Essays on the Production of Ideas

by

Soomi Kim

B.A. English and Economics
Wellesley College (2014)

S.M. Management Research
Massachusetts Institute of Technology (2021)

SUBMITTED TO THE DEPARTMENT OF MANAGEMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN MANAGEMENT

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Authored by: Soomi Kim
MIT Sloan School of Management
March 28, 2023

Certified by: Danielle Li
Professor of MIT Sloan School of Management, Thesis Supervisor

Accepted by: Eric So
Sloan Distinguished Professor of Financial Economics
Professor, Accounting and Finance
Faculty Chair, MIT Sloan PhD Program

Essays on the Production of Ideas

by

Soomi Kim

Submitted to the Department of Management on March 28, 2023
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy in Management

Abstract

Old ideas serve as critical inputs in the production of new ideas. In order to generate knowledge, innovators "stand on the shoulders of giants," the great thinkers who came before, whose ideas serve as the foundation to build on. In this dissertation, I rely on rich empirical data in biomedical settings to identify factors that drive or hinder this cumulative process of knowledge production. The first essay focuses on how knowledge workers innovate in new domains without giants, where there are only few existing ideas to build on. Using the setting of structural biology, I explore how a new technological tool—the automation of analogical reasoning—allowed innovators to import knowledge from an adjacent domain, bypassing the need to build knowledge from the ground up. In the second essay, I turn to how institutions can shape innovative outcomes, particularly when the shoulders of giants rest on a weak foundation. I document that poor communication among different institutional parties of the patent system likely led to the prevalence of biomedical patents based on erroneous or fraudulent science, reducing incentives for innovation. Finally, in the third essay, I highlight the role of private sector polices—specifically, insurance design—in steering the direction of firms' R&D efforts in drug development.

Thesis supervisor: Danielle Li
Title: Associate Professor, Technological Innovation, Entrepreneurship, and Strategic Management, MIT Sloan School of Management

Acknowledgements

My three advisors—Danielle Li, Pierre Azoulay, and Scott Stern—have been extraordinary in their generosity and intellectual guidance. I could not have made it to where I am today without them.

The opportunity to learn from Danielle has been one of the best things that happened to me in graduate school. Danielle's piercing insights and preternatural ability to hone in on the critical strengths and weakness of my work helped me gain clarity whenever I felt lost (which was often). Danielle also taught me the art of presenting; through the countless hours I spent rehearsing my job talk and elevator pitches with Danielle, I learned how to be a storyteller with data and an academic community member who can exchange ideas. But most importantly, I want to thank Danielle for going above and beyond the call of duty. When we first began our co-authored project, Danielle explicitly told me to prioritize my job market paper over our joint project and asked to meet separately to discuss dissertation ideas. This sparked four-years' worth of weekly brainstorming sessions and life lessons that she always shared with her characteristic wit. I have a long way to go, but I hope that I can one day follow in her footsteps.

By taking a chance on my RA application almost eight years ago, Pierre introduced me to the world of innovation—and changed my life. Pierre has provided the foundation behind every step of my PhD. It was Pierre from whom I first encountered the joy of trying to understand the determinants of scientific discoveries. It was Pierre who first showed me how to appreciate both the theory and the details in the data to unearth elegant insights. And it was Pierre who first encouraged me to explore the setting of structural biology, which later formed the basis of my job market paper. Above all, I want to thank Pierre for his unfailing generosity. Whether it would be spontaneous Skype (and later Zoom) conversations at 9pm to weekend emails whenever he had new ideas about my research, Pierre's dedication to not just his own scholarship but to the development of his students was inspiring. Getting a PhD can be a long, lonely journey, and I owe an enormous debt to Pierre, for believing in me when I sometimes did not myself.

I am deeply grateful to Scott for his guidance and tremendous support. Every conversation I had with Scott felt like an adventure that took me to unexpected places. With his boundless knowledge and diverse interests, Scott always uncovered surprising connections between seemingly disparate ideas and shone light on the big picture questions. Scott challenged me to think hard about the overarching implications of my work and left me with fresh perspectives and new puzzles to solve. It feels special to be the next generation of the academic family that leads back to Scott, and the triad model of connecting phenomena, data, and theory that Scott has taught me will serve as the backbone behind all of my work.

Two chapters of this dissertation would not have been possible without my collaborators and mentors, Leila Agha and Janet Freilich. I am very grateful for the opportunities to learn from Leila and Janet on how to transform an initial idea into an empirical project and how to turn a set of regressions into a paper. Above all, I thank them for their guidance and patience, as I was an inexperienced third-year graduate student when I began working with them; our co-authored projects were instrumental in my development as a junior scholar.

I would also like to thank the broader Sloan community. I was amazed by the support network of TIES alumni when I was going through the job market; I am excited to join them and hope to pay it forward. I thank the friends currently in the TIES program for our bond that endured and thrived, even over virtual screens during the past few years. I am particularly grateful to my cohort member, Wes Greenblat. From solving psets together to venting about our failed regressions, Wes has been my PhD buddy since Day 1 and made graduate school so much more fun. Lastly, I thank Natalia Kalas, who is the glue that holds the TIES community together, as well as Hillary Ross and Davin Schnappauf for helping us PhD students meet our milestones.

Outside of Sloan, I was lucky to be a part of the Korean community at MIT, who welcomed me despite my rusty Korean. As I get older, I realize how hard it is to form the kind of effortless friendships that seemed so easy when we were younger, and I especially thank my "92" friends for making me feel like I was back in college (in the best way possible).

Jinyoung, thank you for being my biggest fan.

Finally, to my beloved family: Mom, Dad, Soojin, and Sooyoung. The Kim Family training has always given me the courage and the strength to get back up whenever I fell. For their selfless love and devotion, I dedicate this thesis to my parents.

Table of Contents

Chapter 4. Insurance Design and Pharmaceutical Innovation 112

List of Figures

List of Tables

Chapter 1

Introduction

Old ideas serve as critical inputs in the production of new ideas (Romer 1990). In order to generate knowledge, innovators “stand on the shoulders of giants” (Newton 1675), the great thinkers of the past, whose ideas serve as the foundation to build on.

However, climbing the shoulders of giants can be difficult. For instance, the creation of new ideas does not mean that downstream researchers can automatically build on them (Mokyr 2002). Innovators may be discouraged from producing follow-on research if reward systems only recognize the first discoverer (Scotchmer 1991). Some shoulders of giants may even turn out to be “shaky” and based on ideas that are later proven to be incorrect (Azoulay et al. 2015). Institutional processes therefore play important roles in incentivizing and disseminating knowledge (Furman and Stern 2011; Greenblatt 2021). Moreover, the growing stock of knowledge imposes an increasing educational burden on new generations of innovators, who must learn the prior knowledge before they can reach the frontier (Jones 2009). While new technological tools may mitigate such burden (Teodoridis 2018), studies point to increasing training length and specialization (Jones 2010; Blau and Weinberg 2017).

This dissertation aims to shed light on how technologies, institutions, and policies can influence innovators’ capacity to rely on the giants that came before. Studying innovation processes, however, is empirically challenging. Novel ideas are difficult to define and measure. Above all, in order to assess whether an intervention impacted innovation, one must identify the counterfactual of alternative ideas that could have been pursued in the absence of the intervention. Through a collection of three essays, I overcome these empirical challenges by relying on rich data in biomedical settings to measure and causally identify factors that drive or hinder the cumulative process of knowledge production.

The first essay focuses on how knowledge workers innovate in new domains without giants, where there are only few existing ideas to build on. Using the setting of structural biology, I explore how a new technological tool—the automation of analogical reasoning—allowed innovators to import knowledge from an adjacent domain, bypassing the need to build knowledge from the ground up. In the second essay, I turn to how institutions can shape innovative outcomes, particularly when the shoulders of giants rest on a weak foundation. I document that poor communication among different institutional parties of the patent system likely led to the prevalence of biomedical patents based on erroneous or fraudulent science, reducing incentives for innovation. Finally, in the third essay, I highlight the role of private sector polices—specifically, insurance design—in steering the direction of firms' R&D efforts in drug development.

Shortcuts to Innovation: The Use of Analogies in Knowledge Production

How do knowledge workers innovate when there are only few existing ideas to build on? In the first essay, I explore how analogical reasoning—and technologies that automate it—can serve as "shortcuts" that allow innovators to import knowledge from another domain, instead of building knowledge from scratch.

Although often overlooked, analogical reasoning is ubiquitous in innovation and managerial practice. A wide range of scientific breakthroughs have been sparked by analogies, while managers and entrepreneurs frequently borrow insights from one industry to another. Importantly, by taking analogical reasoning out of an individual mind and outsourcing it to machines, some believe that analogical reasoning can be automated at scale (Kittur et al. 2019). Supervised machine learning, for instance, can be viewed as an analogy-based technology since it discovers patterns from known training templates and apply them in new areas, helping innovators make progress in uncharted terrains.

Yet, one cost of analogies is that they require the availability of other domains as templates. For example, supervised machine learning cannot work without training data. Analogies may therefore restrict the direction of innovation towards areas with available templates, even if those areas are less fruitful. The goals of this essay are to (i) provide a framework of how analogies can serve as shortcuts in innovation and (ii) empirically examine the tradeoffs of relying on analogies.

I leverage the setting of structural biology, a field that studies the 3D structure of proteins. As an important scientific field that has contributed to over a dozen Nobel prizes, structural biology also has empirical features ideally suited for this paper. Using a difference-in-differences design, I examine the introduction of an analogy-based technology and document a tradeoff: while the technology increased the rate of innovation, it also led to workers herding around solving less impactful problems.

Information Quality in the Patent System

The second essay, co-authored with Janet Freilich, examines whether the patent system is sensitive to information quality. Although it is challenging to measure inaccurate information in patents, we develop a novel approach to identify patents with poor-quality information: patent-paper pairs where the paper has been retracted and the corresponding patent—which we term an "unsupported" patent—contains the retracted material.

We find that the participants in the patent system largely appear insensitive to information quality. Even after the material in the unsupported patents was retracted and publicly revealed to be incorrect, applicants invested resources into prosecuting and maintaining these unsupported patents, while examiners failed to reject them. Furthermore, the unsupported patents continued to be cited by downstream patents. Our results raise important concerns. Poor-quality information can damage the disclosure function of the patent system, disseminating incorrect information and potentially decreasing incentives for follow-on innovations.

Private Sector Policies and Pharmaceutical R&D

In the third essay, I explore how private sector policies can shape the R&D strategies of pharmaceutical firms. With co-authors Leila Agha and Danielle Li, I investigate the impact of a major change in insurance policy on upstream drug development.

Private insurance plans traditionally offered coverage for most FDA-approved drugs, but starting in 2012, they began excluding coverage for many drugs, especially those in large disease areas with cheaper alternatives. This policy shifted the R&D incentives of pharmaceutical firms. Prior to the exclusion policy, pharmaceutical firms had strong incentives to develop incremental drugs in proven, historically profitable markets with high prescription volume and already existing therapies. But the new insurance policy excluded these very incremental drugs from coverage, suppressing demand and profitability. We show that pharmaceutical firms adjusted their R&D strategies in response: R&D investments declined in crowded drug classes that faced greater risks of exclusion, highlighting the role that private policies can have on upstream R&D activities.

Conclusion and Future Directions

Taken together, these three essays explore how technologies, institutions, and policies facilitate or slow down the cumulative process of knowledge production. In particular, I hope that this dissertation serves as the prelude to a future research agenda that examines the different types of shortcuts innovators can take and their impact on innovation.

References

Azoulay, Pierre, Jeffrey L Furman, Joshua L Krieger, and Fiona Murray. 2015. "Retractions." *The Review of Economics and Statistics* 97 (5): 1118–36.

Blau, David M., and Bruce A. Weinberg. 2017. "Why the US Science and Engineering Workforce Is Aging Rapidly." *Proceedings of the National Academy of Sciences of the United States of America* 114 (15): 3879–84.

Furman, Jeffrey L., and Scott Stern. 2011. "Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research." *American Economic Review* 101 (5): 1933–63.

Greenblatt, Wesley. 2021. "Building on Solid Ground: Foundational Knowledge and the Dynamics of Innovation." *SSRN Working Paper*. https://ssrn.com/abstract=3919866.

Jones, Benjamin F. 2009. "The Burden of Knowledge and the 'Death of the Renaissance Man': Is Innovation Getting Harder?" *Review of Economic Studies* 76 (1): 283–317.

Jones, Benjamin F. 2010. "Age and Great Invention." *The Review of Economics and Statistics* 92 (1): 1–14. https://doi.org/10.1162/REST.2009.11724.

Kittur, Aniket, Lixiu Yu, Tom Hope, Joel Chan, Hila Lifshitz-Assaf, Karni Gilon, Felicia Ng, Robert E. Kraut, and Dafna Shahaf. 2019. "Scaling up Analogical Innovation with Crowds and AI." *Proceedings of the National Academy of Sciences of the United States of America* 116 (6): 1870–77.

Mokyr, Joel. 2002. *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton, NJ: Princeton University Press.

Newton, Isaac. 1675. "Isaac Newton to Robert Hooke," February 5, 1675.

Romer, Paul M. 1990. "Endogenous Technological Change." *Journal of Political Economy* 98 (5).

Scotchmer, Suzanne. 1991. "Standing on the Shoulders of Giants: Cumulative Research and the Patent Law." *Journal of Economic Perspectives* 5 (1): 29–41.

Teodoridis, Florenta. 2018. "Understanding Team Knowledge Production: The Interrelated Roles of Technology and Expertise." *Management Science* 64 (8): 3625–48.

Chapter 2

Shortcuts to Innovation: The Use of Analogies in Knowledge Production

Abstract

Old ideas serve as critical inputs into new ideas, but how do knowledge workers innovate when there are only few existing ideas to build on? In this paper, I explore how analogical reasoning—and technologies that automate it—can serve as "shortcuts" that allow innovators to import knowledge from an adjacent domain, bypassing the need to build knowledge from the ground up. Yet, because analogies require the availability of other domains as templates, they may also constrain the direction of innovation towards areas with available templates. Using the setting of structural biology, I document a tradeoff: while the arrival of an analogy-based technology increased the rate of innovation, it led to workers herding around solving less impactful problems.

1. Introduction

Knowledge production is cumulative (Romer 1990; Scotchmer 1991). Innovators use existing ideas to produce new ideas—from mechanical engineers relying on the foundation of Newtonian physics to applied economists turning to canonical econometric models. But how do knowledge workers innovate in new domains where there are only few existing ideas to build on?

In this paper, I explore how technologies can be used as "shortcuts" that speed up the process of acquiring foundational knowledge—and even circumvent it altogether. Much of the prior literature on research technologies has focused on vertical shortcuts, those that help innovators more quickly understand and apply existing knowledge to climb to the frontier. What is less obvious is how to innovate in domains where there is no foundation yet. I argue that horizontal shortcuts—specifically, analogies—allow innovators to import knowledge from another domain, bypassing the need to build knowledge from the ground up.

Consider the early history of aviation, when the physics of aerodynamics were not yet discovered. Rather than building up the foundation of aerodynamics, inventors looked to birds as analogies and designed devices that imitate the motion of flapping wings. But all of these attempts failed—until George Cayley, a 19th-centry British inventor, made a breakthrough. In place of flapping wings, Cayley envisioned the first prototype of modern-day airplanes that was later built by the Wright Brothers: a device with fixed wings, which glides to sustain lift.

This history of aviation illustrates both the power and pitfalls of analogies. Not only do analogies identify unexpected connections across knowledge domains, they can also circumvent the need to build foundational knowledge by borrowing insights from another domain. Yet, because analogies require the availability of templates, they may narrow the line of inquiry. Birds provided guidance behind the mechanics of flight, but since they were the only known templates for flight (and the underlying physics were not yet known), early inventors focused on flapping wings and did not consider fixed-wing machines (Spenser 2008; Pollack 2014).

