in Toole (2007).

families. A report by the EveryLife Foundation for Rare Diseases estimates that in 2019 alone, the total costs incurred by patients and their families in the US were approximately US$1 trillion, with direct medical costs alone accounting for nearly US$450 billion on an annual basis (Yang et al., 2022).[6] Moreover, caregiving for rare disease patients also resulted in substantial costs in terms of lost productivity gains, which the authors estimated to be about US$450-500 billion.

Public research funding, in particular, has been deemed pivotal in incentivizing follow-on innovation—oftentimes pioneered by universities and biotech firms—to address these issues (Azoulay et al., 2019; Fleming et al., 2019; Gross and Sampat, 2025). Recent proposed cuts to federal research budget, and especially to agencies such as NIH—have caused concerns that future progress on medical breakthroughs could be either delayed or outright eliminated (Babina et al., 2023; Azoulay et al., 2025). An analysis by Azoulay et al. (2025) shows that in the event NIH funding were to be reduced by up to 40%, the number of drugs that might never appear in the consumer market could drop considerably, in a range somewhere between 12% and 51% from the expected level.[7]

DATA AND ANALYSIS

1 – Data Sources

[6] The analysis is based on the EveryLife team's compilation of medical and insurance claims, as well as other types of expenditure data, for 379 rare diseases. These diseases are estimated to affect approximately 15-16 million people in the US. It is possible that the US$1 trillion estimate may be an undercount, considering that there are 6,000-10,000 rare diseases in the general population, and that up to 10% of Americans may have at least one such disease. A full list of the diseases compiled for EveryLife's analysis is included in the following link: https://everylifefoundation.org/wp-content/uploads/2021/02/The_National_Economic_Burden_of_Rare_Disease_Study_Summary_Report_February_2021.pdf

[7] The authors based their analysis on estimating NIH grants up to fiscal year 2007, and simulating the proposed 40% cut for these grants. They made use of the NIH assessment panel "cutoff scores" to determine, in that hypothetical world, which grants would have been selected and those that would have been eliminated under the 40% cut scenario. Replicating the Azoulay et al. (2019) approach, they analyzed which NIH project-paper-patent linkages would have been affected, and focusing in particular on patents that appear in the FDA Orange Book database.

I relied on several data sources for this study, namely: 1) NORD In-Depth Reports (in order to generate a list of rare diseases and their equivalent medical keywords); 2) Lens.org (to extract publication and patent records based on the disease list); 3) NIH RePORTER (to collect NIH grants, also based on the disease list), and; 4) the Kogan et al. (2017) patent value dataset.

There are several reasons that underlie my motivation to focus on the disease list from NORD In-Depth Reports as the starting point. First, there remains a lack of consensus with regard to the actual number of rare diseases in the general population, with the number widely ranging from over 6,000 (Nguengang-Wakap et al., 2020) to over 10,000 (Haendel et al., 2020), and the number itself is also subject to cross-country differences in defining whether a disease is 'rare' (Richter et al., 2015). Second, the NORD In-Depth Reports have remained a major reference points for patients and their families, as well as doctors and scientists, in trying to better understand the nature of these diseases (Institute of Medicine of the National Academies of Science, 2010). Third, the reports have been utilized in previous studies examining the impacts of policy incentives, e.g. the Orphan Drugs Act (Yin, 2008), as well as disease-specific advocacy organizations (e.g. Best, 2012; Hegde and Sampat, 2015), on incentivizing innovation and the allocation of R&D resources. My conversations with doctors and patients also further corroborate the justifications in making use of the NORD In-Depth Reports.

Regarding the publication and patent data, I choose Lens.org for a few reasons. The first has to do with its usefulness in extracting biological data based on specified keywords, such as mapping science and innovation activities surrounding the SARS-CoV-2 virus (Jefferson et al., 2021a), CRISPR-Cas9 (Jefferson et al., 2021b), or genetic sequence-specific patents (Sampat and Williams, 2019). The second reason is that the database enables a more granular search specification, such as limiting the search process in publications based on the mention of a

requested search term in the "Title, Abstract, Field of Study, or Keyword" section in the publication database, as well as "Title, Abstract, or Claims" in the patent database. Third, Lens.org hosts large-scale publication and patent databases from across the globe that are constantly updated on a biweekly basis.[8]

With regard to information on funding and the economic value of innovations, I turned to NIH RePORTER and the Kogan et al. (2017), or "KPSS", dataset. While NIH RePORTER has been archiving project records dating back to 1970 under the "CRISP" predecessor, it only began including information on project-specific funding amount from 1980 onwards. My observation of the data is that the pre-1995 funding information was incomplete, and thus my preference to limit the period of analysis from 1995 onwards. The KPSS dataset, meanwhile, provides the estimated value of nearly 2 million patent grants from 1926 up to 2010, in both nominal and 1983 US dollars.[9] The latest iteration of the dataset extends the coverage to patents granted up to the year 2023, bringing the total number to over 3.3 million patent grants.[10]

2 – Data Collection Process

I began the search process by specifying the disease name and its equivalent medical terms in the "Title, Abstract, Field of Study, or Keyword" section in the Lens publication database. As an illustration, for the Cystic Fibrosis disease, I included both the 'Cystic Fibrosis' as well as its equivalent key terms (CF, fibrocystic disease of pancreas, mucoviscidosis,

[8] As of early November 2025, the Lens.org database is currently hosting over 292 million publications (dating back to the early 19th century), as well as 167 million patent records from over 100 patent offices globally, including USPTO.
[9] Kogan et al. (2017) estimated the value by focusing on abnormal stock market returns within a three-day period in which a patent application attributed to a focal firm received a grant from USPTO. A major caveat with this approach is that the dataset is solely centered on patents filed by publicly traded firms in the US, and the dataset itself only covers approximately 20% of the entire USPTO patent grant population. This approach also overlooks the potential economic value of innovations attributed to small, innovative, but privately-owned startups, research universities, as well as government research institutions. Hsu et al. (2021) seek to address this issue by estimating the value based on licensing agreement involving patents from US major research universities.
[10] The extended dataset is shared in the following GitHub link: https://github.com/KPSS2017/Technological-Innovation-Resource-Allocation-and-Growth-Extended-Data

pancreatic fibrosis) as generated from the NORD In-Depth Report. The database would then return results with any mention of these keywords in at least one of the four categories specified in the search. In addition to utilizing these terms, I also included boolean terms (e.g. 'AND', 'OR') in cases of diseases with multiple equivalent keywords. Referring again to the Cystic Fibrosis example, I included the following list of terms—namely (("Cystic Fibrosis") OR ("Cystic" AND "Fibrosis") OR ("Cystic" AND "Fibrosis" AND "CF") OR ("fibrocystic disease of pancreas") OR ("mucoviscidosis") OR ("pancreatic fibrosis") OR ("fibrocystic" AND "pancreas") OR ("fibrosis" AND "pancreas"))—ensuring that each of these keywords, whether from the original list or refined with boolean terms, are identified in paper titles, abstracts, 'fields of study', and/or 'keyword' columns in the resulting data.[11] The same process is then repeated, and adjusted, for each of the 1,400+ diseases with NORD In-Depth Reports.[12]

Based on the results returned from each search, I extracted both the publication records as well as the patents citing these papers. Afterwards, I conducted a separate search using the same specified terms for the patent database, with an emphasis on ensuring any of these terms is identified in the titles, abstracts, and/or claims of these patent records.[13] My reasoning is three-fold: (1) there is a possibility that innovators may file patents to protect their inventions independent of scientific linkages to the focal-disease research; (2) inventors may add claims that

---

[11] For the Cystic Fibrosis case, the search terms returned a total of 94,664 unique publications (inclusive of journal articles, conference papers, dissertations, book chapters, books, preprints, and other types of publications) and 35,153 citing patents as of October 29, 2023, the date in which these data were collected. To ensure the validity of the results in the sense that the results are not influenced by false positives due to the inclusion of boolean-refined search terms, I reran the publication data collection for Cystic Fibrosis solely using the terms listed in the NORD In-Depth Report. Focusing on papers published up to 2023, I identified a total of 91,428 unique publications and 35,153 citing patents under the simplified search term. The refined search for this disease was conducted on November 13, 2025.

[12] The full list of diseases with the in-depth reports, equivalent medical keywords, and the adjusted search terms with boolean-refined keywords will be provided in a separate appendix file.

[13] Using the Cystic Fibrosis example again, I showed that the search terms that combine NORD-derived and boolean-revised terms generated a total of 25,012 unique patent records, as of August 19, 2024 (the date in which the search was conducted). A separate search process making use of only NORD-derived terms returned a total of 23,123 unique patent records published up to August 19, 2024. The patent records cover patent applications and grants across all national patent offices across the globe, including USPTO.

a proposed innovation (e.g. a drug compound, medical device, or other pharmaceutical formulation) could be useful in treating conditions associated with these focal diseases, and; (3) this search process also allows me to identify which of these patent records overlaps with the patents citing focal-disease research.

Figure 2 provides a Venn diagram that visualizes the types of innovations that may result from these search processes. The left circle of the diagram ("patents citing rare disease research") represents the universe of patents citing focal-disease research, regardless of their immediate relevance to the research itself. This could serve as a proxy for research spillover. One illustration is the USPTO patent grant "US 10822644 B2", titled "External files for distribution of molecular diagnostic tests and determination of compatibility between tests". This patent, which focuses on diagnostic technologies, cites 229 research articles, only 1 of which is related to Cystic Fibrosis. By contrast, the right circle ("patents with rare disease keywords") represents another universe of patents containing disease-specific keywords that may not necessarily make any references to the focal-disease literature, and which could be another channel of spillover indirectly attributed to the focal-disease research. One example is the USPTO patent grant "US 11707472 B2", titled "Compositions for the treatment of fibrosis". The patent itself does not cite any papers related to Cystic Fibrosis, but it nonetheless contains a claim that the treatment could be potentially useful for patients with 'pancreatic fibrosis', one of the equivalent medical keywords for the disease.

The most important aspect of Figure 2 is the Venn diagram intersection, which represents patents that not only contain disease-specific keywords, but also happen to cite academic works of the same-disease research. One example is "US 10081621 B2", titled "Solid forms of (R)-1(2,2-difluorobenzo[D][1,3]dioxol-5yl)-N-(1-(2,3-dihydroxypropyl)-6-fluoro-2-(1-hydroxy-2-

methylpropan-2-yl)-1H-indol-5-yl)cyclopropanecarboxamide". This patent, which centers on a pharmaceutical formulation, contains claims on cystic fibrosis transmembrane regulator (CFTR), as well as cites 24 research articles related to the same disease.

With regard to the NIH RePORTER database, I also made use of the disease name and its equivalent medical keywords from each of the 1,400+ NORD In-Depth Reports, with a particular emphasis on whether the disease name or its keywords appear in project titles / project abstracts / project terms. Nevertheless, unlike the Lens.org database, I found the RePORTER database to be less amenable to the inclusion of boolean-refined terms. This is illustrated by the Cystic Fibrosis example in Figure 3. In case I made use of *("Cystic Fibrosis") OR ("Cystic" AND "Fibrosis") OR ("Cystic" AND "Fibrosis" AND "CF") OR ("fibrocystic disease of pancreas") OR ("mucoviscidosis") OR ("pancreatic fibrosis") OR ("fibrocystic" AND "pancreas") OR ("fibrosis" AND "pancreas")* as my specified search term, NIH RePORTER will return null results. Instead, the database only returns search results under the following *("Cystic Fibrosis" OR "fibrocystic disease of pancreas" OR "mucoviscidosis" OR "pancreatic fibrosis")*.[14]

One caveat worth noting is that the data collection remains a work in progress. Throughout this study, I focus primarily on the first 700 diseases that are listed alphabetically (from "47, XXY (Klinefelter Syndrome" up to "Larsen Syndrome"). Simultaneously, I am finalizing the data collection and integration for the publications, patents, and NIH grants of the remaining diseases. A total of 3,718,242 publications were collected from the 700-disease sample, and once I narrowed down the data to only papers with at least one US-based co-author,

[14] For each of the diseases, I followed closely the disease name and equivalent keywords provided in the NORD In-Depth Reports. Nevertheless, in some instances, I also allowed some degree of flexibility in using the keywords. As an illustration, certain diseases only have their own abbreviations as their equivalent keywords (e.g. "PML" for Progressive Multifocal Leukoencephalopathy). To reduce the possibility of the inclusion of false positives, I created slightly modified terms, namely "PML (Progressive Multifocal Leukoencephalopathy)" and "Progressive Multifocal Leukoencephalopathy (PML)". Another possibility is to reorder the naming of the disease, such as including "Infant Respiratory Distress Syndrome" in addition to "Respiratory Distress Syndrome, Infant", the original name listed in the NORD In-Depth Report.

the number dropped to 910,507. These 3.7+ million publications, in turn, were cited by 158,386 USPTO patent grants (or 411,382 patent-disease observations after citation double-counting is taken into consideration). Separately, the keyword-based search for patents containing disease-specific terms returned a total of 31,759 USPTO patent grants (or 70,780 patent-disease observations), some of which intersected with the patents citing these disease-specific papers (refer to Figure 1). Once I isolated the subset of patents with disease-specific keywords that also cite same-disease papers, I managed to identify a total of 17,371 USPTO patent grants (or 22,580 patent-disease observations). Additionally, the search process in the NIH RePORTER database generated a total of 229,372 unique project IDs (or 338,029 project-disease observations as certain projects are focused on two or more diseases).[15]

3 – Identification and Measurement Challenges

The ideal experiment in assessing the impacts of public funding on innovation would be for the social planner to instruct any scientists to only cite papers focusing on the same area and acknowledge grants centered on the same topic. Obviously, such experiment is implausible due to the inevitability of spillovers emanating from any research activity (Jones and Summers, 2022; Coburn et al., 2024), as well as serendipitous events where insights from one field unexpectedly lead to important discoveries in another (Yaqub, 2018).

Relatedly, identifying the grant-paper linkages is complicated by the significant presence of cross-disease spillovers. In the analysis of over 300,000 grant-publication pairs for four major diseases across different countries, Coburn et al. (2024) estimated that over 60% of these pairs happen to be spillovers across these diseases and other health conditions not listed in the

[15] The project IDs are identified and counted based on their 14-digit NIH project numbers.

analysis.[16] For the rare disease sample in this study, I expect the heterogeneity in the degree of spillovers to be considerably higher than that in the Coburn et al. study. Furthermore, some papers may acknowledge grants representing multiple diseases, possibly due to the interconnected nature of these diseases.

Figure 4 and 5 visualize the challenges of identifying the grant-paper-patent linkages. This is exemplified by an NIH grant numbered M01RR000750, which was administered by the NIH NCRR (National Center for Research Resources) to the Indiana University School of Medicine.[17] In my preliminary database, I found that this grant was acknowledged by over 146 publications, 8 of which are related to cystic fibrosis, and and the other 138 related to 29 other rare diseases and various conditions not included in the NORD rare disease list. As my interest, at present, is to identify the direct impacts and spillovers at the patent-to-paper linkages, I would focus initially on isolating the grant-paper cross-disease spillovers. Using the M01RR000750 example again, I would only focus on parts of the grant that are centered on Cystic Fibrosis that result in these 8 same-disease publications, which are, in turn, cited by 7 patents.[18] Of these 7 citing patents, only 1 of them (US 8399496 B2, titled "*Inhalable aztreonam lysinate formulation for treatment and prevention of pulmonary bacterial infections*") contains information relevant to Cystic Fibrosis research, whereas the other 6 patents—as a measure of spillover—are focused on

---

[16] The authors focused on four major, common diseases, namely: Chagas disease, malaria, diabetes, and breast cancer. They aggregated grant data across multiple cross-country databases, including NIH RePORTER, UKRI Gateway to Research, EC CORDIS, Wellcome Trust, and Dimensions.

[17] A cursory search in the NIH RePORTER database shows that, as of November 19, 2025, about 1,759 projects are associated with the NIH grant number M01RR000750. These projects provide funding for various research and clinical activities in the IU School of Medicine that cover dozens of diseases, ranging from common diseases (e.g. pulmonary hypertension and autism spectrum disorders) to rare diseases (e.g. cystic fibrosis, amyloidosis, osteogensis imperfecta, Huntington disease, etc).

[18] A major challenge in isolating grant-paper linkages relates to identifying the project number at the most granular level. A vast majority of papers tend to acknowledge NIH grants at the 11-digit level, instead of the more precise 14-digit project number. Using the M01RR000750 example, one approach to identify parts of the grant that are centered on Cystic Fibrosis is to analyze the 14-digit project number columns in the same-disease dataset that I downloaded from the NIH RePORTER. Projects relevant to the disease include "2M01RR000750-23", "2M01RR000750-33", "3M01RR000750-27S1", "3M01RR000750-27S3", etc).

other diseases/medical areas. In case a paper focuses on multiple diseases and happens to acknowledge multiple grants centered on these diseases, I double-count the grant-paper pairs as my analysis will be conducted at a disease-year level of observation.

In the 700-disease sample, I addressed the challenge relating to grant-paper linkage identification by cross-walking the publication IDs from the Lens dataset and disease-specific NIH project IDs with those in the bulk "PUBLNK" file that I downloaded from NIH RePORTER. Using this approach, I extracted a total of 198,325 unique publications acknowledging NIH funding (230,286 paper-disease observations after double-counting), as well as 151,140 unique NIH project numbers (209,367 project-disease observations).[19] It is worth noting, once again, that focal-disease papers may acknowledge funding centered on a different disease. For example, a paper on Cystic Fibrosis may possibly acknowledge an NIH grant focused on Cholera.[20] This is evident when I merged these paper-disease and project-disease pairs, which resulted in a total of 9,699,403 paper-disease-project-disease tetrads. In order to isolate the direct impacts of grants to only focus on focal-disease grants resulting in the same-disease papers, I narrowed down the merged data to 3,063,001 paper-disease-project-disease tetrads for which both paper-disease and project-disease labels are identically matched. These subsetted tetrads consist of 102,970 unique papers (111,923 paper-disease observations) and 123,815 NIH projects (151,236 project-disease observations).

---

[19] In the "PUBLNK" file, the NIH grant-paper linkages are identified by taking into account both the 11-digit NIH grant numbers and the papers' unique PMID identifiers. In the disease-specific project data I collected from NIH RePORTER, the downloaded files contain project numbers at the 14-digit level. One challenge in matching these papers relates to the tendency for scientists to acknowledge the grant numbers at the 11-digit level. One advantage of my approach is that for every file downloaded based on a disease-specific search (including for publications, patents, and NIH grants), I immediately assigned these projects with the disease-name label collected from the NORD In-Depth Reports. I was then able to match the paper IDs with the project IDs by looking at the 11-digit numbers that match with the 14-digit project numbers, as well as the overlapping diseases of focus.

[20] Existing research has shown that patients with Cystic Fibrosis may end up building more resistance to Cholera. One such study is by Gabriel et al. (1994), based on their analysis of CF mice models. Further reading: Gabriel, S. E., Brigman, K. N., Koller, B. H., Boucher, R. C., and Stutts, M. J. (1994). Cystic fibrosis heterozygote resistance to cholera toxin in the cystic fibrosis mouse model. Science, 266(5182): 107-109.

