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

Kevin Chandra

Department of Strategy and Innovation

Questrom School of Business, Boston University

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 list of 1,265 rare diseases, I construct a new dataset linking NIH projects, 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 projects. 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, I find that NIH has been allocating, in nominal terms, over US$113 billion (out of the cumulative total of US$802 billion) within the three-decade period. Nevertheless, some NIH projects encompass multiple diseases, and after weighting disease-based duplicates, the total amount is equivalent to roughly US$171 billion. Second, the funding distribution is heavily concentrated in a small proportion of diseases; the top 5% of funded diseases account for more than half of NIH funding. In contrast, the remaining 250 diseases have never received any NIH funding in the same period. Third, using patent value estimates from Kogan et al. (2017), I find that, on average, every US$1 in disease-specific funding is associated with US$4.7-6.2 in the value of same-disease innovations, and once spillovers are taken into account, the value goes up to about US$31.0-36.4. Fourth, 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 rare cancers, pediatric diseases, and diseases with advocacy organizations.

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.[1] 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'.[2] 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

[1] 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.

[2] 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

cognizant of the fact that every unit of resource invested in a field are inevitably bound to 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 list of 1,265 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$113 billion has been allocated by the NIH to study 1,015 out of 1,265 rare diseases throughout the 1995-2023 period. Second, the share of NIH funding allocated to study rare diseases has been rising over time, from slightly below 8% in fiscal year 1995 to nearly 20% in fiscal year 2023. Third, a significant proportion of NIH projects covers multiple diseases. Once disease duplicates are taken into account (i.e. suppose that a project addresses two different diseases, I double-count the project), the total amount is estimated to be about US$171 billion. Fourth, using the disease duplicate-weighted amount, I observe a substantial degree of funding concentration among a small number of diseases. About 32 diseases with cumulative funding exceeding US$1 billion account for over 52% of rare disease-specific NIH funding. In contrast, there is no recorded NIH funding for about 250 rare diseases, and over 300 diseases have received, on average, cumulative NIH funding either below or slightly above US$1 million throughout the same 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, multiple myeloma, hepatocellular carcinoma, and acute myeloid leukemia. Fifth, rare disease research is significantly and positively associated with the 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 US$3.4 trillion (or about US$13.1 trillion with disease double-counting). Once I account for only patents that cite focal-disease papers acknowledging same-disease NIH projects, the aggregate value of these patents stands at nearly U$1.2 trillion (approximately

US$6.7 trillion with disease double-counting). For patents that share the same disease focus as those in the cited papers, their total nominal value is estimated to be close to US$400 billion (about US$607 billion with disease double-counting), and once accounting for same-disease NIH projects, approximately US$270 billion (or US$452 billion with disease double-counting).

I also run 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. Using both the baseline (correlation between number of NIH projects and number of patents) and refined measures (only focusing on patents citing focal-disease papers acknowledging same-disease NIH projects), I find that, on average, the spillover-to-direct-impact ratio is approximately 10-to-1. In the former, one NIH project is associated with 0.04 same-disease patents and 0.4 'spillover' patents. In the latter approach, the coefficients are approximately 0.18 same-disease patents and 1.7 spillover patents. Additionally, one NIH project is also associated with 0.08-0.37 patents containing the same-disease keyword(s), regardless of their actual linkages to the project itself. 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 spillover-to-direct-impact ratio, remain largely unchanged.

In terms of economic value, my analyses demonstrate some interesting insights. Absent the direct NIH-paper linkages, every US$1 million in focal-disease funding (regardless of whether the funded projects are acknowledged by same-disease papers) does not immediately yield significant value for patented innovations. The strong effects are only observed after a considerable time lag, with every US$1, after six years since the project date, associated with nearly US$5.5 in the value of focal-disease patents citing same-disease research, US$6.9 in the value of patents with focal-disease keywords (regardless of scientific linkages), and about

US$32.4 in the value of spillovers. Once the NIH funding amount is narrowed down to only those eventually acknowledged by the same-disease papers, the effects are more immediate, with the estimates around US$4.7, US$5.1, and US$31.1 for these three categories under the no-lag scenario. The positive association also persists across different time lags.

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 are 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 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).[3]

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

[3] 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.

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).[4] 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).

[4] 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).

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 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).[5] 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.

[5] 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

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.[6]

DATA AND ANALYSIS

1 – Data Sources

I rely 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 projects, 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

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

[6] 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.

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 point 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.[7]

[7] 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.

With regard to information on funding and the economic value of innovations, I turn 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 is 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.[8] 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.[9]

2 – Data Collection Process

I begin the search process by initially collecting the list of disease names and equivalent medical terms from the NORD In-Depth Report database. With December 31, 2025 as the cutoff date for the publication date of these in-depth reports, I manage to collect a total of 1,417 disease names. Afterwards, I specify 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 includ both the 'Cystic Fibrosis' as well as its equivalent key terms (CF, fibrocystic disease of pancreas, mucoviscidosis, pancreatic fibrosis) as generated from the NORD In-Depth Report. The database would then return results with any

[8] 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 agreements involving patents from US major research universities.

[9] The extended dataset is shared in the following GitHub link: https://github.com/KPSS2017/Technological-Innovation-Resource-Allocation-and-Growth-Extended-Data

mention of these keywords in at least one of the four categories specified in the search. In addition to utilizing these terms, I also include boolean terms (e.g. 'AND', 'OR') in cases of diseases with multiple equivalent keywords. Referring again to the Cystic Fibrosis example, I include 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.[10] The same process is then repeated, and adjusted, for each of the 1,417 diseases with NORD In-Depth Reports.[11]

Nevertheless, not all these diseases are necessarily defined as 'rare' due to mulitple reasons. First, some diseases used to be defined as 'rare' until physicians made a consensus to no longer categorize them under this condition, i.e. multiple sclerosis (Talley, 2005). Second, some 'rare' diseases happen to be highly prevalent due to socioeconomic factors, such as neglected tropical diseases (World Health Organization, 2025), and these diseases warrant separate discussions beyond this paper. Third, some diseases may still be defined as 'rare' even as they are highly prevalent in the general population, primarily either because of the lack of scholarly

[10] 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.

[11] 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.

interest or that the genes carrying these diseases (or resulting in severe infections) are not activated in a vast majority of people, i.e. Cytomegalovirus/CMV (Rockefeller University, 2020).

Based on these considerations, I exclude some diseases from the NORD list. In addition, another challenge relates to systematically estimating their prevalence, which for rare diseases has been historically difficult due to limited studies and the lack of consensus about population thresholds delineating prevalent from rare conditions (Richter et al., 2015; Nguengang-Wakap et al., 2020). For each disease, I seek to obtain prevalence estimates using Google Search, with a priority on identifying studies describing disease prevalence in the US. As a rare disease is defined in the US as one affecting 200,000 people or fewer (based on the 1983 Orphan Drugs Act), and assuming the US population of roughly 340 million people as of 2024, I exclude diseases with estimated prevalence above 58.82 per 100,000 (200k/340M multiplied by 100,000).[12] This approach narrows down the number of diseases from 1,417 to 1,265.

Based on the results returned from each search, I extract both the publication records as well as the patents citing these papers. Afterwards, I conduct a separate search using the same specified terms for the patent database, with an emphasis on ensuring any of these terms are 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

[12] A major caveat with the US definition of rare diseases is that the 200,000-people threshold has not been revised since the passage of the Orphan Drugs Act in 1983. In that year, the US population was estimated to be roughly 234 million people, therefore diseases that are deemed as 'prevalent' may have been classified 'rare' in the preceding decades. In contrast, in other jurisdictions, e.g. the European Union, the prevalence threshold used to define a rare disease is more straightforward, at 5 per 10,000 people (or 1/2000) as a cutoff point.

[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 overlap 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 make use of the disease name and its equivalent medical keywords from each of the 1,417 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 find 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 make 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] Of these 1,417 diseases, about 1,200 of them happen to be covered by NIH projects irrespective of time periods. The number drops to 1,015 once I only include diseases with 1) prevalence estimates following the US definition (and assuming 58.82 per 100,000 people as the cutoff), and 2) those covered by NIH projects anytime between 1995 and 2023 fiscal years.

[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.

Initially, the publication search for these 1,417 diseases yields a total of 7,827,906 unique publications (9,513,951 paper-disease observations). Once the list is narrowed down to 1,265 diseases that are considered rare within the US threshold, the sample size shrinks to 4,949,493 unique publications (5,928,141 paper-disease observations). These papers, in turn, are cited by 192,311 unique USPTO patent grants (612,119 patent-disease observations once accounting for disease double-counting). After the list drops to 1,015 diseases to account only for those with NIH funding throughout 1995-2023, the sample size decreases slightly to 4,876,892 unique papers (5,820,420 paper-disease observations). A total of 72,835 unique USPTO patent grants (377,392 patent-disease observations) end up citing these publications. Separately, the keyword-based search for patents containing disease-specific terms regardless of scientific linkages returns a total of 40,843 USPTO patent grants (or 111,654 patent-disease observations), some of which intersect with the patents citing these focal-disease papers (refer to Figure 1). Once I isolate the subset of patents with disease-specific keywords that also cite same-disease papers, I manage to identify a total of 22,809 unique patents (32,113 patent-disease observations), and after accounting for linkages with same-disease NIH projects, the sample size drops to 14,742 unique patents (22,412 patent-disease observations). Additionally, among rare diseases within US prevalence threshold, the search process in the NIH RePORTER database generates a total of 192,130 unique project IDs (or 298,734 project-disease observations as certain projects are focused on two or more diseases).[15]

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

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 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 find that this grant is acknowledged by over 146

[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

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

In the 1,015-funded-disease sample, I address 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 download from NIH RePORTER. Using this approach, I extract a total of 189,450 unique publications acknowledging NIH funding (234,806 paper-disease observations after double-counting), as well as 192,130 unique NIH project numbers (298,734 project-disease observations).[19] It is worth noting, once

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).

[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

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 merge these paper-disease and project-disease pairs, which result in a total of 9,814,708 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 narrow down the merged data to 2,453,855 paper-disease-project-disease tetrads for which both paper-disease and project-disease labels are identically matched. These subsetted tetrads consist of 130,578 unique papers (144,466 paper-disease observations) and 146,116 NIH projects (185,177 project-disease observations).

Overall, these publications (papers focusing on a disease that acknowledge the same-disease grants) are subsequently cited by 72,835 unique USPTO patent grants (377,392 citing patent-disease observations), as shown in Figure 6. Of these citing patents, about 14,742 of them (22,412 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 shrink to 14,731 citing USPTO patents (86,382 citing patent-disease observations) and 2,998 patent grants containing the same-disease keywords (4,741 patent-disease observations). The number of patents with focal-disease keywords, but not necessarily

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.

referencing any disease-specific science, also drops from 40,843 USPTO patent grants (111,654 patent-disease observations) to 9,050 (24,531 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.

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 7.6% in 1995 to nearly 19.4% in 2023. 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 demonstrated a steady increase. This pattern persisted under Obama, first-term Trump, and Biden administrations.[21]

[21] 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.

In aggregate, for rare diseases, NIH has been allocating about US$113 billion within the same period; once disease double-counting is introduced, the amount goes up to US$171 billion. For the remainder of the study, I refer to the disease-double-counted version. Nevertheless, there are two caveats to take into account. First, about 250 out of 1,265 rare diseases have never received any NIH funding. Second, the funding allocation is highly skewed, as highlighted in Figure 8. When I consider any disease-specific funding—regardless of whether these NIH grants result in same-disease publications—I observe that a vast majority of the funding is concentrated in a small number of diseases. Figure 8(a) shows that 32 diseases have cumulative funding in excess of US$1 billion, accounting for nearly US$88.9 billion (approximately 52% of totals) in the aggregate sum. The full list of these billion-dollar-plus diseases is shown in Table 1. Almost half (14 out of 32) of these diseases happen to be cancer-related (e.g. multiple myeloma, glioma, glioblastoma, acute myeloid leukemia). On the other end, 342 diseases are recorded to have a cumulative sum of NIH funding at slightly above US$1.1 billion. On average, this translates to less than US$3.5 million of NIH allocation for each of these diseases.

Figure 8(b) visualizes the funding distribution after accounting for only NIH projects 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$104.7 billion, or approximately 61% 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 result in cross-disease spillovers, with limited insights directly relevant to the focal disease. The number of ‘billion-dollar-plus’ diseases drops from 32 to 19, but in aggregate, they still account

for over 40% of NIH disease-specific allocation.[22] A surprising observation is that the number of "No funding" diseases jumps substantially from 250 to 335, an indication that a large number of these diseases, in spite of NIH funding allocation, ends up with no same-disease publications. The patterns are almost similar once the data are subsetted to NIH projects acknowledged by same-disease papers within five years since the project date (Azoulay et al., 2019), as shown in Figure 8(c). Figure 8(d) visualizes the number of diseases based on cumulative funding categories that result in no acknowledging same-disease publications. A vast majority of these diseases (47 out of 85) 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). The patterns largely hold across different prevalence categories, as displayed in the remaining five figures in 9(a). 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).[23] It appears that the proportion of NIH disease-

[22] 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.

[23] 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.

specific funding acknowledged by same-disease papers rises with the upward shifts in prevalence categories, from about 50% for unknown-prevalence diseases to nearly 67% for diseases with prevalence of 1-5 per 10,000 people.[24]

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, but with a caveat. In part (a) of the figure, the correlation between disease prevalence and cumulative amount of disease-specific funding is weak (at R-squared = 0.002). The correlation is somewhat higher once the funding only takes into consideration the amount eventually acknowledged in same-disease publications. 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, another 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.

Table 2 summarizes the overall descriptive statistics for patents both directly and indirectly related to rare disease research, with the figures mirroring those displayed in Figure 6.

[24] 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 1,265 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.

Panel 1 of Table 3 displays the cumulative value of patents that cite rare disease research, regardless of their linkages to NIH grants. There are a total of 170,636 patent-disease observations, comprising 139,391 observations on the basis of patents citing articles from any of these 1,265 diseases, 24,531 patent-disease observations with any keyword-specific references to at least one of these diseases, and 6,714 observations for focal-disease patents citing same-disease papers. In aggregate, the cumulative value of these patent-disease observations is nearly US$13.1 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 170,636 to 48,044 unique patents, with an aggregate nominal value of over US$3.4 trillion. About 11% of the value is attributed to focal-disease patents citing same-disease publications.

Results for Panel 2 of Table 3 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 86,382 observations (representing 14,731 unique patents), with a cumulative nominal value of US$6.7 trillion (and about US$1.2 trillion once the disease double-counting is taken into account). Among the subset of patents containing same-disease keywords, I identify a total of 4,741 patent-disease observations (or 2,998 unique patents), with an aggregate sum of US$452 billion (or US$270 billion after removing disease double-counting).

Tables 4-7 list out the top 20 diseases based on the cumulative value of their related innovations. Table 4 displays the results based on the total value of disease-specific NIH funding (regardless of its linkages to acknowledging same-disease papers). The top-20 diseases in Table 5 are based only on the total value of focal-disease NIH funding that ends up acknowledged by same-disease publications. One notable observation is that a vast majority of diseases listed in Tables 4 and 5 happen to be rare cancers, due largely to the fact that the National Cancer Institute (NCI) is the most well-funded entity among all other NIH institutes/centers. Tables 6 and 7 repeat the approaches in Tables 4 and 5, but with the exclusion of rare cancers. Some of the most notable non-oncology rare diseases in these tables include cystic fibrosis, ALS, and Duchenne muscular dystrophy. Nonetheless, I suggest two caveats in interpreting these findings. First, the count figures in these tables take into account double-counting across multiple diseases, especially when a patent contains keywords linked to two or more diseases. Another caveat is that the patent value estimates from Kogan et al. (2017) are only available for publicly-funded and mostly incumbent pharmaceutical firms. Given that biotech firms play a major role in the development of orphan drugs, and the fact that a vast majority of them are not publicly traded, the actual value of innovations for these diseases may be substantially larger than estimates provided in these tables.

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 8 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 14.5, but the median is only 2, whereas the 75th-percentile value is 7. 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 variable "NIH funding amount (US$ mil, nominal)", where the average figure US$5.82 million per disease-year observation. Nonetheless, half of the observations in the panel data recorded less than US$0.39 million in NIH funding, and the 75th-percentile value is only about US$2.71 million. Therefore, I would advise initial caution in interpreting the results from the regression analysis. I run the regression analysis under three circumstances: 1) baseline correlation; 2) analysis that only accounts for same-disease NIH-paper linkages, and; 3) repeat analysis of (2) but limited to five-year thresholds following Azoulay et al. (2019).

Table 9 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 0.41-0.68 additional citing patents, and the coefficients are consistently upwards with increasing time lags. The 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 linkages to NIH project-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 0.08-0.15 range, and about 0.04-0.07 for the models in

Panel 3. Based on the results, I estimate that, on average, roughly 9-10% of patents associated with NIH-funded research can be perceived as 'direct impacts', and the remaining 90-91% of the estimates to be spillovers.