Analogical reasoning is ubiquitous in both research and managerial practice. A wide range of scientific and engineering breakthroughs have been sparked by analogies (Gentner, Holyoak, and Kokinov 2001), from Velcros inspired by plant burrs to Ernest Rutherford's model of the atom as a miniature solar system. In strategy, managers often face strategic problems that are well-suited for analogical reasoning (Gavetti, Levinthal, and Rivkin 2005; Bingham and Kahl 2013). Entrepreneurs commonly conceive of new ventures by adopting insights from one industry to another, such as the over hundred startups that claim to be the next "Uber for X" (Madrigal 2019), from Instacart (Uber for grocery deliveries) to Wag (Uber for dog walkers).

Although analogical reasoning has played a central role in innovation, automation has made it increasingly easier to deploy. In particular, algorithms based on supervised machine learning can be seen as the automation of analogical reasoning. Such algorithms mine patterns from known training templates and apply those patterns to new areas, mirroring human ability for finding patterns that explain the unfamiliar in terms of the familiar. For example, drugmakers refer to well-known compounds to identify promising candidates among unexplored compounds, while managers rely on their experiences with past employees to screen applicants—and drug discovery and hiring algorithms now routinely conduct these types of analogical reasoning on their behalf. By taking analogical reasoning out of an individual mind and outsourcing it to machines, some believe that analogies can be harnessed at scale with data-reliant technologies (Kittur et al. 2019).

However, while analogies can help innovators make progress in uncharted terrains, their need for templates can also restrict the direction of innovative activities. The automation of analogical reasoning makes this tradeoff especially salient: analogy-based technologies like machine learning can only be employed in areas with training data. The availability and location of training data can thus shape the direction of innovation towards some areas, while neglecting others.

In addition to providing a framework for how analogies can serve as shortcuts in innovation, the goal of this paper is to empirically study the tradeoff of relying on analogies. Although analogical reasoning has been a topic of great interest in cognitive psychology (Gentner, Holyoak, and Kokinov 2001; Hofstadter and Sander 2013), along with fewer but important studies by management scholars (e.g., Gavetti, Levinthal, and Rivkin 2005), much of the prior work on analogies has been based on laboratory experiments or case studies.

The scarcity of empirical studies based on real-world, large-scale data is perhaps unsurprising. Analogical reasoning, while pervasive, is a mental shortcut that often goes unnoticed (Dunbar 1999). Above all, analogies—and analogy-based technologies—are challenging to study empirically. In order to investigate whether the use of an analogy-based technology shifts the direction of innovation, it is important to be able to observe what ideas could be pursued in the absence of the technology. Finding such a setting is difficult because the counterfactual of ideas that could have been pursued—but were not—is often unobservable. There also needs to be a credible way to measure distance between ideas, as analogies work by identifying similarities between disparate domains. Finally, the analogy-based technology must differentially treat only some areas of the setting, such that outcomes in the areas where the technology was introduced can be compared to the areas without the technology.

I focus on the setting of structural biology, a field with empirical features ideally suited for this paper. Structural biology studies the 3D structures of proteins and has contributed to more than a dozen Nobel prizes, as proteins play vital roles in virtually every biological process. Elucidating a protein structure at atomic resolution—or "solving" the structure—can reveal the protein's function, which helps with applications such as designing vaccines that train human antibodies to recognize the spike proteins of SARS-CoV-2. Importantly, structural biology has several empirical features that allow me to identify how the introduction of an analogy-based technology may have shaped the subsequent rate and direction of innovation.

First, unlike many settings where only realized ideas are observable, structural biology provides a window into the entire idea landscape. Using a database of all known proteins, I observe which proteins structural biologists explored versus could have explored but neglected. For instance, of the approximately 20,000 human proteins, just one-third of them have had their structures experimentally determined as of 2020. In addition, while most settings do not have an easy way to quantify the similarity between each potential idea, the distance between ideas can be measured in structural biology (Hill and Stein 2020; 2021): proteins are composed of sequence of amino acids, so they can be grouped based on their sequence similarity. In other words, it is possible to map out the idea landscape of all known proteins and see which areas of the landscape have been explored (which I term "bright" clusters of proteins) and which areas remain unexplored and thus do not have built-up knowledge (which I term "dark" clusters).

Second, structural biology is a prime setting for studying analogy-based technologies. Solving a protein structure involves deep knowledge of biology, physics, and statistics, but many of the steps have now become automated. The specific technology I examine is the software program Phaser, released in 2003, which automates a method called molecular replacement (MR). Instead of solving a structure from scratch, MR borrows structure information from previously solved proteins that are similar to the unknown structure that the scientist is trying to elucidate. MR can therefore be viewed as an analogy-based technology since it helps knowledge workers make progress in areas of research without existing knowledge (i.e., proteins whose structures are unknown) by importing structure templates from neighboring proteins.

Finally, this analogy-based technology differentially treated some parts of the idea landscape. Since MR needs data on previously solved structures, MR only works for bright clusters of proteins (i.e., clusters with previously solved structures), and does not work for dark clusters. This allows me to employ a difference-in-differences design where bright and dark clusters serve as my treatment and control groups. By matching data from Swiss-Prot (a database of all known proteins) to the Protein Data Bank (a database of all protein structures), I examine the quantity and quality of structures solved in bright clusters after the arrival of MR, relative to dark clusters.

My first set of results focuses on the rate of innovation. Since MR reduced the cost of solving unknown structures in bright clusters, one would expect more structures to be solved in those clusters. Indeed, I find that bright clusters experienced a relative increase in the total number of solved structures after MR was introduced. This effect was sustained throughout the entire sample period: bright clusters got brighter and brighter.

I then turn to how MR impacted quality and distinguish between two dimensions of quality: execution and importance. In any type of innovative activity, the innovation should be well-executed, but it should also solve an important problem. In the case of structural biology, execution refers to how meticulously a structure was solved (e.g., the resolution or the level of detail found in the structure), while importance refers to whether the structure led to a novel understanding of a biological process. The goal of structural biology is not to solve structures for the sake of solving them; the goal is to learn the functional roles the proteins might play by elucidating their structures. I find that while bright clusters received more well-executed structures, these structures were less scientifically important. They provided fewer functional annotations about the proteins and had lower publication and patent footprint.

A potential identification concern is that bright clusters may have been evolving on different trends than dark clusters before the introduction of MR. I conduct several robustness analyses to address this concern. First, I show that there are no pre-trends in the corresponding event studies. Second, I control for predicted brightness; the idea is to compare clusters that share ex-ante similar traits, but some clusters just happened to be actually bright while other clusters happened to be dark. MR only works when the cluster is actually bright regardless of whether the cluster is predicted to be bright or dark, and I verify that only actual brightness matters when estimating the impact of MR.

Taken together, my results suggest a tradeoff: while the arrival of MR increased the rate of innovation, it also led to knowledge workers solving less impactful problems. These results from structural biology illustrate an inherent limitation of analogies: analogies may serve as shortcuts for making progress in new domains (i.e., proteins whose structures are unknown), but they are also constrained by the need for templates and thus may be employed in domains with neighbors—

that is, potentially crowded areas (i.e., bright, already well-explored clusters) that may not be the most impactful.

This paper contributes to several literatures. I first build on a body of evidence that examines how technologies can both advance and constrain knowledge production. Much of this prior work studies technologies that can be classified as those that help innovators digest and apply *existing* knowledge (Teodoridis 2018; Furman and Teodoridis 2020; Anthony 2021; Miric, Ozalp, and Yilmaz 2021; Mannucci 2017). In contrast, by introducing the idea of analogies, this paper aims to shed light on technological shortcuts that help knowledge workers innovate in domains where little is known. In other words, I distinguish between technologies that reduce the "burden of knowledge" (Jones 2009)—the problem of innovators facing an increasing educational burden as knowledge accumulates—and technologies that alleviate a different problem of innovators lacking existing ideas that can serve as inputs in the production of new ideas.

By focusing on analogy-based technologies, I also contribute to the emerging literature on AI and data-driven exploration. While the literature on AI has extensively documented how algorithmic bias can arise from poor-quality training data (Cowgill and Tucker 2020; Cowgill et al. 2020; Choudhury, Starr, and Agarwal 2020), this paper joins a smaller literature that focuses on how the very availability of training data can dictate where innovations take place (Cockburn, Henderson, and Stern 2018; Hoelzemann et al. 2022).

Lastly, I leverage the setting of structural biology, which was first brought to the attention of social scientists by Hill and Stein (2020; 2021). The authors assess the costs of the priority reward system in science, which tends to only recognize the first discoverer.[1] While I do not study priority races and instead examine the impact of an analogy-based technology by exploiting a novel identification strategy, I similarly highlight the strengths of the setting. As a field with both rich scientific achievements and empirical features, structural biology is an attractive setting for investigating broader questions of how to manage innovation.

The rest of the paper proceeds as follows. Section 2 provides a taxonomy of shortcuts and an overview of the key features and tradeoffs of analogies. Section 3 introduces the institutional context and empirical features of structural biology. Section 4 lists the main data sources. Section 5 describes the difference-and-differences design that underpins this study's empirical strategy.

[1] Specifically, Hill and Stein (2020; 2021) document the effects of being "scooped" on subsequent career outcomes, as well as how competition leads to rushing and lower-quality science. Additionally, a recent paper by Zhuo (2022) estimates a model of lab decision-making on resource allocation in structural biology.

Section 6 presents my main results, along with robustness analyses. Section 7 discusses the implications of my results and conclusions.

2. Shortcuts to Innovation

The cumulative nature of knowledge production typically characterizes the innovation process as a sequence of old ideas generating new ideas (Romer 1990; Scotchmer 1991). This section discusses how there can be shortcuts (particularly shortcuts enabled by technologies) that can speed up—or even bypass—this sequence of knowledge accretion.

2.1 Prior Literature on Research Technologies: Vertical Shortcuts

As knowledge accumulates, every new generation of innovators faces a greater educational burden. This "burden of knowledge" has several implications, including increased training length and specialization as innovators struggle to learn the growing body of knowledge (Jones 2009). The prior literature on research technologies has primarily focused on technologies that I consider as "vertical" shortcuts, which mitigate this burden. These are technologies that allow innovators to more quickly acquire existing foundational knowledge, such that they can use the knowledge as stepping stones to climb to the frontier of knowledge and produce new ideas.

As shown in Figure 1, vertical shortcuts can be thought of as aiding in either understanding or applying foundational knowledge. *Summaries*—from textbooks[2] to Wikipedia—help with understanding existing knowledge by providing a short synopsis of a given domain, saving knowledge workers from having to read every research article or replicate every experiment. *Calculators* help with applying foundational knowledge by executing a pre-programmed menu of instructions based on such knowledge. Consider programs like Stata, which is embedded with canonical econometric models. Stata allows even a college first-year with little training in econometrics to run regressions by simply entering "reg y x."

A large body of prior work on research technologies can be conceptualized as vertical shortcuts, particularly calculators of varying sophistication. Calculators are closely related to the idea of modularity, where "information hiding" (Parnas 1972) within modules glued together by standardized interfaces can facilitate a division of innovative labor (Baldwin and Clark 1997; Sanchez and Mahoney 1996; Simcoe 2015). Examples studied in prior work range from financial

[2] A recent work by Greenblatt (2021) on medical guidelines illustrates how summaries can spur innovation.

spreadsheet technology (Anthony 2021) to animation toolkit (Mannucci 2017) to videogame "middleware" (Miric, Ozalp, and Yilmaz 2021). Although this prior literature does not, for the most part, explicitly discuss the burden of knowledge, notable exceptions are Teodoridis (2018), Furman and Teodoridis (2020), and Nagle and Teodoridis (2020). The authors examine the arrival of an automating motion-sensing technology and suggest the role technologies can have in reducing the burden of knowledge.

2.2 Analogies: Horizontal Shortcuts

While vertical shortcuts can assist knowledge workers in innovating in domains with deep foundation, what about in new domains without such foundation? This paper complements the burden of knowledge literature by focusing on a different problem: how to innovate when there are few existing ideas that can serve as inputs into new ideas.

In new domains, knowledge workers face the challenge of having to build foundational knowledge from scratch—and analogical reasoning can be used to circumvent this challenge. In this section, I describe the key features and tradeoffs of analogies.

2.2.1 Key Features of Analogies

Analogical reasoning has been extensively studied by cognitive psychologists as a crucial component of human cognition. Analogy has been broadly described as "the ability to identify similarities in *relations* that hold within domains" (Gentner 1982; Gentner, Holyoak, and Kokinov 2001; Holyoak and Thagard 1996), even if the individual objects are distinct (e.g., how sound propagates through the air is analogous to how water waves travel in a pond, even though sound and water are not alike).

While the concept of analogy has been employed in diverse disciplines, ranging from linguistics to philosophy, I focus on a simple definition of analogy adapted from cognitive science: the importing of patterns from one knowledge domain to another. This definition leads to three key features of analogies, with respect to their role in knowledge production.

(i) Analogies can serve as shortcuts. Analogical reasoning can be viewed as a shortcut because it can serve as an alternative to other ways in which innovators build knowledge in new domains, such as trial-and-error (Thomke 1998) or by generating a theory (Fleming and Sorenson 2004). As an example, suppose a drugmaker is trying to create a drug for a new disease. One approach would be to screen through millions of compounds to detect pharmacological activity through trial-and-error. Another approach would be to start by building a theory of how the disease operates at the molecular level and then design drugs that target the molecular action (Henderson 1994). However, trial-and-error can require extensive resources and does not guarantee a solution, while uncovering underlying causal principles is challenging and not always possible.

Instead of building knowledge from the ground up through trial-and-error or theory development, I focus on how innovators can take a "horizontal" shortcut by importing intuition and insights from a neighboring field. With the case of drug development, in lieu of brute-force screening or rational drug design, drugmakers can rely on pattern recognition by identifying drug candidates for a new disease based on already approved drugs for similar diseases.

This strength of analogical reasoning in helping innovators quickly make progress in new domains has been highlighted in prior work. Psychology studies have shown that scientists frequently substitute slow, iterative experimentation with analogical reasoning to speed up problem solving (Dunbar 2000). In the context of business strategy, using a simulation and a rich set of case studies, Gavetti, Levinthal, and Rivkin (2005) demonstrate that managers often face strategic problems that are best suited for analogical reasoning: problems that are neither too modular (where rational, deductive reasoning can instead be employed) nor too complex (where only trial-and-error can work).[3]

(ii) Analogies are not simple recombinations. Analogical reasoning's reliance on pattern recognition distinguishes analogies from the traditional concept of recombinant innovation (Schumpeter 1934; Weitzman 1998; Uzzi et al. 2013). The key insight in the recombinant literature is that new ideas can be generated from existing, well-understood ideas if they are combined in a novel way.[4] In contrast, rather than mixing well-understood ideas, analogical reasoning involves the borrowing of less-understood patterns from another domain. In the bird-airplane analogy, for example, inventors of flight did not understand the physics of aerodynamics and therefore did not know exactly how birds can fly. But these early inventors speculated that the motion of wings is important and applied this pattern to human-powered flight. In other words, analogical reasoning is the importing of relational patterns—correlations—not causal logic.

(iii) As shown in Figure 1, analogical reasoning has evolved from solely being conceptual in the mind of an individual to being automated and outsourced to

[3] In economics, Gilboa, Samuelson, and Schmeidler (2015) develop a formal model of reasoning, in which economic agents use analogical reasoning when the underlying data generating process is unknown and use rule-based reasoning when the data structure is known.

[4] For instance, Brynjolfsson and McAfee (2014) cite Waze, the navigation app that uses real-time, crowdsourced traffic data, as a classic example of recombination. The individual components of Waze (location sensors, social networks, and smartphones) were all widely known and used, but no one had thought to combine them together to optimize driving routes before Waze.

machines. Conceptual analogies have been fundamental throughout the history of innovation by helping individuals understand a new domain through identifying patterns across domains.[5] The modern field of biomimetics is a prime instance of conceptual analogies as shortcuts; engineers often take inspiration from properties found in the natural world and reverse-engineer them, instead of beginning with first principles or replicating the millennia of trial-and-error experiments that nature conducted through natural selection (Pollack 2014). In managerial practice, conceptual analogical reasoning has also served as critical, if underappreciated, sources of strategic visions and entrepreneurial ventures (Hill and Levenhagen 1995; Gavetti and Rivkin 2005; Gavetti, Levinthal, and Rivkin 2005; Kaplan and Orlikowski 2013; Martins, Rindova, and Greenbaum 2015; Glaser, Fiss, and Kennedy 2016).[6]

With the rise of data-reliant, pattern-recognition algorithms, conceptual analogical reasoning has become increasingly automated. This has several consequences. First, this automation has allowed innovators to apply the borrowed patterns from another domain, without necessarily understanding them. For example, when relying on machine learning software libraries like TensorFlow, knowledge workers often do not know why the algorithm made the prediction it did. Despite not necessarily understanding the correlations that connect the training and test domains, knowledge workers can simply apply those patterns with TensorFlow. Second, unlike conceptual analogical reasoning which is unconstrained in the types of templates it requires, automated analogical reasoning needs specific templates: digitized training data. I discuss the implications below in Section 2.2.2.