Overall, these publications (papers focusing on a disease that acknowledge the same-disease grants) are subsequently cited by 45,020 unique USPTO patent grants (196,232 citing patent-disease observations), as shown in Figure 6. Of these citing patents, about 10,136 of them (14,132 patent-disease observations) happen to contain same-disease keywords in their titles, abstracts, and/or claims. After accounting for patents that appear in the Kogan et al. (2017) dataset, the numbers further shrank to 9,636 citing USPTO patents (45,539 citing patent-disease observations) and 2,155 patent grants containing the same-disease keywords (3,056 patent-disease observations). The number of patents with focal-disease keywords, but not necessarily referencing any disease-specific science, also dropped from 31,759 USPTO patent grants (70,780 patent-disease observations) to 7,474 (16,793 patent-disease observations).

4 – Econometric Specification

Once I have collected and cleaned the data, I run the panel analysis at a disease-year level of observation using the baseline ordinary least squares (OLS) regression. The econometric model is specified in the following equation:

$$Patents_{it} = \alpha_0 + \alpha_1 NIHFund_{it-n} + \theta_i + \gamma_t + \sigma_\tau + \varepsilon_{it} \quad (1)$$

The dependent variable, $Patents_{it}$, comprises two types of outputs. The first type refers to count-based numbers of (i) any patent grants—regardless of the nature of their citations—that reference focal-disease papers, (ii) patent grants containing disease-specific keywords (regardless of the presence of scientific ties), and (iii) patent grants with disease-specific keywords that cite same-disease papers. The latter refers to the KPSS value (in US$ million, nominal terms) of these three types of patents. Patent counts and values are aggregated at the disease-year level on the basis of their application years to USPTO.

On the right-hand side, $NIHFund_{it-n}$ refers to two types of inputs. The former refers to count-based numbers of disease-specific NIH grants that subsequently result in at least one same-disease paper in a disease-year observation. The other type refers to the aggregated amount of funding (in US$ million, nominal terms) of these projects for each disease-year observation. To account for unobserved heterogeneities attributed to disease- and/or time-specific shocks, I included disease (θ_i), year (γ_t), and disease age (σ_τ) fixed effects. My justification for including disease-age fixed effects is that certain diseases have been studied for years, or even decades, preceding the period of analysis; it is important, therefore, to control for the accumulation of resources and knowledge bases over time. The disease-age fixed effects are proxied by a series of binary variables indicating the number of years that have passed since the year in which the first-ever paper about the focal disease was published.

Nevertheless, more often than not, NIH funding does not directly impact innovation, but is rather mediated by the presence of intermediate outcomes, such as scientific publications. To address this concern, I also run a separate instrument-variable regression, with NIH funding as the instrument, and the acknowledging focal-disease papers as the first-stage dependent variable.[21] The OLS model highlighted in the previous paragraphs can, therefore, be seen as a reduced-form regression. The econometric model for the IV regression is specified in the following expression:

$$Patents_{it} = \alpha_0 + \alpha_1 \widehat{SciPubs}_{it} + \theta_i + \gamma_t + \sigma_\tau + \varepsilon_{it} \quad (2)$$

$$SciPubs_{it} = \mathrm{A}_0 + \mathrm{A}_1 NIHFund_{it-1} + \Theta_i + \Upsilon_t + \mathrm{O}_\tau + \varepsilon_{it} \quad (3)$$

[21] Ideally, I would replicate Azoulay et al. (2019) by instrumenting for the panel assessment scores of NIH grant applications. Nevertheless, for this paper, the authors only have available data for NIH grant application scores up to those applied in 2005. In the more recent piece (Azoulay et al., 2025), they only have evaluation scores for applications submitted up to 2007.

In Eq. (2) and (3), $SciPubs_{it}$ is proxied by the count number of disease-specific papers that acknowledge same-disease NIH grants. I run the analysis using two approaches. The first one focuses on the count number of papers irrespective of time distance between the grant period and publication years. The latter limits the count to only papers published within five years of the grant, following Azoulay et al. (2019). The results largely hold with either approach, as I will elaborate in the next section.

PRELIMINARY RESULTS

1 – Some Stylized Facts

This section provides some descriptive statistics about the NIH funding landscape for rare diseases. Figure 7 visualizes the growth of NIH funding, in nominal US$, throughout the 1995-2023 period. The US government, cumulatively, has allocated approximately US$802 billion to the NIH, growing from slightly above US$10 billion in fiscal year 1995 to US$45 billion in fiscal year 2023. Simultaneously, the portion of NIH funding allocated for rare diseases has been rising, but at a more gradual and steady pace. The share of rare disease-specific funding rose from about 6.3% in 1995 to nearly 21.6% in 2023.[22] It is also worth noting that the increase in disease-specific funding did not correspond to a surge in overall funding throughout 1998-2003, the period that coincided with a bipartisan commitment by Clinton and Bush administrations to double federal funding of life science research (Freeman and Van Reenen, 2008). Post-2003, funding allocation to NIH largely plateaued until fiscal year 2015, when funding to the institutes

[22] The figures are provisional as they are based on a preliminary analysis of 700 rare diseases. These figures are subject to further revision once I incorporate the data from the entire diseases listed under the NORD In-Depth Reports.

demonstrated a steady increase. This pattern persisted under Obama, first-term Trump, and Biden administrations.[23]

In aggregate, among the 700-disease sample, NIH has been allocating about US$121.6 billion within the same period. Nevertheless, there are two caveats to take into account. First, about 120 out of these 700 diseases have never received any NIH funding. Second, the funding allocation is highly skewed, as highlighted in Figure 8. When we consider any disease-specific funding—regardless of whether these NIH grants result in same-disease publications—we observe that a vast majority of the funding is concentrated in a small number of diseases. Figure 8(a) shows that 26 diseases have cumulative funding in excess of US$1 billion, accounting for US$75.3 billion (approximately 62.5% of totals) in the aggregate sum. The full list of these billion-dollar-plus diseases is shown in Table 1. Over one-third of these diseases happen to be cancer-related (e.g. Glioma, Glioblastoma, Acute Myeloid Leukemia). In contrast, 194 diseases are recorded to have a cumulative sum of NIH funding at slightly below US$0.7 billion. On average, this translates to less than US$3.6 million of NIH allocation for each of these diseases.

Figure 8(b) visualizes the funding distribution after accounting for only NIH grants that are eventually acknowledged by at least one same-disease publication. Under this scenario, the funding landscape shifts considerably, but the pattern of funding concentration remains largely unchanged. The total amount of NIH funding allocation resulting in such papers is about US$66.7 billion, or approximately 55% of the original estimate in Figure 8(a). This finding somehow aligns with Coburn et al. (2024) in that a large proportion of disease-specific grants

[23] Freeman and Van Reenen (2008) cautioned that the doubling of NIH budget in a short period of time did not result in a substantial increase in the rate of biomedical research in general. As the budget ballooned from US$14 billion in FY1998 up to over US$27 billion in FY2003, both the Congress and Bush administration eventually imposed a spending cap, thus resulting in a years-long stagnation of NIH budget. In real terms, they estimated that NIH spending in 2007 was nearly 7% lower than in 2004.

result in cross-disease spillovers, with limited insights directly relevant to the focal disease. The number of 'billion-dollar-plus' diseases drops from 26 to 14, but in aggregate, they still account for nearly 50% of NIH disease-specific allocation.[24] A surprising observation is that the number of "No funding" diseases jumps substantially from 120 to 204, an indication that a large number of these diseases, in spite of NIH funding allocation, ends up with no same-disease publications. This is evident from Figure 8(c), which visualizes the number of diseases based on cumulative funding categories that result in no acknowledging same-disease publications. A vast majority of these diseases (51 out of 84) have cumulative NIH funding between US$1 million and US$9.99 million throughout the 1995-2023 period.

Could it be that the differences in the amount of disease-specific funding and the amount leading to same-disease papers have to do with discrepancies in how the search terms are specified? I do not rule out these possibilities, but interpreting from Figure 9, the results appear to be largely aligned. Figure 9(a) demonstrates the highly positive correlation between the amount of disease-specific funding (in the x-axis) and the subsequent amount that leads to same-disease papers (y-axis). Figures 9(b) and 9(c) further visualize the results by grouping these diseases based on their prevalence categories, which are obtained from Orphanet (Nguengang-Wakap et al., 2020).[25] It appears that the proportion of NIH disease-specific funding acknowledged by same-disease papers rises with the upward shifts in prevalence categories,

---

[24] Some of these diseases end up with changes to their cumulative funding categories (conditional on the presence of NIH grants subsequently acknowledged by same-disease papers), such as from ">US$1 billion" to "US$500 million – US$1 billion". Figure 8(b) shows revised figures in which I calculated the aggregate amount of funding for projects that eventually lead to same-disease papers.

[25] The disease prevalence data are obtained by consulting various sources. The starting point is Orphanet, which provides systematic information about prevalence estimates (whether for categorical and/or absolute-number figures) for more than 6,000 rare diseases. The categories range from: i) <1 per 1000000, ii) 1-9 per 1000000, iii) 1-9 per 100000, iv) 1-5 per 10000, v) 6-9 per 10000, vi) >1 per 1000, to vii) unknown. In case no figures from Orphanet are available, I also consulted additional sources, such as MedlinePlus, NORD In-Depth Rare Disease Reports, and various webpages as suggested from Google Search results. It is worth noting that some of the diseases in the current list may exceed the 200,000-patient threshold in the rare disease definition by the US government following the Orphan Drugs Act of 1983.

from about 33% for unknown-prevalence diseases to nearly 58% for diseases with prevalence of 1-5 per 10,000 people.[26] Diseases with slightly higher prevalence, however, record a slightly lower percentage of NIH funding acknowledged by same-disease papers.

What could be the possible reason for these findings? One explanation may be that the funding concentration has to do with disease prevalence (Acemoglu and Linn, 2004). Figure 10 appears to provide some support, with a positive correlation between disease prevalence and cumulative amount of disease-specific funding (as shown in part (a) of the figure). This pattern persists when disease prevalence is divided into categories following Orphanet's classification: diseases with either extremely low (e.g. those affecting less than 1 per a million in the general population) or unknown prevalence are more likely to end up with smaller cumulative funding or even no funding at all. Nevertheless, a major caveat with the finding is that disease prevalence is assumed to be time-invariant, and any efforts to systematically document changes in prevalence across diseases over time remain wanting (Nguengang-Wakap et al., 2020). This caveat also has implications for the econometric analysis: the inclusion of prevalence data may override disease fixed effects due to multicollinearity concerns. In the meantime, I will present the regression results using disease fixed effects in place of prevalence categories.

Panel 1 of Table 2 displays the cumulative value of patents that cite rare disease research, regardless of their linkages to NIH grants. There are a total of 119,642 patent-disease observations, comprising 97,980 observations on the basis of patents citing articles from any of these 700 diseases, 16,793 patent-disease observations with any keyword-specific references to

[26] The estimated match rate for same-disease grant-publication pairs is somewhat higher in this study than the estimate provided by Coburn et al. (2024), which is about 23-40%. One caveat is that their estimate is based on an analysis of four diseases, whereas my data covers 700 diseases, at this stage. This could be attributed to the fact that I accounted for any focal-disease NIH grants that end up with at least one acknowledging same-disease paper, while discounting the possibility that the same grant could also be acknowledged by other-disease publications.

at least one of these diseases, and 4,869 observations for focal-disease patents citing same-disease papers. In aggregate, the cumulative value of these patent-disease observations is about US$9.1 trillion in nominal value, but I caution that the figure is inflated due to multiple double-counting across these three categories. As an illustration, a single patent may cite an article from diseases A and B, but contains keywords specific to diseases B and C. The same patent will be counted as an observation in the first category (a patent citing disease-A research), two observations in the second category (a patent containing disease-B and disease-C keywords), and a separate observation in the third category (a patent with disease-B keywords citing disease-B research). Once the disease identifiers are removed, the number of observations drops from 119,642 to 42,222 unique patents, with an aggregate nominal value of nearly US$3 trillion. Nearly 10% of the value is attributed to focal-disease patents citing same-disease publications.

Results for Panel 2 of Table 2 are limited to only patents that cite focal-disease articles acknowledging same-disease NIH grants. The number of patent-disease observations citing focal-disease research drops to 45,539 observations (representing 9,636 unique patents), with a cumulative nominal value of US$3.3 trillion (and about US$704 billion once the disease double-counting is taken into account). Among the subset of patents containing same-disease keywords, I identify a total of 3,056 patent-disease observations (or 2,155 unique patents), with an aggregate sum of US$293 billion (or US$180 billion after removing disease double-counting).

Tables 3 and 4 list out the top 20 diseases based on the cumulative value of their related innovations. For Table 3, the total value is derived from any patents regardless of the presence of their linkages with NIH grants, whereas the cumulative funding amount is for any NIH grants that may or may not be acknowledged by same-disease publications. Separately, for Table 4, the total value and cumulative funding amount have already accounted for only patents that cite a

focal-disease paper acknowledging a same-disease NIH grant. The count figures in both tables also take into consideration double-counting across multiple diseases, especially when a patent contains keywords linked to two or more diseases. One caveat with these findings is that the patent value estimates from Kogan et al. (2017) are only available for publicly-traded firms.

2 – Regression Results

This section provides elaboration on results from regression analyses on the impacts of NIH funding on the value of same-disease innovations and spillovers. At the current stage, the functional form is ordinary least squares (OLS), but for future iterations, I will consider other models, such as Poisson or negative binomial regression. The results have taken into account the inclusion of disease, year, and disease-age fixed effects, including those in the heterogeneity analyses. Table 5 contains descriptive statistics on the variables utilized for the regression. Overall, the statistical distribution for each of the variables exhibits a long right tail, with the mean value consistently surpassing the median. As an illustration, for the variable "NIH projects", the average count number for every disease-year observation is nearly 8, but the median is only 1, whereas the 75th-percentile value is 4. Mirroring the observation in Figure 8, this finding implies the high degree of concentration in NIH funding resources at a small subset of diseases, which I will further elaborate in the heterogeneity analyses. Another example is shown in the "NIH funding" variable (in nominal US$), where the average figure US$3.96 million per disease-year observation. Nonetheless, half of the observations in the panel data recorded less than US$0.12 million in NIH funding, and the 75th-percentile value is only about US$1.66 million. Therefore, I would advise initial caution in interpreting the results from the regression analysis.

Table 6 presents results from the baseline OLS analysis. I run the econometric models across various time lags, up to six years. For Panel 1, the dependent variable is the count number of patents citing disease-specific research, regardless of their relevance to the focal disease. On average, every NIH project is positively associated with 1.06-1.20 additional citing patents, and the coefficients are consistent across different time lags. The consistency pattern is also observed for results in Panel 2 (where the outcome variable is the count number of patents with disease-specific keywords, regardless of their linkages to NIH grant-paper pairs) and Panel 3 (proxied by the count number of patents with disease-specific references citing same-disease research). The coefficients for Panel 2 are within the 0.24-0.29 range, and about 0.104-0.108 for the models in Panel 3. Based on these results, I estimate that, on average, about 9-10% of additional patents associated with NIH-funded research can be perceived as 'direct impacts', and the remaining 90-91% of the estimates to be spillovers.

The analysis is afterwards repeated, with a particular focus on the subsets of patents with available KPSS value estimates, and the results are shown in Table 7. One thing worth noting from the table is that there is a remarkable decline, within the 80-90% range, in the size of the coefficients across the three panels. The decline is attributed to the fact that the patent subset only accounts for those filed to USPTO on behalf of publicly-traded firms, in which Kogan et al. (2017) estimated patent value based on stock price fluctuation within the grant period of the focal patent. In Panel 1, on average, an additional disease-specific NIH project is associated with 0.221 citing patents in the no-lag model, and the coefficient increases steadily with increasing year lags. This is equivalent to an 8-9% increase in citing patents from the baseline average of the outcome variable. Coefficients in Panel 2 recorded the biggest decrease (approximately by 85-90%) in the scale of the coefficients relative to those in Table 6. Each disease-specific NIH

project, on average, is correlated with about 0.03 additional patents containing same-disease keywords that may or may not necessarily be linked to the NIH project. The pattern is remarkably stable across various time lags, with the estimates within the 0.29-0.32 range. From another angle, every additional NIH project is associated with a roughly 3% increase of such patents from the baseline. Finally, for Panel 3, which focuses on the subset of Panel 1 that also contains same-disease keywords, the coefficient is approximately 0.02 (with a slight variation of 0.021-0.023 range across various year-lag specifications). Each project is associated with an additional 0.02 same-disease patents citing same-disease science acknowledging same-disease NIH projects, and approximately 10-11% increase from the baseline averages. Overall, spillovers in these models appear to be roughly 10x bigger for other diseases/medical fields than for direct impacts to the focal disease.

Table 8 displays the estimated aggregate value of patents based on Kogan et al. (2017), with the NIH project variable replaced by NIH funding amount (in US$ million, nominal terms). The aggregate value is calculated by adding the total sum of nominal value of patents related to NIH projects in a filing year. Panel 1 provides results based on the patent count figures in Table 7. In a no-lag scenario, every US$1 million in NIH funding allocated to a disease is, on average, correlated with roughly US$34.6 million in the total value of related patents. Among patents containing disease-specific keywords that may not be directly linked to NIH-funded research (as shown in Panel 2), the estimate—under the no-lag assumption—is about nearly US$6.6 million. For disease-specific patents referencing NIH-funded same-disease research (Panel 3), every US$1 million of focal-disease NIH funding corresponds to about US$5.3 million in the value of same-disease innovations. The effects for same-disease innovations are particularly substantial, given that an extra US$1 million from NIH for a disease is likely associated with an

approximately 30% increase in the patent value from the baseline averages, compared to about 7% in Panel 2 and 16% in Panel 1. The results are, similar to those in Tables 6 and 7, consistent across different time-lag specifications, with the value of associated innovations growing considerably over time. This is demonstrated by the estimates under the six-year-lag scenario, for which the coefficients in Panels 1, 2, and 3 have increased by 50%, 80%, and 40% from the no-lag estimates respectively.

In addition to OLS, I rerun the analysis using an instrument-variable (IV) approach, with NIH funding proxied as the instrument, and the count number of acknowledging same-disease scientific publications as the first-stage dependent variable. Separately, I repeat the analysis by limiting the count number of publication output to those acknowledging the focal NIH project within five years of the project period, following Azoulay et al. (2019). One-year lag period is assumed for both conditions. Results of the IV analysis are displayed in Tables 9, 10, and 11. Overall, judging from F-statistics presented in these tables, the results are reassuring, with the second-stage estimates largely aligned with those in Tables 6, 7, and 8. Moreover, the results are largely unchanged under the five-year limit, as shown in Panel 2 of these tables. On average, an additional disease-specific NIH project is associated with ~0.6 acknowledging same-disease publication. In other words, every extra project correlates with a roughly 12% increase of research output from the baseline averages. Simultaneously, every allocation of US$1 million by NIH to study a specific disease corresponds to approximately 1 additional same-disease paper, or about 20% increase from the baseline level. The spillover-to-direct-impact ratio—proxied by the comparison of patents (of any kinds) citing disease-specific research and those containing same-disease keywords—also remains stable, with an estimation of around 10-to-1 and 11-to-1 (shown in Tables 9 and 10). In terms of private value, every US$1 million in focal-disease research

correlates with US$5.2-5.4 million in the value of same-disease innovations, and about US$34-36 million once the spillovers are also taken into account.