Afterwards, I repeat the baseline analysis with the same set of covariates and outcome variables under the two remaining scenarios. Table 10 displays the results by focusing only on projects that are acknowledged by same-disease papers, and for which these papers are eventually cited by patents (for Panel 1 and Panel 3). I retain the same outcome variable in Panel 2 as the emphasis is on patents with disease-specific keywords that may or may not cite these publications. Interestingly, across the three panels, I observe a downward trend in the coefficients as the time lag grows. As an illustration, in Panel 1, one paper-acknowledged NIH project is associated with 1.71 patents under the no-lag scenario, and drops to 1.46 after a six-year lag. For the remaining panels, the coefficient drops from 0.37 to 0.34 for patents with disease-specific keywords, and from 0.18 to 0.16 for patents containing disease-specific keywords that also cite same-disease papers. The spillover-to-direct-impact ratio, at approximately 10-to-1, remains largely unchanged. Table 11 shows results from the analysis centered on any patents citing focal-disease papers that acknowledge same-disease NIH projects within five years from the project date. Once again, the results are largely robust, with both the estimates and the ratio highly similar to those in Table 11.

I then repeat the analyses for Tables 9, 10, and 11 for the subsets of patents with available KPSS value estimates, the results of which are displayed in Tables 12, 13, and 14. One observation worth noting from these tables is that there is a remarkable decline, by about 80-90%, in the size of the coefficients across different panels. The decline is attributed to the fact that the patent subset only accounts for those filed to USPTO by publicly-traded firms, in which

Kogan et al. (2017) estimated patent value based on their stock price fluctuation within the grant period of the focal patent. As an illustration, in Panel 1 of Table 12, on average, one focal-disease NIH project, regardless of its scientific/technological linkages, is associated with 0.064-0.103 additional citing patents, only about one-sixth of the coefficient estimate range in the same panel in Table 9 (about 0.41-0.68). Coefficients in Panel 2 record the biggest decrease (approximately by 90-92%) in the scale of the coefficients relative to those in Table 9. Each disease-specific NIH project, on average, is correlated with 0.006-0.013 additional patents that contain focal-disease keywords. The coefficient range for patents containing disease-specific keywords that cite same-disease papers (Panel 3) also closely resembles those in Panel 2, at about 0.008-0.014. The spillovers appear to be about 8x larger than direct impacts for this table.

Once I focus only on focal-disease NIH projects resulting in same-disease papers (with results in Table 13) and those acknowledged by papers within five years (Table 14), the coefficients end up considerably larger in comparison to those in Table 12. Similar to Tables 10 and 11, there is a downward trajectory in the observed coefficients with increasing time lags. In Panel 1 of Table 13, one NIH project is associated with 0.35 citing patents under the no-lag scenario, and the coefficient decreases to 0.26 in a six-year lag model. For patents containing focal-disease keywords (Panel 2), the estimates range from 0.048 (no lag) to 0.032 (six-year lag). Whereas for focal-disease patents citing same-disease research (Panel 3), the decrease happens to be largely moderate, from 0.037 (no lag) to 0.033 citing patents (six-year lag). The results are barely unchanged under the five-year threshold in the NIH project-paper linkages, as shown in Table 14. Spillovers, overall, appear to be 8-9x larger than the direct impacts.

Tables 15, 16, and 17 display 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. I first describe the results in Table 15, which describes the correlation between NIH funding (regardless of the amount that yields same-disease publications) and the KPSS value of these patents. Overall, NIH funding is not immediately manifest in the value of subsequent innovations; across the three panels, an estimated 2-3 year lags are needed before the value of citing innovations is significantly correlated with the funding amount. Under the no-lag scenario, every US$1 million in focal-disease NIH funding is associated with US$5.4, US$0.23, and US$0.92 million in the value of overall citing patents, patents with disease-specific keywords, and rare disease patents citing same-disease research. In a 6-year lag, the estimates increase significantly to US$32.4, US$6.9, and US$5.5 million. In monetary terms, for every US$1 generated in the value of direct impacts, the spillovers are worth approximately US$6-7.

Table 16, in contrast, focuses only on the amount of funding (in US$ million) for NIH projects subsequently acknowledged by same-disease publications. Among these projects, the effects are immediately evident from the no-lag scenario onwards. In Column (1), every US$1 million is significantly correlated with US$31.1, US$5.1, and US$4.7 million across these dependent variables. Interestingly, there is a slight dip in the coefficients from 1-year lag to 3-year lag models, before the estimates show a considerable increase from the 4-year lag. Under the 6-year lag, as shown in Column (7), the estimates are US$36.3, US$9.4, and US$6.2 million in the value of associated innovations. The results remain largely robust when I only include the amount of funding for projects acknowledged by same-disease papers within five years of the project award date, as shown in Table 17. Overall, the estimated value of direct impacts of public funding—ranging from US$5 to US$6 per US$1 of R&D—happens to closely approximate

Jones and Summers (2022), in which they show that every US$1 in public research funding may generate up to US$4 in the value of subsequent innovations.

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, nearly 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).[25]

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

[25] 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."

Afterwards, I repeat the OLS regression with a split-sample analysis (both with overall NIH funding amount and only funding amount linked to acknowledging same-disease papers), with the results posted in Tables 18 and 19. The split-sample analysis is performed under six 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 is cancer-related; (v) whether the disease mainly affects child/pediatric patients, and; (vi) whether the disease has at least one representative advocacy organization over time. My reasoning for the inclusion of the last three scenarios—(iv), (v), and (vi)—is three-fold. With regard to 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).[26] Relatedly, a large proportion of rare cancers happen to mainly affect pediatric patients, which often elicit greater emotional responses and subsequent impacts on fundraising (Wood et al., 2013; Kenworthy, 2024). Specific for 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). The analyses are all conducted by assuming one-year time lag across all models.

I first elaborate on results from Panels 1, 2, and 3 of Tables 18 and 19. In Table 18, there are hardly statistically noticeable differences between diseases that are above- versus below-median, above- versus below-average, and above- versus below-90th percentile of the cumulative

[26] 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).

funding amount, with two exceptions. First, among the below-average-cumulative-funding subset in Panel 2, every US$1 in disease-specific funding is associated with roughly US$13 in the value of any patents citing focal-disease research after one year. Second, the value of rare-disease patents citing same-disease research is slightly higher among diseases in the below-90th-percentile-cumulative-funding subset in Panel 3. Another observation worth noting is that the spillover-to-direct-impact ratio is considerably larger among diseases in the below-median / below-average / below-90th-percentile subsets than their remaining counterparts.

Once the focus shifts towards only funding amount leading to same-disease publications, as shown in Table 19, I notice some changing patterns in the results. Diseases on the above-median / above-average / above-90th-percentile of cumulative funding distribution are more likely to demonstrate larger, and statistically significant, results across all three patent categories than those below the cutoffs. In Panel 1, every US$1 million in NIH funding is correlated with nearly US$4.3 million in the value of same-disease innovations, US$4.6 million for any patents (regardless of citation linkages) containing same-disease keywords, and about US$25.9 million for any patents citing focal-disease papers. 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 1.0, 1.2, and 6.5 respectively). Similar patterns are observed for Panel 2, with the exception of no statistical significance in the coefficients for patents with rare disease keywords (3.7 in the above-average and 3.5 in the below-average subsets). Interestingly, in Panel 3, the coefficient for patents with rare disease keywords (5.650) is higher in the below-90th-percentile subset than in the above-90th-percentile subset (3.109). The spillover-to-direct-impact ratio remains somewhat bigger in the below-median / below-average / below-90th-percentile thresholds, although the magnitude is somewhat moderated compared to those in Table 18.

Panels 4, 5, and 6 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. Rare cancers are identified from the cancer list provided by the American Cancer Society (ACS) and the National Cancer Institute (NCI) in their websites.[27] In identifying rare pediatric diseases, I consult various sources, such as disease lists from major hospitals (i.e. Boston Children's Hospital, Children's Hospital of Philadelphia, and Children's Hospital of Pittsburgh), as well as organizations specializing in pediatric cancers (i.e. American Childhood Cancer Organization, Memorial Sloan Kettering Cancer Center, National Cancer Institute, and St. Jude Children's Research Hospital). Data for advocacy organizations are obtained from a separate database maintained by NORD, which contains information about organization names, brief profiles, and diseases represented.

Overall, in both Tables 18 and 19, rare cancers, pediatric diseases, and diseases with advocacy organizations seem to attract more attention from innovators than those not in these subsets. These patterns appear to hold regardless of whether the focal-disease funding is subsequently associated with same-disease publications. Every US$1 million invested by NIH into a rare cancer is associated with approximately US$4.2 million in the value of same-disease patents citing same-disease science (with the amount growing up to US$4.8 million after only accounting for disease funding leading to same-disease papers), US$5.4 million in the value of any patents with same-disease keywords (and up to US$7.6 million in Panel 4 of Table 19), and US$27.4 million for any citing patents (or US$37 million in Table 19). For pediatric diseases, the

[27] NORD has been continuously maintaining a database of advocacy organizations, including those registered as NORD 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/).

returns are somewhat larger, both in the all-funding-inclusive and only-paper-linked-funding scenarios. In the former, every US$1 million in disease-specific NIH investment is associated with US$4.5, US$6.8, and US$21.2 million in the value of these innovations, whereas under the latter, the estimates are US$6.5, US$10.2, and US$36.7 million respectively. It is worth noting that some of the major drivers of high returns in pediatric disease investment happen to be cancers primarily affecting child patients (i.e. acute myeloid leukemia, neuroblastoma, or bone cancers such as osteosarcoma). Lastly, in Panel 6, diseases with advocacy organizations seem to receive more attention from innovators than those without, although the coefficients are not statistically significant as shown in Table 18. When it comes to NIH funding that is subsequently linked to same-disease publications, every US$1 million invested by NIH into a disease with an advocacy organization is associated with US$4.5 million in the value of same-disease patents citing same-disease science, roughly U$5 million in the value of any patents with same-disease keywords, and US$28.8 million for any citing patents.

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 list of 1,265 rare diseases, and specifically a subset of 1,015 such diseases funded by NIH throughout 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$4.7-6.2 in the subsequent value of same-disease innovation, and about US$31.0-36.4 after accounting for spillovers. These estimates largely correspond to the figures based on models developed by Jones and Summers (2022), 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, having accounted for funding amount that leads to same-disease papers, the results either align with or surpass baseline estimates. In contrast, the figures are considerably lower, and with limited statistical significance, for diseases below these thresholds. Rare cancers, pediatric diseases, and diseases with advocacy organizations 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 suggest 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 1,265-disease list—is that no-funding diseases only comprise slightly below 20% (N = 250) 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 be 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 in Table 19, the average figure for focal-disease patents citing same-disease science is about US$7.2 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$1.7 million to the value of same-disease innovations (an increase of nearly 25% from the baseline). In contrast, among the rare-cancer subset, the average figure for the same category is approximately US$120.1 million per disease-year observation. Although every US$1 million would potentially generate US$4.8 million in the value of rare-cancer innovations, this only represents a slightly below 4% 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 described as '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). Moreover, the KPSS estimates—drawn from publicly traded firms—only account for less than one-quarter of patented innovations, an indication that the figures discussed in this study may underestimate the actual value. 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.

Kenworthy, N. J. (2024). *Crowded out: The true costs of crowdfunding healthcare*. Cambridge, MA: MIT 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.

Rockefeller University (January 29, 2020). 'Patient with unusually severe infection leads scientists to a rare type of immune deficiency'. Link retrieved from https://www.rockefeller.edu/news/27419-unusual-patient-case-cmv-rare-immune-deficiency/

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.

Talley, C. L. (2005). The emergence of multiple sclerosis, 1870-1950: A puzzle of historical epidemiology. *Perspectives in Biology and Medicine*, 48(3): 383-395.

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.

World Health Organization (December 14, 2025). 'Neglected tropical diseases'. Link retrieved from https://www.who.int/news-room/questions-and-answers/item/neglected-tropical-diseases

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)).

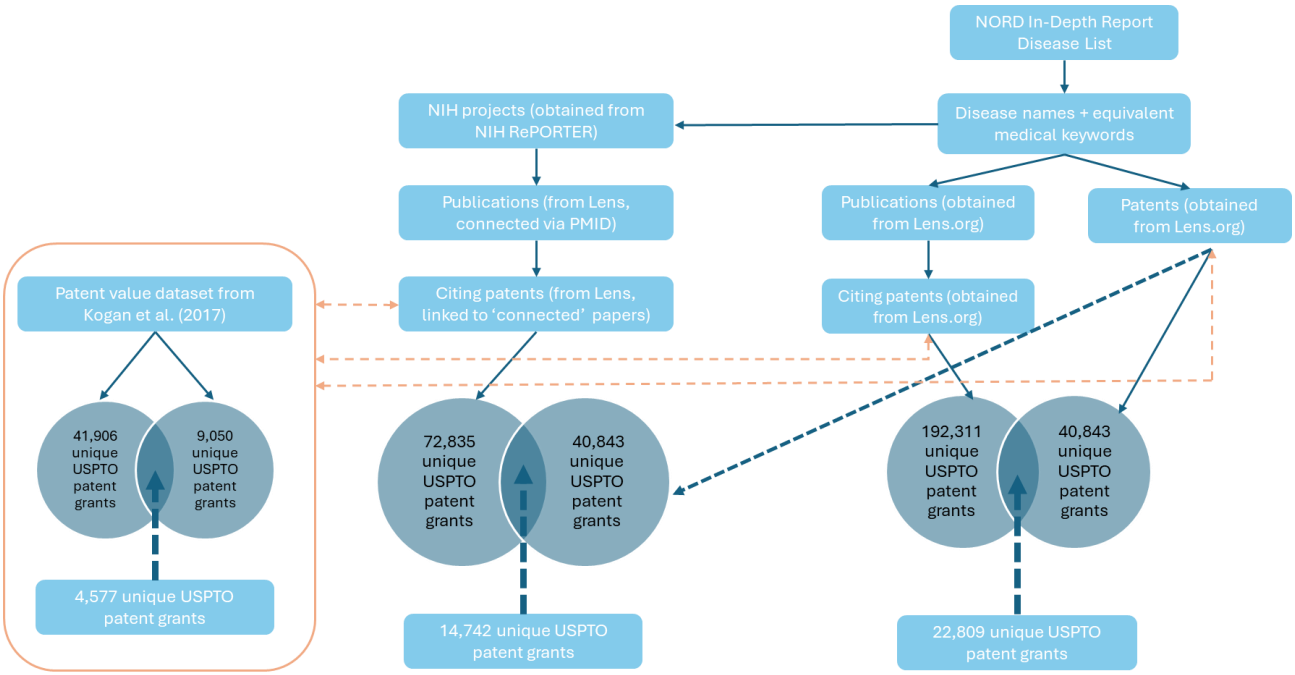


Figure 2. Types of innovation for consideration in the empirical analysis

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.