[5] Hesse (1966) and Holyoak and Thagard (1996) provide numerous examples. One of the first accounts of (conceptual) analogical reasoning was by the ancient Roman architect Vitruvius who proposed the wave-sound model. Since then, analogies have served as the genesis behind many discoveries, from Charles Darwin whose theory of natural selection was based on an analogy to artificial selection by farm breeders to the Nobel laureate Salvador Luria's analogy between slot machines and bacterial mutations. Analogies can even be found in mathematics, a field built on causal logic and thus may seem less amenable to analogical reasoning. In fact, many difficult problems in algebra have been solved by turning them into geometric problems—that is, by finding analogies between algebra and geometry (Hacking 2014).

[6] In addition to these studies on how analogical reasoning can serve as a source of innovation (i.e., create new strategies and ventures), another strand of management literature underscores a different aspect of analogies: a dissemination tool once an innovation has been produced. When introducing novel products or services, analogies can be used to increase legitimacy (Hargadon and Douglas 2001; Bingham and Kahl 2012; Etzion and Ferraro 2010; Cornelissen and Clarke 2010). Apple, for instance, popularized the term "desktop" when launching personal computers. By analogizing between the physical and the digital desks, Apple intuitively drew in customers who were not used to working in the virtual world (Bingham and Kahl 2012).

2.2.2 Tradeoffs in Relying on Analogies

While analogies—and by extension, analogy-based technologies like machine learning—can help knowledge workers innovate in domains where there are no existing ideas yet to build on, analogies can also have several costs. First, because analogies require the availability of templates, they may constrain the direction of innovation towards areas of research with templates, even if those areas are less fruitful. Second, these templates may highlight just superficial similarities between the target and adjacent domains, leading to misleading conclusions (Gentner 1982). Third, because analogies do not build foundational knowledge from scratch, this leads to a weak foundation: knowledge workers may not fully understand the underlying mechanisms of how the target domain works; they can only translate the target in terms of the template.[7]

This paper focuses on the first cost—that analogies may constrain innovation towards areas with templates. This cost of analogies demonstrate an inherent limitation of analogies in balancing exploration and exploitation when knowledge workers search through the idea landscape (Kauffman 1993; Levinthal 1997; Nelson and Winter 1982; Fleming and Sorenson 2004; Kaplan and Vakili 2015). For example, Kneeland, Schilling, and Aharonson (2020) propose that inventors often make "long jumps" across disparate domains to pioneer an uncharted domain; analogical reasoning can be one mechanism that inventors rely on to make such jumps. At the same time, however, its very need for templates from which to import patterns may result in anchoring that leads analogical reasoning astray and bound to local search (Holyoak and Thagard 1996; Gavetti and Rivkin 2005).

The growing automation of analogical reasoning may exacerbate this fixation cost. Although hailed as a tool of exploration (Agrawal, Gans, and Goldfarb 2018; Agrawal, McHale, and Oettl 2018), supervised machine learning has one critical weakness: it cannot work without digitized training data. While the prior literature on AI has focused on algorithmic bias that arises from poor-quality training data (Cowgill and Tucker 2020; Cowgill et al. 2020; Choudhury, Starr, and Agarwal 2020)—for example, how unrepresentative training data can lead to superficial analogies that identify misleading patterns—there has been less attention on how the very availability of training data can restrict the direction of innovation (Cockburn, Henderson, and Stern 2018;

[7] The importing of correlations via analogies brings to the forefront an aspect of knowledge production that is sometimes overlooked: knowledge domains can vary in how "strong" or "shaky" their foundation can be. Prior studies have noted how retractions (Azoulay et al. 2015), institutions (Furman and Stern 2011), or certification (Greenblatt 2021) can impact the strength of foundational knowledge in a given domain. The use of analogical reasoning can be another reason that leads to a weaker foundation due to the accumulation of correlations (in lieu of causal theories).

Hoelzemann et al. 2022). Given the growing importance of analogy-based technologies, understanding this cost is important as the locus of digitized data can have long-term implications for what gets innovated.

3. Empirical Setting

An ideal empirical setting needs three ingredients: (i) an observable idea landscape, where I can track which ideas get explored versus unexplored, as well as a measure of distance between ideas, (ii) the arrival of an analogy-based technology, and (iii) specifically, the differential arrival of the technology, such that it only arrives in some parts (treated) but not other parts of the setting (control). In this section, I first introduce the setting of structural biology and its scientific importance. I then describe the empirical features of structural biology that make it an ideal setting to study analogies.

3.1 Structural Biology: The Study of Proteins

Structural biology is a field that studies the 3D structures of proteins and aims to uncover the functional roles of proteins by elucidating their structures. As Francis Crick (who had discovered the helical structure of DNA) remarked, “If you want to understand function, study structure” (Crick 1990). Since proteins are responsible for carrying out most functions in cells, insights from structural biology has helped with a broad range of applications, from identifying targets for new drugs to understanding disease progression. As one evidence of its wide-reaching impact, structural biology has been recognized with more than a dozen Nobel prizes.

Structural biology has also played an important role in the current fight against the coronavirus pandemic. As shown in Figure 2A, researchers solved the structure of the spike proteins that stud the surface of SARS-CoV-2—that is, determined the 3D coordinates of individual atoms in the protein. Through this direct visualization, researchers learned how these proteins latch onto receptors on human cells like “a key to a lock” (Patel, Lucet, and Roy 2020), enabling the development of vaccines that are designed to block these proteins.

3.2 Structural Biology as an Empirical Setting

3.2.1 Observable Idea Landscape

In order to investigate whether an analogy-based technology changes the direction of innovation, it is important to be able to observe the entire idea landscape. In most settings, however, researchers can only observe ideas that were realized, while alternative ideas that could have been pursued (but were not) remain invisible. For example, in scientific research, not all papers that can be written ultimately get written. Only papers that actually became published can be observed, while other papers that may have been in consideration (but were not written) cannot be observed. Structural biology, however, provides a unique window into the idea landscape. As I explain in Section 4, I leverage a database of all known proteins, and I can observe which proteins structural biologists chose to explore versus could have explored but neglected.

Furthermore, in most settings, it is difficult to quantify the similarity between each potential idea. For instance, in the case of scientific publishing, measuring the distance between each paper is challenging; text similarity is often used, but this is an imperfect metric for measuring intellectual distance. In contrast, structural biology provides an objective measure: proteins are composed of sequences of amino acids—which are given by nature—and therefore they can be grouped based on their sequence similarity (Hill and Stein 2020; 2021).[8] Since analogies work by identifying similarities, this measure of distance enables me to track the use of analogies.

3.2.2. Arrival of an Analogy-Based Technology

Since proteins are too small to be seen through an optical microscope, structural biologists had to develop various experimental techniques to reveal the atomic structure of proteins—or "solve" the structure. Solving a protein structure involves deep knowledge of biology, physics, and statistics, and this used to be—and remains—extremely challenging. A complex structure could take months, even years, to solve. For instance, determining the structure of the ribosome (a macromolecular machine responsible for translating DNA code to produce proteins) took over two decades, culminating in the 2009 Nobel Prize in Chemistry (Ramakrishnan 2018).

[8] Hill and Stein (2020; 2021) study the effect of competition in structural biology and cluster structures based on their sequence similarity to identify scientists engaged in "priority races" (i.e., competing teams that worked on structures in the same cluster, unbeknownst to each other). In this paper, rather than focusing on just proteins whose structures have been characterized, I look at instead the entire universe of proteins—both structurally characterized and uncharacterized—and cluster this universe of proteins based on their sequence similarity.

The dominant method of solving a structure is called X-ray crystallography,[9] which proceeds in three main steps (Figure 2B). First, the protein sample must be produced in a very specific way, which is to crystallize it—packing multiple copies of the protein in a well-ordered crystal lattice. Second, once the crystal is obtained, X-ray beams are shot at the crystal, which produces diffraction patterns, as electrons in the crystal diffract the X-ray. Third, using a combination of physical laws, statistics, and intuition, structural biologists construct a density map of electrons from the diffraction patterns and build up a 3D atomic model of the protein structure. This paper focuses on this third step of interpreting the diffraction data. Unlike the days of Max Perutz (co-winner of the 1962 Nobel Prize in Chemistry) who solved the first protein structure (hemoglobin) through painstaking hand-calculations, many of the steps of interpreting the diffraction data have become automated.

The specific technology I examine is a software program called Phaser. Phaser was released in September 2003, which automates a method called molecular replacement, or MR. Figure 3 shows the rise in the number of structures solved by MR at the Protein Data Bank, a global repository of all solved structures.[10] One of the biggest challenges in interpreting the diffraction data is called the "phase problem," a problem difficult enough that one method to solve it resulted in a Nobel prize.[11] Prior to MR, structural biologists had to resort to time-consuming experimental methods to solve the phase problem, but MR allowed structural biologists to bypass experimental phasing. Instead of solving the phase problem from scratch, MR uses other previously solved structures that share similar sequence similarity to the unknown structure and use them as templates to solve the phase problem of the unknown structure. One structural biologist I interviewed noted that MR could be up to 100 times faster than experimental phasing methods,[12] potentially saving a few years of work.

[9] In addition to X-ray crystallography, two other methods can be used to solve a structure: nuclear magnetic resonance spectroscopy and cryo-EM. However, crystallography is by far the dominant method used, with over 95% of all protein structures solved using this method.

[10] The method of MR was first proposed in 1962, but MR was not put into wide practice until decades later due to lack of available structures, as well as lack of ready-made software programs (Doerr 2014). While Phaser was not the first software program to implement MR, it is the most user-friendly, automated, efficient, and widely-used program (Scapin 2013).

[11] X-ray reflections have both amplitudes and phases, but the phase cannot be measured from the diffraction patterns. Without knowing the phase, a model of the protein structure cannot be constructed.

[12] There are two experimental methods that solve the phase problem from scratch. The first method is called isomorphous replacement, which involves producing a "native" target crystal and a "derivative" crystal with a heavy metal ion introduced. By measuring the difference in diffraction patterns between the native and the derivative crystals, structural biologists can recover the phase. The second method is called anomalous dispersion, where structural biologists vary the X-ray wavelength to induce atoms of specific elements to produce anomalous scattering. By locating these anomalous scattering atoms, the missing phase

In other words, MR can be viewed as an analogy-based technology. The key insight behind MR is that sequence similarity has been observed to be highly correlated with structural similarity (although little is understood regarding the causal mechanism of why sequences of amino acids cause proteins to fold into their particular 3D shapes). By taking advantage of this pattern, structural biologists use MR to import phase information from neighboring proteins that share sequence similarity, rather than solving the phase problem de novo.[13]

3.2.3 Differential Arrival of an Analogy-Based Technology

Finally, this analogy-based technology of MR arrived in some parts of structural biology but not others. As mentioned earlier, I observe the entire map of known proteins and the distance between each protein in terms of their sequence similarity. While some clusters of proteins received attention from structural biologists before the arrival of MR (which I term "bright" clusters of proteins), other clusters of proteins did not get any attention ("dark" clusters). Since MR needs data on previously solved structures, MR can be applied in bright clusters but is not useful for dark clusters, so bright and dark clusters serve as my treatment and control groups, respectively. This paves the way for a difference-in-differences design, as described in Section 5.

4. Data

In order to construct my map of proteins, I use two main datasets: UniProtKB/Swiss-Prot, a database of all known proteins, and the Protein Data Bank, a database of all protein structures. I then cluster proteins based on their sequence similarity to construct my final sample.

4.1 UniProt Knowledgebase/Swiss-Prot

The Universal Protein Resource Knowledgebase (UniProtKB) is a comprehensive database of proteins. A protein is composed of sequence of organic compounds called amino acids. Information for making a protein is stored in a gene's DNA, and by translating the DNA sequence

of the rest of the protein can be backed out. While isomorphous replacement and anomalous dispersion do not rely on the availability of prior solved structures, they can require arduous experimental efforts.

[13] In November 2020, a technology that supersedes MR was introduced: the AI program AlphaFold, created by Google's DeepMind team. AlphaFold can predict the structure of a protein based on purely its sequence of amino acids. While MR helps with specifically the phase problem of experimental structure solving, AlphaFold bypasses the need to conduct experiments at all. While AlphaFold's success falls outside of the time period studied in this paper, I discuss potential implications in Section 7.

of a gene, scientists can determine the protein's existence and the sequence of amino acids that will appear in the protein. Protein sequences in UniProtKB are thus sourced by translating genes from major genome sequence databases.

To define the complete set of proteins at risk of being structurally characterized, I focus on the Swiss-Prot section of UniProtKB.[14] Created in 1986, the Swiss-Prot database is extensively reviewed, maintained, and annotated by experts based on experimental results and literature review. I also use data constructed by Perdigão et al. (2015), which provides additional characteristics on each protein in Swiss-Prot. As of October 2020, Swiss-Prot contains 563,552 protein entries.

4.2 Protein Data Bank

Established in 1971, the Protein Data Bank (PDB) is a repository of protein structures and contains over 170,000 structures as of October 2020. Since 1989, most journals have required authors to deposit their structures at the PDB as a requirement for publication, and therefore the PDB contains the universe of all publicly available structures. The PDB provides detailed descriptions about each structure, as well as crosswalks to Swiss-Prot.

4.3 Sample Construction: Clustering Proteins

After identifying which proteins in Swiss-Prot were found to be structurally characterized in the PDB, the final step is to measure the distance between each protein and cluster proteins that share sequence similarity.

I rely on MMseqs2,[15] an algorithm used by both Swiss-Prot and the PDB to cluster similar proteins (Steinegger and Söding 2018; Hauser, Steinegger, and Söding 2016). Given that molecular replacement will likely be successful if the template and the target proteins share at least 30% sequence identity (Schmidberger et al. 2010; Phenix), I chose a threshold of 30% sequence identity to group all proteins in Swiss-Prot into mutually exclusive clusters. I then restricted the sample to clusters with at least one human protein and clusters that had at least one protein discovered

[14] In addition to Swiss-Prot, UniProtKB has a database called TrEMBL, which is larger but contains computationally annotated proteins whose existence are largely not proven. More details are provided in the Data Appendix.

[15] MMseqs2 can be downloaded from https://github.com/soedinglab/MMseqs2. More details on MMseqs2 are provided in the Data Appendix.

by 1998, the year before my panel begins. More details on sample construction can be found in the Data Appendix.

5. Empirical Strategy

5.1 Main Specification

As described in Section 3.2, since MR relies on having similar, previously solved structures as templates, MR only works on clusters of proteins with previously solved structures (i.e., bright clusters) and does not work for clusters of proteins that have not yet been structurally characterized (i.e., dark clusters).

This enables me to employ a difference-in-differences approach and estimate the following regression equation to examine the impact of MR:

$$Y_{ct} = \beta_0 + \beta_1 PostMR_t \times Bright_c + \delta_t + \gamma_c + \varepsilon_{ct} \quad (1)$$

Y_{ct} is the total number of structures that gets solved in cluster *c* in year *t. $PostMR_t$* is an indicator variable that turns one after the arrival of MR in 2003, and $Bright_c$ is an indicator variable for bright clusters, defined as whether the cluster had at least one structure by 1998.[16] δ_t are calendar-year fixed effects, and γ_c are cluster fixed effects. β_1 is the coefficient of interest and can be interpreted as the impact of MR on the number of solved structures. Standard errors are clustered at the cluster level.