3 – Heterogeneity Analyses

In the previous section, I have shown that NIH funding, on average, is positively and significantly associated with the value of subsequent innovations, whether related to the same disease or the broader spillovers. Nonetheless, one limitation with the analysis is that it masks substantial heterogeneities in the nature of the data. As I have pointed out in the descriptive statistics, more than 70% of NIH funding for rare diseases has been concentrated in only about 5% of diseases in the sample. Another possibility is that funding for certain diseases may result in returns that either vastly exceed the current estimates, or no returns at all (with no identifiable patents citing any of the resulting disease-specific research).[27]

Figure 11 displays histograms on the distribution of returns of NIH funding across these rare diseases. For figure parts (a), (b), and (c), the denominator is proxied by the total amount of disease-specific NIH funding, regardless of its linkage to same-disease scientific output. The denominator in (d), (e), and (f) is proxied by the total amount of NIH funding that is linked to same-disease publications. The x-axis, in turn, represents the 'multipliers' (the cumulative value of innovations divided by disease-specific funding). Overall, there is a uniform pattern of right-

[27] One thing I overheard from an NBER workshop is that a panelist shared with the attendees his conversation with members of Congress about the potential returns of investment on research. He recalled the lawmakers asking about what the 'actual' returns are, and in turn responded that, "the level of ROI [to federal research funding] may range from 7% to 7,000%, and it depends on how you ask."

skewed distribution of returns across these six histograms, with a small proportion of diseases recording 'multipliers' of 100 and above.

Afterwards, I repeat the OLS regression with a split-sample analysis, with the results posted in Table 12. The split-sample analysis is performed under five scenarios: (i) whether a disease is above or below the median cumulative funding throughout 1995-2023 period; (ii) whether the focal disease is above or below the average cumulative funding; (iii) whether the disease is in the 90th percentile in terms of cumulative funding; (iv) whether the disease has at least one representative advocacy organization, and; (v) whether the disease is cancer-related. My reasoning for the inclusion of the last two scenarios—(iv) and (v)—is two-fold. With regard to advocacy, existing studies have shown that advocacy groups often engage in competition to attract policymakers' attention, and possibly to redirect the allocation of resources to support their causes (Best, 2012; Hegde and Sampat, 2015). Specific for oncology, historically there has been more funding—as well as drug development and other related industry activities—in this area (Bagley et al., 2019; Rollet, Lemoine, and Dunoyer, 2012).[28] The analyses are all conducted by, once again, assuming one-year time lag across all models.

I first elaborate on results from Panels 1, 2, and 3 of Table 12. Overall, a general interpretation is that diseases that are above-median, above-average, and/or at the 90th percentile (or above) of the cumulative funding distribution are more likely to demonstrate the statistically significant results across all three patent categories than those below these cutoffs. Diseases with above-median cumulative funding, in particular, exhibit coefficients that are nearly identical to

[28] A descriptive search of all NIH projects in the NIH RePORTER database shows that the National Cancer Institute (NCI) is, compared to the remaining NIH institutes, the largest recipient of federal research funding. Throughout the 1995-2023 period, the federal government has cumulatively invested over US$86.4 billion (in nominal terms) in NCI, followed by NIAID/National Institute of Allergy and Infectious Diseases (about US$76 billion), NHLBI/National Heart, Lung, and Blood Institute (US$67.4 billion), NIGMS/National Institute of General Medical Sciences (US$57.5 billion), and NIDDK/National Institute of Diabetes and Digestive and Kidney Diseases (approximately US$41.4 billion).

those presented in Table 8. Under this circumstance, every US$1 million in NIH funding is correlated with about US$5.2 million in the value of same-disease innovations, US$6.7 million for any patents (regardless of citation linkages) containing same-disease keywords, and about US$34.3 million once spillovers are taken into account. In contrast, among below-median diseases, although the coefficients are positive, the estimates are statistically noisy and largely insignificant (with the results shown to be 3.6, 10.4, and 22.6 respectively). The discrepancies are more considerable once the sample is split into above or below average and 90th percentile figures. The contrast is particularly evident in the case of rare-disease patents citing same-disease research (referring to columns (3) and (6) in Panels 2 and 3), with above-average and above-90th-percentile diseases showing coefficients that are 10-11x more than those below these thresholds. In other words, every US$1 invested by NIH in already well-funded diseases may potentially generate US$4.5-4.8 in direct impacts (in terms of focal-disease patents citing same-disease science), but among diseases below the average and/or 90th-percentile threshold, the estimate is only about US$0.4-0.7. The direct-impact results also constitute a substantially smaller proportion compared to the value that goes to spillovers (which are within the range of US$9.4-US$13.9), as proxied by any patents that cite these papers.

Panels 4 and 5 display results based on the split-sample analysis by taking into account the presence of advocacy organizations and whether the focal disease is a rare cancer. Data for advocacy organizations are obtained from a separate database maintained by NORD, which contains information about organization names, brief profiles, and diseases represented. Rare cancers are identified from the cancer list provided by the American Cancer Society (ACS) and the National Cancer Institute (NCI) in their websites.[29] Overall, diseases with advocacy

[29] NORD has been continuously maintaining a database of advocacy organizations, including those registered as NORD

organizations and rare cancers seem to attract more attention from innovators than those not in the two subsets. Every US$1 million invested by NIH into a disease with a representative advocacy group is associated with roughly US$5.4 million in the value of same-disease patents citing same-disease science, US$7.0 million in the value of any patents with same-disease keywords (regardless of citation linkages), and US$36.9 million for any citing patents. These gains are roughly double the amount for coefficients in the no-advocacy disease subset (about US$2.3 million, US$5.7 million, and US$19.2 million respectively). The contrast is even bigger after controlling for the presence of rare cancers. Every additional allocation of US$1 million in supporting rare cancer research is correlated with US$8.5 million in the aggregate value of subsequent same-disease innovations, almost three times larger compared to US$3.0 million for the non-rare cancer subset. After including the spillover, the value significantly increases to about US$53.4 million for the former, nearly double the estimated US$28.9 million in the latter.

CONCLUSION AND FUTURE DIRECTIONS

What are the economic impacts of public research funding on science and innovation? And how large is the magnitude of spillovers from field-specific funding relative to direct impacts to the focal field itself? Using a preliminary list of 700 rare diseases—and spanning almost three decades (1995-2023)—as my empirical context, I have shown that, on average, every US$1 invested by NIH into a disease is associated with US$5.4 in the subsequent value of same-disease innovation, and about US$34.3 after accounting for spillovers. These estimates largely correspond to the figures based on models developed by Jones and Summers (2022),

---

members, as well as those who are listed based on patient/caregiver input. Over 2,000 organizations (nearly 90% of which are based in the US) are currently listed in the database. Link: National Organization for Rare Disorders (NORD) – Find a Rare Disease Patient Organization (https://rarediseases.org/organizations/). Separately, rare cancer lists are shown in the following links for the American Cancer Society (ACS) (https://www.cancer.org/cancer/types/rare-cancers.html) and the National Cancer Institute (NCI) (https://seer.cancer.gov/seerstat/variables/seer/raresiterecode/).

which assume about US$4 in direct returns for every US$1 of research funding (and even bigger estimates for spillovers after accounting for different discount rates). Nevertheless, heterogeneity analyses reveal significant discrepancies in the level of direct impacts and spillovers across different diseases. Among diseases with above-median, above-average, and/or above-90th-percentile cumulative funding, the results either align with or surpass baseline estimates, and the figures are considerably lower, and with limited statistical significance, for diseases below these thresholds. Diseases with advocacy organizations and rare cancers, in particular, appear to capture a significant share of direct impacts and spillovers from NIH funding.

At present, the results are descriptive and correlational. I have yet to examine causal mechanisms underlying the funding-output linkages, and would therefore several possible directions to enhance this study. To begin with, one limitation with this study is that I have not figured out what the ideal counterfactual would be. While no-NIH-funding diseases could be a possible counterfactual, the issue—based on the 700-disease list—is that no-funding diseases only comprise slightly below 20% (N = 120) of the sample, rendering any efforts for comparison difficult. One possibility is to compare principal investigators (PIs) of these disease-specific NIH projects, and to specifically differentiate those that only receive NIH funding vis-à-vis those that also receive funding from private sources, especially from biotech/pharma firms. Another possibility—replicating Azoulay et al. (2019)—is to take into account NIH funding proposals that either barely made or missed the cutoff points set by the institutes' study sections. One challenge with conducting this analysis, however, is that data for such applications are only available up to those filed in FY2007 (Azoulay et al., 2025), the period by which rare disease research was still, by and large, emerging. The other possibilities include comparing and contrasting scientific communities whose star scientists experienced unexpected, premature

deaths (Azoulay, Fons-Rosen, and Graff-Zivin, 2019), or to focus on NIH-funded scientists that ended up as patent inventors versus those not engaged in inventive activities (Scharfmann, Marx, and Fleming, 2025).

With regard to spillovers, the analysis, at present, is centered on identifying spillovers across diseases. However, existing literature has shown that spillovers could also occur across geographical regions (e.g. Jaffe, Trajtenberg, and Henderson, 1993) or carried out via inter-organizational career mobility of these scientists (Marx, 2011; Chandra and Xu, 2025). The next steps for this study are to identify 'hubs'—to be proxied either by most-active institutions or observing the geographical agglomeration of firms and other organizations on specific diseases—and to assess whether research emanating from these hubs is potentially more impactful, and, in turn, generating higher returns, than similar research taking place outside of these hubs (Bikard and Marx, 2020; Balsmeier, Manso, and Fleming, 2025). Another possibility is to extend Chandra and Xu (2025) by investigating whether, and in what ways, scientist movement influences the impact of NIH funding on ROI.

In addition, the existing models in this paper are based on 'flow' approach (using the annual allocation of NIH funding) instead of 'stock' (proxied by the accumulation of capital stock over time). One potential extension is to replicate the growth model from Solow (1957) to examine the long-term productivity of NIH investments across different diseases. Insights from this additional analysis could be especially meaningful in discerning whether an increased inflow in public research funding corresponds to more medical innovations in the long run, as well as informing policymakers about the optimal mode of resource allocation without crowding-out private-sector research (Toole, 2007; Freeman and Van Reenen, 2008).

Findings from heterogeneity analyses also pose a question on the current state of resource allocation. While highly-funded diseases capture a large share of positive returns from NIH investments, diseases below such funding thresholds exhibit a larger increase in the amount of direct returns and spillovers relative to their baseline averages. As an illustration, among non-rare-cancer diseases, the average figure for focal-disease patents citing same-disease science is about US$10.3 million per disease-year observation. An additional US$1 million investment from NIH to a disease under this subset would be associated with an increase of US$3 million to the value of same-disease innovations (an increase of nearly 30% from the baseline). In contrast, among the rare-cancer subset, the average figure for the same category is approximately US$174.2 million per disease-year observation. Although every US$1 million would potentially generate US$8.5 million in the value of rare-cancer innovations, this only represents a slightly below 5% increase from the baseline level. Nevertheless, it is worth noting that reallocating resources from well-funded diseases to their less-well-funded counterparts is not always guaranteed to result in a significant, drastic increase in the future amount of innovation. Under some circumstances, the predicted returns could be statistically noisy and deemed less reliable. Moreover, scientific infrastructure in studying certain diseases, e.g. cancer, has been more well-established than for other diseases.

This challenge harkens back to a major insight from scholars in public policy/political economy/sociology: a vast majority of issues (rare diseases in this setting) have not received sufficient attention from policymakers (Hilgartner and Bosk, 1988; Baumgartner et al., 2009; Best, 2012; Hegde and Sampat, 2015). One possible implication is that simply allocating additional resources to less-well-studied diseases may not be sufficient to incentivize more scientists and innovators to investigate these diseases. A multi-stakeholder partnership approach

involving patient advocacy groups, industry, universities, medical centers, and government agencies will be increasingly essential in developing the necessary infrastructure to better understand these diseases, and in turn to incentivize the entry of scientists (Moors and Faber, 2007; Belousova, Groen, and Ouendag, 2020; Patterson et al., 2023). Some practices emerging under this approach include the development of patient registries by advocacy groups (Patterson et al., 2023), creation of disease-specific research consortia encompassing different stakeholders (Moors and Faber, 2007), or pooled cross-country 'mega-funds' to study specific disease groups (Belousova et al., 2020).

Lastly, the current study focuses only on patents as the main outcome variable. While advances have been recently made in better estimating the potential economic value of patents (i.e. Kogan et al., 2017), not all patents are considered commercializable (Sichelman, 2009; Frakes and Wasserman, 2015). Indeed, only about 40-50% patents are considered 'sufficiently worthy' for commercialization prospects (Sichelman, 2009), and concerns have been raised about the tendency by USPTO to 'over-grant' patent applications (Frakes and Wasserman, 2015). Future extensions to this study may encompass linking NIH funding and publication data to other innovation-related indicators, such as biologics (Frakes and Wasserman, 2026), VC-funded startups (Ewens and Marx, 2025), clinical trials, orphan drugs, and/or medical devices.

REFERENCES

Acemoglu, D., and Linn, J. (2004). Market size in innovation: Theory and evidence from the pharmaceutical industry. *Quarterly Journal of Economics*, 119(3): 1049-1090.

Arrow, K. J. (1962). Economic welfare and the allocation of resources for invention. In *The rate and direction of inventive activity: Economic and social factors*. Princeton, NJ: Princeton University Press, pp. 609-626.

Azoulay, P., Clancy, M., Li, D., and Sampat, B. N. (2025). What if NIH had been 40% smaller? *Science*, 389(6767): 1303-1305.

Azoulay, P., Fons-Rosen, C., and Graff-Zivin, J. S. (2019). Does science advance one funeral at a time? *American Economic Review*, 109(8): 2889-2920.

Azoulay, P., Graff-Zivin, J. S., Li, D., and Sampat, B. N. (2019). Public R&D investments and private-sector patenting: Evidence from NIH funding rules. *Review of Economic Studies*, 86: 117-152.

Babina, T., He, A. X., Howell, S. T., Perlman, E. R., and Staudt, J. (2023). Cutting the innovation engine: How federal funding shocks affect university patenting, entrepreneurship, and publications. *Quarterly Journal of Economics*, 138(2): 895-954.

Bagley, N., Berger, B., Chandra, A., Garthwaite, C., and Stern, A. D. (2019). The Orphan Drug Act at 35: Observations and an outlook for the twenty-first century. *Innovation Policy and the Economy*, 19(1): 97-137.

Balsmeier, B., Manso, G., and Fleming, L. (2025). Do technology clusters distort innovation? NBER Working Paper, link retrieved from https://conference.nber.org/conf_papers/f230192.pdf

Ban, T. A. (2006). The role of serendipity in drug discovery. *Dialogues in Clinical Neuroscience*, 8(3): 335-344.

Baumgartner, F. R., Berry, J. M., Hojnacki, M., Kimball, D. C., and Leech, B. L. (2009). *Lobbying and policy change: Who wins, who loses, and why.* Chicago, IL: The University of Chicago Press.

Belenzon, S., and Cioaca, L. (2025). Guaranteed demand and corporate R&D. *Management Science*, DOI: 10.1287/mnsc.2023.00087.

Belousova, O. A., Groen, A. J., and Ouendag, A. M. (2020). Opportunities and barriers for innovation and entrepreneurship in orphan drug development. *Technological Forecasting and Social Change*, 161: 120333.

Best, R. K. (2012). Disease politics and medical research funding: Three ways advocacy shapes policy. *American Sociological Review*, 77(5): 780-803.

Bikard, M., and Marx, M. (2020). Bridging academia and industry: How geographic hubs connect university science and corporate technology. *Management Science*, 66(8): 3425-3443.

Bryan, K. A., and Williams, H. L. (2021). 'Innovation: Market failures and public policies'. In K. Ho, A. Hortacsu, and A. Lizzeri (eds.), *Handbook of industrial organization*, pp. 281-388. Amsterdam: North-Holland.

Bryan, K. A., Ozcan, Y., and Sampat, B. N. (2020). In-text patent citations: A user's guide. *Research Policy*, 49(4): 103946.

Busch, C. (2024). Towards a theory of serendipity: A systematic review and conceptualization. *Journal of Management Studies*, 61(3): 1110-1151.

Carnehl, C., and Schneider, J. (2025). A quest for knowledge. *Econometrica*, 93(2): 623-659.

Chandra, A., Kao, J., and Stern, A. D. (2019). Regulatory incentives for pharmaceutical innovation: The FDA's Breakthrough Therapy Designation. Working paper.

Chandra, A., and Xu, C. (2025). Where discovery happens: Research institutions and fundamental knowledge in the life sciences. NBER Working Paper 33996, link retrieved from https://www.nber.org/system/files/working_papers/w33996/w33996.pdf

Coburn, J., Yaqub, O., Rafols, I., and Chataway, J. (2024). Cross-disease spillover from research funding: Evidence from four diseases. *Social Science & Medicine*, 349: 116883.

David, P. A., Hall, B. H., and Toole, A. A. (2000). Is public R&D a complement or substitute for private R&D? A review of the econometric evidence. *Research Policy*, 29(4-5): 497-529.

Diamond Jr., A. M. (1999). Does federal funding 'crowd in' private funding of science? *Contemporary Economic Policy*, 17(4): 423-431.

Dyevre, A. (2023). Public R&D Spillovers and Productivity Growth. European Central Bank (ECB) Working Paper, link retrieved from https://www.ecb.europa.eu/press/conferences/ecbforum/shared/pdf/2024/EFCB_2024_Dyevre_paper.pl.pdf

Ewens, M., and Marx, M. (2025). 'Startup patenting'. Link retrieved from https://relianceonscience.org/startup-patenting

Fleming, L., Greene, H., Li, G., Marx, M., and Yao, D. (2019). Government-funded research increasingly fuels innovation. *Science*, 364(6446): 1139-1141.

Frakes, M. D., and Wasserman, M. (2015). Does the US Patent and Trademark Office grant too many bad patents? Evidence from a quasi-experiment. *Stanford Law Review*, 67: 613.

Frakes, M. D., and Wasserman, M. (2026). Unveiling the patent landscape of biologic drugs. *Northwestern Law Review*, 120, forthcoming.

Freeman, R. B., and Van Reenen, J. (2008). Be careful what you wish for: A cautionary tale about budget doubling. *Issues in Science and Technology*, 14(1), link retrieved from https://issues.org/p_freeman/

Garber, A. M., and Romer, P. M. (1996). Evaluating the federal role in financing health-related research. *Proceedings of the National Academy of Sciences*, 93(23): 12717-12724.

Goolsbee, A. (1998). Does government R&D policy mainly benefit scientists and engineers? NBER Working Paper 6532, link retrieved from https://www.nber.org/system/files/working_papers/w6532/w6532.pdf

Griliches, Z. (1958). Research costs and social returns: Hybrid corn and related innovations. *Journal of Political Economy*, 66(5): 419-431.

Gross, D. P., and Sampat, B. N. (2025). The Government Patent Register: A new resource for measuring US government-funded patenting. *Research Policy*, 54(1): 105142.