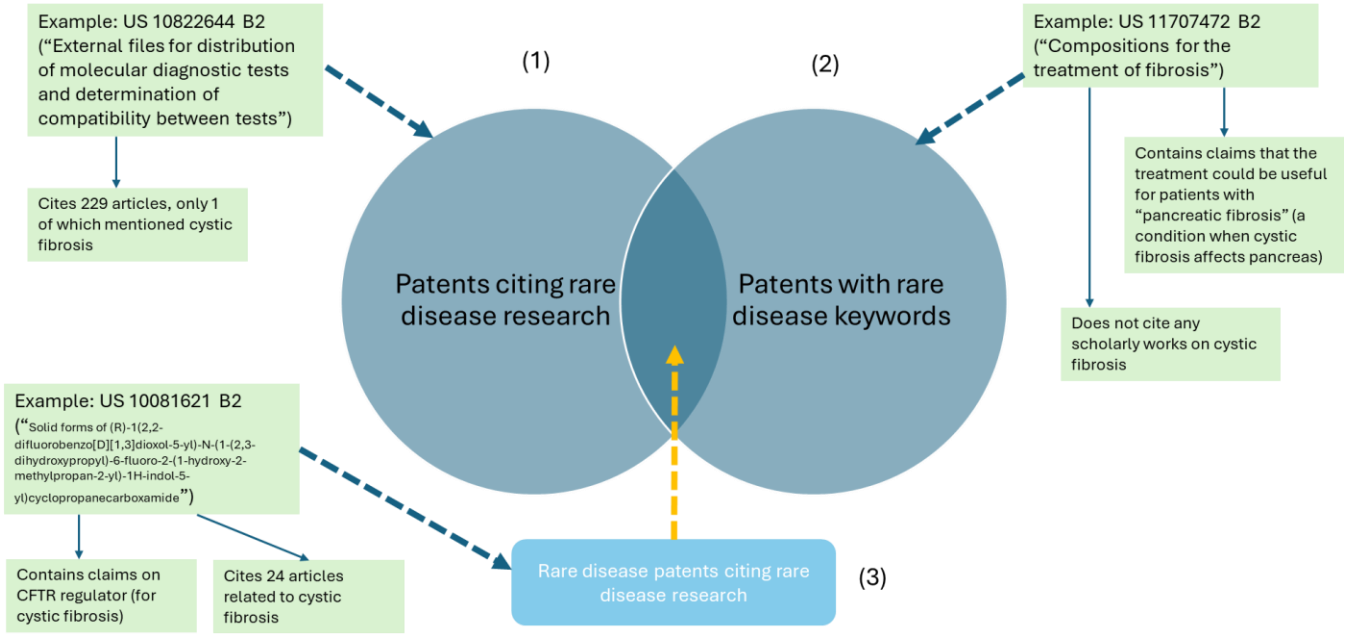
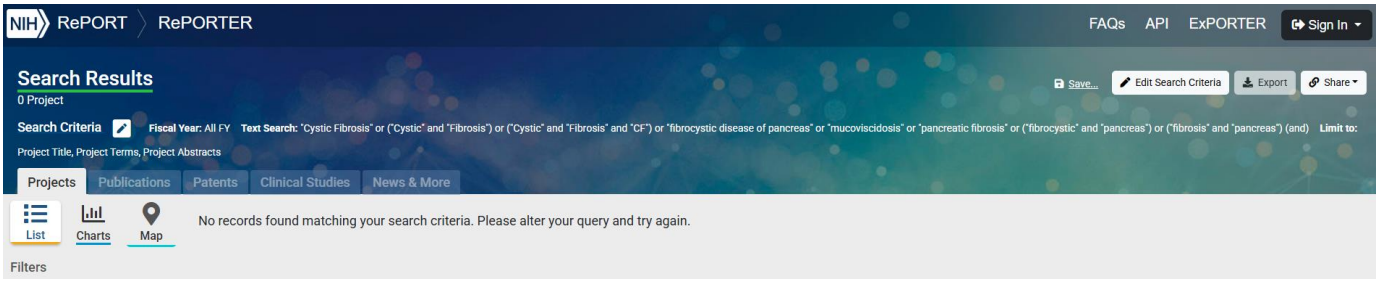
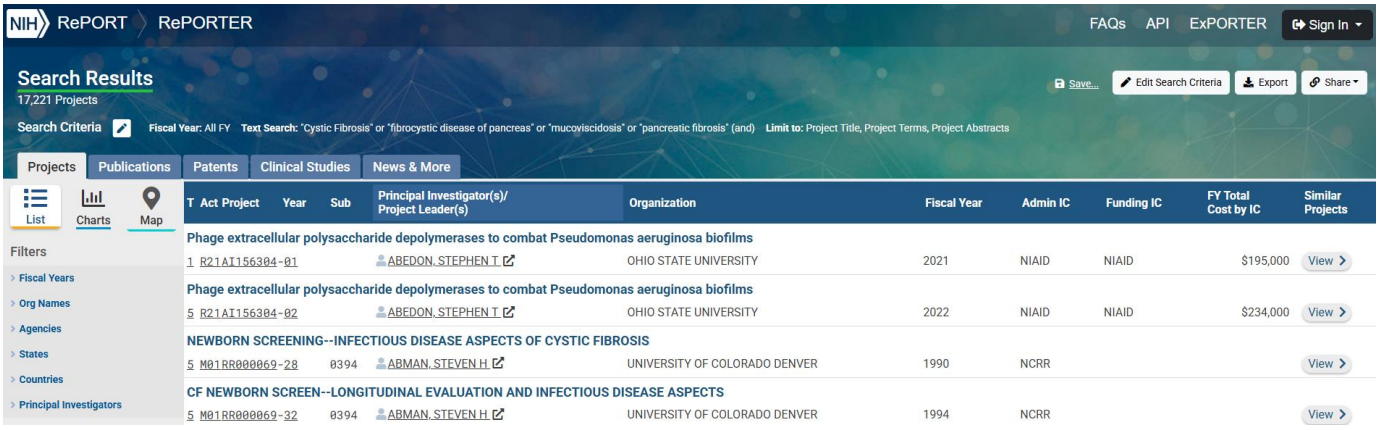


Figure 3. Comparison of search results in NIH RePORTER with (a) NORD and boolean-refined terms, and (b) only NORD terms, using the example of Cystic Fibrosis disease

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.

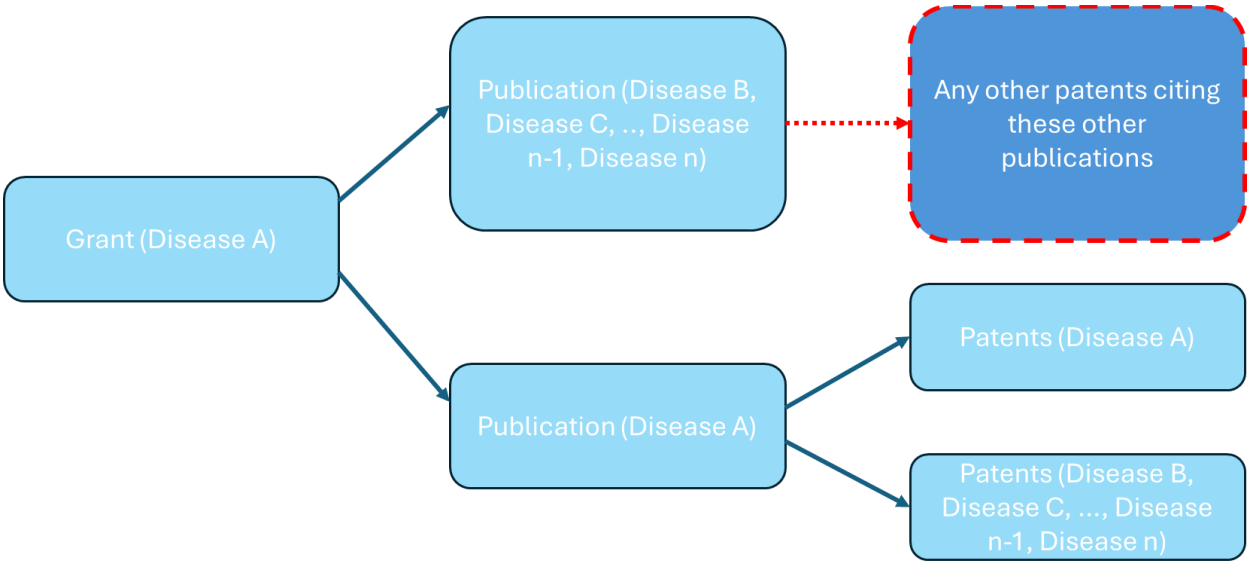


Figure 5. Isolating cross-disease research spillovers for empirical analysis

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'.

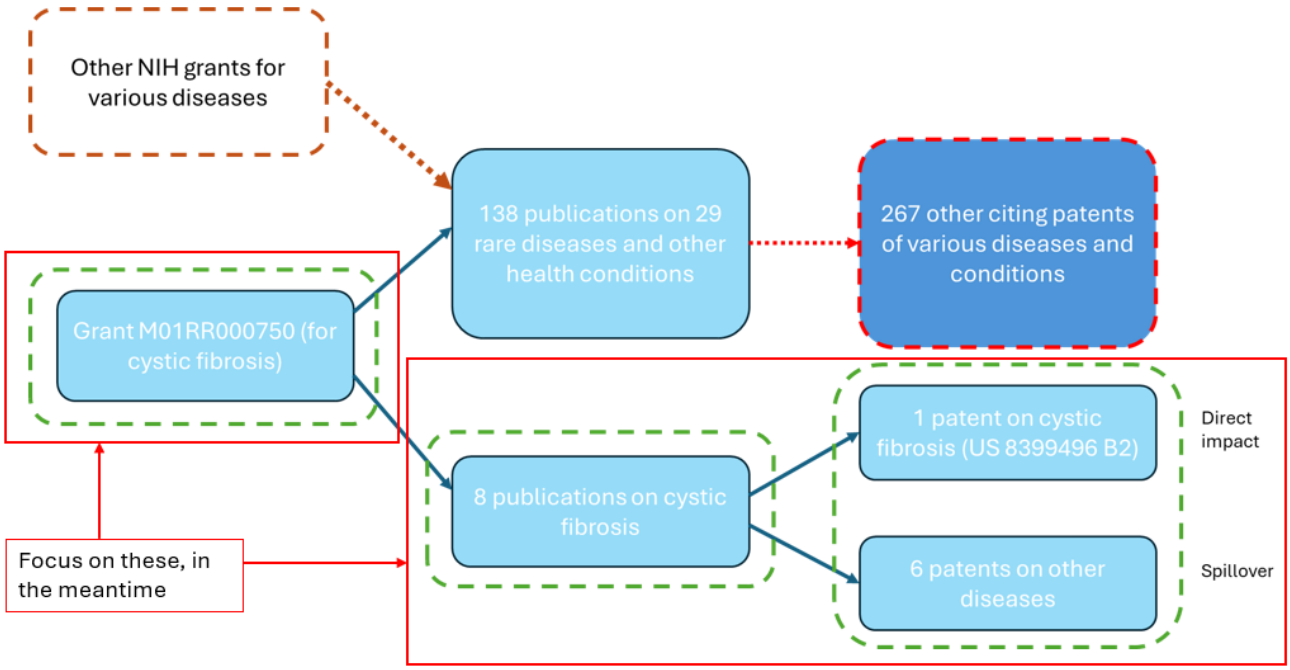


Figure 6. Integration of NIH grants, acknowledging publications, and citing patents

Based on the data collected from Lens.org and NIH RePORTER, and a list of 1,015 rare diseases, I identified a total of 189,450 unique publications (234,806 paper-disease observations after taking into account double-counting) acknowledging NIH funding, as well as 192,130 unique NIH project IDs (298,734 project-disease observations) that are acknowledged by these papers. I afterwards merged these two sets of data, resulting in a total of 9,814,708 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 = 2,453,855). In turn, these fully matched observations represent 130,578 unique papers (144,466 paper-disease observations) and 146,116 unique NIH project IDs (185,177 project-disease observations). They are also associated with 72,835 unique patent grants citing the same-disease NIH-paper pairs, as well as a subset of 14,742 unique patent grants containing same-disease keywords.

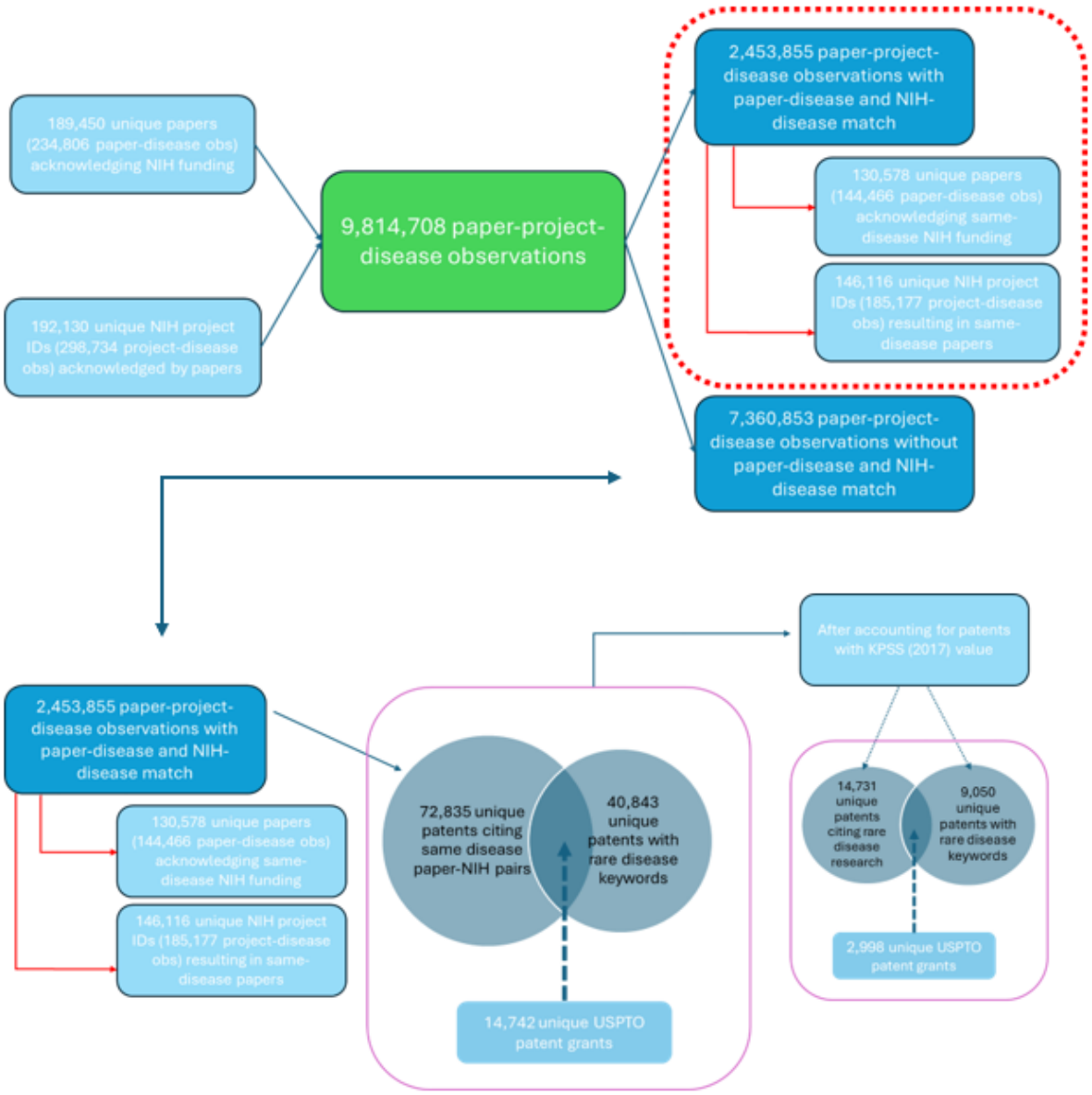


Figure 7. Changes in NIH funding, and allocation towards rare disease research, over time (in billion US$, nominal terms, for 1995-2023 period)

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 1,015 funded 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 7.6% in FY1995 to 19.4% in 2023.