In order for the coefficient β_1 to capture the causal impact of MR, parallel trends assumption must hold: in the absence of MR, trends in outcomes between bright and dark clusters must have been the same, conditional on cluster fixed effects and year fixed effects (as well as time-varying controls that I use in some specifications). I discuss this concern in detail in Sections 6.1 and 6.4.

[16] The treatment variable, $Bright_c$, is defined as whether the cluster had a structure by 1998 (the year before my panel begins) instead of 2003 (when MR arrived). If $Bright_c$ is defined using the year 2003, then the treatment is mechanically correlated with the outcome variable (the number of structures being solved each year) in the pre-period from 1999 to 2003 since the treatment is a lagged outcome of the pre-period. The panel was chosen to begin in 1999 because this is (i) early enough to yield at least five years of pre-period before the introduction of MR, but (ii) late enough that there has been some accumulation of prior solved structures in the PDB (6% of structures that will eventually be deposited at the PDB by 2019 had accumulated by 1998).

5.2 Descriptive Statistics

As shown in Table 1, my sample consists of 6,944 clusters, with 9% of the clusters classified as bright. Not surprisingly, in terms of levels, bright and dark clusters differ on several characteristics when MR arrived. First, bright clusters are on average older and bigger. Second, a higher share of the proteins in the bright clusters have characteristics that make them more amenable to crystallization (as discussed in Section 3.2, in order for a protein's structure to be solved, the protein must be first crystallized). Bright clusters contain proteins that are less likely to be membrane, disordered, or have compositional bias.[17] Proteins in the bright clusters are also on average shorter in sequence length. Third, while bright clusters had more publications and drugs related to their proteins, dark clusters are higher in one measure of biological significance: the share of human proteins in the cluster (human proteins are of high interest to drug developers). This reassuringly suggests that dark clusters are not devoid of biological importance.

While these differences in levels do not threaten my difference-in-difference strategy (as long as there are no differences in trends in outcomes in the pre-period), in Section 6.4, I revisit these characteristics in a robustness analysis to develop a "predicted brightness" measure, where I control for pre-period traits related to crystallization feasibility and biological significance.

6. Main Results

6.1 Impact of MR on the Number of Solved Structures

I begin by examining how MR impacted the number of solved structures. As shown in Table 2, bright clusters got brighter (i.e., received more structures) after MR, relative to dark clusters. The outcome is the total number of solved structures in a cluster each year. Columns 1-2 report the outcome after Log(+1) transformation,[18] while Columns 3-4 report the results in levels (scaled

[17] Membrane proteins are proteins that are found in (or interact with) cell membranes; these proteins tend to be flexible and partially hydrophobic, which make crystallization challenging. Proteins with intrinsically disordered regions (i.e., regions that do not adopt a well-defined structure) or extreme sequence length (very short or long) can also impede crystallization (Slabinski et al. 2007). Finally, compositional bias refers to whether the protein contains regions with overrepresented subsets of amino acids. Proteins are typically composed of twenty amino acids, but not all amino acids may show up equally. For example, QHQQQGQHHQHHHQQQQHH has a bias for the amino acids Q (glutamine) and H (histidine) (Harrison 2017). Compositional bias is associated with decreased crystallization potential.

[18] In Appendix Table 1, I provide a robustness analysis using inverse hyperbolic sine transformation. Results remain similar.

by the standard deviation). As reported in Column 1, bright clusters experienced a 7% increase in the number of solved structures after the arrival of MR, relative to dark clusters. Results in levels also indicate that bright clusters got brighter. Bright clusters received an increase of 0.744 annual number of structures after the arrival of MR, which translates to a 30.1% increase relative to the baseline standard deviation of 2.47.

I conduct several analyses to ensure that these results are being driven by MR. First, one concern is that bright clusters may be getting more structures not necessarily due to MR but because it is also getting increasingly bigger (i.e., more protein sequences are being discovered) relative to dark clusters. In Columns 2 and 4, I additionally control for time-varying cluster size while estimating Equation 1; the magnitude of the impact of MR remains similar and significant.

Second, the impact of MR should only show up in structures that were actually solved by MR. If I observe that bright clusters experienced an increase in both structures that were solved by MR and non-MR methods, this raises the concern that factors unrelated to MR may be causing bright clusters to get brighter. In Appendix Table 2, I confirm that MR only impacts structures that were solved by MR and does not impact structures that were not solved by MR.

Third, since MR needs just one previously solved structure in order to work, the impact of MR should be stronger when comparing dark clusters versus bright clusters with a single previously solved structure, and weaker when comparing bright clusters with a single structure versus bright clusters with multiple structures. Appendix Table 3 shows this exact result. I split the bright clusters into whether they had just a single or multiple previously solved structures. I then compare the impact of MR, comparing dark versus bright clusters with just a single structure (Column 2) and comparing bright clusters with just a single structure versus multiple structures (Column 3). The impact of MR is stronger in Column 2 relative to Column 3.

Fourth, and most importantly, to asses pre-period trends, I show an event studies version of Equation 1, replacing the single $PostMR_t$ indicator with indicators for every year before and after the introduction of MR. Figure 4 plots the dynamic effects of MR on the number of solved structures. Reassuringly, in both Panels A (Log(+1) transformation) and B (levels), there appears to be no difference in trends between bright and dark clusters in the number of solved structures before MR. Moreover, the impact of MR is sustained over the entire sample period: bright clusters got brighter and brighter.

6.2 Impact of MR on the Quality of Solved Structures

MR deceased the cost of solving structures in well-explored, bright clusters, and, not surprisingly, increased the volume of structures in those areas. But what about quality? In the next set of results, I investigate how MR impacted the quality of solved structures.

I distinguish between two dimensions of quality: execution (how meticulously a project was completed) versus importance (whether a project led to a novel insight). While I provide below measures of execution and importance in the specific setting of structural biology, these are general dimensions of quality that apply to any innovative activity: the innovation should be well-executed, but it should also solve an important problem.

The impact of analogical reasoning on quality is not immediately obvious. On one hand, analogical reasoning has the power to make "long jumps" (Kauffman 1993; Kneeland, Schilling, and Aharonson 2020) to explore a novel domain and discover creative opportunities. But it may be more challenging to rigorously execute innovations stemming from analogies because analogies are rooted in correlations, not precise causal logic. On the other hand, one of the pitfalls of analogies is that their need for templates could cause fixation and steer the direction of innovation towards areas with templates, even if they are less fruitful. This may be particularly true when analogical reasoning becomes automated, as analogy-based technologies like MR and supervised machine learning require specifically digitized training data of past successful innovations. Analogies may thus lead to "short jumps," landing on unexplored but near potentially crowded areas where it may be harder to unearth new insights.

6.2.1 Execution

The first dimension of quality I examine is execution, and I take advantage of measures provided in the PDB called the R-free and resolution.[19] These are objective metrics used by the structural biology community to assess the technical execution—specifically, the accuracy and precision—of the structures (Kleywegt and Jones 1997).

The R-free refers to accuracy or goodness-of-fit: how well the model of the protein structure fits the observed experimental data. As discussed in Section 3.2, structural biologists build the atomic model of their protein structure from experimentally observed diffraction data. They then simulate diffraction patterns based on the model and compare the simulated diffractions to the

[19] Hill and Stein (2021) use the R-free and resolution as their main quality measures in their study of how competition affects the quality of scientific research. I interpret the R-free and resolution as indicating specifically the execution level of the structure.

experimentally observed patterns. The R-free can be improved as researchers undergo iterative refinement process of their model to better fit the experimental data.

Resolution refers to precision or the level of detail that can be found in the structure. Figure 2C shows an example of a protein (tyrosine 103 from myoglobin) at different resolutions, from a poor resolution where only general contours are visible to a resolution where individual atoms can be plotted. Resolution depends on the degree of order in the crystallized protein. Researchers can improve the resolution by obtaining high-ordered crystals (proteins that are packed and aligned identically in the crystal), which produce diffraction patterns with fine details.

With these measures, I investigate the impact of MR on execution in Table 3. Column 1 reports the same result as Column 1 of Table 2 and shows the impact of MR on the total number of structures solved in a cluster. Columns 2-4 decompose this result into terciles based on the structure's R-free values and investigate the impact of MR on the number of structures in the bottom tercile (Column 2), middle tercile (Column 3), and top tercile (Column 4) of R-free values. Columns 5-7 similarly report results using the resolution of the structures.

A clear pattern emerges: bright clusters especially received more structures that were well-executed. For the R-free, there was no difference between bright and dark clusters in the number of structures that were solved in the bottom tercile. In contrast, bright clusters received 14% more structures that were solved in the top tercile, relative to dark clusters. Likewise, for resolution, bright clustered received just 3% more structures from the bottom tercile, but 10% more structures from the top tercile.

6.2.2 Importance

The second dimension of quality is the scientific importance of the structure: did the structure lead to a novel insight about a biological process? When the PDB was established in 1971 with only seven structures in its database, every new structure provided valuable information. However, as the PDB grew, it became no longer enough to just solve structures for the sake of solving them. As early as 1994, the editors of *Nature Structural Biology* advocated in their inaugural issue, "[T]he static image of the molecule is rarely an end in itself, but rather a beginning of comprehension" (Nature Structural Biology 1994). Through additional biochemistry or cell biology experiments, structural biologists try to explicitly link a protein's function to its potential function to understand the role the protein plays in various biological processes (Cassiday 2014). To evaluate whether a structure led to a new biological understanding, I present below three measures.

First, did the structure lead to a publication? Structures that were simply deposited at the PDB without a corresponding publication to explain their biological significance are structures that contributed very little, if at all, to revealing biological insights. As a prominent researcher at Yale once declared, "The fact is that protein structures come alive intellectually only when they are connected with [other data] indicating what they do" (Moore 2007). There could be several reasons why some structures do not have accompanying publications. One structural biologist I interviewed noted that a "stamp collection" of structures are sometimes needed to win grants from funding agencies. Some of these structures are also from structural genomics consortiums, whose goals are to catalogue as many types of structures as possible without necessarily explicating them (Petsko 2007; Hill and Stein 2021).

Second, was the structure cited by Swiss-Prot? The Swiss-Prot database contains extensive annotations about a protein's function and provides references behind each functional annotation. Importantly, the references are added manually by experts who follow well-defined curation protocols, undergo quality checks, and are updated as new data becomes available, ensuring that selection of these references are impartial.

Finally, I measure whether the structure got cited by a patent, with the assumption that the protein must have led to enough functional insights in order for an inventor to develop commercial applications. I leverage data from Marx and Fuegi (2020) on patent citations to scientific articles to identify structures with papers that were cited by at least one patent.

Bright clusters especially received more structures that did *not* reveal functional insights. As shown in Table 4, bright clusters received 9% more *unpublished* structures, which have no accompanying articles that describe their function (Column 2). In contrast, bright clusters experienced a relative decrease in the number of structures that were cited by a patent (Column 4), and there were no differences between bright and dark clusters in the number of structures that were cited by Swiss-Prot (Column 6)—which are the set of structures that are the most likely to have yielded functional insights.

6.3 How Did the Scientific Community Receive MR?

How did the scientific community receive this shift in research direction as a result of MR? Did the scientific community value the fact that MR led to more structures that are well-executed? Or did the community find these types of structures less valuable since they ultimately did not lead to new functional insights?

To examine this question, I use standard measures of publication impact (citations and journal impact factor).[20] A natural question is why I interpret these publication measures as different from my measures on functional insights. My earlier measures (whether the structure has an accompanying publication and whether it was cited by Swiss-Prot or a patent) assess a specific fact: did the structure lead to some kind of function insight? In contrast, journal impact factor and publication citation serve as proxies for how the scientific community typically rewards a piece of research and can be due to either the technical execution or functional insights of the research, which is hard to disentangle. For instance, it is unclear whether a structural biology paper was published in a prestigious journal because the structure was technically well-executed or because it led to a new biological understanding (or a combination).

The idea behind this analysis is that if the scientific community valued execution more, they would have rewarded the well-executed structures in bright clusters by publishing them in prestigious journals and highly citing them. In contrast, if the scientific community valued functional insights more, they would have punished the structures in bright clusters by publishing them in less prestigious journals and citing them less.

In Table 5, I investigate the effect of MR on the publication impact of the structures, in terms of the mean number citations the structure's publication received and the journal impact factor (i.e., journal prestige). In Columns 3-5, I decomposed the total number of solved structures in a cluster into terciles based on citations, while in Columns 8-10, I decomposed the total number of structures into terciles based on the journal impact factor.[21]

Bright clusters especially received more structures with *less* publication impact. After the arrival of MR, bright clusters received approximately 7% more structures that were either unpublished or published with very few citations, relative to dark clusters. In contrast, bright clusters had a 2% decline in the number of structures with the highest number of citations. In terms of journal impact factor, bright clusters received 7% more structures that were published in the least prestigious journals, relative to dark clusters. However, there was no difference between bright and dark clusters in the number of solved structures that were published in the most prestigious journals. This suggests that the scientific community appears to believe that there has been a decline in the quality of research conducted as a result of a MR.

[20] I linked the primary paper associated with each structure in the PDB to the PubMed data and obtained citation data from the Web of Science (specifically, the mean annual number of citations received by each structure, within the first five years of paper publication).

[21] Due to data availability, I restricted the panel to end in 2012 for the citation analyses and 2017 for the journal impact factor analyses.

Taken together, these results indicate that bright clusters received more structures overall and particularly structures that were well-executed. However, these structures tended to not lead to new functional insights or have a high publication impact.

6.4 Predicting Brightness

6.4.1 Specification with Predicted Brightness

The identification underpinning my difference-in-differences framework hinges on parallel trends assumption. While there was no evidence of pre-trends in the event studies as well as in other robustness analyses, there may still be concerns over whether bright and dark clusters were evolving on different trends for factors unrelated to the introduction of MR. For example, a particular concern is that proteins in the dark clusters cannot be crystallized and thus cannot be structurally characterized to begin with.

This concern is mitigated by the evidence in the bioinformatics literature, where it has been documented that while there are certain traits (e.g., membrane or disordered proteins) that indeed make a protein challenging to crystallize, the crystallization process remains an unpredictable art, rather than a science. For instance, Perdigão et al. (2015) surveyed the Swiss-Prot data to understand the features of dark proteins; they find that most of the dark proteins cannot be explained by the "usual suspects," that a majority of the dark proteins are in fact not membrane or disordered proteins. Other studies also point to the difficulty in predicting which proteins will crystallize (Elbasir et al. 2019; Terwilliger, Stuart, and Yokoyama 2009).

Since there are still some characteristics that are known to confound crystallization, I develop a predicted brightness measure, *Predicted_Bright$_c$*, where I measure whether a cluster was predicted to be bright in 1998,[22] using the pre-period characteristics of the proteins in the cluster. I then modify Equation 1 to estimate the following:

$$Y_{ct} = \beta_0 + \beta_1 PostMR_t \times Bright_c + \beta_2 PostMR_t \times Predicted\_Bright_c + \delta_t + \gamma_c + \varepsilon_{ct} \qquad (2)$$

The thought experiment in Equation 2 is that I compare clusters of proteins that are similarly predicted to have their structures characterized by 1998 because they are ex-ante similar in traits related to biological importance and crystallization feasibility, but some clusters just happened to

[22] Recall that (actually) bright clusters are defined as whether they had a structure by 1998.

actually have characterized structures before the arrival of MR while other clusters did not. If I observe that only being predicted bright has an impact on the number of structures solved after MR (i.e., β_1 is non-significant but β_2 is significant), then there would be concern that unobserved characteristics are driving bright clusters to both have had their structures characterized in 1998 and subsequent structure characterization after MR. However, if there is an added effect of being actually bright in addition to being predicted bright (i.e., β_1 is significant), then this reduces the concern of omitted variable bias.