Gullo, T., Page, B., Weiner, D., and Williams, H. (2025). Estimating the economic and budgetary effects of research investments. *Tax Policy and the Economy*, 39(1): 53-81.

Hall, B. H. (1996). The private and social returns to research and development. *NBER Working Paper No. 2092*.

Hall, B. H., Mairesse, J., and Mohnen, P. (2010). Measuring the returns to R&D. In *Handbook of the Economics of Innovation*, 2: 1033-1082. North-Holland.

Hegde, D., and Sampat, B. N. (2015). Can private money buy public science? Disease group lobbying and federal funding for biomedical research. *Management Science*, 61(10): 2281-2298.

Hilgartner, S., and Bosk, C. L. (1988). The rise and fall of social problems: A public arenas model. *American Journal of Sociology*, 94(1): 53-78.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics*, 108(3): 577-598.

Jefferson, O. A., Koellhofer, D., Warren, B., Ehrich, T., Lang, S., Williams, K., Ballagh, A., Schellberg, B., Sharma, R., and Jefferson, R. A. (2021). Mapping innovation trajectories on SARS-CoV-2 and its variants. *Nature Biotechnology*, 39(4): 401-403.

Jefferson, O. A., Lang, S., Williams, K., Koellhofer, D., Ballagh, A., Warren, B., Schellberg, B., Sharma, R., and Jefferson, R. (2021). Mapping CRISPR-Cas9 public and commercial innovation using The Lens institutional toolkit. *Transgenic Research*, 30(4): 585-599.

Jones, B. F., and Summers, L. H. (2022). 'A calculation of the social returns to innovation'. In A. Goolsbee and B. F. Jones (eds.), *Innovation and public policy*, Chapter I: 13-59. Chicago, IL: The University of Chicago Press.

Kogan, L., Papanikolaou, D., Seru, A., and Stoffman, N. (2017). Technological innovation, resource allocation, and growth. *Quarterly Journal of Economics*, 132(2): 665-712.

Kucukkeles, B., Ben-Menahem, S. M., and von Krogh, G. (2019). Small numbers, big concerns: Practices and organizational arrangements in rare disease drug repurposing. *Academy of Management Discoveries*, 5(4): 415-437.

Ma, T., Vetriviel, L., Yang, H., Pedemonte, N., Zegarra-Moran, O., Galietta, L. J. V., and Verkman, A. S. (2002). High-affinity activators of cystic fibrosis transmembrane conductance regulator (CFTR) chloride conductance identified by high-throughput screening. *Journal of Biological Chemistry*, 277(40): 37235-37241.

Marx, M. (2011). The firm strikes back: Non-compete agreements and the mobility of technical professionals. *American Sociological Review*, 76(5): 695-712.

Marx, M., and Fuegi, A. (2020). Reliance on science: Worldwide front-page citations to scientific articles. *Strategic Management Journal*, 41(9): 1572-1594.

Millum, J. (2023). Should health research funding be proportional to the burden of disease? *Politics, Philosophy & Economics*, 22(1): 76-99.

Moors, E. H. M., and Faber, J. (2007). Orphan drugs: Unmet societal needs for non-profitable privately supplied new products. *Research Policy*, 36: 336-354.

Myers, K. (2020). The elasticity of science. *American Economic Journal: Applied Economics*, 12(4): 103-134.

Myers, K., and Lanahan, L. (2022). Estimating spillovers from publicly-funded R&D: Evidence from the US Department of Energy. *American Economic Review*, 112(7): 2393-2423.

Nguengang-Wakap, S., Lambert, D. M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., Murphy, D., Le Cam, Y., and Rath, A. (2020). Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *European Journal of Human Genetics*, 28(2): 165-173.

Patterson, A. M., O'Boyle, M., VanNoy, G. E., and Dies, K. A. (2023). Emerging roles and opportunities for rare disease patient advocacy groups. *Therapeutic Advances in Rare Diseases*, 4: 1-18.

Rezaei, R., and Yao, Y. (2024). Government-backed science and venture capital. UNSW Business School Working Paper, link retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5044008

Richter, T., Nestler-Parr, S., Babela, R., Khan, Z. M., Tesoro, T., Molsen, E., and Hughes, D. A. (2015). Rare disease terminology and definitions—A systematic global review: report of the ISPOR Rare Disease Special Interest Group. *Value in Health*, 18: 906-914.

Rolley, P., Lemoine, A., and Dunoyer, M. (2013). Sustainable rare disease business and drug access: no time for misconceptions. *Orphanet Journal of Rare Diseases*, 8: 109.

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, 98(5): 71-102.

Sampat, B. N., and Williams, H. L. (2019). How do patents affect follow-on innovation? Evidence from the Human Genome. *American Economic Review*, 109(1): 203-236.

Scharfmann, E., Marx, M., and Fleming, L. (2025). Pasteur's quadrant researchers bring novelty, impact to publishing, and patenting. *Science*, 390(6776): 891-893.

Sichelman, T. (2009). Commercializing patents. *Stanford Law Review*, 62: 341.

Solow, R. M. (1957). Technical change and the aggregate production function. *Review of Economics and Statistics*, 39(3): 312-320.

Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy*, 15(6): 285-305.

Thibodeaux, L. K., Burnett, K. G., and Burnett, L. E. (2009). Energy metabolism and metabolic depression during exercise in Callinectes sapidus, the Atlantic blue crab: effects of the bacterial pathogen Vibrio campbellii. *Journal of Experimental Biology*, 212(21); 3428-3439.

Toole, A. A. (2007). Does public scientific research complement private investment in research and development in the pharmaceutical industry? *Journal of Law and Economics*, 50: 81-104.

Varmus, H. (1999). Evaluating the burden of disease and spending the research dollars of the National Institutes of Health. *New England Journal of Medicine*, 340: 1914-1915.

Wood, J., Sames, L., Moore, A., and Ekins, S. (2013). Multifaceted roles of ultra-rare and rare disease patients/parents in drug discovery. *Drug Discovery Today*, 18(21-22): 1043-1051.

Yaqub, O. (2018). Serendipity: Towards a taxonomy and a theory. *Research Policy*, 47(1): 169-179.

Yin, W. (2008). Market incentives and pharmaceutical innovation. *Journal of Health Economics*, 27: 1060-1077.

TABLES AND FIGURES

Figure 1. Summary of data-generating process

Figure 1 visualizes the processes by which the data are collected for analysis. The starting point of the data collection is the disease, specifically the disease name and its equivalent keywords. Each of these diseases is retrieved from the National Organization for Rare Disorders (NORD) In-Depth Rare Disease Reports. I then utilize the disease name and its keywords to extract publications, patents citing these papers, patents containing the focal disease name and/or the keywords, as well as the relevant NIH grants. The resulting patent records are then matched with the Kogan et al. (2017) dataset in order to generate the estimated private value of these innovations. The Venn diagram highlights the possible types of innovations as follows: (1) patents that cite focal-disease research, but not necessarily related to the disease; (2) patents that contain focal-disease name and/or keywords, but not necessarily referencing focal-disease science, and; (3) patents with focal-disease names/keywords that cite focal-disease research (where (1) ∩ (2)).

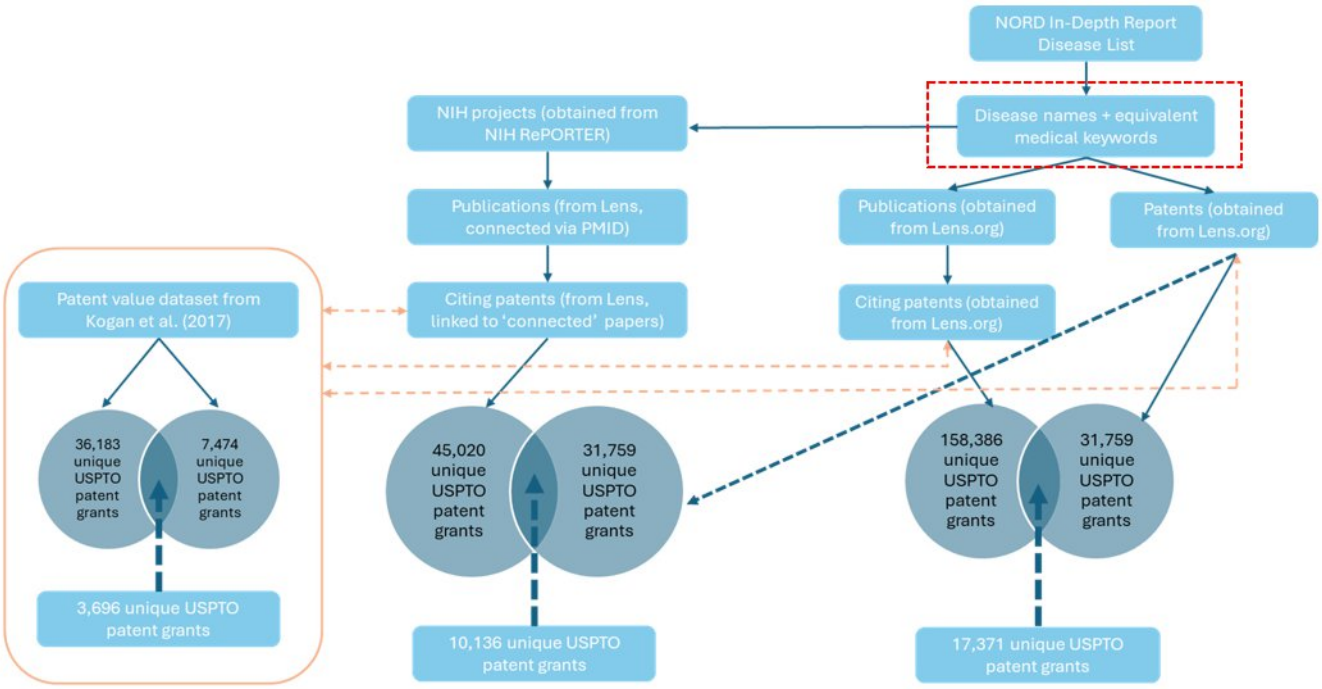


Figure 2. Types of innovation for consideration in the empirical analysis

Figure 2 further illustrates the types of innovation that may be related to rare disease research, as a follow-up to the explanation in Figure 1. An example of type (1) innovation is the USPTO patent grant number 10822644, for which one of its 229 references is related to cystic fibrosis research, but the patent itself is about diagnostic testing. An example of type (2) innovation is the USPTO patent grant number 11707472, which contains claims that the proposed treatment could be useful for patients with pancreatic fibrosis (a subset of cystic fibrosis), but does not cite any scientific articles on the focal disease. An example of type (3) innovation (basically (1) ∩ (2)) is the patent grant number 10081621, which cites 24 papers on cystic fibrosis, and contains several claims on CFTR regulator, a pathway in which the treatment is intended to address the disease.

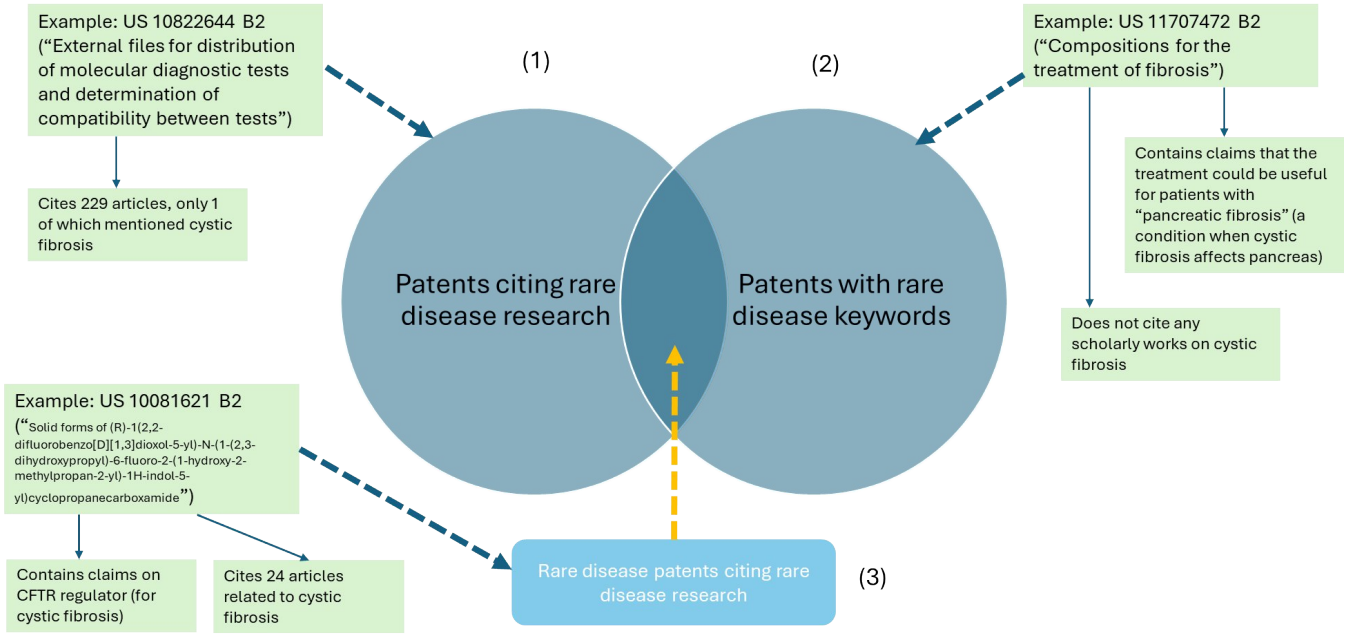
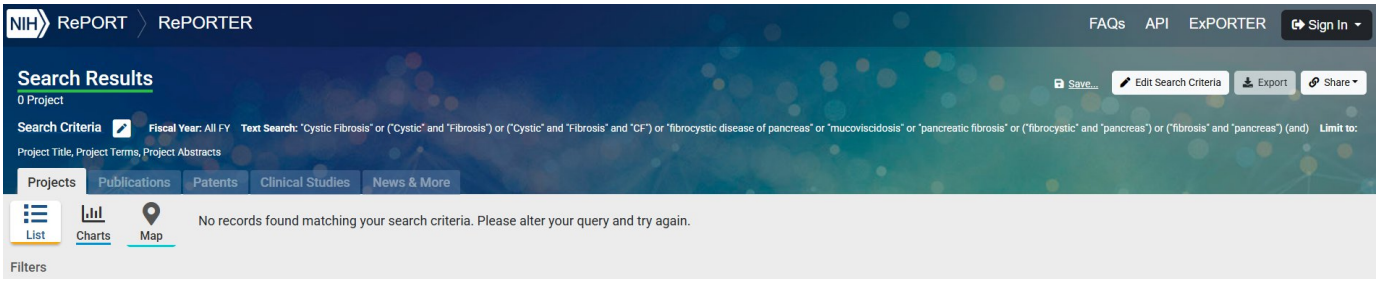
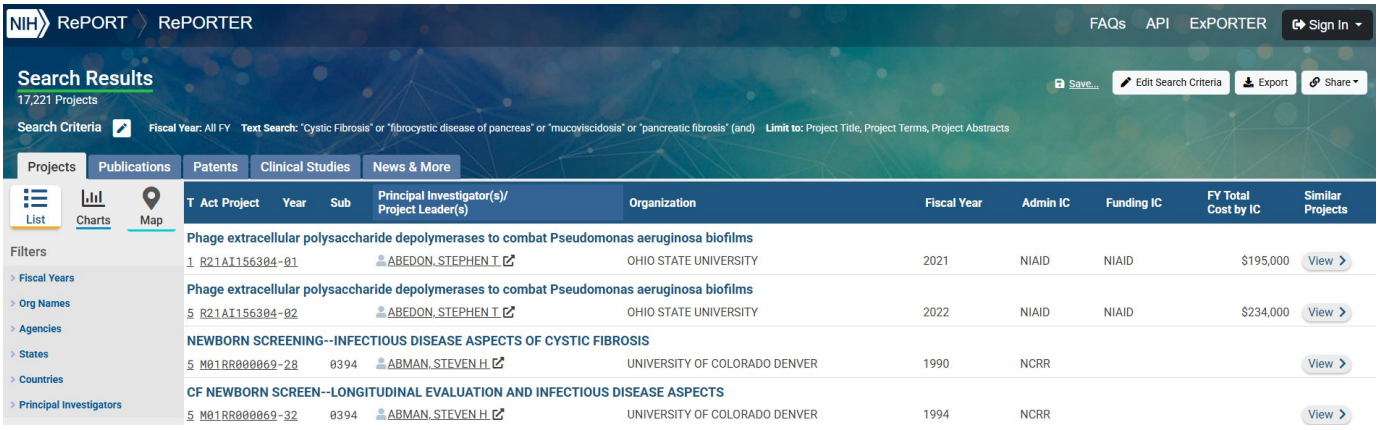


Figure 3. Comparison of search results in NIH RePORTER with (a) NORD and boolean-refined terms, and (b) only NORD terms, using the example of Cystic Fibrosis disease

Figure 3 provides screenshots of search results from NIH RePORTER under the two following scenarios: (a) where both NORD and boolean-refined terms are included in the search, and; (b) where only NORD terms (disease-specific keywords) are utilized in the search. Using the Cystic Fibrosis case as an illustration, I show that under scenario (a), the database is not able to return any search results, in contrast to scenario (b). One possible reason is that the search process may be highly sensitive to the additional boolean terms inside the specified keyword list, which already contains boolean terms separating each of these keywords. In contrast, for the Lens.org website, which I utilized to collect publication and patent data, the inclusion of boolean-refined search terms does not appear to generate such concerns.



(a) With NORD and boolean-refined terms



(b) With only NORD terms

Figure 4. Challenges in differentiating same-disease impacts of research funding from spillovers

Figure 4 visualizes the challenges in identifying, as well as differentiating, direct impacts and spillovers resulting from a grant. Cross-disease spillovers, especially at the grant-publication linkages, are highly common, and Coburn et al. (2024) show that cross-disease spillovers often comprise more than half of grant-publication pairs emanating from disease-specific grants. As an illustration, a grant specific for disease A may end up acknowledged not only by a publication specific for disease A, but also for other diseases (disease B, C, …, n). In turn, these papers may be cited by various patents (from whether for disease A, B, C, …, n), which complicate the analysis to identify direct impacts and spillovers.