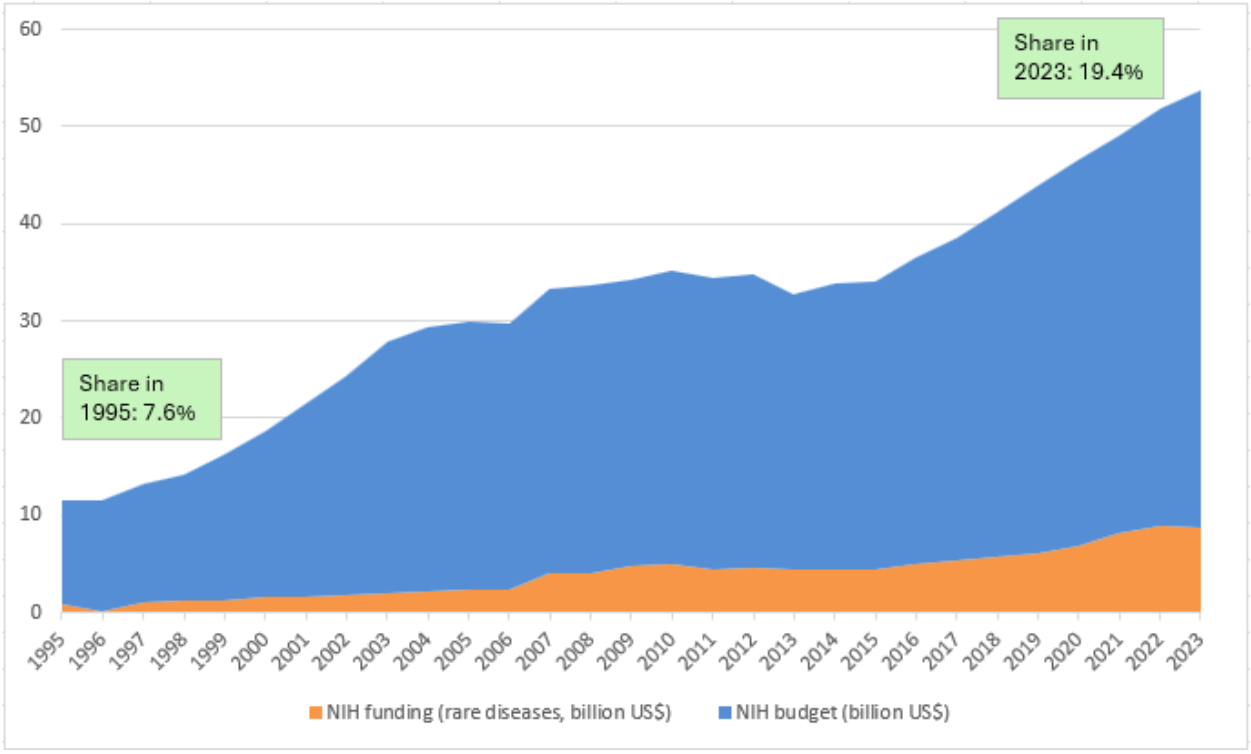
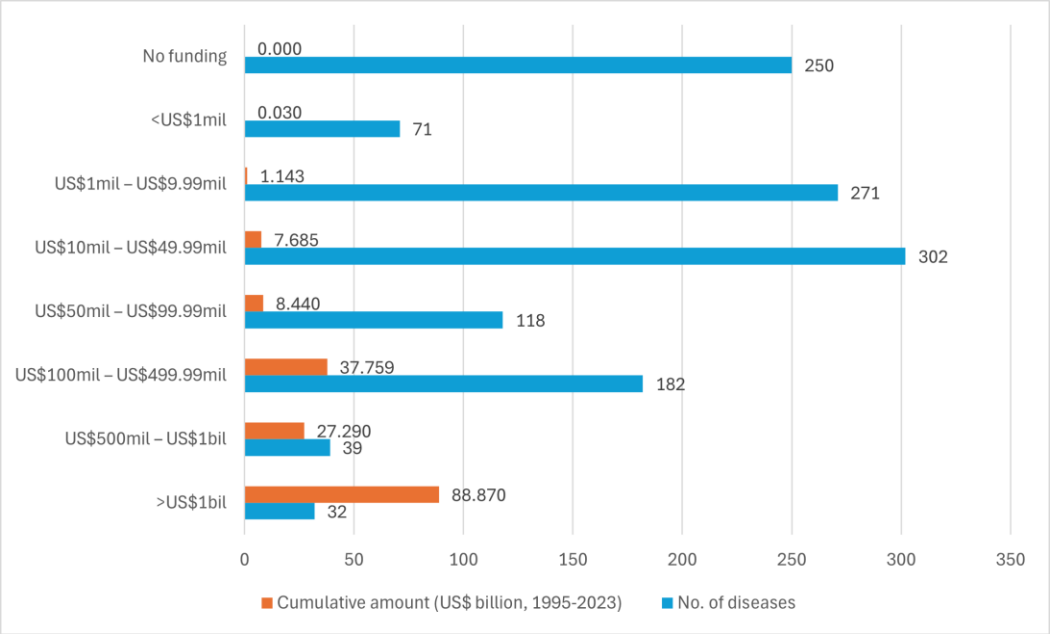
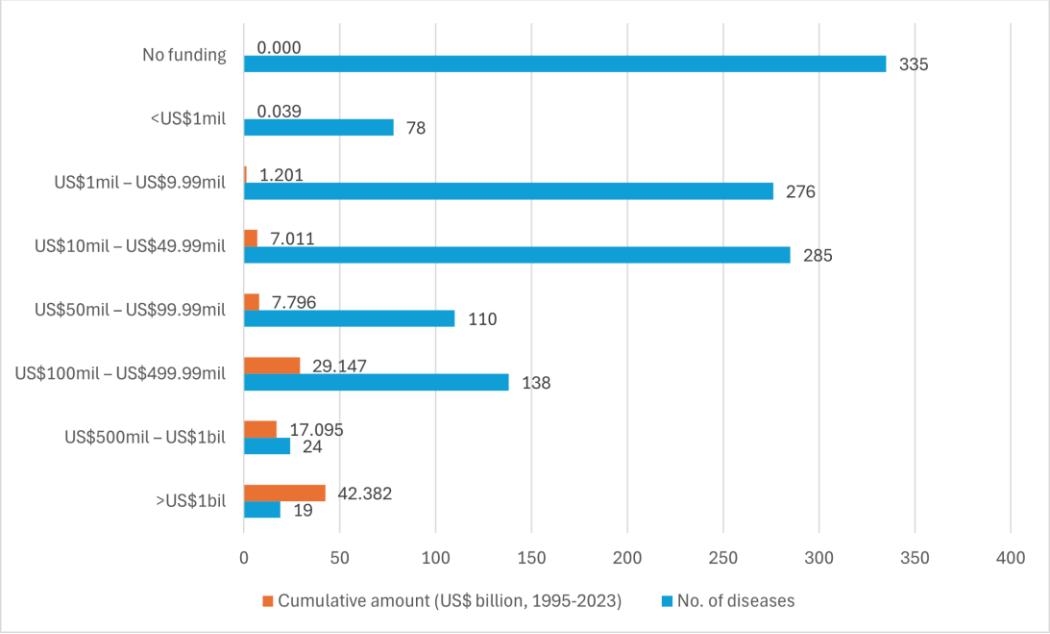


Figure 8. Number of rare diseases based on tiers of cumulative NIH funding amount (1995-2023)

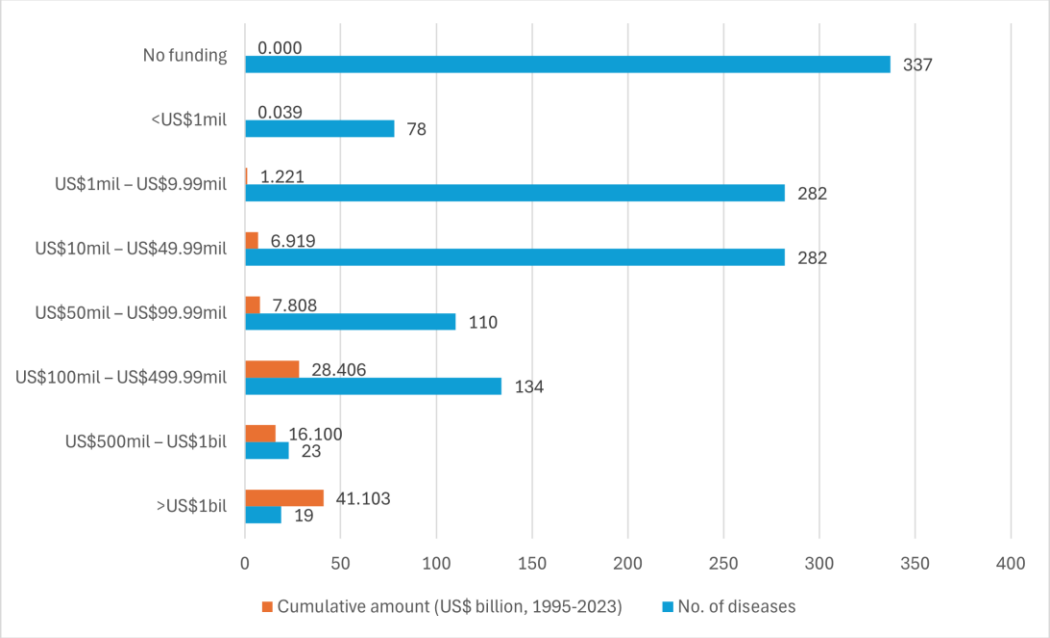
This figure visualizes the distribution of cumulative funding (1995-2023) across a list of 1,265 rare diseases. Part (a) summarizes the number of diseases based on cumulative funding categories. Part (b) is a follow-up to part (a) after adjusting only for disease-specific NIH projects that lead to at least 1 acknowledging same-disease publication within the time period. Part (c) narrows down the linkages in Part (b) by focusing only on papers citing same-disease NIH projects within five years of the project award, following Azoulay et al. (2019). Part (d) describes diseases based on their cumulative funding categories in Part (a) that do not end up with any same-disease publications.



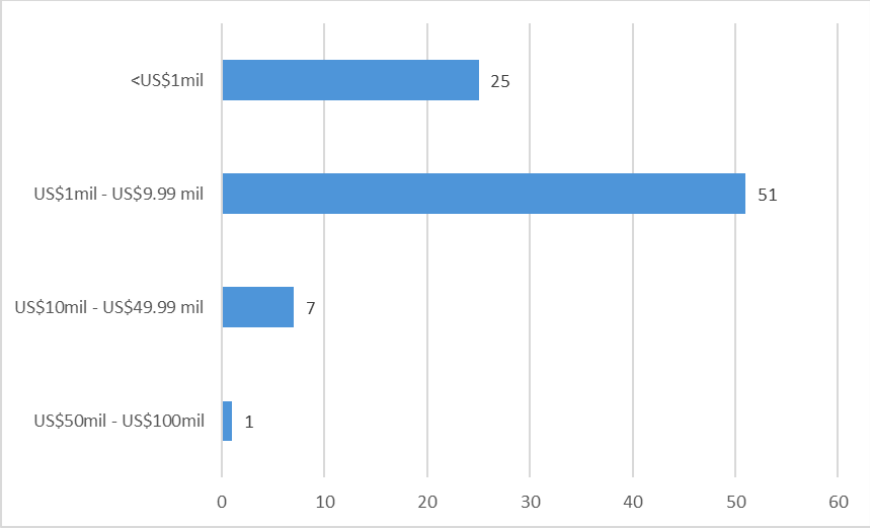
(a) Any NIH projects, regardless of linkages with acknowledging same-disease papers



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



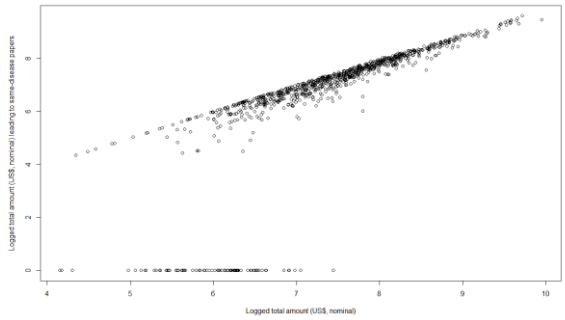
(c) Only NIH projects that are acknowledged by 1+ same-disease paper(s) within 5 years following Azoulay et al. (2019)



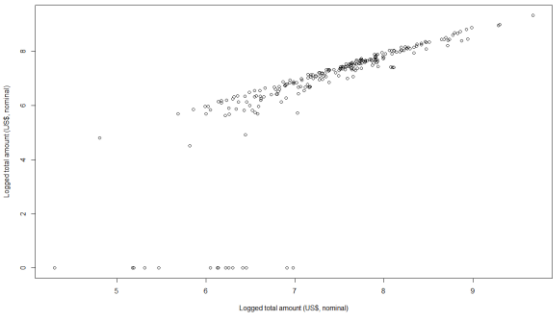
(d) Diseases by cumulative funding categories that do not result in same-disease publications

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

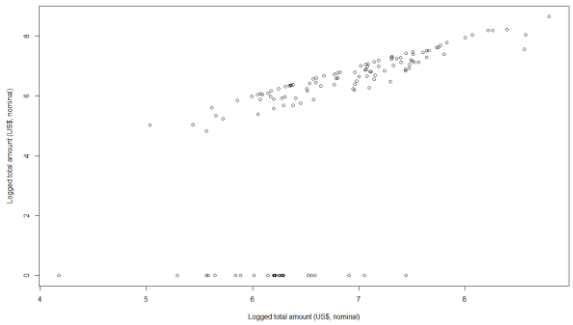
This figure provides additional summaries on the linkages between NIH disease-specific funding and the subsequent amount of scientific literature attributed to the same diseases. Part (a) visualizes the correlation between the cumulative funding amount across diseases and the actual amount that leads 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.



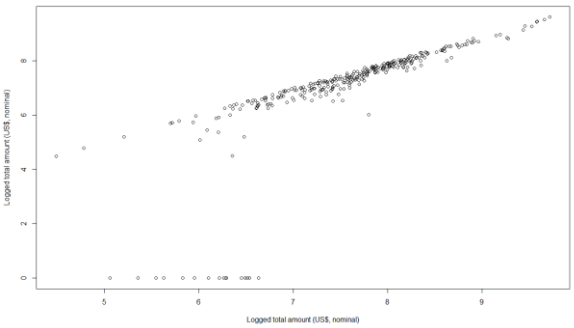
(a) All funded rare diseases



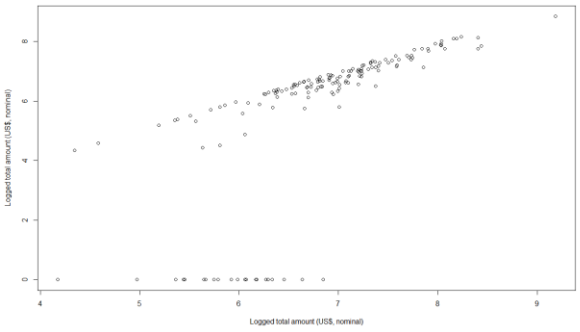
(d) Prevalence: 1-9 / 1 000 000



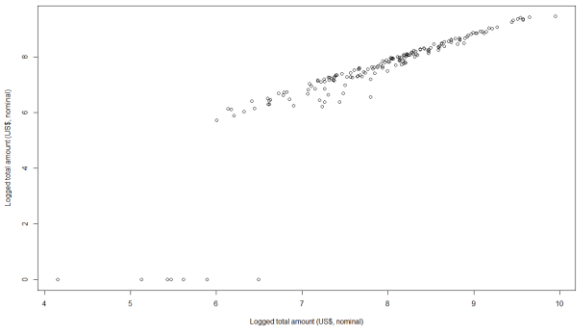
(b) Prevalence: Unknown



(e) Prevalence: 1-9 / 100 000

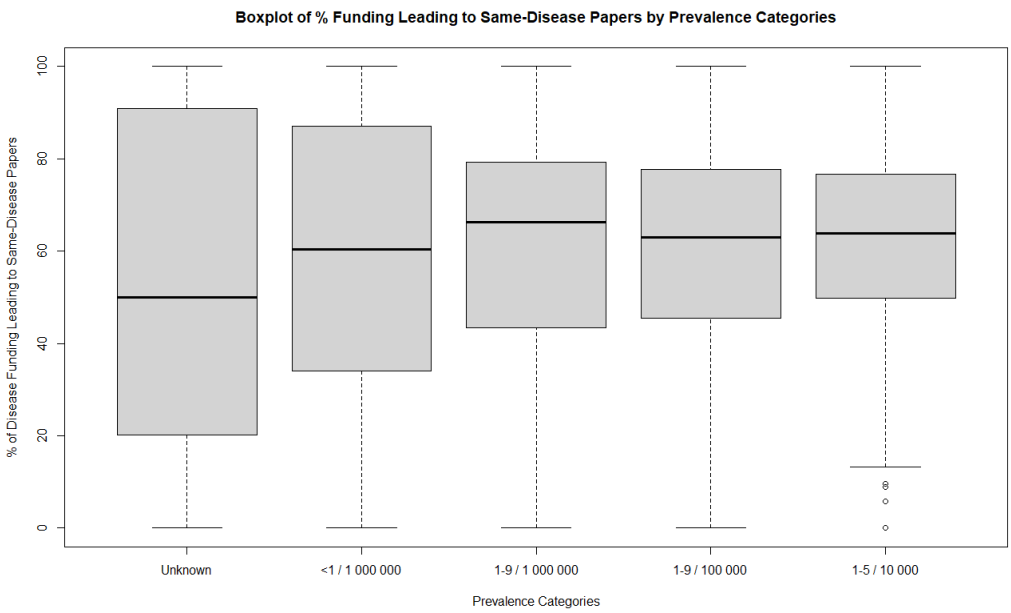


(c) Prevalence: <1 / 1 000 000



(f) Prevalence: 1-5 / 10 000

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



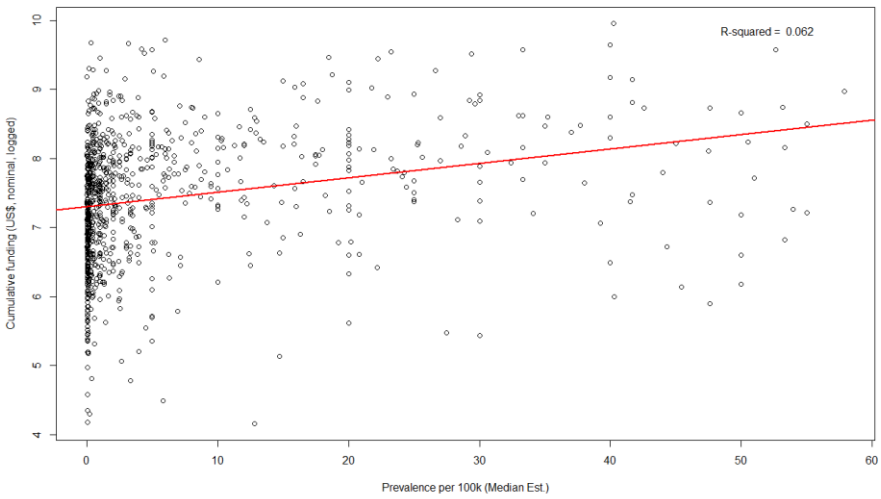
(b) Percentage of NIH disease-funding leading to same-disease papers and prevalence categories

	0	1.0-24.99	25.0-49.99	50.0-74.99	75.0-99.99	100	Median (%)
Unknown	25	11	30	19	29	17	49.94
<1 / 1 000 000	21	12	29	42	35	21	60.42
1-9 / 1 000 000	16	13	47	82	66	12	66.29
1-9 / 100 000	16	12	67	128	78	14	63.01
1-5 / 10 000	7	8	29	81	46	2	67.10

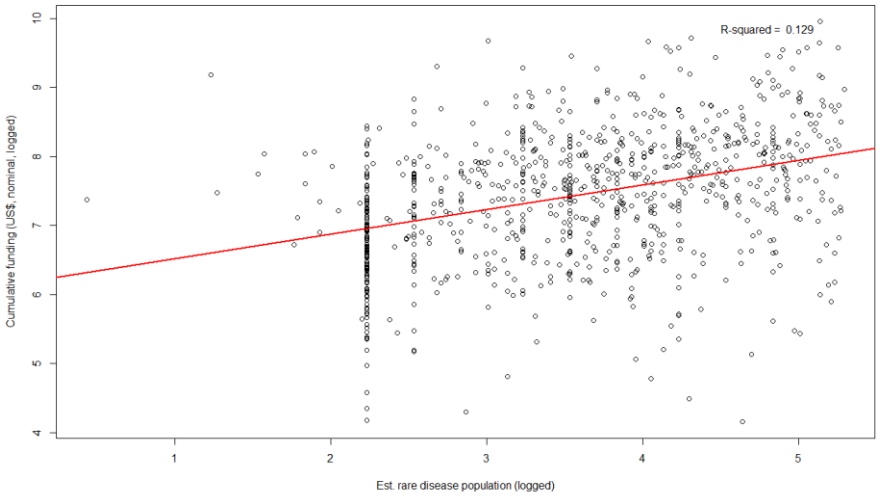
(c) Summary of diseases grouped by percentage of NIH funding leading to same-disease papers and Orphanet prevalence categories

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 1,015 funded 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) visualizes the results based on the estimated focal disease patient population, assuming the US population of 340 million. Part (c) expands upon part (a) by visualizing the distribution of cumulative funding across Orphanet prevalence categories. Plots in (a) and (b) are drawn based on 884 out of 1,015 diseases, as the remaining diseases have no available prevalence estimates. Following the US government's definition, I only focus on diseases estimated to affect 200,000 people or fewer, and assuming the 340-million figure, this translates to a threshold of approximately 58.82 per 100,000 people.