6.4.2 Constructing Predicted Brightness

I first restrict the sample to proteins that were discovered by 1998 (n = 41,781 proteins), and, for each protein, I focus on several sets of characteristics as of 1998. First, I have characteristics on how hard it is to crystallize the protein: whether the protein is a membrane protein, disordered protein, has compositional bias, and has long sequence length. Second, I have characteristics on the biological importance of the proteins: which species the protein is from (2,814 indictors), the number of publications written about the protein (10 indicators), and the number of approved drugs that target the protein (8 indicators).[23] Finally, I have indicators for the year of when the protein was discovered (29 indicators). After dropping collinear variables, this translates to a total of 817 predictors.

To avoid overfitting and increase prediction performance, I use Lasso to predict whether a protein is predicted to be bright (i.e., structurally characterized by 1998). Appendix Figure 1A shows the receiver operating characteristic (ROC) curve of the resulting prediction. The ROC curve plots the True Positive Rate (what share of actually bright proteins were correctly predicted to be bright?) against the False Positive Rate (what share of actually dark proteins were incorrectly predicted to be bright?). The area under the curve (AUC) of the ROC curve evaluates the performance of the prediction and can be interpreted as the probability that a random actually bright protein will have a higher predicted brightness than a random actually dark protein. The AUC can range from 0 to 1, and a general rule of thumb considers an AUC above 0.8 to be indicating high performance; the AUC of my prediction exercise is 0.91.

[23] Information on drugs is provided by DrugBank. This dataset provides comprehensive information on drugs at various development phases and their targets (i.e., proteins.) and is freely available for academic use. A limitation of the free version of the data is that it only provides marketing dates for approved drugs, and there are no dates on when a drug entered pre-clinical or clinical trial phases.

I then use the fitted values to predict each protein's brightness in 1998. Appendix Figure 1B shows the distribution of this predicted brightness, by whether the protein was actually bright by 1998. There is variation and overlap in the distributions of predicted brightness between actually bright and dark proteins, suggesting that while there are some characteristics that may make some proteins more likely to be structurally determined, some proteins just happened to have structures actually determined by 1998, while other proteins with similarly predicted brightness did not. This supports the findings in the bioinformatics literature which also notes that many of the dark proteins cannot be explained by the usual factors that defy crystallization and that it is in fact difficult to predict crystallization. From this protein-level prediction, I aggregate up to the cluster-level by taking the sum of the predicted brightness of all proteins in each cluster to construct $\textit{Predicted\_Bright}_c$.

6.4.3 Results with Predicted Brightness

I then estimate Equation 2 that additionally controls for $\textit{Post-MR}_t \times \textit{Predicted\_Bright}_c$, modifying my baseline difference-in-differences framework. Appendix Table 4 shows the results from estimating Equation 2. The coefficients on $\textit{Post-MR}_t \times \textit{Bright}_c$ remain positive and significant; among clusters predicted to be similarly bright, there is still an effect of being actually bright. While it may seem surprising that the coefficients on $\textit{Post-MR}_t \times \textit{Predicted\_Bright}_c$ are not significant, MR only works when there are actually solved prior structures in the cluster and should not work if the cluster is only predicted to be bright. Appendix Table 4 therefore supports the evidence that the arrival of MR indeed caused the number of solved structures in (actually) bright clusters to increase, relative to dark clusters, even among clusters that were ex-ante similarly predicted to be bright due to their traits related to biological importance or crystallization feasibility.

7. Discussion & Conclusion

This paper provides, to my knowledge, the first empirical study of how the automation of analogical reasoning may shape the direction of knowledge production. Using the setting of structural biology, which provides a unique window into the entire idea landscape, I study the introduction of an analogy-based technology, MR, which solves protein structures by relying on data of prior structures.

I find that MR increased the number of solved structures, specifically in bright clusters with already solved structure templates. This result from structural biology highlights the power of analogies in reducing the cost of producing innovation in certain areas, thereby shifting the direction of innovation. In particular, rather than building knowledge from scratch, analogies can provide shortcuts that enable knowledge workers to innovate in domains where there is no existing knowledge yet (e.g., structurally uncharacterized proteins) by importing knowledge from neighboring domains (e.g., borrowing structure information from similar, already solved proteins). Yet, because of their very need for templates, analogies could restrict knowledge workers into focusing on just domains with neighbors (e.g., bright, already well-explored clusters).

One may argue that this shift in research direction due to analogies does not necessarily imply a decline in the quality of innovation, as analogies allow knowledge workers to quickly make progress in previously unexplored domains. An important question then is what quality of innovation is ultimately produced. My results suggest that structural biologists used the extra time gained from taking a shortcut with MR to improve the technical execution—the resolution and the goodness-of-fit—of the structures. These structures, however, had low scientific importance and publication impact. That is, at least in this specific setting, knowledge workers appear to use the imported knowledge from analogies to focus on incremental execution, rather than attempting to discover a fundamental insight.

Understanding this tradeoff of analogies can have crucial implications for the management of knowledge production, given the growing automation of analogical reasoning and, more broadly, data-driven exploration. A recent work by Hoelzemann et al. (2022) is close in spirit to this paper: using a laboratory experiment, the authors document the "streetlight" effect of data,[24] that when data reveals a satisfactory—but not the best—option, data can discourage workers from exploring further to reach the best. Extending this idea of the streetlight effect, I suggest a "snowballing" effect: analogies may steer the direction of innovation towards areas with templates, which in turn, gain more templates and thus become more amenable to analogies, while neglected (and potentially fruitful) areas without templates may never get attention.

This snowballing effect can manifest in several ways. In the case of structural biology, bright clusters got brighter and brighter after the arrival of MR; among clusters that were predicted to have similar crystallization potential and biological importance, clusters that happened to have structures before MR took off, while clusters without structures remained dark. This increasing

[24] The authors draw from the aphorism of the drunk looking for his keys under the streetlight, despite dropping the keys on the other side of the street, because "this is where the light is."

returns to training data can have profound strategic consequences for firms as well. For firms that produce products and services based on data, early entrants may use their control over data to crowd out competitors (Cockburn, Henderson, and Stern 2018; Bessen et al. 2022). At the frontier of AI, there are now even machine learning models (trained on real data) that generate synthetic data, which will be fed into other machine learning models, raising the possibility of amplifying the influence of the original training data (Zewe 2022).

A question still remains on what is the net value of increased innovative activities in the bright clusters. This is difficult to assess. First, my difference-in-differences framework is limited to reporting only the relative increase in solved structures between bright and dark clusters. This relative increase can mean either an overall increase in bright clusters or a reallocation of innovative efforts from dark to bright clusters. Second, while structures in the bright clusters may not have led to novel biological insights, the increased number of (well-executed) structures may still be valuable, especially for drug development (which requires precisely solved structures) as well as for serving as training data for machine learning algorithms.

Finally, while this paper focuses on how analogies can potentially constrain the direction of innovation, future work can explore other costs of analogies. In particular, because analogies do not build foundational knowledge from scratch, innovators may not fully understand the underlying mechanisms of how the target domain works. Overreliance on analogies may lead to knowledge domains with weak foundation, where only correlations accumulate and causal theories are neglected (Zittrain 2019). Popular rhetoric often warns the danger of this "black box" nature of AI (and by extension, analogies). For instance, in drug discovery, analogical reasoning and pattern recognition can be employed to identify promising drug candidates based on prior approved drugs. A downside to this approach is that the drugs' mechanisms of action remain unknown, preventing drugmakers from anticipating side effects or applying the drugs for other diseases that may share the same mechanisms.

The setting of structural biology is currently facing its own AI revolution. In November 2020, Google's DeepMind team cracked a 50-year-old grand challenge in biology: to predict how a protein folds into its 3D structure from purely its sequence of amino acids. While MR helps with only one part of experimental structure solving, AlphaFold bypasses the need to conduct experiments at all. Celebrated as perhaps the most important application of AI in science as of date, the source code of AlphaFold became publicly available in July 2021, followed by the release of a database of predicted structures a year later.

Despite its breakthrough, however, AlphaFold also serves as a reminder of the limitations of analogies. DeepMind's claim that AlphaFold has produced enough structures to cover the "entire protein universe" (Walsh 2022) must be qualified with important caveats. First, these are only *predicted* structures. When comparing AlphaFold's structures to experimentally determined structures, researchers have found that while many of these predicted structures can be remarkably accurate, some are still too inaccurate to be useful (Mullard 2021). Second, and most importantly, AlphaFold is limited by its training data, the PDB, and can only populate the protein structural space based on analogies to known PDB structures. In particular, protein structures can change in the presence of small molecule drugs. Because there could be up to one novemdecillion[25] small molecules (Reymond and Awale 2012), the PDB does not contain enough information on structures bound to small molecules for AlphaFold to predict how proteins might interact with drugs (Callaway 2022). Furthermore, diseases are often caused by mutations to proteins. Since these mutated proteins have no evolutionarily-related sequences, it is difficult for AlphaFold to predict their structures (Buel and Walters 2022; Callaway 2022). This is why, at least for now, many remain skeptical that AlphaFold will dramatically impact drug development.[26]

Finally, AlphaFold illustrates that analogies work by identifying patterns, not causal theories. While AlphaFold has advanced the ability to predict structures, scientists still have little understanding of the physics of *why* proteins fold into their shapes (Lowe 2022). The rise of AlphaFold may signal the trend of hypothesis-driven science turning into "data science." As one of the scientists in the field lamented, "We've focused too much on data and not enough on understanding . . [we may be] going away from human-conceived theories and models of natural phenomena to more data-driven methods and models" (Samuel 2019).

[25] A novemdecillion is equivalent to million billion billion billion billion billion billion (American Chemical Society 2012).

[26] While not about AlphaFold, Lou and Wu (2022) demonstrates the limits of AI in drug development; AI is less useful for developing drugs that are radically novel and have no known mechanisms of actions. In addition, a recent paper by Cavalli (2022) investigates how AlphaFold changed the organizational structure of academic labs in computational biology.

References

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Boston, MA: Harvard Business Review Press.

Agrawal, Ajay, John McHale, and Alex Oettl. 2018. "Finding Needles in Haystacks: Artificial Intelligence and Recombinant Growth." *NBER Working Paper* #24541.

American Chemical Society. 2012. "1 Million Billion Billion Billion Billion Billion Billion: Number of Undiscovered Drugs." *Phys.Org*, June 6, 2012. https://phys.org/news/2012-06-million-billion-undiscovered-drugs.html.

Anthony, Callen. 2021. "When Knowledge Work and Analytical Technologies Collide: The Practices and Consequences of Black Boxing Algorithmic Technologies." *Administrative Science Quarterly* 66 (4): 1173–1212.

Azoulay, Pierre, Jeffrey L Furman, Joshua L Krieger, and Fiona Murray. 2015. "Retractions." *The Review of Economics and Statistics* 97 (5): 1118–36.

Baldwin, Carliss Y., and Kim B. Clark. 1997. "Managing in an Age of Modularity." *Harvard Business Review*, 1997.

Bessen, James, Stephen Michael Impink, Lydia Reichensperger, and Robert Seamans. 2022. "The Role of Data for AI Startup Growth." *Research Policy* 51 (5).

Bingham, Christopher B., and Steven J. Kahl. 2012. "How to Use Analogies to Introduce New Ideas." *MIT Sloan Management Review* 56 (2): 10–12.

———. 2013. "The Process of Schema Emergence: Assimilation, Deconstruction, Unitization and the Plurality of Analogies." *Academy of Management Journal* 56 (1): 14–34.

Brynjolfsson, Erik, and Andrew McAfee. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York, NY: W. W. Norton & Company.

Buel, Gwen R., and Kylie J. Walters. 2022. "Can AlphaFold2 Predict the Impact of Missense Mutations on Structure?" *Nature Structural & Molecular Biology* 29 (1): 1–2.

Callaway, Ewen. 2022. "What's Next for AlphaFold and the AI Protein-Folding Revolution." *Nature*, April 13, 2022. https://www.nature.com/articles/d41586-022-00997-5.

Cassiday, Laura. 2014. "Structural Biology: More than a Crystallographer." *Nature* 505 (7485): 711–13.

Cavalli, Gabriel. 2022. "How Scientific Organizations React to Novel Methodological Advances: The Impact of AlphaFold V1." *Working Paper*.

Choudhury, Prithwiraj, Evan Starr, and Rajshree Agarwal. 2020. "Machine Learning and Human Capital Complementarities: Experimental Evidence on Bias Mitigation." *Strategic Management Journal* 41 (8): 1381–1411.

Cockburn, Iain M, Rebecca Henderson, and Scott Stern. 2018. “The Impact of Artificial Untelligence on Innovation.” *NBER Working Paper* #24449.

Cornelissen, Joep P., and Jean S. Clarke. 2010. “Imagining and Rationalizing Opportunities: Inductive Reasoning and the Creation and Justification of New Ventures.” *Academy of Management Review* 35 (4): 539–57.

Cowgill, Bo, Fabrizio Dell’acqua, Samuel Deng, Daniel Hsu, Nakul Verma, and Augustin Chaintreau. 2020. “Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics.” *Proceedings of the 21st ACM Conference on Economics and Computation*, 679–81.

Cowgill, Bo, and Catherine E. Tucker. 2020. “Algorithmic Fairness and Economics.” *Columbia Business School Research Paper*. https://ssrn.com/abstract=3361280.

Crick, Francis. 1990. *What Mad Pursuit: A Personal View of Scientific Discovery*. London, UK: Penguin.

Doerr, Allison. 2014. “A Method Ahead of Its Time.” *Nature* 511 (Suppl 7509): 13.

Dunbar, Kevin. 1999. “How Scientists Build Models In Vivo Science as a Window on the Scientific Mind.” *Model-Based Reasoning in Scientific Discovery*, 85–99.

———. 2000. “How Scientists Think in the Real World: Implications for Science Education.” *Journal of Applied Developmental Psychology* 21 (1): 49–58.

Elbasir, Abdurrahman, Balasubramanian Moovarkumudalvan, Khalid Kunji, Prasanna R Kolatkar, Raghvendra Mall, Halima Bensmail, and John Hancock. 2019. “DeepCrystal: A Deep Learning Framework for Sequence-Based Protein Crystallization Prediction.” *Bioinformatics* 35 (13): 2216–25.

Etzion, Dror, and Fabrizio Ferraro. 2010. “The Role of Analogy in the Institutionalization of Sustainability Reporting.” *Organization Science* 21 (5): 1092–1107.

Fleming, Lee, and Olav Sorenson. 2004. “Science as a Map in Technological Search.” *Strategic Management Journal* 25 (8–9): 909–28.

Furman, Jeffrey L., and Scott Stern. 2011. “Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research.” *American Economic Review* 101 (5): 1933–63.

Furman, Jeffrey L., and Florenta Teodoridis. 2020. “Automation, Research Technology, and Researchers’ Trajectories: Evidence from Computer Science and Electrical Engineering.” *Organization Science* 31 (2): 330–54.

Gavetti, Giovanni, Daniel A. Levinthal, and Jan W. Rivkin. 2005. “Strategy Making in Novel and Complex Worlds: The Power of Analogy.” *Strategic Management Journal* 26 (8): 691–712.

Gavetti, Giovanni, and Jan W. Rivkin. 2005. “How Strategists Really Think: Tapping the Power of Analogy.” *Harvard Business Review* 83 (4): 54–63.

Gentner, Dedre. 1982. “Structure Mapping: A Theoretical Framework for Analogy.” *Cognitive*

Science 7 (2): 155–70.

Gentner, Dedre, Keith J. Holyoak, and Boicho N. Kokinov. 2001. *The Analogical Mind: Perspectives from Cognitive Science*. Cambridge, MA: MIT Press.

Gilboa, Itzhak, Larry Samuelson, and David Schmeidler. 2015. *Analogies and Theories: Formal Models of Reasoning*. New York, NY: Oxford University Press.

Glaser, Vern L., Peer C. Fiss, and Mark Thomas Kennedy. 2016. "Making Snowflakes like Stocks: Stretching, Bending, and Positioning to Make Financial Market Analogies Work in Online Advertising." *Organization Science* 27 (4): 1029–48.

Greenblatt, Wesley. 2021. "Building on Solid Ground: Foundational Knowledge and the Dynamics of Innovation." *SSRN Working Paper*. https://ssrn.com/abstract=3919866.

Hacking, Ian. 2014. *Why Is There Philosophy of Mathematics at All?* Cambridge, UK: Cambridge University Press.