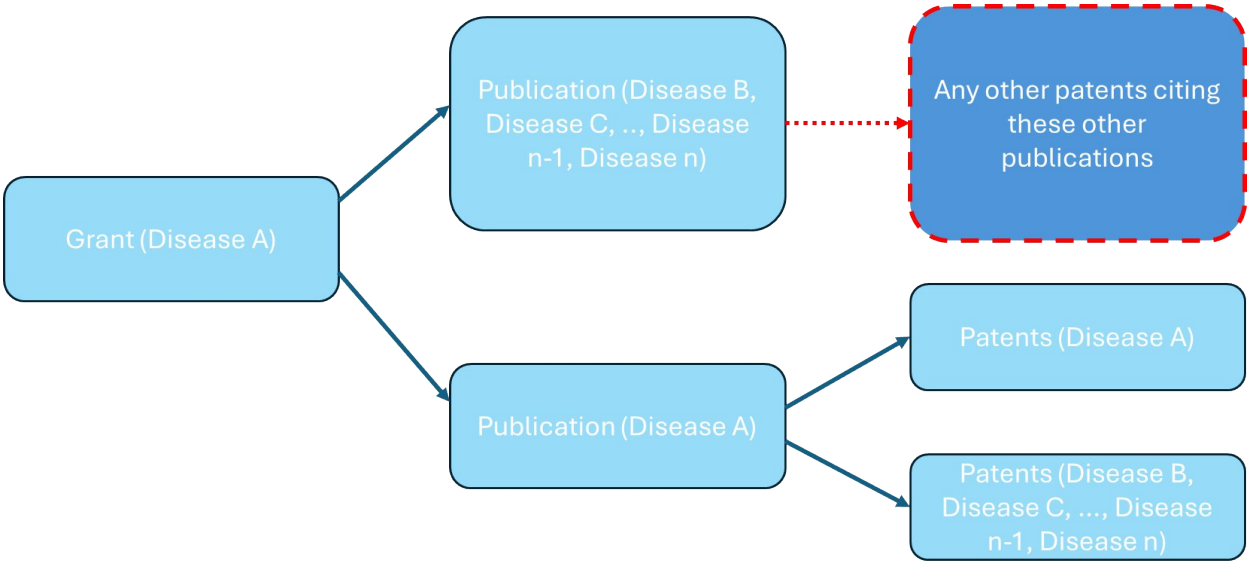


Figure 5. Isolating cross-disease research spillovers for empirical analysis

Figure 5 is a follow-up to Figure 4, in which I explained about the challenge in differentiating direct impacts from spillovers, especially in linking research grants to acknowledging publications to subsequently citing patents. This is attributed to the fact that spillovers may exist at both the paper-to-grant and patent-to-paper linkages. One approach to addressing this issue is to isolate the research spillovers for grant-paper linkages, and focusing on spillovers only among patent-to-paper citations. This is demonstrated by NIH grant number M01RR000750, specifically projects under the grant linked to Cystic Fibrosis. A total of 8 publications related to the disease acknowledge this grant, but the grant is also acknowledged by 138 publications focusing on 29 rare diseases and other health conditions. For this study, I isolate only 8 same-disease publications, and for other-disease publications that acknowledge projects under the grant for other diseases, I will double count them separately. In turn, these 8 Cystic Fibrosis publications are cited by a total of 7 USPTO patent grants, in which only 1 of them (US patent grant no. 8399496) is related to the same disease. This patent is proxied as a 'direct impact' under the same-disease paper-to-grant matching, and the other 6 patents are proxied as 'spillover'.

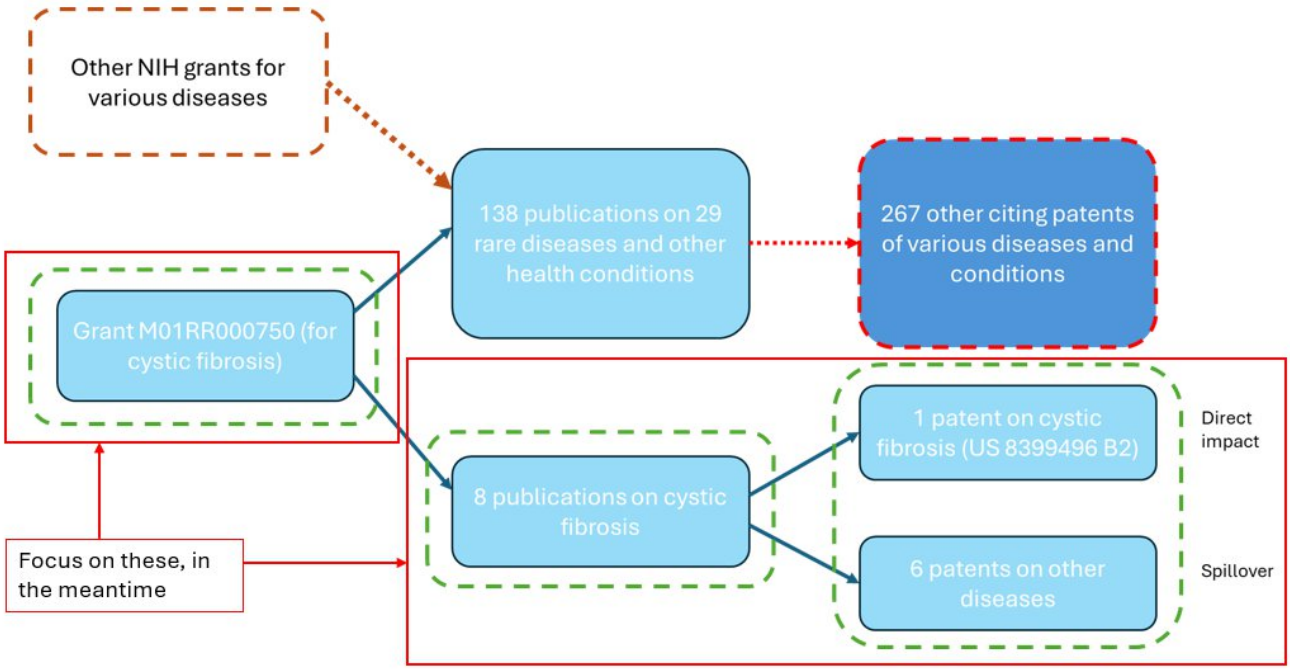


Figure 6. Integration of NIH grants, acknowledging publications, and citing patents

Based on the data collected from Lens.org and NIH RePORTER, and for a preliminary list of 700 rare diseases, I identified a total of 198,325 unique publications (230,286 paper-disease observations after taking into account double-counting) acknowledging NIH funding, as well as 151,140 unique NIH project IDs (209,367 project-disease observations) that are acknowledged by these papers. I afterwards merged these two sets of data, resulting in a total of 9,699,403 tetrads at the paper-disease-project-disease level. I selected only subsets of data with perfect matching for both the paper-disease and project-disease observations (N = 3,063,001). In turn, these fully matched observations represent 102,970 unique papers (111,923 paper-disease observations) and 123,815 unique NIH project IDs (151,236 project-disease observations). They are also associated with 45,020 unique patent grants citing the same-disease NIH-paper pairs, as well as a subset of 10,136 unique patent grants containing same-disease keywords.

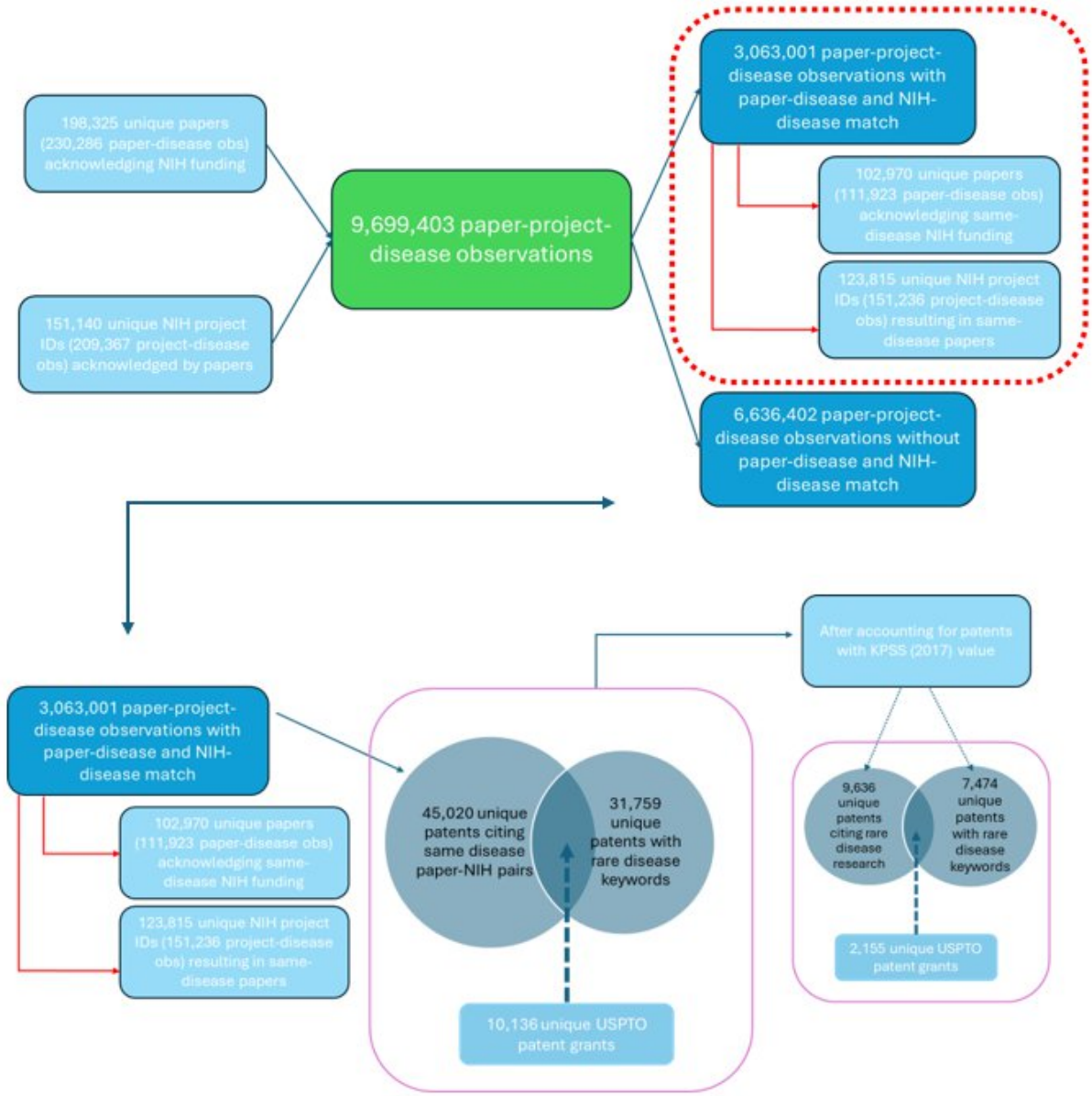


Figure 7. Changes in NIH funding, and allocation towards rare disease research, over time (in billion US$, nominal terms, for 1995-2023 period)

Figure 7 illustrates changes in NIH funding throughout the last three decades and the proportion of funding allocated to rare diseases (based on the analysis of a preliminary list of 700 diseases). The figures are shown in nominal US$. Overall, NIH funding from the federal government has been growing considerably from slightly above US$10 billion in FY1995 to over US$45 billion in FY2023. There was an initial 'doubling' of NIH budget from 1998 to 2003 due to the bipartisan consensus under the Clinton and Bush administrations about the need to invest in biomedical research. Funding subsequently stalled from FY2003 due to fiscal tightening and the difficulties faced by researchers in absorbing the large increase in federal research funding. NIH funding, in nominal terms, was largely flat until FY2015. The funding increase has remained at a steady pace under the Obama, first-term Trump, and Biden administrations. Simultaneously, the proportion of NIH funding to rare diseases has also been growing steadily, from 6.3% in FY1995 to 21.6% in 2023. The figures presented for rare diseases are provisional and subject to further adjustments once additional diseases are included in the analysis.

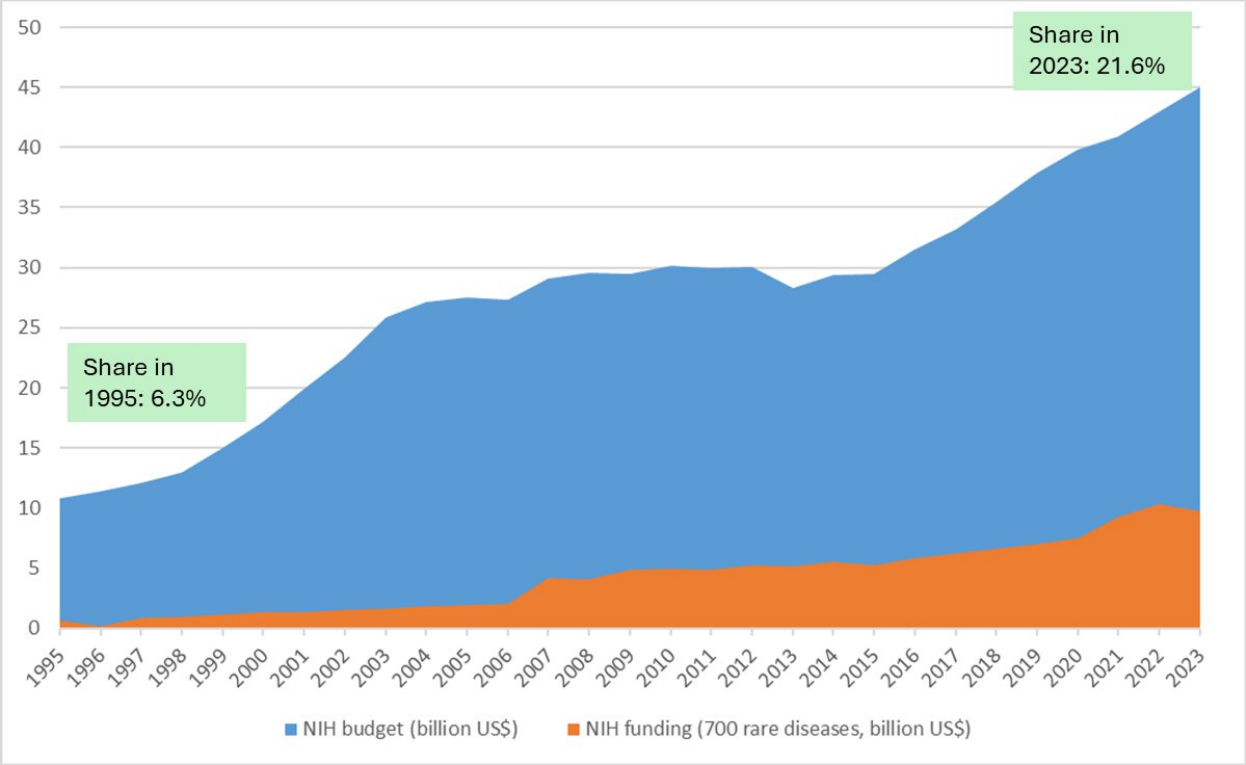
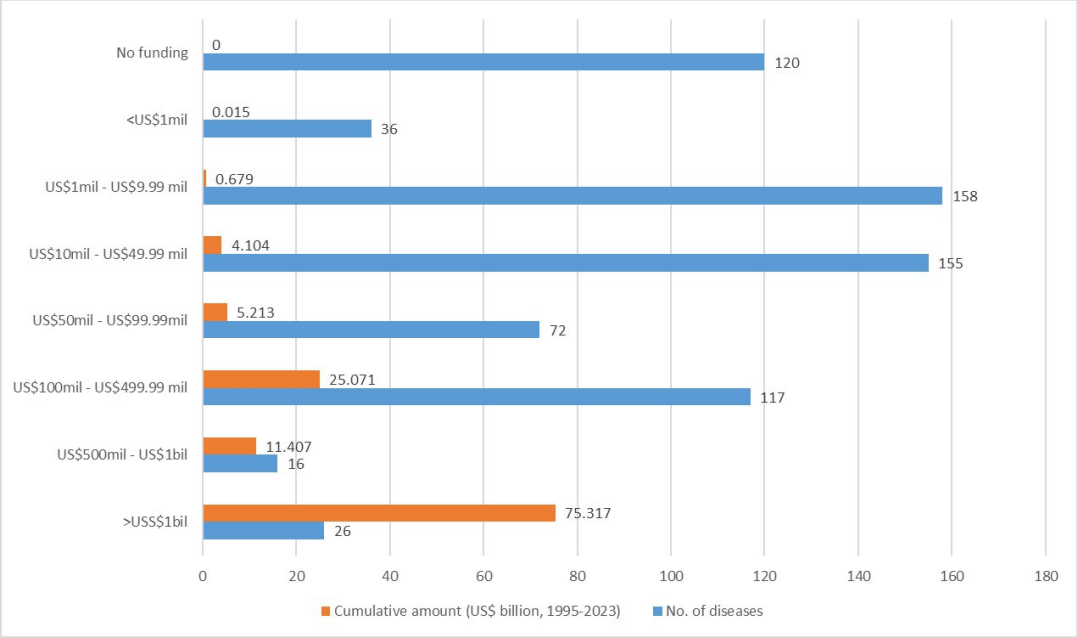
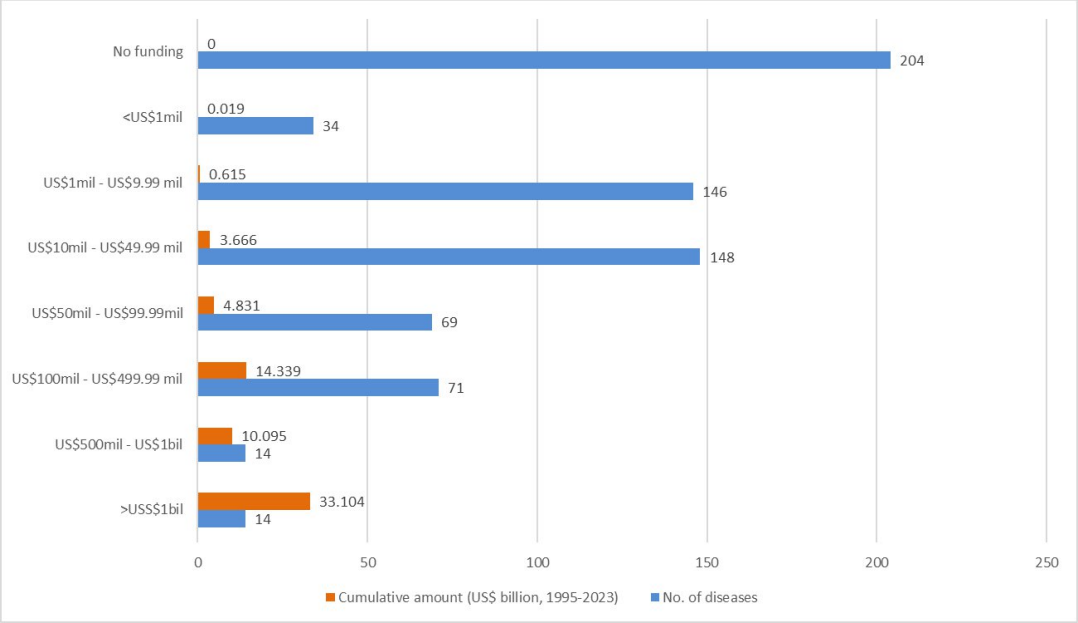


Figure 8. Number of rare diseases based on tiers of cumulative NIH funding amount (1995-2023)

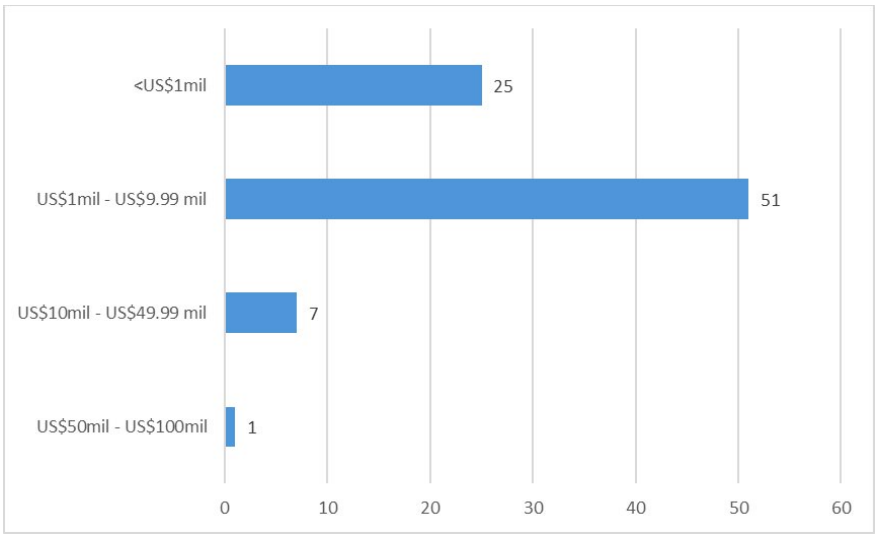
This figure visualizes the distribution of cumulative funding (1995-2023) across a preliminary list of 700 rare diseases. Part (a) summarizes the number of diseases based on cumulative funding categories. Part (b) is a follow-up to part (a) after adjusting only for disease-specific NIH projects that lead to at least 1 acknowledging same-disease publication within the time period. Part (c) describes diseases based on their cumulative funding categories in part (a) that do not end up with any same-disease publications.