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



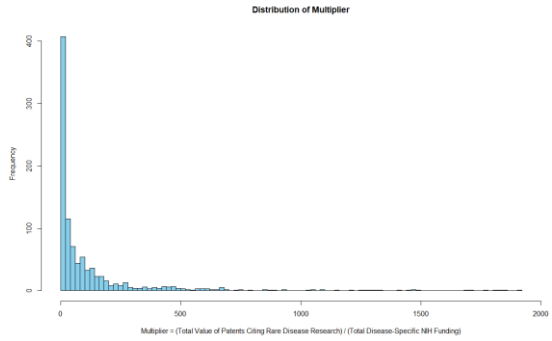
(b) Descriptive correlation of US rare disease patient population estimates and amount of disease-specific NIH funding (both logged)

Cumulative funding \ Prevalence	Unknown	<1 / 1 000 000	1-9 / 1 000 000	1-9 / 100 000	1-5 / 10 000
No funding	99	72	28	36	15
<US$1mil	17	24	10	14	6
US$1mil - US$9.99mil	57	67	67	61	19
US$10mil - US$49.99mil	44	47	74	95	42
US$50mil - US$99.99mil	5	10	38	47	18
US$100mil - US$499.99mil	7	11	33	75	56
US$500mil - US$1bil	1	0	11	12	15
>US$1bil	0	1	3	11	17

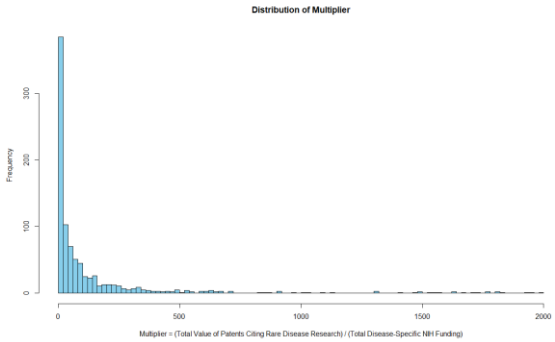
(c) Summary of disease-specific NIH funding by Orphanet prevalence categories

Figure 11. Distribution of ROI to disease-specific research

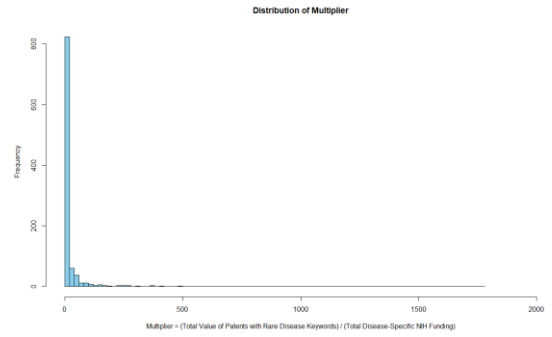
This figure visualizes the distribution of returns on investment (ROI) of NIH funding across various diseases. With regard to three figures on the left—mainly (a), (b), and (c)—the X-axis is proxied by the total value of patents divided by the cumulative amount of disease-specific funding. The remaining 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 associated with same-disease publications, and the numerator adjusted for the total value of patents citing these acknowledging publications.



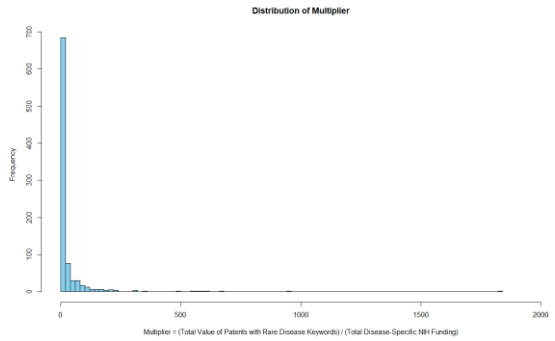
(a) X-axis: Total value of patents citing rare disease research (1995-2023) / Total disease-specific NIH funding



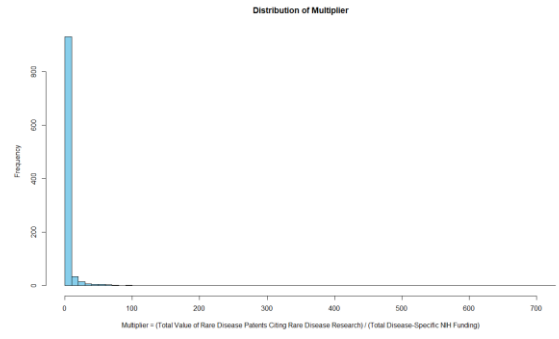
(d) X-axis: Total value of patents citing rare disease research (1995-2023) / Total disease-specific NIH funding resulting in same-disease papers



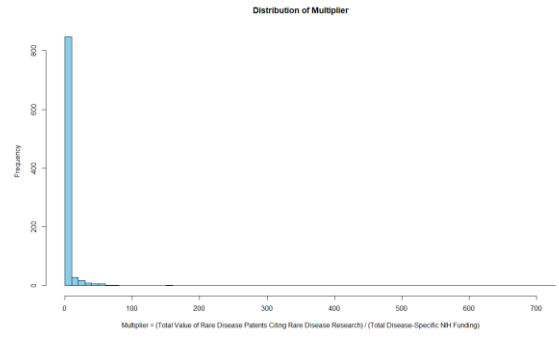
(b) X-axis: Total value of patents with rare disease keywords (1995-2023) / Total disease-specific NIH funding



(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	Cumulative amount (US$, nominal), only funding that leads to same-disease papers within 5 years
Acute Respiratory Distress Syndrome	N	14,094	8,971,212,911	2,893,649,511	2,801,823,371
Glioma	Y	13,379	5,242,800,370	4,074,119,855	4,007,961,511
WAS Related Disorders	N	11,121	4,749,634,374	2,098,482,605	1,803,387,389
Glioblastoma	Y	11,487	4,567,810,977	3,290,168,169	3,255,093,855
Cystic Fibrosis	N	14,402	4,438,910,685	2,737,555,414	2,652,970,725
Graft versus Host Disease	N	8,134	3,846,506,792	2,852,255,449	2,788,885,643
Astrocytoma	Y	10,605	3,774,337,996	2,755,608,937	2,728,942,223
Multiple Myeloma	Y	8,318	3,767,986,727	2,127,757,867	2,068,577,792
Hepatocellular Carcinoma	Y	10,521	3,747,482,273	2,299,479,717	2,194,080,549
Acute Myeloid Leukemia	Y	9,257	3,480,120,351	2,572,117,745	2,532,186,380
Amyotrophic Lateral Sclerosis	N	9,472	3,355,080,197	1,867,664,808	1,815,042,838
Sickle Cell Disease	N	7,783	3,241,063,040	2,334,476,019	2,219,725,463
Frontotemporal Degeneration	N	5,598	2,881,866,511	2,090,949,699	2,082,274,093
Neuroblastoma	Y	7,364	2,859,620,255	1,913,702,705	1,861,886,495
Small Cell Lung Cancer	Y	8,301	2,771,209,611	1,770,104,447	1,722,680,314
Huntington’s Disease	N	7,395	2,725,500,536	1,361,790,598	1,324,964,307
Anthrax	N	3,364	1,997,648,675	960,155,035	920,681,587
Radiation Sickness	N	4,132	1,930,192,697	899,385,560	865,903,075
Osteosarcoma	Y	5,353	1,889,020,487	657,510,706	622,848,254
Retinitis Pigmentosa	N	4,632	1,875,076,310	1,182,248,661	1,132,718,613
Amyloidosis	N	4,103	1,830,596,822	695,471,321	665,802,200
Fragile X Syndrome	N	4,090	1,644,119,431	1,045,108,482	1,005,001,557
Retinoblastoma	Y	4,966	1,546,878,853	914,621,448	877,223,032
Hutchinson-Gilford Progeria Syndrome	N	3,665	1,534,044,216	700,293,538	682,182,101
Myelodysplastic Syndromes	Y	3,651	1,499,596,803	1,114,265,829	1,105,151,192

Cone Dystrophy	N	3,121	1,410,321,247	854,666,132	833,319,777
Stomach Cancer	Y	3,670	1,395,034,263	810,077,569	796,243,003
Neuroleptic Malignant Syndrome	N	4,447	1,316,236,300	735,942,596	700,490,243
Hemophilia A	N	3,025	1,264,431,725	805,074,052	772,183,234
Idiopathic Pulmonary Fibrosis	N	3,126	1,195,226,898	857,586,799	843,659,629
Esophageal Cancer	Y	2,106	1,073,026,769	722,141,485	717,700,913
Chronic Myelogenous Leukemia	Y	3,150	1,047,405,122	738,149,460	726,217,399

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 32 diseases are identified from the list of 1,015 funded diseases. 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. Summary statistics 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 (with disease double-counting)**	**Total count (without disease double-counting)**	**No. of diseases covered**
All patents across 3 categories	755,886	220,153	1,171
Patents citing rare disease research (regardless of relevance)	612,119	192,311	1,143
Patents with rare disease keywords (regardless of scientific linkages)	111,654	40,843	1,010
Rare disease patents citing rare disease research (citing same-disease papers)	32,113	22,809	693
Panel 2: Patents that cite focal-disease papers acknowledging same-disease NIH projects			
Category	**Total count (with disease double-counting)**	**Total count (without disease double-counting)**	**No. of diseases covered**
Patents citing rare disease research (regardless of relevance)	377,392	72,835	1,102
Rare disease patents citing rare disease research (citing same-disease papers)	22,412	14,742	626

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. Summary statistics of USPTO patent grants directly/indirectly related to rare disease research with KPSS value

Panel 1: Patents regardless of linkages to disease-specific NIH projects					
Category	**Total count (with disease double-counting)**	**Total value (US$ mil, nominal)**	**Total count (excl. disease double-counting)**	**Total value (US$ mil, nominal, excl. disease double-counting)**	**No. of diseases covered**
All patents across 3 categories	170,636	13,096,824	48,044	3,407,980	1,004
Patents citing rare disease research (regardless of relevance)	139,391	10,217,808	41,906	2,884,525	959
Patents with rare disease keywords (regardless of scientific linkages)	24,531	2,272,036	9,050	785,815	740
Rare disease patents citing rare disease research (citing same-disease papers)	6,714	606,980	4,577	394,488	411
Panel 2: Patents that cite focal-disease papers acknowledging same-disease NIH projects					
Category	**Total count (with disease double-counting)**	**Total value (US$ mil, nominal)**	**Total count (excl. disease double-counting)**	**Total value (US$ mil, nominal, excl. disease double-counting)**	**No. of diseases covered**
Patents citing rare disease research (regardless of relevance)	86,382	6,680,414	14,731	1,178,559	883
Rare disease patents citing rare disease research (citing same-disease papers)	4,741	452,194	2,998	270,399	349

Note: this table summarizes the total count of patents directly/indirectly related to disease-specific NIH projects with available estimated private value from the Kogan et al. (2017), or KPSS, dataset. On average, patents with KPSS value in this table comprise roughly 20-22% of the figures displayed in Table 2, and belong to publicly traded firms for which the KPSS dataset derives from. 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 4. 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 NIH funding (in US$ million, nominal)**
Neuroblastoma	18,726	4,154	326,718.0	2,859.6
Small Cell Lung Cancer	18,612	3,702	303,852.1	2,771.2
Acute Myeloid Leukemia	12,839	3,564	278,700.9	3,480.1
Glioma	20,283	4,192	266,619.4	5,242.8
Multiple Myeloma	10,968	2,615	255,993.1	3,768.0
Hepatocellular Carcinoma	15,284	3,237	255,752.0	3,747.5
Glioblastoma	15,738	3,544	254,668.9	4,567.8
Cystic Fibrosis	17,215	3,785	218,755.8	4,438.9
Chronic Myelogenous Leukemia	9,064	2,593	186,407.6	1,047.4
C3 Glomerulopathy: Dense Deposit Disease and C3 Glomerulonephritis	6,405	1,878	160,066.8	34.2
Severe Combined Immunodeficiency	8,109	1,977	159,607.6	657.2
Stomach Cancer	10,580	2,017	156,882.2	1,395.0
WAS Related Disorders	8,627	1,849	148,909.5	4,749.6
Retinoblastoma	8,341	1,972	148,165.6	1,546.9
Graft versus Host Disease	7,640	1,755	146,927.6	3,846.5
Osteosarcoma	8,688	2,039	141,584.6	1,889.0
Acute Respiratory Distress Syndrome	5,381	1,419	124,970.7	8,971.2
Follicular Lymphoma	5,013	1,563	123,170.8	419.9
Hemophilia A	7,236	1,509	117,412.7	1,264.4
Astrocytoma	5,374	1,355	100,799.2	3,774.3
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 NIH funding (in US$ million, nominal)**
WAS Related Disorders	3,291	1,061	95,842.0	4,749.6
Stomach Cancer	4,041	609	63,569.7	1,395.0
Arthritis, Infectious	1,627	537	57,473.0	168.2
Small Cell Lung Cancer	2,608	528	55,561.0	2,771.2
Huntington's Disease	2,249	553	55,258.2	2,725.5
Chronic Myelogenous Leukemia	1,741	472	53,948.0	1,047.4
Amyotrophic Lateral Sclerosis	1,978	455	52,280.9	3,355.1
Esophageal Cancer	2,453	403	46,602.0	1,073.0
Glioblastoma	1,979	444	44,457.3	4,567.8
Multiple Myeloma	2,413	505	42,770.5	3,768.0
Acute Respiratory Distress Syndrome	1,323	394	42,559.9	8,971.2

Guillain-Barre Syndrome	893	275	41,528.8	248.2
Cystic Fibrosis	1,662	383	36,437.5	4,438.9
Acute Myeloid Leukemia	1,402	358	36,044.3	3,480.1
Glioma	1,726	320	33,412.7	5,242.8
Myelodysplastic Syndromes	954	235	30,882.6	1,499.6
Vasculitis	992	227	27,140.4	766.8
Hepatocellular Carcinoma	1,364	273	25,534.9	3,747.5
Scleroderma	1,332	309	24,582.0	856.9
Graft versus Host Disease	1,119	284	24,366.6	3,846.5

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 NIH funding (in US$ million, nominal)**
Small Cell Lung Cancer	1,773	338	48,173.1	2,771.2
Multiple Myeloma	1,320	339	39,614.3	3,768.0
Acute Myeloid Leukemia	1,025	293	35,832.8	3,480.1
Glioblastoma	1,176	210	34,577.9	4,567.8
Cystic Fibrosis	1,236	350	30,995.6	4,438.9
Stomach Cancer	1,105	193	18,172.1	1,395.0
Chronic Myelogenous Leukemia	531	146	15,109.0	1,047.4
Hepatocellular Carcinoma	741	145	14,979.7	3,747.5
Huntington's Disease	824	183	14,390.2	2,725.5
Glioma	1,007	153	12,055.1	5,242.8
Graft versus Host Disease	543	102	11,548.2	3,846.5
Amyotrophic Lateral Sclerosis	807	137	10,834.0	3,355.1
Myelodysplastic Syndromes	360	106	10,707.4	1,499.6
Desmoid Tumor	17	9	10,210.2	14.0
WAS Related Disorders	229	82	10,090.7	4,749.6
Follicular Lymphoma	262	97	8,610.1	419.9
Neuroblastoma	683	112	7,822.5	2,859.6
Pulmonary Aerial Hypertension	321	82	7,505.6	963.4
C3 Glomerulopathy: Dense Deposit Disease and C3 Glomerulonephritis	294	80	7,420.9	34.3
Cholangiocarcinoma	123	24	7,321.0	420.3

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 refers to the cumulative funding amount, regardless of whether the funding leads to same-disease papers.