Hargadon, Andrew B., and Yellowlees Douglas. 2001. "When Innovations Meet Institutions: Edison and the Design of the Electric Light." *Administrative Science Quarterly* 46 (3): 476–501.

Harrison, Paul M. 2017. "FLPS: Fast Discovery of Compositional Biases for the Protein Universe." *BMC Bioinformatics* 18 (1): 1–9.

Hauser, Maria, Martin Steinegger, and Johannes Söding. 2016. "MMseqs Software Suite for Fast and Deep Clustering and Searching of Large Protein Sequence Sets." *Bioinformatics* 32 (9): 1323–30.

Henderson, Rebecca. 1994. "The Evolution of Integrative Capability: Innovation in Cardiovascular Drug Discovery." *Industrial and Corporate Change* 3 (3): 607–30.

Hesse, Mary B. 1966. *Models and Analogies in Science*. Notre Dame, IN: Univ Notre Dame Press.

Hill, Robert C., and Michael Levenhagen. 1995. "Metaphors and Mental Models: Sensemaking and Sensegiving in Innovative and Entrepreneurial Activities." *Journal of Management* 21 (6): 1057–74.

Hill, Ryan, and Carolyn Stein. 2020. "Scooped! Estimating Rewards for Priority in Science." *Working Paper*.

———. 2021. "Race to the Bottom: Competition and Quality in Science." *Working Paper*.

Hoelzemann, Johannes, Gustavo Manso, Abhishek Nagaraj, and Matteo Tranchero. 2022. "The Streetlight Effect in Data-Driven Exploration." *Working Paper*.

Hofstadter, Douglas R, and Emmanuel Sander. 2013. *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*. New York, NY: Basic Books.

Holyoak, Keith J., and Paul Thagard. 1996. *Mental Leaps: Analogy in Creative Thought*. Cambridge, MA: MIT Press.

Jones, Benjamin F. 2009. "The Burden of Knowledge and the 'Death of the Renaissance Man': Is Innovation Getting Harder?" *Review of Economic Studies* 76 (1): 283–317.

Kaplan, Sarah, and Wanda J Orlikowski. 2013. “Temporal Work in Strategy Making” 24 (4): 965–95.

Kaplan, Sarah, and Keyvan Vakili. 2015. “The Double-Edged Sword of Recombination in Breakthrough Innovation.” *Strategic Management Journal* 36 (10): 1435–57.

Kauffman, Stuart A. 1993. *The Origins of Order: Self-Organization and Selection in Evolution*. New York, NY: Oxford University Press.

Kittur, Aniket, Lixiu Yu, Tom Hope, Joel Chan, Hila Lifshitz-Assaf, Karni Gilon, Felicia Ng, Robert E. Kraut, and Dafna Shahaf. 2019. “Scaling up Analogical Innovation with Crowds and AI.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (6): 1870–77.

Kleywegt, G. J., and T. A. Jones. 1997. “Model Building and Refinement Practice.” *Methods in Enzymology* 277: 208–30.

Kneeland, Madeline K., Melissa A. Schilling, and Barak S. Aharonson. 2020. “Exploring Uncharted Territory: Knowledge Search Processes in the Origination of Outlier Innovation.” *Organization Science* 31 (3): 535–57.

Koonin, Eugene V., Yuri I. Wolf, and Georgy P. Karev. 2002. “The Structure of the Protein Universe and Genome Evolution.” *Nature* 420 (6912): 218–23.

Levinthal, Daniel A. 1997. “Adaptation on Rugged Landscapes.” *Management Science* 43 (7): 934–50.

Lou, Bowen, and Lynn Wu. 2022. “AI on Drugs: Can Artificial Intelligence Accelerate Drug Development? Evidence from a Large-Scale Examination of Bio-Pharma Firms.” *Working Paper*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3524985.

Lowe, Derek. 2022. “The Law of Conservation of Data.” *Chemistry World*, January 11, 2022. https://www.chemistryworld.com/opinion/the-law-of-conservation-of-data/4014927.article.

Madrigal, Alexis C. 2019. “The Servant Economy.” *The Atlantic*, March 6, 2019. https://www.theatlantic.com/technology/archive/2019/03/what-happened-uber-x-companies/584236/.

Mannucci, Pier Vittorio. 2017. “Drawing Snow White and Animating Buzz Lightyear: Technological Toolkit Characteristics and Creativity in Cross-Disciplinary Teams.” *Organization Science* 28 (4): 711–28.

Martins, Luis L., Violina P. Rindova, and Bruce E. Greenbaum. 2015. “Unlocking the Hidden Value of Concepts: A Cognitive Approach to Business Model Innovation.” *Strategic Entrepreneurship Journal* 9: 99–117.

Marx, Matt, and Aaron Fuegi. 2020. “Reliance on Science: Worldwide Front-Page Patent Citations to Scientific Articles.” *Strategic Management Journal* 41 (9): 1572–94.

Mill, John Stuart. 1974. *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. Toronto, Canada:

University of Toronto Press.

Mirdita, Milot, Lars Von Den Driesch, Clovis Galiez, Maria J. Martin, Johannes Soding, and Martin Steinegger. 2017. “Uniclust Databases of Clustered and Deeply Annotated Protein Sequences and Alignments.” *Nucleic Acids Research* 45 (D1): D170–76.

Miric, Milan, Hakan Ozalp, and Dogukan Yilmaz. 2021. “Tradeoffs to Using Standardized Tools: An Innovation Enabler or Creativity Constraint?” *USC Marshall School of Business Research Paper*. https://ssrn.com/abstract=3358801.

Moore, Peter B. 2007. “Let’s Call the Whole Thing Off: Some Thoughts on the Protein Structure Initiative.” *Structure* 15 (11): 1350–52.

Mullard, Asher. 2021. “What Does AlphaFold Mean for Drug Discovery?” *Nature Reviews Drug Discovery* 20 (10): 725–27.

Nagle, Frank, and Florenta Teodoridis. 2020. “Jack of All Trades and Master of Knowledge: The Role of Diversification in New Distant Knowledge Integration.” *Strategic Management Journal* 41 (1): 55–85.

Nature Structural Biology. 1994. “The Changing Structure of Biology.” *Nature Structural Biology* 1 (1).

Nelson, Richard R, and Sidney G Winter. 1982. “The Schumpeterian Tradeoff Revisited.” *The American Economic Review* 72 (1): 114–32.

Parnas, D. L. 1972. “On the Criteria to Be Used in Decomposing Systems into Modules.” *Communications of the ACM* 15 (12): 1053–58.

Patel, Onisha, Isabelle Lucet, and Michael Roy. 2020. “‘Like a Key to a Lock’: How Seeing the Molecular Machinery of the Coronavirus Will Help Scientists Design a Treatment.” *The Conversation*, March 24, 2020. https://theconversation.com/like-a-key-to-a-lock-how-seeing-the-molecular-machinery-of-the-coronavirus-will-help-scientists-design-a-treatment-134135.

Perdigão, Nelson, Julian Heinrich, Christian Stolte, Kenneth S. Sabir, Michael J. Buckley, Bruce Tabor, Beth Signal, et al. 2015. “Unexpected Features of the Dark Proteome.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (52): 15898–903.

Petsko, Gregory A. 2007. “An Idea Whose Time Has Gone.” *Genome Biology* 8 (6): 1–3.

Phenix. n.d. “Overview of Molecular Replacement in Phenix.” Accessed September 1, 2022. https://phenix-online.org/documentation/reference/mr_overview.html.

Pollack, John. 2014. *Shortcut: How Analogies Reveal Connections, Spark Innovation, and Sell Our Greatest Ideas*. New York, NY: Gotham Books.

Ramakrishnan, Venki. 2018. *Gene Machine: The Race to Decipher the Secrets of the Ribosome*. New York, NY: Basic Books.

Reymond, Jean Louis, and Mahendra Awale. 2012. “Exploring Chemical Space for Drug Discovery

Using the Chemical Universe Database." *ACS Chemical Neuroscience* 3 (9): 649–57.

Romer, Paul M. 1990. "Endogenous Technological Change." *Journal of Political Economy* 98 (5).

Samuel, Sigal. 2019. "How One Scientist Coped When AI Beat Him at His Life's Work." *Vox*, February 15, 2019. https://www.vox.com/future-perfect/2019/2/15/18226493/deepmind-alphafold-artificial-intelligence-protein-folding.

Sanchez, Ron, and Joseph T. Mahoney. 1996. "Modularity, Flexibility, and Knowledge Management in Product and Organization Design." *Strategic Management Journal* 17 (Suppl Winter): 63–76.

Scapin, Giovanna. 2013. "Molecular Replacement Then and Now." *Acta Crystallographica Section D: Biological Crystallography* 69 (11): 2266.

Schmidberger, Jason W., Mark A. Bate, Cyril F. Reboul, Steve G. Androulakis, Jennifer M.N. Phan, James C. Whisstock, Wojtek J. Goscinski, David Abramson, and Ashley M. Buckle. 2010. "MrGrid: A Portable Grid Based Molecular Replacement Pipeline." *PLOS ONE* 5 (4): e10049.

Schumpeter, Joseph. 1934. *The Theory of Economic Development*. Cambridge, MA: Harvard University Press.

Scotchmer, Suzanne. 1991. "Standing on the Shoulders of Giants: Cumulative Research and the Patent Law." *Journal of Economic Perspectives* 5 (1): 29–41.

Simcoe, Timothy. 2015. "Modularity and the Evolution of the Internet." In *Economic Analysis of the Digital Economy*, 21–47. University of Chicago Press.

Slabinski, Lukasz, Lukasz Jaroszewski, Ana P.C. Rodrigues, Leszek Rychlewski, Ian A. Wilson, Scott A. Lesley, and Adam Godzik. 2007. "The Challenge of Protein Structure Determination--Lessons from Structural Genomics." *Protein Science : A Publication of the Protein Society* 16 (11): 2472–82.

Spenser, Jay. 2008. *The Airplane: How Ideas Gave Us Wings*. New York, NY: HarperCollins.

Steinegger, Martin, and Johannes Söding. 2018. "Clustering Huge Protein Sequence Sets in Linear Time." *Nature Communications* 9 (1): 1–8.

Teodoridis, Florenta. 2018. "Understanding Team Knowledge Production: The Interrelated Roles of Technology and Expertise." *Management Science* 64 (8): 3625–48.

Terwilliger, Thomas C, David Stuart, and Shigeyuki Yokoyama. 2009. "Lessons from Structural Genomics." *Annual Review of Biophysics* 38: 371–83.

Thomke, Stefan H. 1998. "Managing Experimentation in the Design of New Products." *Management Science* 44 (6): 743–62.

Uzzi, Brian, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. "Atypical Combinations and Scientific Impact." *Science* 342 (6157): 468–72.

Walsh, Bryan. 2022. "Finally, an Answer to the Question: AI — What Is It Good For?" *Vox*, August 3, 2022. https://www.vox.com/future-perfect/2022/8/3/23288843/deepmind-alphafold-artificial-intelligence-biology-drugs-medicine-demis-hassabis.

Weitzman, Martin L. 1998. “Recombinant Growth.” *The Quarterly Journal of Economics* 113 (2): 331–60.

Zewe, Adam. 2022. “When It Comes to AI, Can We Ditch the Datasets?” *MIT News*, March 15, 2022. https://news.mit.edu/2022/synthetic-datasets-ai-image-classification-0315.

Zhuo, Ran. 2022. “Exploit or Explore? An Empirical Study of Resource Allocation in Scientific Labs.” *Working Paper*.

Zittrain, Jonathan. 2019. “The Hidden Costs of an Automated Thinking.” *The New Yorker*, July 23, 2019. https://www.newyorker.com/tech/annals-of-technology/the-hidden-costs-of-automated-thinking.

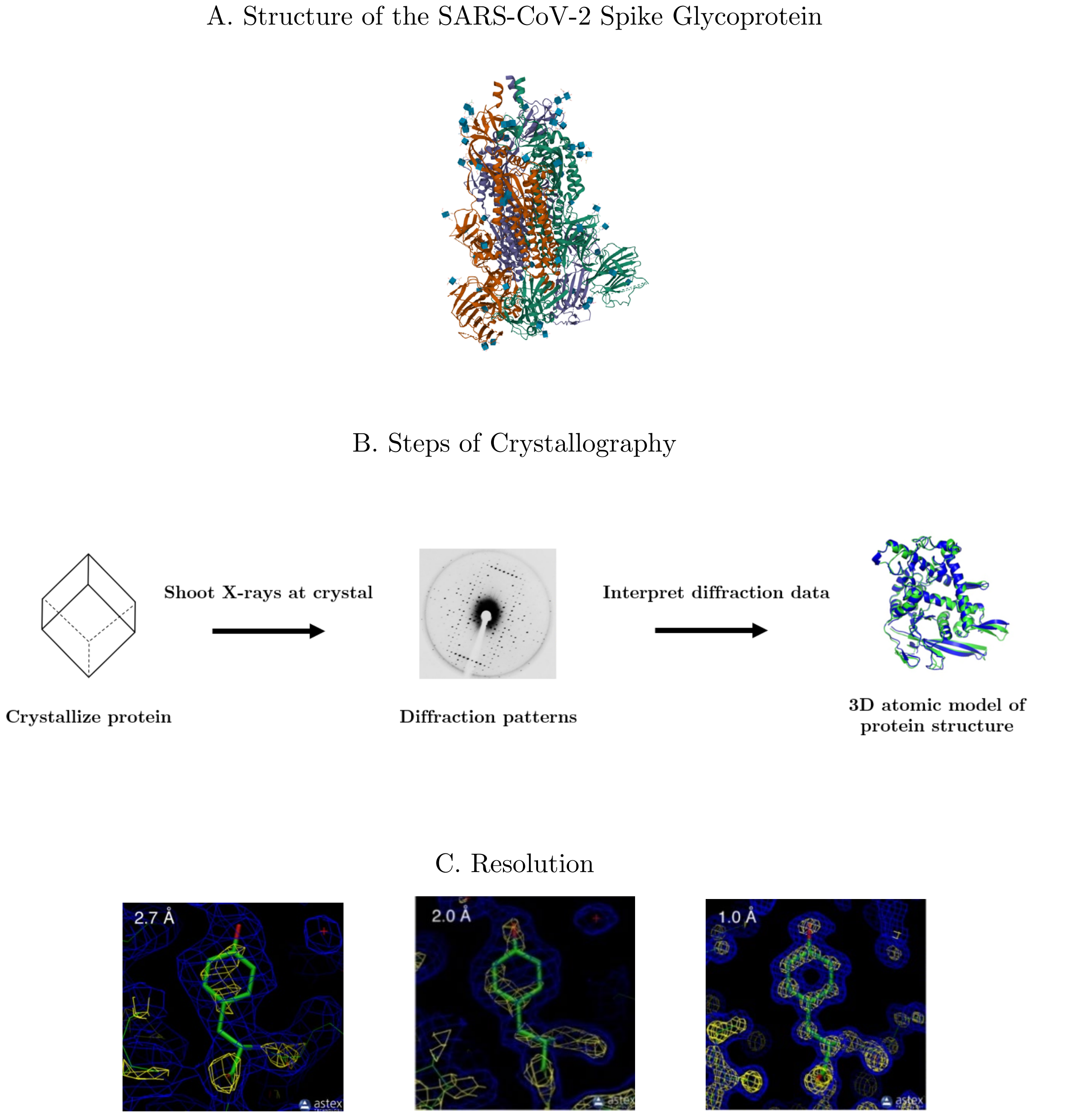
Figures & Tables

Figure 1. A Taxonomy of Shortcuts

	Vertical Shortcuts (Used in older domains with foundational knowledge)	**Horizontal Shortcuts (Used in newer domains w/out foundational knowledge)**
Understanding	**Summaries** Provide synopses of foundational knowledge underlying a domain e.g., Wikipedia	**Conceptual Analogies** Understand a new domain by importing patterns from a known domain e.g., biomimetics
Application	**Calculators** Execute instructions based on foundational knowledge underlying a domain e.g., Stata	**Automated Analogies** Apply patterns to a new domain from a known domain e.g., TensorFlow

Notes: This figure provides a taxonomy of shortcuts that can be used in cumulative knowledge production.