(a) Any NIH projects, regardless of linkages with acknowledging same-disease papers



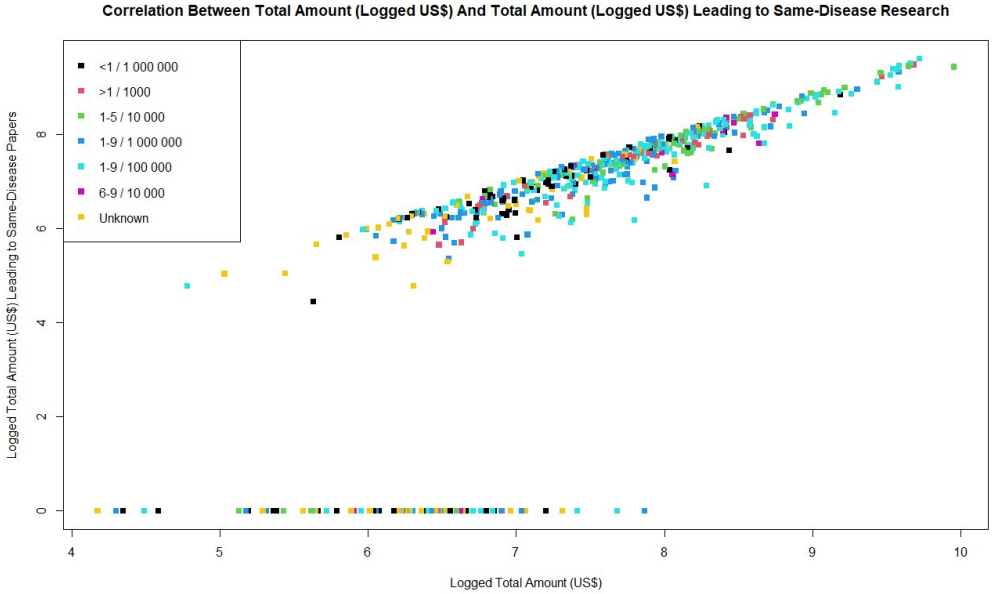
(b) Only NIH projects that are acknowledged by 1+ same-disease paper(s)



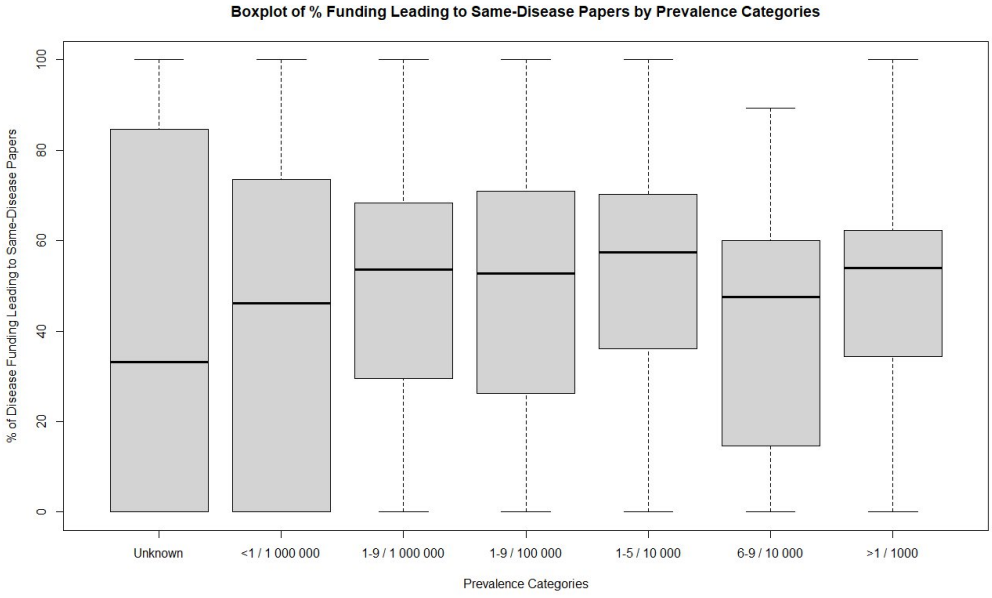
(c) Diseases by cumulative funding categories that do not result in same-disease publications

Figure 9. NIH disease-specific funding and the proportion that leads to same-disease papers

This figure provides additional summaries on the linkages between NIH disease-specific funding and the subsequent amount of scientific literature attributed to the same diseases. Part (a) visualizes the correlation between the cumulative funding amount across diseases and the actual amount that leads to same-disease publications. Part (b) visualizes boxplots of the proportion of disease-specific funding leading to same-disease papers across various prevalence categories set by Orphanet. Part (c) further expands the visualization in part (b), together with information about the median estimates across Orphanet prevalence categories.



(a) Correlation between NIH disease-specific funding amount (logged) and the amount leading to same-disease papers (logged)



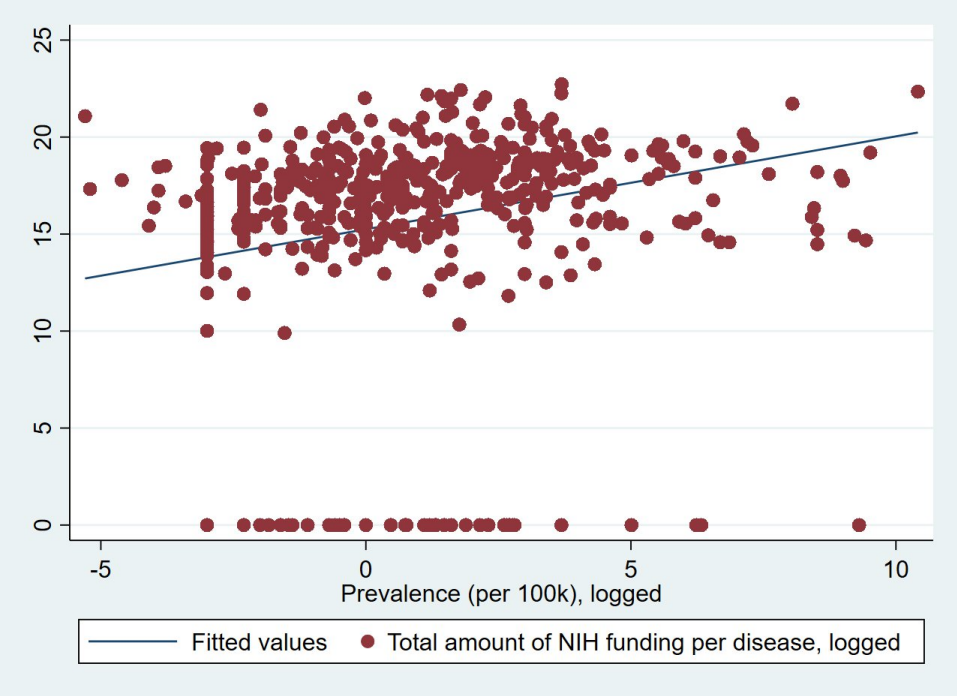
(b) Percentage of NIH disease-funding leading to same-disease papers and prevalence categories

	0	1.0-24.99	25.0-49.99	50.0-74.99	75.0-99.99	100	Median (%)
Unknown	23	11	13	6	13	8	33.11
<1 / 1 000 000	19	6	14	16	13	4	46.19
1-9 / 1 000 000	11	16	30	45	18	7	53.63
1-9 / 100 000	17	19	40	57	24	5	52.79
1-5 / 10 000	7	6	20	38	13	2	57.48
6-9 / 10 000	2	2	4	4	2	0	47.52
>1 / 1000	5	5	11	21	2	1	53.89

(c) Summary of diseases grouped by percentage of NIH funding leading to same-disease papers and Orphanet prevalence categories

Figure 10. NIH rare-disease funding and disease prevalence

This figure plots the distribution of NIH funding (including those that do not lead to same-disease papers) across 700 rare diseases. It is worth noting that the disease list is preliminary and is subject to further analysis after all diseases from NORD In-Depth Reports are accounted for. Some diseases have disproportionately high prevalence rates but are considered rare in the US due to a few reasons, notably due to their tendency to be neglected tropical diseases. The figures for prevalence rates and US$ amount have been log-transformed to simplify the visualization, as shown in part (a). Part (b) expands upon part (a) by visualizing the distribution of cumulative funding across Orphanet prevalence categories.



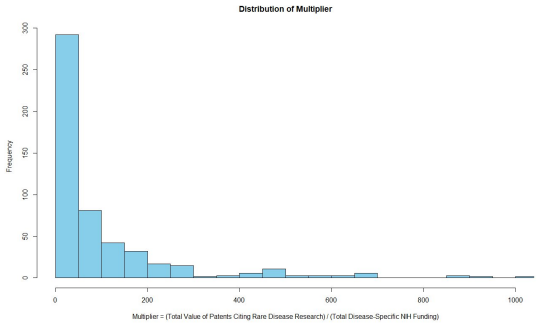
(a) Descriptive correlation of prevalence and amount of disease-specific NIH funding (logged)

Cumulative funding \ Prevalence	Unknown	<1 / 1 000 000	1-9 / 1 000 000	1-9 / 100 000	1-5 / 10 000	6-9 / 10 000	>1 / 1000
>US$1bil	0	1	2	13	8	0	2
US$500mil - US$1bil	0	0	4	5	5	1	1
US$100mil - US$499.99 mil	3	8	18	39	30	5	14
US$50mil - US$99.99mil	4	3	25	21	10	1	8
US$10mil - US$49.99 mil	21	20	38	47	21	3	5
US$1mil - US$9.99 mil	34	31	36	31	8	3	15
<US$1mil	12	9	4	6	4	1	0
No funding	41	33	16	17	9	0	4

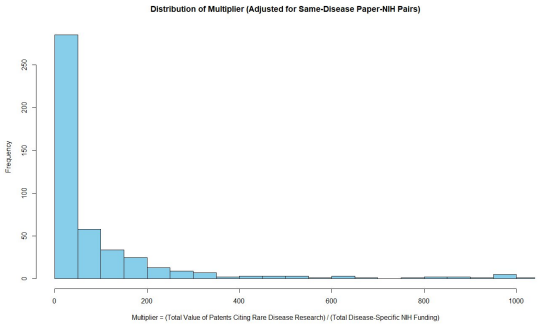
(b) Summary of disease-specific NIH funding by Orphanet prevalence categories

Figure 11. Distribution of ROI to disease-specific research

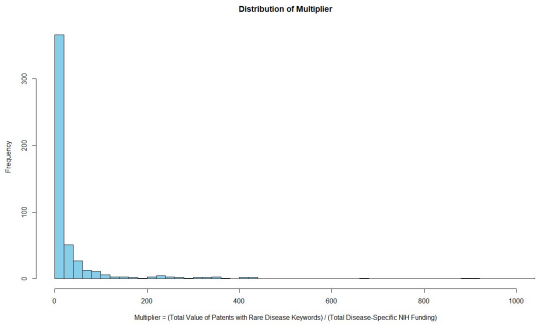
This figure visualizes the distribution of returns on investment (ROI) of NIH funding across various diseases. With regard to three figures on the left—mainly (a), (b), and (c)—the X-axis is proxied by the total value of patents divided by the cumulative amount of disease-specific funding. The remaining three figures on the right—(d), (e), and (f)—somewhat differs from the previous three figures as the denominator is proxied by the cumulative amount of disease-specific funding that leads to same-disease publications.



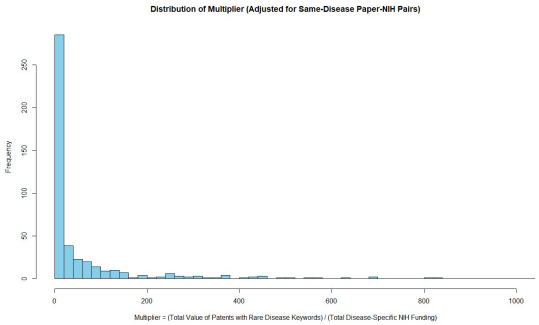
(a) X-axis: Total value of patents citing rare disease research (1995-2023) / Total disease-specific NIH funding



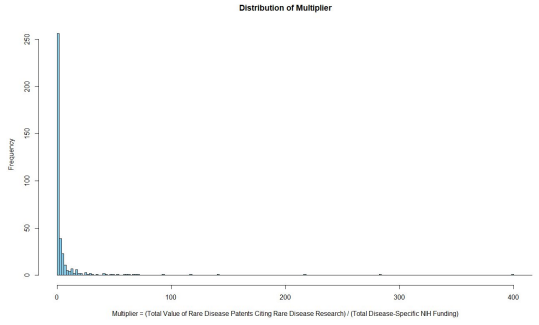
(d) X-axis: Total value of patents citing rare disease research (1995-2023) / Total disease-specific NIH funding resulting in same-disease papers



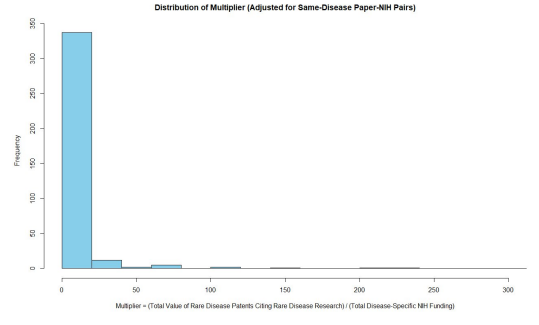
(b) X-axis: Total value of patents with rare disease keywords (1995-2023) / Total disease-specific NIH funding



(e) X-axis: Total value of patents with rare disease keywords (1995-2023) / Total disease-specific NIH funding resulting in same-disease papers



(c) X-axis: Total value of rare disease patents citing rare disease research (1995-2023) / Total disease-specific NIH funding



(f) X-axis: Total value of rare disease patents citing rare disease research (1995-2023) / Total disease-specific NIH funding resulting in same-disease papers

Table 1. List of diseases with cumulative NIH funding amount exceeding US$1 billion, in nominal terms (1995-2023)

Disease Name	Cancer-related?	Total no. of NIH projects	Cumulative amount (US$, nominal)	Cumulative amount (US$, nominal), only funding that leads to same-disease papers
Acute Respiratory Distress Syndrome	N	11461	8,971,212,911	2,658,525,051
Glioma	Y	12679	5,242,800,370	4,019,380,752
Cytomegalovirus Infection	N	8382	4,810,142,292	2,985,337,064
Glioblastoma	Y	11186	4,567,810,977	3,212,280,670
Cystic Fibrosis	N	10489	4,438,910,685	2,774,991,743
Graft versus Host Disease	N	7554	3,846,506,792	2,821,008,327
Hodgkin's Disease	Y	7417	3,805,267,202	2,109,917,311
Astrocytoma	Y	8585	3,774,337,996	1,013,033,675
Hepatocellular Carcinoma	Y	9973	3,747,482,273	2,381,710,827
Acute Myeloid Leukemia	Y	7505	3,480,120,351	2,410,742,668
Amyotrophic Lateral Sclerosis	N	7888	3,355,080,197	1,772,726,331
Dengue Fever	N	4286	2,927,039,304	1,634,095,221
Frontotemporal Degeneration	N	5454	2,881,866,511	2,026,132,355
Huntington's Disease	N	7294	2,725,500,536	1,283,841,125
Anthrax	N	2598	1,997,648,675	882,486,014
Amyloidosis	N	3260	1,830,596,822	709,467,068
Fragile X Syndrome	N	3931	1,644,119,431	958,876,495
Hutchinson-Gilford Progeria Syndrome	N	3585	1,534,044,216	682,030,450
Chronic Lymphocytic Leukemia	Y	2765	1,505,026,448	793,812,314
Cone Dystrophy	N	2939	1,410,321,247	286,238,509
Hemophilia A	N	2797	1,264,431,725	768,940,739
Idiopathic Pulmonary Fibrosis	N	2614	1,195,226,898	858,894,779
Cholera	N	3130	1,151,915,417	696,733,028
Chikungunya	N	1335	1,089,404,777	462,227,204
Esophageal Cancer	Y	1951	1,073,026,769	672,526,206
Chronic Myelogenous Leukemia	Y	2438	1,047,405,122	622,017,571

Note: this table describes the list of most highly-funded diseases, proxied by the cumulative amount of funding (1995-2023) with at least US$1 billion in nominal terms, as shown in the fourth column. A total of 26 diseases are identified from the preliminary list. The second column queries whether a focal disease is related to cancer. The third column summarizes the total number of NIH projects (proxied by their unique 14-digit numbers) related to a focal disease within the 1995-2023 period. The fifth column focuses specifically on the amount of funding that is subsequently associated with same-disease publications.

Table 2. Total value of USPTO patent grants directly/indirectly related to rare disease research

Panel 1: patents regardless of linkages to disease-specific NIH projects

Category	Total count	Total value (US$ mil, nominal)
All patents across 3 categories (with disease double-counting)	119,642	9,116,896
All patents across 3 categories (without disease double-counting)	42,222	2,999,116
Patents citing rare disease research (with disease double-counting)	97,980	7,158,402
Patents citing rare disease research (no disease double-counting)	36,183	2,487,550
Patents with rare disease keywords (with disease double-counting)	16,793	1,525,953
Patents with rare disease keywords (no disease double-counting)	7,474	641,336
Rare disease patents citing rare disease research (with disease double-counting)	4,869	432,542
Rare disease patents citing rare disease research (no disease double-counting)	3,696	297,493

Panel 2: patents that cite focal-disease papers acknowledging same-disease NIH projects

Category	Total count	Total value (US$ mil, nominal)
Patents citing rare disease research (with disease double-counting)	45,539	3,316,225
Patents citing rare disease research (no disease double-counting)	9,636	704,531
Rare disease patents citing rare disease research (with disease double-counting)	3,056	293,337
Rare disease patents citing rare disease research (no disease double-counting)	2,155	180,277

Note: this table summarizes the total count of patents that are related, whether directly or indirectly, to disease-specific NIH projects. Panel 1 counts the number of patents across different categories regardless of the nature of their linkages to these NIH projects. A major caveat with the count figures is that the total sum for each of these three categories may exceed the combined sum when these three categories are combined, especially after removing disease double-counting. As an illustration, a patent on disease A may cite research from the same disease and disease B, but also contains keywords from disease C in its claims. Panel 2 only takes into account patents that cite focal-disease papers (whether of the same disease or not) that also acknowledge same-disease NIH projects.