Table 5. 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 NIH funding (in US$ million, nominal)**
Neuroblastoma	11,805	2,610	224,488.9	1,913.7
Small Cell Lung Cancer	12,420	2,465	206,822.2	1,770.1
Acute Myeloid Leukemia	8,653	2,469	201,631.0	2,572.1
Glioblastoma	10,882	2,562	190,683.4	3,290.2
Multiple Myeloma	8,138	1,930	181,527.2	2,127.8
Glioma	12,690	2,675	171,162.3	4,074.1
Hepatocellular Carcinoma	8,837	1,742	144,230.0	2,299.5
Chronic Myelogenous Leukemia	6,308	1,884	139,386.4	738.1
Cystic Fibrosis	9,206	1,972	119,694.2	2,737.6
C3 Glomerulopathy: Dense Deposit Disease and C3 Glomerulonephritis	3,931	1,271	111,400.7	9.3
Retinoblastoma	5,460	1,329	109,683.3	914.6
Stomach Cancer	6,559	1,262	103,988.0	810.1
Graft versus Host Disease	5,491	1,247	103,136.0	2,852.3
Follicular Lymphoma	3,910	1,298	102,844.4	319.5
Severe Combined Immunodeficiency	5,776	1,350	99,101.2	392.2
Osteosarcoma	5,290	1,274	97,433.1	657.5
Acute Respiratory Distress Syndrome	3,082	937	87,371.7	2,893.6
Myelodysplastic Syndromes	3,364	929	77,512.3	1,114.3
WAS Related Disorders	4,652	963	74,985.2	2,098.5
Mantle Cell Lymphoma	2,643	867	71,213.3	193.0
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 NIH funding (in US$ million, nominal)**
Small Cell Lung Cancer	1,283	263	35,882.8	1,770.1
Multiple Myeloma	1,067	287	33,086.3	2,127.8
Glioblastoma	911	178	30,728.1	3,290.2
Acute Myeloid Leukemia	797	237	28,859.7	2,572.1
Cystic Fibrosis	891	261	25,696.1	2,737.6
Stomach Cancer	714	158	16,097.4	810.1
Hepatocellular Carcinoma	525	112	13,449.7	2,299.5
Chronic Myelogenous Leukemia	416	116	12,835.7	738.1
Glioma	747	126	10,531.7	4,074.1
Graft versus Host Disease	440	83	10,168.1	2,852.3

Myelodysplastic Syndromes	301	89	9,142.9	1,114.3
Desmoid Tumor	5	4	7,998.9	5.9
Amyotrophic Lateral Sclerosis	545	89	7,493.6	1,867.7
Huntington's Disease	500	112	7,063.0	1,361.8
Soft Tissue Sarcoma	58	10	6,452.0	291.3
Frontotemporal Degeneration	270	46	6,345.4	2,090.9
Mantle Cell Lymphoma	206	56	6,129.4	193.0
Cholangiocarcinoma	102	19	5,950.6	292.9
Follicular Lymphoma	213	75	5,798.8	319.5
Progressive Multifocal Leukoencephalopathy	74	44	5,215.6	145.5

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 refers to the cumulative funding amount that eventually results in same-disease publications.

Table 6. Top 20 diseases based on the value of innovations related to focal-disease research (excluding rare cancers)

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 NIH funding (in US$ million, nominal)**
Cystic Fibrosis	17,215	3,785	218,755.8	4,438.9
C3 Glomerulopathy: Dense Deposit Disease and C3 Glomerulonephritis	6,405	1,878	160,066.8	34.3
Severe Combined Immunodeficiency	8,109	1,977	159,607.6	657.2
WAS Related Disorders	8,627	1,849	148,909.5	4,749.6
Graft versus Host Disease	7,640	1,755	146,927.6	3,846.5
Acute Respiratory Distress Syndrome	5,381	1,419	124,970.7	8,971.2
Hemophilia A	7,236	1,509	117,412.7	1,264.4
Amyotrophic Lateral Sclerosis	5,958	1,203	100,013.9	3,355.1
Vasculitis	4,084	872	96,700.5	766.8
Neuroleptic Malignant Syndrome	8,555	1,733	88,113.2	1,316.2
Huntington's Disease	6,489	1,116	85,024.2	2,725.5
Arthritis, Infectious	4,443	1,082	81,658.8	168.2
Amyloidosis	3,719	879	81,116.9	1,830.6
Scleroderma	2,764	603	74,463.3	856.9
POMC Deficiency	1,596	688	58,503.3	1.8
Immune Thrombocytopenia	2,191	577	57,686.2	39.7
Sarcoidosis	1,911	517	56,037.0	393.3
Duchenne Muscular Dystrophy	4,085	993	53,584.2	974.3
Meningitis, Bacterial	2,052	557	52,947.5	155.6
Antithrombin Deficiency	3,698	975	50,649.6	30.1
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 NIH funding (in US$ million, nominal)**
WAS Related Disorders	3,291	1,061	95,842.0	4,749.6
Arthritis, Infectious	1,627	537	57,473.0	168.2
Huntington's Disease	2,249	553	55,258.2	2,725.5
Amyotrophic Lateral Sclerosis	1,978	455	52,280.9	3,355.1
Acute Respiratory Distress Syndrome	1,323	394	42,559.9	8,971.2
Guillain-Barre Syndrome	893	275	41,528.8	248.2
Cystic Fibrosis	1,662	383	36,437.5	4,438.9
Vasculitis	992	227	27,140.4	766.8
Scleroderma	1,332	309	24,582.0	856.9
Graft versus Host Disease	1,119	284	24,366.6	3,846.5

Myasthenia Gravis	923	252	22,758.7	385.5
C3 Glomerulopathy: Dense Deposit Disease and C3 Glomerulonephritis	1,146	366	22,594.8	34.3
Sarcoidosis	859	217	19,512.3	393.3
Idiopathic Pulmonary Fibrosis	743	181	19,138.8	1,195.2
Progressive Supranuclear Palsy	452	112	18,074.3	552.7
Henoch-Schonlein Purpura	555	177	18,006.4	20.1
Toxic Shock Syndrome	317	104	17,535.7	177.8
Autoimmune Hepatitis	833	242	17,340.4	134.9
Epidermolysis Bullosa	233	81	16,958.8	138.5
Pulmonary Arterial Hypertension	648	154	16,593.6	963.4

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 NIH funding (in US$ million, nominal)**
Cystic Fibrosis	1,236	350	30,995.6	4,438.9
Huntington's Disease	824	183	14,390.2	2,725.5
Graft versus Host Disease	543	102	11,548.2	3,846.5
Amyotrophic Lateral Sclerosis	807	137	10,834.0	3,355.1
WAS Related Disorders	229	82	10,090.7	4,749.6
Pulmonary Arterial Hypertension	321	82	7,505.6	963.4
C3 Glomerulopathy: Dense Deposit Disease and C3 Glomerulonephritis	294	80	7,420.9	34.3
Frontotemporal Degeneration	329	54	6,940.6	2,881.9
Hemophilia A	442	104	6,105.1	1,264.4
Arthritis, Infectious	142	45	6,004.0	168.2
Narcolepsy	194	77	5,537.9	300.4
Progressive Multifocal Leukoencephalopathy	78	44	5,215.6	190.7
Sickle Cell Disease	270	59	4,568.4	3,241.1
Progressive Supranuclear Palsy	92	20	4,292.7	552.7
Amyloidosis	266	47	3,991.9	1,830.6
Myasthenia Gravis	145	30	3,837.3	385.5
Primary Sclerosing Cholangitis	107	46	3,797.1	162.2
Systemic Scleroderma	135	27	3,393.9	450.7
Primary Biliary Cholangitis	81	37	3,360.7	44.7
Precocious Puberty	40	27	3,328.5	150.1

Note: this table provides descriptive statistics of the count number of patented innovations and the cumulative value of these patents across different patenting categories, excluding rare cancers. 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 refers to the cumulative funding amount, regardless of whether the funding leads to same-disease papers.

Table 7. 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 (excluding rare cancers)

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 NIH funding (in US$ million, nominal)**
Cystic Fibrosis	9,206	1,972	119,694.2	2,737.6
C3 Glomerulopathy: Dense Deposit Disease and C3 Glomerulonephritis	3,931	1,271	111,400.7	9.3
Graft versus Host Disease	5,491	1,247	103,136.0	2,852.3
Severe Combined Immunodeficiency	5,776	1,350	99,101.2	392.2
Acute Respiratory Distress Syndrome	3,082	937	87,371.7	2,893.6
WAS Related Disorders	4,652	963	74,985.2	2,098.5
Hemophilia A	4,262	888	70,543.4	805.1
Amyotrophic Lateral Sclerosis	3,896	844	68,992.3	1,867.7
Vasculitis	2,264	526	52,465.6	490.4
Huntington's Disease	4,088	657	51,771.1	1,361.8
Amyloidosis	2,326	524	49,383.1	695.5
Immune Thrombocytopenia	1,531	450	46,649.3	25.6
Duchenne Muscular Dystrophy	2,871	725	42,786.4	759.5
Arthritis, Infectious	2,352	597	42,185.0	101.3
Sarcoidosis	1,089	310	41,100.4	217.1
Neuroleptic Malignant Syndrome	3,152	583	35,839.9	735.9
Sickle Cell Disease	2,977	679	34,431.4	2,334.5
Scleroderma	1,632	358	34,288.5	610.7
Listeriosis	2,027	405	33,069.2	138.2
Meningitis, Bacterial	1,141	325	32,977.9	92.4
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 NIH funding (in US$ million, nominal)**
Cystic Fibrosis	891	261	25,696.1	2,737.6
Graft versus Host Disease	440	83	10,168.1	2,852.3
Amyotrophic Lateral Sclerosis	545	89	7,493.6	1,867.7
Huntington's Disease	500	112	7,063.0	1,361.8
Frontotemporal Degeneration	270	46	6,345.4	2,090.9
Progressive Multifocal Leukoencephalopathy	74	44	5,215.6	145.5
WAS Related Disorders	104	39	4,687.8	2,098.5
C3 Glomerulopathy: Dense Deposit Disease and C3 Glomerulonephritis	145	33	4,092.2	9.3
Sickle Cell Disease	221	53	3,926.7	2,334.5

Progressive Supranuclear Palsy	63	15	3,806.1	375.0
Hemophilia A	288	52	2,930.9	805.1
Acid Sphingomyelinase Deficiency	52	8	2,731.7	107.3
Precocious Puberty	17	12	2,659.8	72.9
Idiopathic Pulmonary Fibrosis	233	44	2,530.6	857.6
Myasthenia Gravis	111	21	2,421.2	229.6
Amyloidosis	167	35	2,410.5	695.5
Duchenne Muscular Dystrophy	236	48	2,379.3	759.5
Sitosterolemia	18	14	2,371.4	47.9
Spinal Muscular Atrophy	150	38	2,370.2	373.2
Atypical Hemolytic Uremic Syndrome	33	13	2,350.5	23.2

Note: this table re-estimates the number of patented innovations and the value of patents that directly acknowledges NIH funding, excluding rare cancers. 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 refers to the cumulative funding amount that eventually results in same-disease publications.

Table 8. Summary statistics

Variable	Mean	p25	p50	p75	Min	Max
NIH projects	14.504	0	2	7	0	2,005
NIH projects (same-disease paper match)	5.622	0	0	3	0	423
NIH projects (same-disease paper match within 5 years)	5.284	0	0	3	0	423
NIH funding amount (US$ mil, nominal)	5.817	0	0.391	2.708	0	2,504.719
NIH funding amount (US$ mil, nominal, same-disease paper match)	3.556	0	0.175	1.684	0	461.166
NIH funding amount (US$ mil, nominal, same-disease paper match within 5 years)	3.452	0	0.130	1.610	0	461.166
Patents citing rare disease-specific research	20.079	0	3	12	0	1,665
Patents with rare disease keywords (regardless of paper links)	3.632	0	0	2	0	467
Rare disease patents citing rare disease research	1.054	0	0	0	0	208
Patents citing rare disease-specific research (same-disease NIH-paper links)	12.547	0	1	7	0	1,259
Rare disease patents citing rare disease research (same-disease NIH-paper links)	0.740	0	0	0	0	163
Patents citing rare disease-specific research (same-disease NIH-paper links within 5 years)	12.421	0	1	7	0	1,254
Rare disease patents citing rare disease research (same-disease NIH-paper links within 5 years)	0.738	0	0	0	0	162
Patents citing rare disease-specific research with KPSS value	4.576	0	0	2	0	394
Patents with rare disease keywords (regardless of paper links) with KPSS value	0.794	0	0	0	0	64
Rare disease patents citing rare disease research with KPSS value	0.223	0	0	0	0	51
Patents citing rare disease-specific research (same-disease NIH-paper links) with KPSS value	2.886	0	0	1	0	321
Rare disease patents citing rare disease-specific research (same-disease NIH-paper links) with KPSS value	0.159	0	0	0	0	43

Patents citing rare disease-specific research (same-disease NIH-paper links within 5 years) with KPSS value	2.858	0	0	1	0	321
Rare disease patents citing rare disease-specific research (same-disease NIH-paper links within 5 years) with KPSS value	0.158	0	0	0	0	43
KPSS value (US$ mil, nominal) of patents citing rare disease-specific research	340.156	0	0	99.690	0	33,852.330
KPSS value (US$ mil, nominal) of patents with rare disease keywords	71.896	0	0	0	0	9,889.384
KPSS value (US$ mil, nominal) of rare disease patents citing rare disease research	20.281	0	0	0	0	9,882.777
KPSS value (US$ mil, nominal) of patents citing rare disease-specific research (same-disease NIH-paper links)	223.945	0	0	24.580	0	31,577.720
KPSS value (US$ mil, nominal) of rare disease patents citing rare disease-specific research (same-disease NIH-paper links)	15.271	0	0	0	0	7,325.696
KPSS value (US$ mil, nominal) of patents citing rare disease-specific research (same-disease NIH-paper links within 5 years)	221.894	0	0	24.124	0	31,216.170
KPSS value (US$ mil, nominal) of rare disease patents citing rare disease-specific research (same-disease NIH-paper links within 5 years)	15.204	0	0	0	0	7,325.696

Note: the descriptive statistics are displayed at the disease-year level of observation. As the data encompasses 1,015 diseases with NIH funding and 29 fiscal years (1995-2023), the total number of observations is 1,015 * 29 = 29,435. 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 9. 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	0.409*	0.444**	0.500***	0.574***	0.581***	0.624***	0.678***
	(0.17)	(0.15)	(0.12)	(0.10)	(0.10)	(0.11)	(0.12)
Constant	14.146***	14.134***	14.053***	13.754***	14.392***	14.538***	14.591***
	(2.42)	(2.16)	(1.66)	(1.39)	(1.36)	(1.42)	(1.55)
R-squared	0.784	0.789	0.802	0.814	0.822	0.832	0.840
Adjusted R-squared	0.774	0.779	0.792	0.804	0.812	0.822	0.831
Within R-squared	0.100	0.103	0.112	0.126	0.124	0.135	0.147
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358	23,343
Mean DV	20.080	20.449	21.029	21.583	22.147	22.678	23.209
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	0.080*	0.089**	0.101***	0.118***	0.120***	0.131***	0.145***
	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)
Constant	2.469***	2.459***	2.412***	2.313***	2.424***	2.420***	2.356***
	(0.50)	(0.46)	(0.37)	(0.31)	(0.30)	(0.31)	(0.33)
R-squared	0.670	0.680	0.693	0.706	0.714	0.724	0.735
Adjusted R-squared	0.655	0.665	0.678	0.691	0.698	0.709	0.719
Within R-squared	0.070	0.076	0.083	0.096	0.093	0.102	0.115
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358	23,343
Mean DV	3.632	3.721	3.827	3.927	4.031	4.124	4.205
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	0.043*	0.047**	0.052***	0.060***	0.061***	0.065***	0.071***
	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.425	0.414	0.380*	0.327	0.359*	0.357*	0.333
	(0.25)	(0.23)	(0.19)	(0.17)	(0.17)	(0.18)	(0.19)
R-squared	0.640	0.649	0.663	0.678	0.687	0.701	0.718
Adjusted R-squared	0.624	0.633	0.647	0.661	0.670	0.685	0.702
Within R-squared	0.106	0.107	0.114	0.125	0.122	0.130	0.144
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358	23,343
Mean DV	1.054	1.078	1.111	1.142	1.173	1.207	1.240