Figure 2. Structural Biology



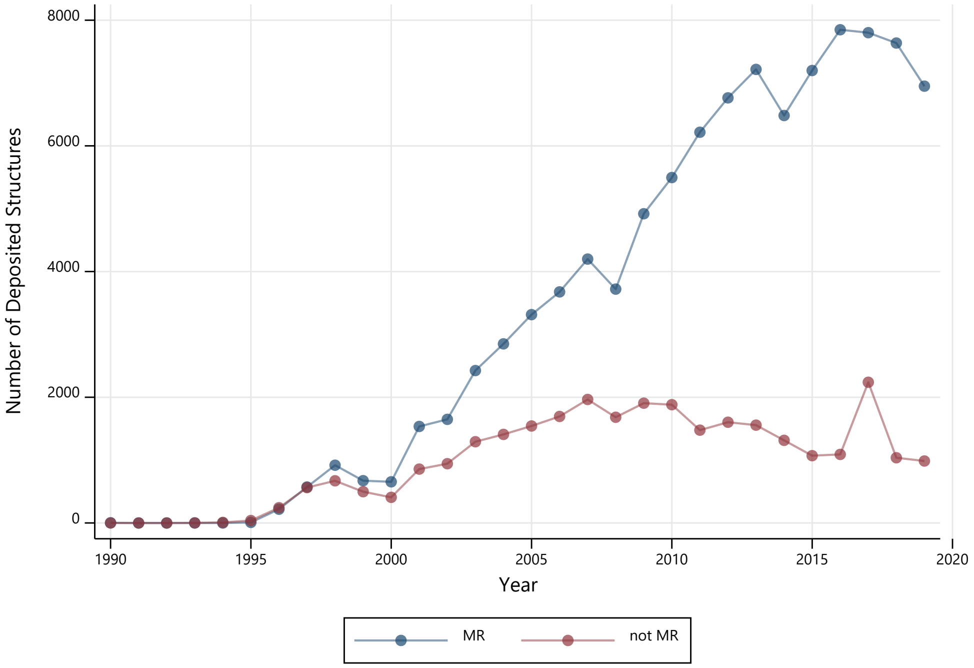
A. Structure of the SARS-CoV-2 Spike Glycoprotein

B. Steps of Crystallography

C. Resolution

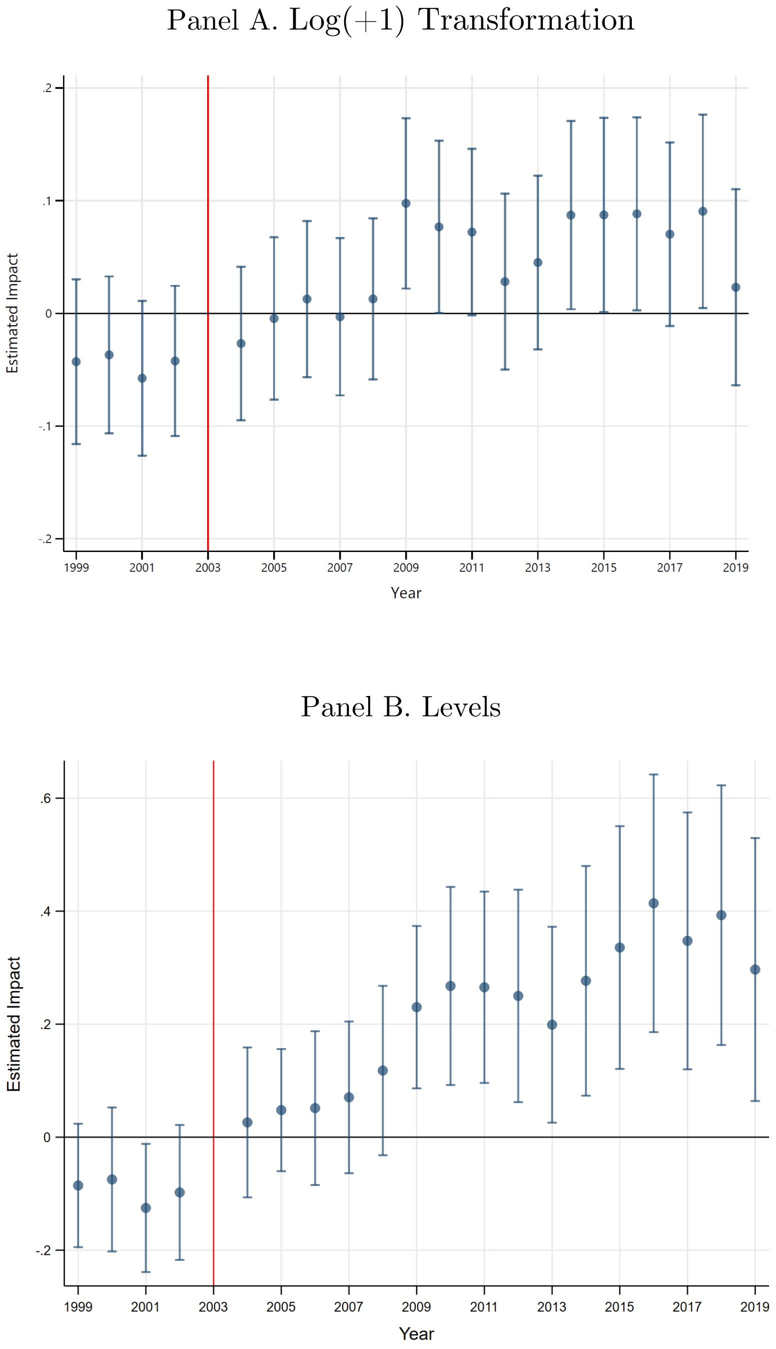
Notes: Panel A shows the structure of a spike protein on the surface of the coronavirus (PDB entry 6VYB; source: https://www.rcsb.org/structure/6VYB). Panel B shows the three main steps of crystallography; this paper focuses on the automation of solving the "phase problem" that occurs during the interpretation of the diffraction data. Panel C shows an example of the electron density map behind the structure of tyrosine 103 from myoglobin, at three different resolutions; lower resolution is better and shows finer details (source: https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/resolution).

Figure 3. Number of Structures Solved by Molecular Replacement



Notes: This figure plots the number of X-ray crystallography structures in the Protein Data Bank that were solved by molecular replacement (MR) vs. non-MR methods.

Figure 4. Event Study: Impact of MR on Number of Solved Structures



Panel A. Log(+1) Transformation

Panel B. Levels

Notes: This figure shows the impact of MR on the number of solved structures. The figure plots the coefficients and 95% confidence intervals from estimating a modified, event studies version of Equation 1 that replaces the pooled $PostMR_t$ indicator with separate indicators for every year before and after the arrival of MR. The outcome is the total annual number of solved structures in a cluster; Panel A reports the outcome after Log(+1) transformation, while Panel B reports the outcome in levels. The unit of analysis is a cluster × year, and the sample consists of 6,944 clusters, which translates to 145,824 cluster-years.

Table 1. Summary Statistics

	Bright			Dark		
	Mean	Median	SD	Mean	Median	SD
Discovery Year	1983.47	1986.00	7.7	1993.11	1995.00	5.1
Cluster Size	39.73	15	81.6	7.66	4	15
Protein Production Feasibility						
Disorder	0.15	0.1	0.2	0.24	0.1	0.2
Membrane	0.02	0	0.1	0.04	0	0.1
Compositional Bias	0.01	0	0	0.03	0	0.1
Sequence Length	513.61	339.1	1,438.50	634.97	457.3	675.5
Biological Importance						
% of Cluster that is Human	0.18	0.1	0.2	0.36	0.3	0.2
N of Publications	142.47	75	199.2	26.79	12	57.4
N of Approved Drugs	3.63	0	14.5	1.23	0	21.8
N of Solved Structures per Year	1.99	0	5.64	0.23	0	1.78
N of Clusters		649			6,295	

Notes: This table provides the summary characteristics of clusters when MR was introduced. The sample consists of 6,944 clusters, of which 649 are classified as “bright” (i.e., had at least one structure in 1998) and 6,295 are classified as “dark.”

Table 2. Impact of MR on Number of Solved Structures

	(1)	(2)	(3)	(4)
	Log(+1)	Log(+1)	Levels	Levels
VARIABLES	N Structures	N Structures	N Structures	N Structures
Post-MR × Bright	0.071***	0.065***	0.301***	0.292***
	(0.018)	(0.018)	(0.052)	(0.052)
Cluster Size		0.013**		0.019*
		(0.006)		(0.011)
R-squared	0.471	0.471	0.400	0.400
Calendar-year FE	YES	YES	YES	YES
Cluster FE	YES	YES	YES	YES
N of clusters	6,944	6,944	6,944	6,944
N of cluster-years	145,824	145,824	145,824	145,824

Notes: This table reports results from estimating Equation 1 and shows the impact of MR on the number of solved structures. The unit of analysis is a cluster × year, and the panel spans from 1999-2019. The outcome variable is the total annual number of solved structures in a cluster, reported after Log(+1) transformation (Columns 1-2) or in levels scaled by the standard deviation (Columns 3-4). The treatment variable "Bright" is defined as clusters that had at least one structure by 1998, while "Post-MR" includes years 2004 and onwards. All columns include calendar-year and cluster fixed effects; Columns 2 and 4 additionally control for time-varying (standardized) cluster size. Standard errors are clustered at the cluster level. Statistical significance is indicated as: *** p<0.01, ** p<0.05, * p<0.1.

Table 3. Impact of MR on Execution

VARIABLES	(1) All Structures	(2) *R-Free* Bottom Tercile	(3) *R-Free* Middle Tercile	(4) *R-Free* Top Tercile	(5) *Resolution* Bottom Tercile	(6) *Resolution* Middle Tercile	(7) *Resolution* Top Tercile
Post-MR × Bright	0.071***	-0.001	0.077***	0.133***	0.034***	0.047***	0.096***
	(0.018)	(0.011)	(0.012)	(0.013)	(0.011)	(0.012)	(0.013)
R-squared	0.471	0.381	0.397	0.403	0.365	0.414	0.427
Calendar-year FE	YES	YES	YES	YES	YES	YES	YES
Cluster FE	YES	YES	YES	YES	YES	YES	YES
N of clusters	6,944	6,944	6,944	6,944	6,944	6,944	6,944
N of cluster-years	145,824	145,824	145,824	145,824	145,824	145,824	145,824

Notes: This table reports results from estimating Equation 1 and shows the impact of MR on the number of solved structures at different terciles of execution level (a structure's level of execution can be defined in terms of its R-free value and resolution). The unit of analysis is a cluster × year, and the panel spans from 1999-2019. The outcomes of all columns are the annual number of solved structures in a cluster, with Log(+1) transformation. Column 1 parallels Column 1 in Table 2 and reports the total number of solved structures. Columns 2-4 decompose this result by examining the number of solved structures in the bottom (Column 2), middle (Column 3), and top terciles (Column 4) with respect to the structures' R-free values. Columns 5-6 similarly decompose the number of solved structures into terciles based on their resolution. The treatment variable "Bright" is defined as clusters that had at least one structure by 1998, while "Post-MR" includes years 2004 and onwards. All columns include calendar-year and cluster fixed effects. Standard errors are clustered at the cluster level. Statistical significance is indicated as: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

TABLE 4. IMPACT OF MR ON FUNCTIONAL INSIGHTS

VARIABLES	(1) All Structures	(2) Unpublished Structures	(3) Published Structures *Not Cited by* *Patent*	(4) Published Structures *Cited by* *Patent*	(5) Published Structures *Not Fxn* *Annotated*	(6) Published Structures *Fxn* *Annotated*
Post-MR × Bright	0.071***	0.085***	0.117***	-0.064***	0.039**	0.006
	(0.018)	(0.009)	(0.016)	(0.011)	(0.017)	(0.004)
R-squared	0.471	0.221	0.389	0.350	0.473	0.117
Calendar-year FE	YES	YES	YES	YES	YES	YES
Cluster FE	YES	YES	YES	YES	YES	YES
N of clusters	6,944	6,944	6,944	6,944	6,944	6,944
N of cluster-years	145,824	145,824	145,824	145,824	145,824	145,824

NOTES: This table reports results from estimating Equation 1 and shows the impact of MR on the number of solved structures at different levels of functional insights (a structure is considered to have contributed to new insights about a protein's function if it is published in a scientific article and additionally cited by a patent or by the functional summary section of Swiss-Prot). The unit of analysis is a cluster × year, and the panel spans from 1999-2019. The outcomes of all columns are the annual number of solved structures in a cluster, with Log(+1) transformation. Column 1 parallels Column 1 in Table 2 and reports the total number of solved structures. Columns 2-4 decompose this result into the number of solved structures that do not get published in a scientific article (Column 2), the number of solved structures that are published but not cited by a patent (Column 3), and the number of solved structures that are both published and cited by a patent (Column 4). Columns 5 and 6 parallel Columns 3 and 4 but decompose the number of solved structures based on whether they were cited by the functional summary section of Swiss-Prot. The treatment variable "Bright" is defined as clusters that had at least one structure by 1998, while "Post-MR" includes years 2004 and onwards. All columns include calendar-year and cluster fixed effects. Standard errors are clustered at the cluster level. Statistical significance is indicated as: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

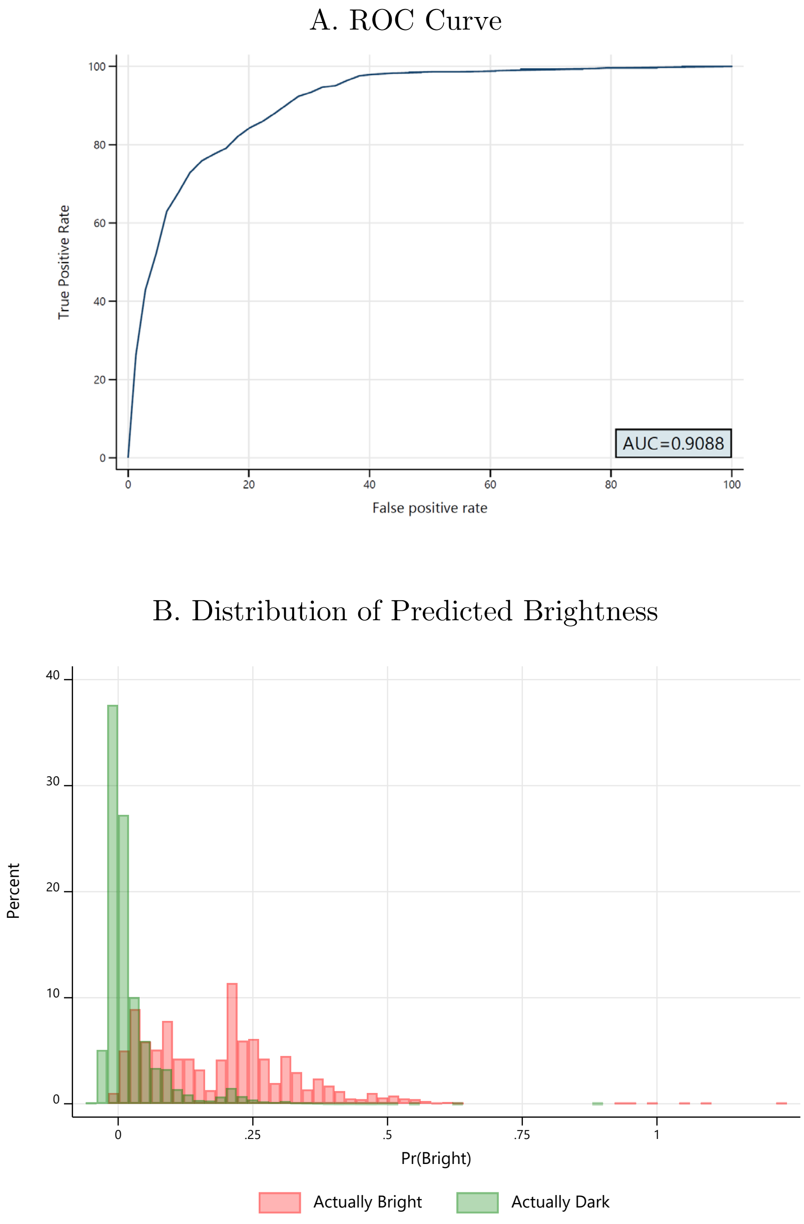
Table 5. Impact of MR on Publication Impact