Table 3. Top 20 diseases based on the value of innovations related to focal-disease research

Category 1: Any patents citing focal-disease research					
Disease Name	**USPTO patent grant count**	**USPTO patent grant count (with KPSS value)**	**Total value (in US$ million, nominal)**	**Total value (in 1983 US$ million)**	**Total NIH funding (in US$ million, nominal)**
Hodgkin's Disease	13,826	3,673	345,867.4	135,077.2	3,805.2
Acute Myeloid Leukemia	13,045	3,616	280,600.1	111,459.4	3,480.1
Glioma	20,836	4,333	271,229.6	111,453.1	5,242.8
Glioblastoma	16,049	3,631	259,061.7	103,849.2	4,567.8
Hepatocellular Carcinoma	15,583	3,306	257,817.6	100,068.1	3,747.5
Chronic Lymphocytic Leukemia	9,553	2,666	223,475.1	89,212.5	1,505.0
Cystic Fibrosis	17,541	3,878	221,312.1	91,253.7	4,438.9
Chronic Myelogenous Leukemia	9,410	2,655	189,887.9	78,739.0	1,047.4
Cytomegalovirus Infection	10,166	2,599	188,546.9	76,003.3	4,810.1
C3 Glomerulopathy: Dense Deposit Disease and C3 Glomerulonephritis	6,594	1,931	162,734.9	66,967.5	3,425.3
Graft versus Host Disease	7,772	1,785	148,241.2	58,544.0	3,846.5
Cholera	8,407	2,008	134,801.7	57,123.4	1,151.9
Acute Respiratory Distress Syndrome	5,536	1,485	129,683.6	53,923.0	8,971.2
Follicular Lymphoma	5,072	1,574	123,274.2	49,916.7	4,199.5
Hemophilia A	7,489	1,590	119,491.4	50,233.8	1,264.4
HTLV Type I and Type II	6,811	1,528	115,035.4	45,471.5	4,686.8
Astrocytoma	5,546	1,395	102,606.0	41,831.5	3,774.3
Amyotrophic Lateral Sclerosis	6,015	1,211	100,140.3	40,472.7	3,355.1
Familial Hypercholesterolemia	4,477	1,000	92,118.9	38,258.2	3,772.7
Huntington's Disease	6,612	1,150	86,530.4	35,415.3	2,725.5
Category 2: Patents containing disease-specific keywords (regardless of linkages with focal-disease research)					
Disease Name	**USPTO patent grant count**	**USPTO patent grant count (with KPSS value)**	**Total value (in US$ million, nominal)**	**Total value (in 1983 US$ million)**	**Total NIH funding (in US$ million, nominal)**
Arthritis, Infectious	1,686	558	58,708.5	29,430.9	1,682.3
Huntington's Disease	2,328	587	56,113.9	26,485.4	2,725.5
Chronic Myelogenous Leukemia	1,768	483	54,169.2	21,092.9	1,047.4
Amyotrophic Lateral Sclerosis	2,023	469	53,034.7	24,223.5	3,355.1
Esophageal Cancer	2,466	406	46,620.8	16,874.6	1,073.0
Glioblastoma	2,001	446	44,808.9	17,209.9	4,567.8
Acute Respiratory Distress Syndrome	1,418	441	44,184.0	22,175.0	8,971.2
Hodgkin's Disease	2,213	518	44,006.0	18,436.4	3,805.3
Guillain-Barré Syndrome	899	277	41,528.9	18,439.5	2,482.1
Cluster Headache	437	211	38,067.7	18,368.3	7,984.9
Chronic Lymphocytic Leukemia	1,596	373	36,924.4	14,615.3	1,505.0

Cystic Fibrosis	1,735	397	36,494.7	16,240.9	4,438.9
Acute Myeloid Leukemia	1,408	360	36,224.9	13,411.5	3,480.1
Glioma	1,763	329	33,503.7	13,781.0	5,242.8
HTLV Type I and Type II	1,084	262	32,876.9	12,479.8	468.7
Hepatocellular Carcinoma	1,377	273	25,534.9	9,414.1	3,747.5
Graft versus Host Disease	1,181	306	25,299.8	12,022.8	3,846.5
C3 Glomerulopathy: Dense Deposit Disease and C3 Glomerulonephritis	1,176	374	22,774.8	10,431.8	34.3
Follicular Lymphoma	540	158	22,549.1	8,555.9	420.0
Cholangiocarcinoma	819	132	19,156.5	6,775.5	420.3

Category 3: Patents with disease-specific keywords that cite same-disease research

Disease Name	USPTO patent grant count	USPTO patent grant count (with KPSS value)	Total value (in US$ million, nominal)	Total value (in 1983 US$ million)	Total NIH funding (in US$ million, nominal)
Acute Myeloid Leukemia	1,032	294	35,837.4	13,154.3	3,480.1
Glioblastoma	1,188	210	34,577.9	12,450.8	4,567.8
Cystic Fibrosis	1,290	359	31,087.9	11,971.0	4,438.9
Hodgkin's Disease	898	292	28,061.0	10,602.4	3,805.3
Chronic Lymphocytic Leukemia	740	208	19,973.2	7,808.2	1,505.0
Chronic Myelogenous Leukemia	552	149	15,120.1	5,717.5	1,047.4
Hepatocellular Carcinoma	752	145	14,979.7	5,243.5	3,747.5
Huntington's Disease	850	189	14,669.0	6,148.5	2,725.5
HTLV Type I and Type II	349	84	13,067.7	4,532.1	468.7
Glioma	1,031	157	12,234.1	4,807.8	5,242.8
Cytomegalovirus Infection	563	146	12,141.8	5,727.0	4,810.1
Graft versus Host Disease	560	109	11,859.5	4,710.1	3,846.5
Amyotrophic Lateral Sclerosis	825	139	10,884.8	4,334.8	3,355.1
Desmoid Tumor	17	9	10,210.2	3,427.6	14.0
Follicular Lymphoma	272	98	8,611.0	3,291.2	420.0
C3 Glomerulopathy: Dense Deposit Disease and C3 Glomerulonephritis	306	81	7,425.5	3,099.1	34.3
Cholangiocarcinoma	124	24	7,321.0	2,509.1	420.3
Arthritis, Infectious	143	46	6,992.1	3,515.6	168.2
Frontotemporal Degeneration	332	54	6,940.6	2,498.5	2,881.9
Hemophilia A	464	110	6,241.4	2,633.0	1,264.4

Note: this table provides descriptive statistics of the count number of patented innovations and the cumulative value of these patents across different patenting categories. The second column is proxied by the total number of USPTO patent grants related to a disease. The third column focuses only subsets of these patents with available KPSS value estimates from Kogan et al. (2017). The fourth column provides the aggregate value of patents among the KPSS subset in nominal terms. The fifth column provides the value in 1983-adjusted US$. The sixth column refers to the cumulative funding amount, regardless of whether the funding leads to same-disease papers.

Table 4. Top 20 diseases based on the value of innovations related to focal-disease research, only after accounting for NIH funding acknowledged by same-disease publications

Category 1: Any patents citing focal-disease research					
Disease Name	**USPTO patent grant count**	**USPTO patent grant count (with KPSS value)**	**Total value (in US$ million, nominal)**	**Total value (in 1983 US$ million)**	**Total NIH funding (in US$ million, nominal)**
Hodgkin's Disease	13826	1771	157,762.7	60,104.9	2,109.9
Acute Myeloid Leukemia	13045	1824	136,568.0	52,676.4	2,410.7
Glioma	20836	2154	135,701.7	52,285.5	4,019.4
Glioblastoma	16049	1867	133,403.4	50,823.9	3,212.3
Chronic Lymphocytic Leukemia	9553	1519	129,466.9	48,741.0	793.8
Hepatocellular Carcinoma	15583	1365	119,133.1	43,821.1	2,381.7
Cystic Fibrosis	17541	1630	99,699.7	39,202.5	2,775.0
Chronic Myelogenous Leukemia	9410	1378	97,444.5	38,374.2	622.0
Cytomegalovirus Infection	10166	1216	81,846.8	31,750.6	2,985.3
Graft versus Host Disease	7772	1025	80,229.9	30,709.4	2,821.0
HTLV Type I and Type II	6811	931	70,136.4	26,490.8	335.8
Hemophilia A	7489	831	65,668.7	25,379.5	768.9
Amyotrophic Lateral Sclerosis	6015	741	59,098.3	22,739.8	1,772.7
C3 Glomerulopathy: Dense Deposit Disease and C3 Glomerulonephritis	6594	720	57,683.1	22,397.2	8.6
Follicular Lymphoma	5072	811	57,544.9	21,951.3	277.0
Astrocytoma	5546	656	54,817.0	20,886.4	1,013.0
Familial Hypercholesterolemia	4477	476	48,472.3	19,265.9	255.8
Acute Respiratory Distress Syndrome	5536	527	46,746.4	17,865.8	2,658.5
Amyloidosis	3802	466	41,316.4	16,001.7	709.5
Huntington's Disease	6612	601	40,352.1	15,305.5	1,283.8
Category 2: Patents with disease-specific keywords that cite same-disease research					
Disease Name	**USPTO patent grant count**	**USPTO patent grant count (with KPSS value)**	**Total value (in US$ million, nominal)**	**Total value (in 1983 US$ million)**	**Total NIH funding (in US$ million, nominal)**
Glioblastoma	1188	158	28,141.3	9,992.0	3,212.3
Acute Myeloid Leukemia	1032	223	27,026.9	9,831.9	2,410.7
Cystic Fibrosis	1290	246	24,524.7	9,433.9	2,775.0
Hodgkin's Disease	898	188	19,323.5	7,181.6	2,109.9
Chronic Lymphocytic Leukemia	740	161	15,713.7	6,093.1	793.8
Chronic Myelogenous Leukemia	552	121	14,016.2	5,256.8	622.0
HTLV Type I and Type II	349	59	11,858.0	3,999.4	335.8
Hepatocellular Carcinoma	752	91	11,011.5	3,811.4	2,381.7
Glioma	1031	102	9,748.5	3,598.6	4,019.4
Graft versus Host Disease	560	75	8,200.4	3,206.1	2,821.0
Desmoid Tumor	17	4	7,998.9	2,622.9	4.0

Amyotrophic Lateral Sclerosis	825	82	7,265.9	2,863.7	1,772.7
Huntington's Disease	850	108	6,440.1	2,620.4	1,283.8
Cholangiocarcinoma	124	15	5,818.0	1,957.0	299.4
Follicular Lymphoma	272	62	5,316.6	1,981.1	277.0
Cytomegalovirus Infection	563	99	5,005.8	1,982.9	2,985.3
Adenoid Cystic Carcinoma	6	3	4,076.0	1,374.7	19.5
Cholera	330	27	3,578.3	1,421.2	696.7
Frontotemporal Degeneration	332	34	3,514.8	1,292.0	2,026.1
Esophageal Cancer	291	42	3,319.5	1,220.9	672.5

Note: this table re-estimates the number of patented innovations and the value of patents that directly acknowledges NIH funding. The second column refers to the total number of patents related to a disease, regardless of the nature of linkages to NIH funding. The third column shows count figures for patents that already acknowledge NIH funding. The fourth column provides the aggregate value of patents among the KPSS subset in nominal terms. The fifth column provides the value in 1983-adjusted US$. The sixth column refers to the cumulative funding amount that eventually results in same-disease publications.

Table 5. Summary statistics

Variable	Mean	p25	p50	p75	Min	Max
NIH projects (same-disease paper match)	7.878	0	1	4	0	546
NIH funding amount (same-disease paper match, US$ mil, nominal)	3.964	0	0.123	1.662	0	299.190
Patents citing rare disease-specific research	11.442	0	1	6	0	953
Patents with rare disease keywords	4.005	0	0	2	0	320
Rare disease patents citing rare disease research	0.822	0	0	0	0	110
Patents citing rare disease-specific research with KPSS value	2.655	0	0	1	0	215
Patents with rare disease keywords with KPSS value	0.936	0	0	0	0	63
Rare disease patents citing rare disease research with KPSS value	0.179	0	0	0	0	39
KPSS value (US$ mil, nominal) of patents citing rare disease-specific research	194.660	0	0	17.784	0	21,599.630
KPSS value (US$ mil, nominal) of patents with rare disease keywords	88.344	0	0	0	0	9,854.467
KPSS value (US$ mil, nominal) of rare disease patents citing rare disease research	17.366	0	0	0	0	6,167.088

Note: the descriptive statistics are displayed at the disease-year level of observation. As the data encompasses 580 diseases with NIH funding and 29 fiscal years (1995-2023), the total number of observations is 580 * 29 = 16,280. These figures already take into account only patents that cite focal-disease papers acknowledging same-disease grants, as well as same-disease grants associated with at least one same-disease publication.

Table 6. Baseline analysis with different time lags

Panel 1: Patents citing rare disease-specific research							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	No lag	1-year lag	2-year lag	3-year lag	4-year lag	5-year lag	6-year lag
NIH projects	1.057***	1.059***	1.063***	1.069***	1.102***	1.166***	1.200***
	(0.11)	(0.11)	(0.10)	(0.09)	(0.09)	(0.10)	(0.10)
Constant	3.517***	3.866***	4.347***	4.857***	5.215***	5.414***	5.846***
	(0.88)	(0.83)	(0.76)	(0.69)	(0.66)	(0.70)	(0.69)
R-squared	0.787	0.789	0.796	0.801	0.810	0.822	0.830
Adjusted R-squared	0.775	0.777	0.784	0.789	0.798	0.810	0.818
Within R-squared	0.334	0.328	0.322	0.314	0.317	0.331	0.330
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	580	580	580	580	580	580	580
Disease FE	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y
Disease age FE	Y	Y	Y	Y	Y	Y	Y
N	16,023	15,459	14,895	14,331	13,767	13,201	12,634
Mean DV	11.845	12.203	12.649	13.078	13.531	13.997	14.429
Panel 2: Patents with rare disease-specific keywords							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	No lag	1-year lag	2-year lag	3-year lag	4-year lag	5-year lag	6-year lag
NIH projects	0.238***	0.242***	0.243***	0.246***	0.256***	0.275***	0.288***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)
Constant	2.282***	2.369***	2.505***	2.624***	2.709***	2.728***	2.784***
	(0.26)	(0.27)	(0.26)	(0.26)	(0.27)	(0.27)	(0.27)
R-squared	0.734	0.741	0.749	0.756	0.763	0.773	0.780
Adjusted R-squared	0.720	0.727	0.734	0.741	0.748	0.758	0.765
Within R-squared	0.186	0.189	0.185	0.183	0.186	0.199	0.204
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	580	580	580	580	580	580	580
Disease FE	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y
Disease age FE	Y	Y	Y	Y	Y	Y	Y
N	16,023	15,459	14,895	14,331	13,767	13,201	12,634
Mean DV	4.155	4.271	4.400	4.517	4.641	4.752	4.844
Panel 3: Rare disease patents citing rare disease-specific research							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	No lag	1-year lag	2-year lag	3-year lag	4-year lag	5-year lag	6-year lag
NIH projects	0.104***	0.103***	0.102***	0.102***	0.105***	0.108***	0.108***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.031	0.069	0.114	0.155+	0.183*	0.214**	0.265***
	(0.10)	(0.09)	(0.09)	(0.08)	(0.08)	(0.07)	(0.07)
R-squared	0.732	0.734	0.737	0.740	0.746	0.753	0.756
Adjusted R-squared	0.718	0.719	0.722	0.724	0.730	0.737	0.742
Within R-squared	0.305	0.291	0.279	0.267	0.262	0.257	0.243
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	580	580	580	580	580	580	580
Disease FE	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y
Disease age FE	Y	Y	Y	Y	Y	Y	Y

N	16,023	15,459	14,895	14,331	13,767	13,201	12,634
Mean DV	0.850	0.877	0.911	0.940	0.971	1.005	1.038

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: this table displays the initial OLS regression results for the impacts of NIH funding on the subsequent amount of innovation. 'NIH projects' are proxied by the total number of disease-specific NIH projects per disease-year observation that lead to at least one same-disease publication in the subsequent years. The dependent variables across the three panels are proxied by the count number of patents related to a disease. The analysis is run across different models, from no-lag model in Column (1) to six-year lags in Column (7). Standard errors are clustered at the disease year. The results have taken into account the presence of disease, year, and disease-age fixed effects.

Table 7. Analysis with different time lags and only on patent counts with KPSS value

Panel 1: Patents citing rare disease-specific research							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	No lag	1-year lag	2-year lag	3-year lag	4-year lag	5-year lag	6-year lag
NIH projects	0.221***	0.226***	0.229***	0.232***	0.239***	0.257***	0.273***
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	1.019***	1.050***	1.148***	1.250***	1.337***	1.351***	1.382***
	(0.21)	(0.20)	(0.19)	(0.18)	(0.17)	(0.17)	(0.17)
R-squared	0.720	0.726	0.734	0.742	0.751	0.765	0.776
Adjusted R-squared	0.705	0.711	0.719	0.726	0.735	0.749	0.760
Within R-squared	0.224	0.229	0.227	0.224	0.224	0.239	0.253
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	580	580	580	580	580	580	580
Disease FE	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y
Disease age FE	Y	Y	Y	Y	Y	Y	Y
N	16,023	15,459	14,895	14,331	13,767	13,201	12,634
Mean DV	2.758	2.828	2.932	3.031	3.138	3.240	3.335
Panel 2: Patents with rare disease-specific keywords							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	No lag	1-year lag	2-year lag	3-year lag	4-year lag	5-year lag	6-year lag
NIH projects	0.029**	0.030**	0.029*	0.028*	0.030**	0.032**	0.032**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.752***	0.765***	0.794***	0.821***	0.828***	0.828***	0.837***
	(0.09)	(0.09)	(0.09)	(0.09)	(0.08)	(0.08)	(0.07)
R-squared	0.639	0.643	0.649	0.656	0.659	0.665	0.672
Adjusted R-squared	0.619	0.622	0.628	0.635	0.638	0.643	0.649
Within R-squared	0.039	0.041	0.037	0.035	0.037	0.040	0.039
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	580	580	580	580	580	580	580
Disease FE	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y
Disease age FE	Y	Y	Y	Y	Y	Y	Y
N	16,023	15,459	14,895	14,331	13,767	13,201	12,634
Mean DV	0.979	0.998	1.019	1.040	1.054	1.067	1.069
Panel 3: Rare disease patents citing rare disease-specific research							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	No lag	1-year lag	2-year lag	3-year lag	4-year lag	5-year lag	6-year lag
NIH projects	0.021***	0.021***	0.021***	0.021***	0.022***	0.023***	0.023***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.022	0.027	0.038	0.042	0.047	0.052*	0.061**
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)
R-squared	0.577	0.585	0.591	0.600	0.609	0.621	0.630
Adjusted R-squared	0.554	0.562	0.567	0.575	0.584	0.595	0.605
Within R-squared	0.127	0.125	0.117	0.117	0.116	0.115	0.112
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	580	580	580	580	580	580	580
Disease FE	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y
Disease age FE	Y	Y	Y	Y	Y	Y	Y

N	16,023	15,459	14,895	14,331	13,767	13,201	12,634
Mean DV	0.186	0.191	0.198	0.204	0.211	0.218	0.226

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: this table provides estimates from another OLS regression, with the dependent variables proxied by the subset of patent count figures with available KPSS value estimates from Kogan et al. (2017). The 'NIH projects' variable is still proxied by the total count number of disease-specific NIH projects leading to same-disease publications. In general, the coefficient size across these models shrinks by 80-90% after taking into consideration only patents with KPSS value. The analysis is run across various models from no-lag to six-year-lag conditions. Standard errors are clustered at the disease level. These results have taken into account the presence of disease, year, and disease-age fixed effects.