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

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 10. Rerun of above analysis, but filtered only to NIH grants acknowledged by same-disease scientific publications

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	1.709***	1.608***	1.486***	1.410***	1.366***	1.380***	1.459***
	(0.22)	(0.20)	(0.18)	(0.17)	(0.16)	(0.17)	(0.20)
Constant	2.936*	3.695**	4.786***	5.691***	6.457***	6.926***	7.097***
	(1.22)	(1.13)	(1.04)	(0.95)	(0.89)	(0.91)	(1.07)
R-squared	0.770	0.763	0.761	0.764	0.769	0.778	0.788
Adjusted R-squared	0.760	0.752	0.750	0.752	0.757	0.765	0.776
Within R-squared	0.302	0.264	0.229	0.204	0.187	0.183	0.189
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358	23,343
Mean DV	12.547	12.846	13.250	13.661	14.085	14.514	14.954
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	0.365***	0.349***	0.323***	0.312***	0.302***	0.309***	0.340***
	(0.06)	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)	(0.06)
Constant	1.578***	1.736***	1.986***	2.161***	2.346***	2.427***	2.374***
	(0.32)	(0.32)	(0.31)	(0.30)	(0.29)	(0.28)	(0.30)
R-squared	0.704	0.706	0.708	0.714	0.718	0.725	0.736
Adjusted R-squared	0.691	0.692	0.694	0.699	0.703	0.710	0.720
Within R-squared	0.166	0.150	0.130	0.119	0.107	0.106	0.118
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358	23,343
Mean DV	3.632	3.721	3.827	3.927	4.031	4.124	4.205
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	0.176***	0.166***	0.153***	0.146***	0.142***	0.143***	0.155***
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Constant	-0.249	-0.184	-0.086	-0.019	0.041	0.073	0.050
	(0.16)	(0.15)	(0.13)	(0.12)	(0.12)	(0.12)	(0.14)
R-squared	0.687	0.678	0.671	0.670	0.673	0.683	0.699
Adjusted R-squared	0.673	0.663	0.655	0.653	0.656	0.665	0.682
Within R-squared	0.300	0.263	0.224	0.200	0.182	0.176	0.191
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358	23,343
Mean DV	0.740	0.759	0.785	0.809	0.833	0.861	0.888

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

Note: this table provides estimates from another OLS regression, with the dependent variables proxied by the subset of patent count figures that cite rare-disease publications acknowledging same-disease NIH projects. 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 11. Rerun of above analysis, but with a 5-year threshold following Azoulay et al. (2019)

Panel 1: Patents citing rare disease-specific research						
	(1)	(2)	(3)	(4)	(5)	(6)
	No lag	1-year lag	2-year lag	3-year lag	4-year lag	5-year lag
NIH projects	1.750***	1.627***	1.491***	1.403***	1.356***	1.374***
	(0.21)	(0.19)	(0.18)	(0.16)	(0.15)	(0.16)
Constant	3.174**	4.030***	5.170***	6.125***	6.904***	7.384***
	(1.11)	(1.02)	(0.93)	(0.85)	(0.80)	(0.82)
R-squared	0.777	0.767	0.763	0.764	0.768	0.776
Adjusted R-squared	0.767	0.755	0.751	0.751	0.755	0.764
Within R-squared	0.326	0.280	0.238	0.207	0.188	0.183
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358
Mean DV	12.421	12.718	13.117	13.525	13.946	14.372
Panel 2: Patents with rare disease-specific keywords						
	(1)	(2)	(3)	(4)	(5)	(6)
	No lag	1-year lag	2-year lag	3-year lag	4-year lag	5-year lag
NIH projects	0.374***	0.352***	0.322***	0.309***	0.297***	0.305***
	(0.06)	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)
Constant	1.656***	1.844***	2.108***	2.299***	2.486***	2.571***
	(0.30)	(0.30)	(0.29)	(0.27)	(0.26)	(0.26)
R-squared	0.708	0.708	0.709	0.713	0.717	0.725
Adjusted R-squared	0.695	0.694	0.695	0.699	0.702	0.709
Within R-squared	0.177	0.156	0.132	0.118	0.105	0.104
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358
Mean DV	3.632	3.721	3.827	3.927	4.031	4.124
Panel 3: Rare disease patents citing rare disease-specific research						
	(1)	(2)	(3)	(4)	(5)	(6)
	No lag	1-year lag	2-year lag	3-year lag	4-year lag	5-year lag
NIH projects	0.177***	0.165***	0.151***	0.144***	0.140***	0.142***
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	-0.198	-0.124	-0.024	0.047	0.102	0.133
	(0.14)	(0.13)	(0.12)	(0.11)	(0.11)	(0.11)
R-squared	0.694	0.682	0.673	0.670	0.673	0.683
Adjusted R-squared	0.680	0.667	0.657	0.653	0.656	0.665
Within R-squared	0.316	0.272	0.228	0.201	0.182	0.177
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358
Mean DV	0.738	0.757	0.782	0.806	0.830	0.858

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

Note: this table provides estimates from another OLS regression, with the first and third dependent variables proxied by the subset of patent count figures that cite rare-disease publications acknowledging same-disease NIH projects within five years of the project date, following Azoulay et al. (2019). 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 five-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 12. Correlation between NIH projects & disease-specific USPTO patents with KPSS value

Panel 1: Patents citing rare disease-specific research with KPSS value							
	(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	0.064*	0.072*	0.081***	0.093***	0.091***	0.099***	0.103***
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	3.643***	3.604***	3.619***	3.598***	3.762***	3.779***	3.857***
	(0.43)	(0.40)	(0.30)	(0.24)	(0.23)	(0.25)	(0.28)
R-squared	0.683	0.683	0.692	0.699	0.704	0.710	0.713
Adjusted R-squared	0.668	0.667	0.677	0.684	0.688	0.693	0.696
Within R-squared	0.032	0.034	0.037	0.040	0.036	0.039	0.038
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358	23,343
Mean DV	4.576	4.628	4.753	4.863	4.978	5.077	5.170
Panel 2: Patents with rare disease-specific keywords with KPSS value							
	(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	0.006	0.007	0.008	0.010*	0.010*	0.011**	0.013**
	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.712***	0.707***	0.711***	0.701***	0.720***	0.719***	0.698***
	(0.09)	(0.08)	(0.07)	(0.06)	(0.06)	(0.05)	(0.05)
R-squared	0.643	0.647	0.655	0.663	0.666	0.669	0.675
Adjusted R-squared	0.627	0.630	0.638	0.646	0.649	0.650	0.656
Within R-squared	0.007	0.009	0.010	0.014	0.013	0.015	0.019
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358	23,343
Mean DV	0.794	0.808	0.826	0.844	0.857	0.868	0.869
Panel 3: Rare disease patents citing rare disease-specific research with KPSS value							
	(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	0.008*	0.009**	0.010***	0.011***	0.012***	0.013***	0.014***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.103*	0.102*	0.094*	0.084*	0.090**	0.086*	0.077*
	(0.05)	(0.05)	(0.04)	(0.03)	(0.03)	(0.03)	(0.04)
R-squared	0.524	0.532	0.544	0.555	0.566	0.582	0.598
Adjusted R-squared	0.502	0.510	0.522	0.533	0.543	0.559	0.575
Within R-squared	0.052	0.052	0.056	0.060	0.059	0.066	0.076
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358	23,343
Mean DV	0.223	0.228	0.235	0.241	0.247	0.254	0.261

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

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 with KPSS value. 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 13. Rerun of above analysis, but filtered only to NIH grants acknowledged by same-disease scientific publications

Panel 1: Patents citing rare disease-specific research with KPSS value							
	(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	0.349***	0.329***	0.300***	0.279***	0.261***	0.262***	0.264***
	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)
Constant	0.923**	1.069***	1.329***	1.556***	1.769***	1.882***	1.996***
	(0.28)	(0.26)	(0.24)	(0.22)	(0.20)	(0.21)	(0.26)
R-squared	0.596	0.591	0.591	0.593	0.596	0.604	0.609
Adjusted R-squared	0.578	0.571	0.571	0.573	0.575	0.582	0.586
Within R-squared	0.133	0.116	0.095	0.080	0.067	0.064	0.058
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358	23,343
Mean DV	2.886	2.944	3.038	3.132	3.229	3.323	3.418
Panel 2: Patents with rare disease-specific keywords with KPSS value							
	(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	0.048***	0.042**	0.037*	0.033*	0.029*	0.029*	0.032**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.523***	0.568***	0.618***	0.658***	0.693***	0.710***	0.697***
	(0.08)	(0.08)	(0.08)	(0.08)	(0.07)	(0.07)	(0.06)
R-squared	0.660	0.658	0.662	0.667	0.668	0.670	0.675
Adjusted R-squared	0.645	0.642	0.646	0.650	0.651	0.652	0.656
Within R-squared	0.055	0.042	0.031	0.025	0.019	0.018	0.020
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358	23,343
Mean DV	0.794	0.808	0.826	0.844	0.857	0.868	0.869
Panel 3: Rare disease patents citing rare disease-specific research with KPSS value							
	(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	0.037***	0.033***	0.031***	0.030***	0.028***	0.029***	0.033***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	-0.047	-0.026	-0.007	0.008	0.022	0.024	0.016
	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
R-squared	0.544	0.537	0.537	0.540	0.545	0.559	0.575
Adjusted R-squared	0.523	0.515	0.515	0.516	0.521	0.535	0.551
Within R-squared	0.160	0.130	0.111	0.098	0.086	0.088	0.099
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358	23,343
Mean DV	0.159	0.164	0.170	0.175	0.180	0.186	0.192

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

Note: this table provides estimates from another OLS regression, with the dependent variables proxied by the subset of patent count figures with KPSS value that cite rare-disease publications acknowledging same-disease NIH projects. The 'NIH projects' variable is still proxied by the total count number of disease-specific NIH projects leading to same-disease publications. 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 14. Rerun of above analysis, but with a 5-year threshold following Azoulay et al. (2019)

Panel 1: Patents citing rare disease-specific research with KPSS value						
	(1)	(2)	(3)	(4)	(5)	(6)
	No lag	1-year lag	2-year lag	3-year lag	4-year lag	5-year lag
NIH projects	0.359***	0.335***	0.302***	0.276***	0.259***	0.262***
	(0.05)	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)
Constant	0.960***	1.126***	1.398***	1.646***	1.851***	1.960***
	(0.26)	(0.24)	(0.22)	(0.20)	(0.18)	(0.18)
R-squared	0.600	0.592	0.590	0.591	0.594	0.601
Adjusted R-squared	0.581	0.573	0.570	0.570	0.572	0.579
Within R-squared	0.145	0.124	0.100	0.080	0.067	0.064
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358
Mean DV	2.858	2.916	3.008	3.102	3.199	3.292
Panel 2: Patents with rare disease-specific keywords with KPSS value						
	(1)	(2)	(3)	(4)	(5)	(6)
	No lag	1-year lag	2-year lag	3-year lag	4-year lag	5-year lag
NIH projects	0.048***	0.041**	0.035*	0.031*	0.027*	0.026*
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.539***	0.588***	0.637***	0.681***	0.716***	0.733***
	(0.08)	(0.08)	(0.08)	(0.07)	(0.07)	(0.07)
R-squared	0.661	0.658	0.662	0.666	0.668	0.669
Adjusted R-squared	0.645	0.642	0.645	0.649	0.650	0.651
Within R-squared	0.056	0.041	0.030	0.023	0.017	0.015
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358
Mean DV	0.794	0.808	0.826	0.844	0.857	0.868
Panel 3: Rare disease patents citing rare disease-specific research with KPSS value						
	(1)	(2)	(3)	(4)	(5)	(6)
	No lag	1-year lag	2-year lag	3-year lag	4-year lag	5-year lag
NIH projects	0.037***	0.033***	0.031***	0.029***	0.028***	0.029***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	-0.038	-0.016	0.005	0.020	0.033	0.036
	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
R-squared	0.551	0.542	0.541	0.542	0.548	0.562
Adjusted R-squared	0.531	0.520	0.518	0.518	0.524	0.537
Within R-squared	0.170	0.136	0.114	0.099	0.087	0.089
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358
Mean DV	0.158	0.163	0.169	0.174	0.179	0.185

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

Note: this table provides estimates from another OLS regression, with the first and third dependent variables proxied by the subset of patent count figures with KPSS value that cite rare-disease publications acknowledging same-disease NIH projects within five years of the project date, following Azoulay et al. (2019). The 'NIH projects' variable is still proxied by the total count number of disease-specific NIH projects leading to same-disease publications. The analysis is run across various models from no-lag to five-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 15. Correlation between NIH funding and the total KPSS value of patents (US$ million)

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)	5.444	5.347	9.512	21.701***	26.026***	30.763***	32.433***
	(4.38)	(4.87)	(6.09)	(3.82)	(3.92)	(4.24)	(4.62)
Constant	308.511***	317.821***	307.932***	258.054***	250.414***	242.045***	249.052***
	(25.51)	(27.11)	(32.25)	(19.21)	(18.87)	(19.51)	(20.38)
R-squared	0.545	0.548	0.562	0.580	0.589	0.602	0.609
Adjusted R-squared	0.524	0.527	0.540	0.559	0.567	0.580	0.587
Within R-squared	0.014	0.011	0.019	0.048	0.057	0.071	0.072
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358	23,343
Mean DV	340.177	347.578	358.291	367.295	375.562	383.693	392.126
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)	0.229	0.492	1.232	3.011*	4.077**	5.642***	6.876***
	(0.91)	(1.05)	(1.42)	(1.44)	(1.46)	(1.41)	(1.36)
Constant	70.571***	71.180***	69.142***	62.142***	58.152***	50.536***	45.044***
	(5.28)	(5.86)	(7.50)	(7.23)	(7.01)	(6.50)	(6.00)
R-squared	0.449	0.455	0.462	0.469	0.475	0.494	0.507
Adjusted R-squared	0.424	0.429	0.436	0.442	0.447	0.466	0.478
Within R-squared	0.0003	0.001	0.004	0.012	0.018	0.034	0.049
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358	23,343
Mean DV	71.901	73.919	75.665	77.297	77.758	76.513	75.378
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)	0.923	1.052	1.815+	3.715***	4.299***	4.838***	5.471***
	(0.75)	(0.81)	(1.02)	(0.95)	(1.00)	(1.10)	(1.19)
Constant	14.911***	15.073***	11.943*	3.500	2.091	1.079	-0.089
	(4.36)	(4.51)	(5.40)	(4.76)	(4.81)	(5.06)	(5.27)
R-squared	0.330	0.339	0.353	0.377	0.390	0.404	0.418
Adjusted R-squared	0.299	0.307	0.321	0.345	0.357	0.371	0.384
Within R-squared	0.015	0.016	0.025	0.050	0.056	0.062	0.072
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358	23,343
Mean DV	20.282	20.925	21.555	22.203	22.761	23.353	24.044

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

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 disease-specific NIH funding (in US$ million, nominal), regardless of whether the funding 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, regardless of linkages to same-disease NIH-funded research. 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 16. Rerun of above analysis, but filtered to NIH grants resulting in same-disease papers

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)	31.060***	27.030***	26.728***	27.565***	32.155***	35.760***	36.353***
	(6.12)	(5.70)	(4.83)	(4.60)	(3.89)	(4.56)	(5.24)
Constant	113.504***	134.385***	145.170***	153.036***	149.973***	151.534***	163.046***
	(21.75)	(20.13)	(16.67)	(15.38)	(12.52)	(14.07)	(15.59)
R-squared	0.494	0.487	0.492	0.500	0.511	0.523	0.530
Adjusted R-squared	0.471	0.463	0.467	0.474	0.485	0.497	0.503
Within R-squared	0.086	0.064	0.056	0.054	0.059	0.065	0.061
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358	23,343
Mean DV	223.953	229.806	237.409	245.270	253.367	262.016	271.109
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)	5.105*	4.871*	4.827*	4.633*	5.786**	7.825***	9.389***
	(2.26)	(2.26)	(2.09)	(2.11)	(2.20)	(2.12)	(2.01)
Constant	53.745***	56.723***	59.008***	61.796***	59.154***	52.338***	47.468***
	(8.05)	(7.96)	(7.22)	(7.07)	(7.07)	(6.54)	(5.98)
R-squared	0.460	0.464	0.469	0.470	0.474	0.493	0.504
Adjusted R-squared	0.435	0.439	0.442	0.443	0.446	0.464	0.475
Within R-squared	0.020	0.018	0.016	0.014	0.017	0.031	0.044
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358	23,343
Mean DV	71.900	73.918	75.664	77.297	77.759	76.513	75.378
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)	4.742***	4.329***	4.452***	4.542***	5.208***	5.831***	6.175***
	(1.29)	(1.18)	(1.15)	(1.14)	(1.18)	(1.29)	(1.34)
Constant	-1.593	0.490	0.980	1.710	0.654	-0.017	0.278
	(4.59)	(4.15)	(3.95)	(3.81)	(3.80)	(4.00)	(3.98)
R-squared	0.354	0.352	0.358	0.363	0.374	0.387	0.397
Adjusted R-squared	0.325	0.321	0.326	0.331	0.341	0.353	0.362
Within R-squared	0.071	0.058	0.055	0.052	0.054	0.060	0.061
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358	23,343
Mean DV	15.271	15.772	16.345	16.908	17.400	17.997	18.635

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

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 disease-specific 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, with the first and third categories focusing on patents that cite focal-disease publications acknowledging same-disease NIH-funded research. Standard errors are clustered at the disease level, and the results take into account the inclusion of disease, year, and disease-age fixed effects.