Table 8. Estimation of the value of direct impacts and spillovers of NIH funding on the KPSS value of patented innovations

Panel 1: KPSS value (US$ mil, nominal) of patents citing rare disease-specific research							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	No lag	1-year lag	2-year lag	3-year lag	4-year lag	5-year lag	6-year lag
NIH amount (US$ mil)	34.566***	35.477***	37.265***	40.558***	45.495***	47.526***	50.554***
	(4.68)	(4.49)	(4.28)	(4.33)	(4.93)	(5.43)	(5.92)
Constant	67.080***	72.367***	76.947***	77.235***	73.183***	80.918***	86.435***
	(18.53)	(17.32)	(15.99)	(15.63)	(17.08)	(18.05)	(18.91)
R-squared	0.611	0.614	0.620	0.634	0.649	0.659	0.670
Adjusted R-squared	0.590	0.592	0.598	0.611	0.626	0.636	0.647
Within R-squared	0.183	0.177	0.175	0.186	0.204	0.206	0.212
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	580	580	580	580	580	580	580
Disease FE	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y
Disease age FE	Y	Y	Y	Y	Y	Y	Y
N	16,023	15,459	14,895	14,331	13,767	13,201	12,634
Mean DV (US$ mil)	204.088	210.867	218.950	226.982	235.294	243.623	252.877
Panel 2: KPSS value (US$ mil, nominal) of patents with rare disease-specific keywords							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	No lag	1-year lag	2-year lag	3-year lag	4-year lag	5-year lag	6-year lag
NIH amount (US$ mil)	6.572*	7.085*	7.172*	7.120*	8.191*	10.453**	11.823***
	(2.84)	(3.00)	(3.15)	(3.43)	(3.55)	(3.44)	(3.27)
Constant	66.518***	67.787***	70.196***	73.321***	70.506***	61.405***	56.347***
	(11.25)	(11.60)	(11.77)	(12.35)	(12.28)	(11.44)	(10.45)
R-squared	0.458	0.465	0.470	0.473	0.481	0.513	0.534
Adjusted R-squared	0.428	0.435	0.438	0.441	0.448	0.481	0.501
Within R-squared	0.028	0.030	0.027	0.024	0.029	0.052	0.067
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	580	580	580	580	580	580	580
Disease FE	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y
Disease age FE	Y	Y	Y	Y	Y	Y	Y
N	16,023	15,459	14,895	14,331	13,767	13,201	12,634
Mean DV (US$ mil)	92.569	95.447	97.524	99.608	99.694	97.193	95.274
Panel 3: KPSS value (US$ mil, nominal) of rare disease patents citing rare disease-specific research							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	No lag	1-year lag	2-year lag	3-year lag	4-year lag	5-year lag	6-year lag
NIH amount (US$ mil)	5.326***	5.298***	5.715***	6.124***	6.910***	7.306***	7.424***
	(1.22)	(1.22)	(1.25)	(1.32)	(1.43)	(1.44)	(1.46)
Constant	-2.845	-1.633	-1.857	-1.910	-3.104	-2.723	-1.211
	(4.83)	(4.71)	(4.67)	(4.77)	(4.97)	(4.81)	(4.67)
R-squared	0.396	0.397	0.405	0.412	0.424	0.433	0.438
Adjusted R-squared	0.363	0.363	0.370	0.376	0.388	0.396	0.399
Within R-squared	0.086	0.077	0.080	0.081	0.090	0.091	0.084
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	580	580	580	580	580	580	580
Disease FE	Y	Y	Y	Y	Y	Y	Y

Year FE	Y	Y	Y	Y	Y	Y	Y
Disease age FE	Y	Y	Y	Y	Y	Y	Y
N	16,023	15,459	14,895	14,331	13,767	13,201	12,634
Mean DV (US$ mil)	18.267	19.050	19.919	20.701	21.517	22.290	23.231

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: this table summarizes results from another OLS regression, with "NIH amount (US$ mil" proxied by, at the disease-year level of observation, the total amount of same-disease NIH funding (in US$ million, nominal) that eventually leads to same-disease publications. The dependent variables are proxied by the total value of patents (in US$ million, nominal) across these three patenting categories. Standard errors are clustered at the disease level, and the results have taken into account the inclusion of disease, year, and disease-age fixed effects.

Table 9. Instrument variable regression on the impacts of NIH disease-specific funding on the count number of patents

Panel 1: Any same-disease publications				
	1st Stage		2nd Stage	
	(1)	(2) No. of patents citing rare disease research	(3) No. of patents with rare disease-specific keywords	(4) No. of rare disease patents citing rare disease research
L1.NIH projects	0.606***			
	(0.022)			
L1.Same-disease papers		1.747***	0.398***	0.169***
		(0.160)	(0.050)	(0.017)
Constant	0.393*			
	(0.173)			
N	15,459	15,459	15,459	15,459
Kleibergen-Paap Wald F	769.042			
Stock-Yogo F-test (10%)	16.38			
R-squared	0.951			
Adjusted R-squared	0.948			
Within R-squared	0.770			
Clustered standard error	Disease			
Disease FE	Y			
Year FE	Y			
Disease age FE	Y			
Mean DV	5.166	9.092	2.074	0.880
Panel 1: Any same-disease publications within 5 years (Azoulay et al., 2019)				
	1st Stage		2nd Stage	
	(1)	(2) No. of patents citing rare disease research	(3) No. of patents with rare disease-specific keywords	(4) No. of rare disease patents citing rare disease research
L1.NIH projects	0.590***			
	(0.021)			
L1.Same-disease papers		1.795***	0.409***	0.174***
		(0.165)	(0.052)	(0.017)
Constant	0.319+			
	(0.162)			
N	15,459	15,459	15,459	15,459
Kleibergen-Paap Wald F	824.316			
Stock-Yogo F-test (10%)	16.38			
R-squared	0.951			
Adjusted R-squared	0.948			
Within R-squared	0.773			

Clustered standard error	Disease			
Disease FE	Y			
Year FE	Y			
Disease age FE	Y			
Mean DV	4.964	8.975	2.047	0.869

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: this table displays the results of instrument-variable (IV), two-stage least-squares regression, with NIH projects as the instrument, and the total number of disease-specific publications as the first-stage dependent variable. The dependent variables are proxied by patent count figures across different patenting categories. The analysis is run by assuming a one-year time lag between the NIH projects and the second-stage dependent variables. Results remain robust after the first-stage dependent variable is replaced by only publication count within five years of the focal-disease NIH project, following Azoulay et al. (2019). Standard errors are clustered at the disease level, and the results have taken into account the inclusion of disease, year, and disease-age fixed effects.

Table 10. Instrument variable regression on the impacts of NIH disease-specific funding on the count number of patents with KPSS value

Panel 1: Any same-disease publications				
	1st Stage		2nd Stage	
	(1)	(2) No. of patents citing rare disease research with KPSS value	(3) No. of patents with rare disease-specific keywords with KPSS value	(4) No. of rare disease patents citing rare disease research with KPSS value
L1.NIH projects	0.606***			
	(0.022)			
L1.Same-disease papers		0.373***	0.049**	0.034***
		(0.038)	(0.018)	(0.007)
Constant	0.393*			
	(0.173)			
N	15,459	15,459	15,459	15,459
Kleibergen-Paap Wald F	769.042			
Stock-Yogo F-test (10%)	16.38			
R-squared	0.951			
Adjusted R-squared	0.948			
Within R-squared	0.770			
Clustered standard error	Disease			
Disease FE	Y			
Year FE	Y			
Disease age FE	Y			
Mean DV	5.166	1.939	0.255	0.178
Panel 1: Any same-disease publications within 5 years (Azoulay et al., 2019)				
	1st Stage		2nd Stage	
	(1)	(2) No. of patents citing rare disease research with KPSS value	(3) No. of patents with rare disease-specific keywords with KPSS value	(4) No. of rare disease patents citing rare disease research with KPSS value
L1.NIH projects	0.590***			
	(0.021)			
L1.Same-disease papers		0.383***	0.050**	0.035***
		(0.039)	(0.018)	(0.007)
Constant	0.319+			
	(0.162)			
N	15,459	15,459	15,459	15,459
Kleibergen-Paap Wald F	824.316			

Stock-Yogo F-test (10%)	16.38			
R-squared	0.951			
Adjusted R-squared	0.948			
Within R-squared	0.773			
Clustered standard error	Disease			
Disease FE	Y			
Year FE	Y			
Disease age FE	Y			
Mean DV	4.964	1.914	0.251	0.176

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: this table provides results from the instrument-variable (IV), two-stage least-squares regression in a way similar to those provided in Table 9. The difference is that the dependent variables in Table 9 only account for the presence of KPSS values provided by Kogan et al. (2017). Results remain robust after the first-stage dependent variable is replaced by the count number of same-disease publications only within five years of the NIH project period, following Azoulay et al. (2019). Standard errors are clustered at the disease level, and results have taken into account the inclusion of disease, year, and disease-age fixed effects. The analysis assumes a one-year time lag between the NIH projects and the dependent variables.

Table 11. Instrument variable regression on the impacts of NIH disease-specific funding on the aggregate KPSS value of patents

Panel 1: Any same-disease publications				
	1st Stage		2nd Stage	
	(1)	(2) KPSS value of patents citing rare disease research	(3) KPSS value of patents with rare disease-specific keywords	(4) KPSS value of rare disease patents citing rare disease research
L1.NIH amount (US$ mil)	1.012***			
	(0.077)			
L1.Same-disease papers		35.050***	7.000*	5.234***
		(4.140)	(2.763)	(1.003)
Constant	1.280***			
	(0.295)			
N	15,459	15,459	15,459	15,459
Kleibergen-Paap Wald F	174.928			
Stock-Yogo F-test (10%)	16.38			
R-squared	0.930			
Adjusted R-squared	0.926			
Within R-squared	0.674			
Clustered standard error	Disease			
Disease FE	Y			
Year FE	Y			
Disease age FE	Y			
Mean DV	5.231	182.452	36.439	27.247
Panel 1: Any same-disease publications within 5 years (Azoulay et al., 2019)				
	1st Stage		2nd Stage	
	(1)	(2) KPSS value of patents citing rare disease research	(3) KPSS value of patents with rare disease-specific keywords	(4) KPSS value of rare disease patents citing rare disease research
L1.NIH amount (US$ mil)	0.983***			
	(0.076)			
L1.Same-disease papers		36.097***	7.209*	5.391***
		(4.360)	(2.839)	(1.033)
Constant	1.190***			
	(0.292)			
N	15,459	15,459	15,459	15,459
Kleibergen-Paap Wald F	169.296			
Stock-Yogo F-test (10%)	16.38			
R-squared	0.930			

Adjusted R-squared	0.926			
Within R-squared	0.673			
Clustered standard error	Disease			
Disease FE	Y			
Year FE	Y			
Disease age FE	Y			
Mean DV	5.027	180.496	36.048	26.955

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: this table provides results from the instrument-variable (IV), two-stage least-squares regression, with NIH funding proxied by the total amount of disease-specific funding (in US$ million, nominal) that subsequently leads to same-disease publications. The first-stage dependent variable is proxied by the total number of disease-specific publications, with and without the five-year limit following Azoulay et al. (2019). The second-stage dependent variables are proxied by the total value (in US$ million, nominal) of focal-disease patents. Standard errors are clustered at the disease level, and the results have taken into account the inclusion of disease, year, and disease-age fixed effects. The analysis assumes a one-year lag between the NIH funding and the second-stage dependent variables.

Table 12. Heterogeneity analyses of NIH funding and impacts on the value of same-disease innovations and spillovers

Panel 1: above vs below median comparison						
	Above median cumulative funding (median = US$13.3 million)			Below median cumulative funding (median = US$13.3 million)		
	(1) Patents citing rare disease research	(2) Patents with rare disease keywords	(3) Rare disease patents citing rare disease research	(4) Patents citing rare disease research	(5) Patents with rare disease keywords	(6) Rare disease patents citing rare disease research
L1.NIH amount (US$ mil)	34.304***	6.743*	5.178***	22.643	10.406	3.647
	(4.45)	(3.08)	(1.23)	(16.66)	(8.23)	(2.41)
Constant	111.000**	113.294***	-3.571	43.374***	23.650***	1.455***
	(33.58)	(23.22)	(9.30)	(1.81)	(0.89)	(0.26)
R-squared	0.629	0.474	0.427	0.407	0.373	0.122
Adjusted R-squared	0.602	0.435	0.384	0.364	0.328	0.060
Within R-squared	0.173	0.029	0.078	0.0009	0.0004	0.0002
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	290	290	290	290	290	290
Disease FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Disease age FE	Y	Y	Y	Y	Y	Y
N	7,804	7,804	7,804	7,632	7,632	7,632
Mean DV (US$ million)	375.075	165.201	36.290	45.861	24.793	1.856
Panel 2: above vs below average						
	Above average cumulative funding (average = US$115 million)			Below average cumulative funding (average = US$115 million)		
	(1) Patents citing rare disease research	(2) Patents with rare disease keywords	(3) Rare disease patents citing rare disease research	(4) Patents citing rare disease research	(5) Patents with rare disease keywords	(6) Rare disease patents citing rare disease research
L1.NIH amount (US$ mil)	26.470***	5.501	4.565**	9.439*	4.945	0.406
	(4.71)	(3.45)	(1.39)	(3.92)	(4.61)	(0.69)
Constant	398.394***	247.648**	5.840	65.673***	42.433***	3.266***
	(99.22)	(72.78)	(29.28)	(2.75)	(3.23)	(0.48)
R-squared	0.684	0.513	0.492	0.370	0.397	0.116
Adjusted R-squared	0.635	0.439	0.414	0.333	0.361	0.063
Within R-squared	0.107	0.021	0.055	0.002	0.0006	0.000
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	89	89	89	491	491	491
Disease FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Disease age FE	Y	Y	Y	Y	Y	Y
N	2,397	2,397	2,397	13,055	13,055	13,055
Mean DV (US$ million)	968.854	366.206	104.211	72.329	45.920	3.552

Panel 3: above vs below 90th percentile						
	Above 90th percentile cumulative funding (90th percentile = US$192 million)			Below 90th percentile cumulative funding (90th percentile = US$192 million)		
	(1) Patents citing rare disease research	(2) Patents with rare disease keywords	(3) Rare disease patents citing rare disease research	(4) Patents citing rare disease research	(5) Patents with rare disease keywords	(6) Rare disease patents citing rare disease research
L1.NIH amount (US$ mil)	25.821***	5.735	4.831**	13.910**	7.828*	0.666
	(5.67)	(4.03)	(1.65)	(4.23)	(3.86)	(0.42)
Constant	574.242**	324.857**	7.824	68.731***	42.703***	3.271***
	(168.49)	(119.66)	(49.01)	(4.02)	(3.67)	(0.40)
R-squared	0.705	0.525	0.510	0.375	0.390	0.114
Adjusted R-squared	0.640	0.422	0.402	0.338	0.355	0.062
Within R-squared	0.099	0.022	0.054	0.006	0.003	0.0002
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	58	58	58	522	522	522
Disease FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Disease age FE	Y	Y	Y	Y	Y	Y
N	1,563	1,563	1,563	13,892	13,892	13,892
Mean DV (US$ million)	1360.141	499.423	154.859	82.026	50.185	3.908
Panel 4: has advocacy organization vs no advocacy organization						
	Focal disease has advocacy organization(s)			Focal disease has no advocacy organization(s)		
	(1) Patents citing rare disease research	(2) Patents with rare disease keywords	(3) Rare disease patents citing rare disease research	(4) Patents citing rare disease research	(5) Patents with rare disease keywords	(6) Rare disease patents citing rare disease research
L1.NIH amount (US$ mil)	36.909***	7.034*	5.371***	19.212*	5.746*	2.268
	(4.71)	(3.29)	(1.30)	(8.96)	(2.83)	(1.71)
Constant	82.624***	77.981***	-1.255	57.718***	45.556***	1.690
	(22.33)	(15.58)	(6.17)	(15.16)	(4.78)	(2.90)
R-squared	0.637	0.477	0.434	0.494	0.518	0.208
Adjusted R-squared	0.614	0.444	0.398	0.442	0.469	0.126
Within R-squared	0.186	0.031	0.080	0.053	0.011	0.010
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	412	412	412	168	168	168
Disease FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Disease age FE	Y	Y	Y	Y	Y	Y
N	11,018	11,018	11,018	4,422	4,422	4,422
Mean DV (US$ million)	257.430	111.297	24.184	90.222	55.277	5.527

Panel 5: disease is rare cancer vs not rare cancer						
	Rare cancer			Not rare cancer		
	(1) Patents citing rare disease research	(2) Patents with rare disease keywords	(3) Rare disease patents citing rare disease research	(4) Patents citing rare disease research	(5) Patents with rare disease keywords	(6) Rare disease patents citing rare disease research
L1.NIH amount (US$ mil)	53.432***	23.718***	8.501***	28.910***	-0.684	2.999***
	(11.32)	(4.50)	(1.80)	(5.82)	(2.98)	(0.78)
Constant	369.556**	173.096***	38.485*	62.403***	74.922***	1.028
	(102.09)	(40.62)	(16.23)	(18.03)	(9.23)	(2.43)
R-squared	0.675	0.659	0.503	0.592	0.477	0.424
Adjusted R-squared	0.589	0.569	0.372	0.569	0.447	0.391
Within R-squared	0.156	0.154	0.057	0.138	0.0003	0.047
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	32	32	32	546	546	546
Disease FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Disease age FE	Y	Y	Y	Y	Y	Y
N	845	845	845	14,525	14,525	14,525
Mean DV (US$ million)	1222.863	551.878	174.240	151.986	72.804	10.320

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: this table provides results from heterogeneity analyses with various split-sample analyses. Panel 1 examines whether a disease is above or below the median cumulative funding for the 1995-2023 period. Panel 2 investigates whether a disease is above or below the average cumulative funding. Panel 3 focuses on whether a disease is above or below the 90th percentile of the cumulative funding distribution. Panel 4 separates the sample into diseases with at least one advocacy organization (for which data are obtained from the NORD Organization Database) and those without such entities. Panel 5 considers whether a disease is a rare cancer or not (for which data are consulted by reviewing the lists from American Cancer Society and the National Cancer Institute). Analyses are conducted by assuming a one-year lag between the NIH funding and the subsequent dependent variables. Standard errors are clustered at the disease level, and the results have taken into account the presence of disease, year, and disease-age fixed effects.