Table 17. Rerun of above analysis, but with a 5-year threshold following Azoulay et al. (2019)

Panel 1: KPSS value (US$ mil, nominal) of patents citing rare disease-specific research						
	(1)	(2)	(3)	(4)	(5)	(6)
	No lag	1-year lag	2-year lag	3-year lag	4-year lag	5-year lag
NIH amount (US$ mil)	31.338***	27.154***	26.697***	27.377***	31.986***	35.628***
	(6.16)	(5.71)	(4.81)	(4.56)	(3.82)	(4.46)
Constant	113.737***	134.784***	146.093***	154.601***	152.061***	154.086***
	(21.25)	(19.55)	(16.08)	(14.72)	(11.82)	(13.22)
R-squared	0.493	0.485	0.489	0.497	0.507	0.520
Adjusted R-squared	0.470	0.461	0.464	0.471	0.481	0.494
Within R-squared	0.088	0.065	0.056	0.053	0.058	0.063
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358
Mean DV	221.902	227.706	235.231	243.018	251.039	259.664
Panel 2: KPSS value (US$ mil, nominal) of patents with rare disease-specific keywords						
	(1)	(2)	(3)	(4)	(5)	(6)
	No lag	1-year lag	2-year lag	3-year lag	4-year lag	5-year lag
NIH amount (US$ mil)	5.157*	4.949*	4.903*	4.691*	5.763**	7.774***
	(2.29)	(2.27)	(2.09)	(2.12)	(2.21)	(2.17)
Constant	54.100***	56.981***	59.295***	62.147***	59.924***	53.476***
	(7.90)	(7.78)	(6.98)	(6.83)	(6.85)	(6.43)
R-squared	0.460	0.465	0.469	0.470	0.474	0.492
Adjusted R-squared	0.436	0.439	0.443	0.443	0.446	0.464
Within R-squared	0.021	0.019	0.017	0.014	0.017	0.030
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015
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	29,433	28,418	27,403	26,388	25,373	24,358
Mean DV	71.900	73.918	75.664	77.297	77.759	76.513
Panel 3: KPSS value (US$ mil, nominal) of rare disease patents citing rare disease-specific research						
	(1)	(2)	(3)	(4)	(5)	(6)
	No lag	1-year lag	2-year lag	3-year lag	4-year lag	5-year lag
NIH amount (US$ mil)	5.159***	4.713***	4.788***	4.845***	5.570***	6.194***
	(1.44)	(1.33)	(1.30)	(1.29)	(1.33)	(1.43)
Constant	-2.605	-0.426	0.285	1.188	0.086	-0.439
	(4.96)	(4.54)	(4.35)	(4.17)	(4.11)	(4.24)
R-squared	0.320	0.316	0.320	0.325	0.337	0.351
Adjusted R-squared	0.295	0.290	0.294	0.298	0.308	0.322
Within R-squared	0.086	0.070	0.065	0.060	0.064	0.070
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	1,015	1,015	1,015	1,015	1,015	1,015
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	29,435	28,420	27,405	26,390	25,375	24,360
Mean DV	15.204	15.702	16.272	16.834	17.323	17.917

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

Note: this table provides estimates from another OLS regression, with the first and third dependent variables proxied by the total KPSS value of patents that cite rare-disease publications acknowledging same-disease NIH projects within five years of the project date, following Azoulay et al. (2019). The 'NIH funding' variable is still proxied by the total amount of disease-specific NIH funding that is acknowledged by same-disease publications within five years of the project date. The analysis is run across various models from no-lag to five-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 18. 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$24.114 million)			Below median cumulative funding (median = US$24.114 million)		
	(1) Patents citing rare disease research	(2) Patents with rare disease keywords	(3) Rare disease patents citing rare disease research	(4) Patents citing rare disease research	(5) Patents with rare disease keywords	(6) Rare disease patents citing rare disease research
L1.NIH amount (US$ mil)	4.935	0.441	1.024	6.959	2.046	-0.019
	(4.61)	(1.02)	(0.80)	(8.64)	(1.76)	(0.47)
Constant	579.634***	133.652***	29.191***	60.602***	8.995***	1.550***
	(50.17)	(11.09)	(8.69)	(2.17)	(0.44)	(0.12)
R-squared	0.562	0.464	0.353	0.217	0.233	0.121
Adjusted R-squared	0.536	0.433	0.315	0.174	0.190	0.072
Within R-squared	0.010	0.001	0.015	0.0001	0.0001	0.000
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	507	507	507	508	508	508
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	14,195	14,195	14,195	14,222	14,222	14,222
Mean DV (US$ million)	633.377	138.457	40.344	62.347	9.508	1.545
Panel 2: above vs below average						
	Above average cumulative funding (average = US$168.686 million)			Below average cumulative funding (average = US$168.686 million)		
	(1) Patents citing rare disease research	(2) Patents with rare disease keywords	(3) Rare disease patents citing rare disease research	(4) Patents citing rare disease research	(5) Patents with rare disease keywords	(6) Rare disease patents citing rare disease research
L1.NIH amount (US$ mil)	3.204	0.144	0.836	12.993***	2.165	0.330
	(3.49)	(0.85)	(0.69)	(3.25)	(1.89)	(0.41)
Constant	1381.312***	303.641***	79.456***	117.778***	24.480***	4.743***
	(99.20)	(24.12)	(19.54)	(3.72)	(2.16)	(0.47)
R-squared	0.613	0.483	0.408	0.332	0.327	0.121
Adjusted R-squared	0.574	0.431	0.348	0.299	0.294	0.078
Within R-squared	0.005	0.0001	0.011	0.002	0.0005	0.0001
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	162	162	162	853	853	853
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	4,517	4,517	4,517	23,879	23,879	23,879
Mean DV (US$ million)	1473.710	307.781	103.578	132.655	26.959	5.120

Panel 3: above vs below 90th percentile						
	Above 90th percentile cumulative funding (90th percentile = US$298.369 million)			Below 90th percentile cumulative funding (90th percentile = US$298.369 million)		
	(1) Patents citing rare disease research	(2) Patents with rare disease keywords	(3) Rare disease patents citing rare disease research	(4) Patents citing rare disease research	(5) Patents with rare disease keywords	(6) Rare disease patents citing rare disease research
L1.NIH amount (US$ mil)	2.900	0.036	0.729	10.162***	3.319*	0.880*
	(2.96)	(0.74)	(0.60)	(2.62)	(1.39)	(0.38)
Constant	1932.547***	415.025***	123.739***	140.877***	27.904***	4.700***
	(121.60)	(30.28)	(24.62)	(4.11)	(2.17)	(0.60)
R-squared	0.648	0.489	0.439	0.308	0.349	0.132
Adjusted R-squared	0.600	0.420	0.363	0.275	0.317	0.090
Within R-squared	0.005	0.000	0.009	0.001	0.002	0.001
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	101	101	101	914	914	914
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	2,792	2,792	2,792	25,590	25,590	25,590
Mean DV (US$ million)	2053.670	416.517	154.195	156.782	33.098	6.079
Panel 4: 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)	27.404**	5.443**	4.152+	1.068	-0.938***	0.257
	(10.33)	(1.84)	(2.46)	(2.08)	(0.20)	(0.25)
Constant	1318.770***	266.552***	57.218	217.793***	53.621***	9.636***
	(199.40)	(35.55)	(47.54)	(9.02)	(0.85)	(1.10)
R-squared	0.648	0.573	0.488	0.533	0.504	0.354
Adjusted R-squared	0.590	0.503	0.404	0.510	0.480	0.323
Within R-squared	0.043	0.030	0.031	0.001	0.008	0.004
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	72	72	72	942	942	942
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	2,005	2,005	2,005	26,374	26,374	26,374
Mean DV (US$ million)	1905.876	383.153	146.162	222.422	49.552	10.749

Panel 5: whether a disease mainly affects pediatric patients						
	A disease mainly affects pediatric patients			A disease does not mainly affect pediatric patients		
	(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)	21.228***	6.816***	4.495**	3.877	-0.104	0.735
	(5.08)	(1.71)	(1.63)	(4.24)	(0.69)	(0.64)
Constant	258.881***	1.544	-6.042	316.489***	83.303***	17.031***
	(29.93)	(10.09)	(9.59)	(23.20)	(3.80)	(3.53)
R-squared	0.658	0.595	0.429	0.510	0.461	0.340
Adjusted R-squared	0.630	0.561	0.381	0.486	0.434	0.306
Within R-squared	0.048	0.173	0.073	0.007	0.0001	0.010
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	218	218	218	797	797	797
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	6,097	6,097	6,097	22,314	22,314	22,314
Mean DV (US$ million)	383.929	41.695	20.435	337.716	82.734	21.058
Panel 6: 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)	5.175	0.400	1.001	0.955	0.492	0.195*
	(4.96)	(1.02)	(0.81)	(1.58)	(0.49)	(0.10)
Constant	397.488***	95.812***	22.378***	182.438***	26.268***	4.315***
	(35.35)	(7.25)	(5.74)	(4.16)	(1.28)	(0.26)
R-squared	0.556	0.472	0.358	0.552	0.446	0.233
Adjusted R-squared	0.533	0.444	0.324	0.522	0.408	0.182
Within R-squared	0.012	0.0009	0.014	0.0001	0.0003	0.0006
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	662	662	662	353	353	353
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	18,531	18,531	18,531	9,880	9,880	9,880
Mean DV (US$ million)	434.376	98.662	29.515	184.953	27.563	4.828

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

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 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). Panel 5 considers whether a disease mainly affects child/pediatric patients (including cancers that predominantly affect pediatric patients). Panel 6 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. 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.

Table 19. Heterogeneity analyses of NIH funding and impacts on the value of same-disease innovations and spillovers, filtered to NIH grants resulting in same-disease papers

Panel 1: above vs below median comparison						
	Above median cumulative funding (median = US$15.095 million)			Below median cumulative funding (median = US$15.095 million)		
	(1) Patents citing rare disease research	(2) Patents with rare disease keywords	(3) Rare disease patents citing rare disease research	(4) Patents citing rare disease research	(5) Patents with rare disease keywords	(6) Rare disease patents citing rare disease research
L1.NIH amount (US$ mil)	25.869***	4.570*	4.258***	6.475	1.197	1.007
	(5.68)	(2.26)	(1.17)	(10.74)	(2.05)	(1.31)
Constant	229.620***	104.537***	0.432	50.260***	10.980***	1.502***
	(39.35)	(15.65)	(8.13)	(1.50)	(0.28)	(0.18)
R-squared	0.507	0.473	0.372	0.277	0.296	0.118
Adjusted R-squared	0.478	0.443	0.335	0.237	0.257	0.068
Within R-squared	0.062	0.016	0.058	0.000	0.000	0.000
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	507	507	507	508	508	508
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	14,195	14,195	14,195	14,199	14,199	14,199
Mean DV (US$ million)	408.816	136.192	29.931	51.166	11.148	1.643
Panel 2: above vs below average						
	Above average cumulative funding (average = US$103.124 million)			Below average cumulative funding (average = US$103.124 million)		
	(1) Patents citing rare disease research	(2) Patents with rare disease keywords	(3) Rare disease patents citing rare disease research	(4) Patents citing rare disease research	(5) Patents with rare disease keywords	(6) Rare disease patents citing rare disease research
L1.NIH amount (US$ mil)	21.875***	3.724	3.951**	4.818	3.549	0.219
	(5.57)	(2.36)	(1.21)	(3.79)	(2.45)	(0.39)
Constant	558.737***	226.138***	8.928	79.574***	23.686***	2.971***
	(94.76)	(40.10)	(20.66)	(2.56)	(1.66)	(0.26)
R-squared	0.556	0.486	0.412	0.256	0.334	0.105
Adjusted R-squared	0.514	0.438	0.356	0.219	0.301	0.061
Within R-squared	0.048	0.013	0.047	0.0002	0.0006	0.000
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	176	176	176	839	839	839
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	4,909	4,909	4,909	23,487	23,487	23,487
Mean DV (US$ million)	933.684	289.961	76.645	82.827	26.082	3.119

Panel 3: above vs below 90th percentile						
	Above 90th percentile cumulative funding (90th percentile = US$198.324 million)			Below 90th percentile cumulative funding (90th percentile = US$198.324 million)		
	(1) Patents citing rare disease research	(2) Patents with rare disease keywords	(3) Rare disease patents citing rare disease research	(4) Patents citing rare disease research	(5) Patents with rare disease keywords	(6) Rare disease patents citing rare disease research
L1.NIH amount (US$ mil)	19.288***	3.109	3.609**	8.572**	5.650**	1.164*
	(5.19)	(2.36)	(1.17)	(2.92)	(2.08)	(0.50)
Constant	865.028***	331.902***	24.079	93.764***	27.510***	3.199***
	(134.99)	(61.38)	(30.44)	(2.91)	(2.08)	(0.49)
R-squared	0.603	0.495	0.445	0.245	0.349	0.119
Adjusted R-squared	0.550	0.427	0.370	0.209	0.318	0.076
Within R-squared	0.043	0.010	0.039	0.0007	0.003	0.001
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	102	102	102	913	913	913
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	2,817	2,817	2,817	25,562	25,562	25,562
Mean DV (US$ million)	1370.476	413.376	118.657	102.310	33.143	4.359
Panel 4: 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)	37.045**	7.615*	4.845+	13.162**	-2.043	1.704**
	(11.49)	(3.22)	(2.86)	(4.99)	(1.70)	(0.54)
Constant	831.784***	270.190***	48.601	102.997***	54.986***	2.673
	(152.97)	(42.83)	(38.12)	(13.27)	(4.52)	(1.44)
R-squared	0.595	0.573	0.477	0.442	0.502	0.360
Adjusted R-squared	0.529	0.503	0.391	0.415	0.478	0.330
Within R-squared	0.056	0.030	0.030	0.031	0.004	0.033
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	72	72	72	942	942	942
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	2,005	2,005	2,005	26,374	26,374	26,374
Mean DV (US$ million)	1378.398	382.548	120.097	138.012	49.553	7.206

Panel 5: whether a disease mainly affects pediatric patients						
	A disease mainly affects pediatric patients			A disease does not mainly affect pediatric patients		
	(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)	36.727***	10.225***	6.478**	23.494***	2.593	3.498**
	(6.15)	(2.75)	(2.13)	(6.94)	(2.40)	(1.23)
Constant	122.297***	1.396	-9.308	139.363***	73.871***	3.692
	(24.25)	(10.86)	(8.40)	(23.72)	(8.21)	(4.19)
R-squared	0.600	0.613	0.465	0.444	0.463	0.335
Adjusted R-squared	0.566	0.581	0.420	0.416	0.436	0.301
Within R-squared	0.104	0.210	0.115	0.049	0.004	0.040
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	218	218	218	797	797	797
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	6,097	6,097	6,097	22,314	22,314	22,314
Mean DV (US$ million)	267.047	41.694	16.224	219.658	82.734	15.646
Panel 6: 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)	28.847***	4.991*	4.540***	4.961	0.722	0.298*
	(6.31)	(2.45)	(1.27)	(3.65)	(0.59)	(0.14)
Constant	161.959***	76.032***	1.931	103.677***	26.372***	2.637***
	(28.61)	(11.11)	(5.77)	(6.01)	(0.97)	(0.23)
R-squared	0.505	0.482	0.371	0.445	0.446	0.219
Adjusted R-squared	0.479	0.455	0.338	0.408	0.408	0.166
Within R-squared	0.075	0.019	0.058	0.001	0.0003	0.001
Clustered standard errors	Disease	Disease	Disease	Disease	Disease	Disease
Num. of diseases	662	662	662	353	353	353
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	18,531	18,531	18,531	9,880	9,880	9,880
Mean DV (US$ million)	292.736	98.657	22.515	111.855	27.562	3.129

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

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 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). Panel 5 considers whether a disease mainly affects child/pediatric patients (including cancers that predominantly affect pediatric patients). Panel 6 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. 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.