	Citations					*Journal Impact Factor*				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	All Structures	Unpublished Structures	*Citations* Published Structures (Bottom Tercile)	*Citations* Published Structures (Middle Tercile)	*Citations* Published Structures (Top Tercile)	All Structures	Unpublished Structures	*JIF* Published Structures (Bottom Tercile)	*JIF* Published Structures (Middle Tercile)	*JIF* Published Structures (Top Tercile)
Post-MR × Bright	0.056***	0.075***	0.067***	0.012	-0.022**	0.070***	0.083***	0.070***	0.006	0.014
	(0.017)	(0.009)	(0.012)	(0.012)	(0.010)	(0.018)	(0.009)	(0.012)	(0.012)	(0.009)
R-squared	0.487	0.227	0.401	0.330	0.331	0.477	0.223	0.378	0.363	0.285
Calendar-year FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Cluster FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
N of clusters	6,944	6,944	6,944	6,944	6,944	6,944	6,944	6,944	6,944	6,944
N of cluster-years	97,216	97,216	97,216	97,216	97,216	131,936	131,936	131,936	131,936	131,936

Notes: This table reports results from estimating Equation 1 and shows the impact of MR on the number of solved structures at different terciles of publication impact. A structure's publication impact is measured as the mean number of citations and the journal impact factor (JIF). The unit of analysis is a cluster × year. Due to data availability, the panel ends in 2012 for the citation analyses and in 2017 for the JIF analyses. The outcomes of all columns are the annual number of solved structures in a cluster, with Log(+1) transformation. Column 1 parallels Column 1 in Table 2 and reports the total number of solved structures. Columns 2-5 decompose this result into the number of solved structures that do not get published in a scientific article (Column 2) and the number of solved structures that are published and in bottom (Column 3), middle (Column 4), or top (Column 5) terciles in terms of citation impact. Columns 6-7 similarly decompose the number of solved structures into terciles based on JIF. The treatment variable "Bright" is defined as clusters that had at least one structure by 1998, while "Post-MR" includes years 2004 and onwards. All columns include calendar-year and cluster fixed effects. Standard errors are clustered at the cluster level. Statistical significance is indicated as: *** p<0.01, ** p<0.05, * p<0.1.

Appendix Figures & Tables

Appendix Figure 1. Predicting Brightness



A. ROC Curve

B. Distribution of Predicted Brightness

Notes: Panel A plots the ROC curve of a prediction exercise, where I predict whether a protein is bright by 1998. Panel B plots the distribution of the resulting predicted brightness by whether the protein was actually bright (i.e., had a structure by 1998) or dark (i.e., did not have a structure by 1998). The unit of analysis is a protein. The sample consists of 41,781 proteins that were discovered by 1998. Each protein's predicted brightness was constructed by using the fitted values from estimating a Lasso model that predicted whether the protein had a structure by 1998.

Appendix Table 1. Impact of MR on Number of Solved Structures, Inverse Hyperbolic Sine Transformation

	(1)	(2)
	IHS	IHS
VARIABLES	N of Structures	N of Structures
Post-MR × Bright	0.083***	0.075***
	(0.022)	(0.022)
Cluster Size		0.017**
		(0.007)
R-squared	0.464	0.464
Calendar-year FE	YES	YES
Cluster FE	YES	YES
N of clusters	6,944	6,944
N of cluster-years	145,824	145,824

Notes: This table parallels Table 2 but presents a robustness analysis using inverse hyperbolic sine transformation of the outcome. The table shows the impact of MR on the number of solved structures. The unit of analysis is a cluster × year, and the panel spans from 1999-2019. The outcome variable is the total annual number of solved structures in a cluster after inverse hyperbolic sine transformation. The treatment variable "Bright" is defined as clusters that had at least one structure by 1998, while "Post-MR" includes years 2004 and onwards. All columns include calendar-year and cluster fixed effects; Columns 2 additionally control for time-varying (standardized) cluster size. Standard errors are clustered at the cluster level. Statistical significance is indicated as: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Appendix Table 2. Impact of MR on MR vs. Non-MR Structures

VARIABLES	(1) All Structures	(2) MR Structures	(3) Non-MR Structures
Post-MR × Bright	0.071***	0.143***	-0.007*
	(0.018)	(0.017)	(0.004)
R-squared	0.471	0.475	0.101
Calendar-year FE	YES	YES	YES
Cluster FE	YES	YES	YES
N of clusters	6,944	6,944	6,944
N of cluster-years	145,824	145,824	145,824

Notes: This table parallels Column 1 from Table 2. The table reports results from estimating Equation 1 and shows the impact of MR on the total number of solved structures (Column 1) and decomposes this into number of solved MR structures (Column 2) and non-MR structures (Column 3). All of the outcomes are Log(+1) transformed. The unit of analysis is a cluster × year, and the panel spans from 1999-2019. The treatment variable "Bright" is defined as clusters that had at least one structure by 1998, while "Post-MR" includes years 2004 and onwards. All columns include calendar-year and cluster fixed effects. Standard errors are clustered at the cluster level. Statistical significance is indicated as: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Appendix Table 3. Impact of MR, Splitting Bright Clusters

	(1)	(2)	(3)
	Dark vs. All Bright Clusters	Dark vs. Bright Clusters with 1 Structure	Bright Clusters with 1 Structure vs. Bright Clusters with >1 Structures
Post-MR × Bright (1 or more structure)	0.071*** (0.018)	0.072*** (0.022)	
Post-MR × Bright (more than 1 structure)			-0.002 (0.034)
R-squared	0.471	0.325	0.595
Calendar-year FE	YES	YES	YES
Cluster FE	YES	YES	YES
N of clusters	6,944	6,558	649
N of cluster-years	145,824	137,718	13,629

Notes: Column 1 of this table parallels Column 1 of Table 2 and shows the impact of MR on the total number of solved structures in the full sample. Column 2 investigates the impact of MR on the sample of dark clusters and bright clusters with just 1 structure solved by 1998; the treatment variable "Bright" is defined as clusters that had just one structure by 1998. Column 3 investigates the impact of MR on the sample of bright clusters with 1 or more structures solved by 1998; the treatment variable "Bright" is defined as clusters that had more than 1 structure by 1998. All of the outcomes are Log(+1) transformed. The unit of analysis is a cluster × year, and the panel spans from 1999-2019. All columns include calendar-year and cluster fixed effects. Standard errors are clustered at the cluster level. Statistical significance is indicated as: *** p<0.01, ** p<0.05, * p<0.1.

Appendix Table 4. Impact of MR on Number of Solved Structures with Predicted Brightness

	(1)	(2)	(3)	(4)
	Log(+1)	Log(+1)	Levels	Levels
VARIABLES	N Structures	N Structures	N Structures	N Structures
Post-MR × Bright	0.049**	0.044**	0.218***	0.213***
	(0.021)	(0.021)	(0.055)	(0.054)
Post-MR × Predicted Bright	0.024	0.023	0.089	0.087
	(0.017)	(0.016)	(0.056)	(0.055)
Cluster Size		0.012**		0.014
		(0.006)		(0.011)
R-squared	0.472	0.472	0.401	0.401
Calendar-year FE	YES	YES	YES	YES
Cluster FE	YES	YES	YES	YES
N of clusters	6,878	6,878	6,878	6,878
N of cluster-years	144,438	144,438	144,438	144,438

Notes: This table reports results from estimating Equation 2 and shows the impact of MR on the number of solved structures, controlling for predicted brightness. The unit of analysis is a cluster × year, and the panel spans from 1999-2019. The outcome variable is the total annual number of solved structures in a cluster, reported after Log(+1) transformation (Columns 1-2) or in levels scaled by the standard deviation (Columns 3-4). The treatment variable "Bright" is defined as clusters that had at least one structure by 1998, while "Post-MR" includes years 2004 and onwards. "Predicted Bright" was constructed after predicting whether a protein was structurally characterized by 1998, using its pre-period characteristics and aggregating to the cluster-level. All columns include calendar-year and cluster fixed effects; Columns 2 and 4 additionally control for time-varying standardized cluster size. Standard errors are clustered at the cluster level. Statistical significance is indicated as: *** p<0.01, ** p<0.05, * p<0.1.

Data Appendix

A.1 UniProtKB/Swiss-Prot

The Universal Protein Resource Knowledgebase (UniProtKB) is a comprehensive database of known proteins. A protein is composed of sequence of organic compounds called amino acids. Information for making a protein is stored in a gene's DNA, and, therefore, by translating the DNA sequence of a gene, scientists can determine the protein's existence and the sequence of amino acids that will appear in the protein. Protein sequences in UniProtKB are thus sourced by translating genes from major genome sequence databases.

UniProtKB is divided into two parts: Swiss-Prot (manually reviewed) and TrEMBL (computationally reviewed). Created in 1986, the Swiss-Prot database is extensively reviewed, maintained, and annotated by experts based on experimental results and literature review. As of October 2020, Swiss-Prot contains 563,552 protein entries. In contrast, TrEMBL was created in 1996 and houses computationally annotated protein entries. Once a protein from TrEMBL becomes manually reviewed, it is removed from TrEMBL and enters Swiss-Prot. TrEMBL was established in recognition that manual curation efforts cannot keep pace with the increased number of protein sequences resulting from genome sequence projects and contains nearly two hundred million entries.

To define the complete set of proteins at risk of being structurally characterized by structural biologists, I follow Perdigão et al. (2015) —a bioinformatics paper that descriptively mapped out which proteins' structures have been determined—and focus on the proteins in the Swiss-Prot database. While smaller than TrEMBL, using the Swiss-Prot database has several advantages. First, Swiss-Prot is one of the best datasets of proteins whose existence is experimentally proven (Perdigão et al. 2015); TrEMBL primarily contains proteins whose existence is only predicted. Second, since Swiss-Prot tends to include more well-described proteins, this allows me to ensure that I examine proteins that share a similar baseline level of documentation and thus similarly at risk of catching the attention of structural biologists, instead of looking at unreviewed proteins that may not even be real proteins. Third, Swiss-Prot's expertly curated annotation provides rich description of each protein, including its function, clinical impact, and sequence features, that allows me to develop a "predicted brightness" measure, as described in Section 6.4.

A.2 Linking Swiss-Prot to the Protein Data Bank

The PDB provides crosswalks to Swiss-Prot, allowing me to observe which proteins in Swiss-Prot have had their structures characterized in the PDB. However, the level of the crosswalk between an entry in the PDB and an entry in Swiss-Prot is not a many-to-one crosswalk as one might expect (a many-to-one, since a protein in Swiss-Prot can have its structure solved multiple times), but rather a many-to-many crosswalk (i.e., a single protein structure in the PDB can also be linked to multiple Swiss-Prot entries). This is because in the PDB, large protein structures are composed of discrete regions called "entities"; the crosswalk between the PDB and Swiss-Prot is at this entity level. Approximately 80% of the structures in the PDB has a single entity, while the remaining 20% has multiple entities and therefore linked to multiple Swiss-Prot entries. Whenever a single protein structure from the PDB links to multiple Swiss-Prot entries, I split the protein structure into fractions based on the percentage of amino acids each Swiss-Prot entry contributes to the protein structure.

A.3 MMseqs2

MMseqs2 is a software package that clusters databases of proteins and can be downloaded at https://github.com/soedinglab/MMseqs2 (Steinegger and Söding 2018; Hauser, Steinegger, and Söding 2016). MMseqs2 uses a greedy set cover algorithm and aims to create the fewest number of mutually exclusive clusters, given a set of proteins at a user-specified sequence similarity. In this paper, I chose the threshold of 30% sequence similarity, given that MR will likely be successful if the template and the target proteins share at least 30% sequence identity. If the sequence similarity falls below 30%, MR will be usually challenging, if at all possible, to implement (Schmidberger et al. 2010; Phenix). The algorithm takes the following steps:

1. MMseqs2 first computes all pairwise sequence identities between proteins in Swiss-Prot
2. MMseqs2 chooses a "representative" sequence, which is the protein with the highest number of neighbors that share at least 30% sequence similarity
3. MMseqs2 forms the first cluster with this representative sequence and all of its neighbors
4. MMseqs2 then looks at the remaining sequences and chooses the next representative sequence with the highest number of neighbors
5. MMseqs2 iterates through Steps 2-4 until all sequences belong in a cluster

This ensures that each member of a cluster shares at least 30% sequence similarity with the representative sequence of the cluster.[27] MMseqs2 is used by both Swiss-Prot and the PDB to cluster similar proteins.

A.4 Sample Construction

Using the MMseqs2 algorithm, I grouped all 563,552 proteins in Swiss-Prot into 74,017 mutually exclusive clusters, using 30% sequence identity threshold.

Restricting to clusters with at least one human protein: I then restricted the sample to clusters with at least one human protein (n = 13,150 clusters, which is equivalent to 161,392 proteins). There are two reasons for this restriction. First, restricting to clusters with at least one human protein ensures that all of the clusters in the final sample have a minimum baseline level of biological importance; one of the main goals of structural biology is to understand human biological processes and thus structural biologists are interested in human proteins and proteins similar to human proteins. Second, a nice feature of focusing on human proteins (and their similarity neighbors) is that this mitigates some of the concern of growing cluster size. The total number of possible proteins in the universe is essentially infinite,[28] and new protein sequences are continuously being discovered. However, all human proteins have been discovered by the early 2000s when the human genome project was completed; since one gene encodes one protein, and humans have approximately 20,000 genes, they also have 20,000 proteins.[29] Since the number of newly discovered human proteins have plateaued since the early 2000s when MR arrived, this alleviates the concern of whether human proteins are getting more structures because of MR or because there are simply more human proteins being discovered. To additionally address the

[27] A caveat is that while it is likely that all possible pairs of sequences within the cluster also share at least 30% sequence similarity with each other (since they are all similar to the representative sequence), this is not guaranteed. Mirdita et al. (2017) performed a cluster quality check that mitigates this concern; the authors computed the mean sequence identity among all possible pairs of sequences in a cluster and found that MMseqs2 indeed yielded clusters where all possible pairs of sequences shared on average >30% sequence similarity.

[28] Given that there are 20 different amino acids and an average protein has a sequence length of 200 amino acids, this amounts to 20^{200} possible proteins, which is larger than the number of electrons in the universe (Koonin, Wolf, and Karev 2002).

[29] This is called the "one gene, one protein" rule, which contributed to the 1941 Nobel Prize in Medicine. As explained in Section 4.1, by translating the DNA sequence of a gene, scientists can determine the protein's existence, and the sequence of amino acids that will appear in the final protein. Recently, the "one gene, one protein" rule has been challenged, as one gene may produce multiple proteins through, for instance, alternative splicing. Nonetheless, this paper follows the "one gene, one protein" rule since Swiss-Prot provides a non-redundant set of proteins, in that all proteins that are encoded by one gene in a species is folded into a single entry (including alternative splicing isoforms).

concern of changing cluster size, I also control for time-varying cluster size in some of my specifications.

Restricting to clusters born on or before 1998: For each cluster, I compute its discovery year by taking the earliest discovery year among the proteins in the cluster. (Discovery year is defined as the earliest known documentation of the protein's existence.) Since my panel starts in 1999, I only keep clusters that were born on or before 1998 in my sample.[30] This led to my final sample of 6,944 clusters of proteins.

[30] Alternatively, in a robustness analysis, I restricted the sample to clusters born on or before 2003 when MR arrived. This unbalanced sample consists of 12,294 clusters. Results remained similar.

Chapter 3

Is the Patent System Sensitive to Incorrect Information?

(with Janet Freilich)

Abstract

We investigate whether participants in the patent system are sensitive to information quality by examining how they treat inaccurate information. We use a novel approach to identify patents with inaccurate information: patent-paper pairs where the paper has been retracted and the corresponding patent contains the retracted material. Despite containing inaccurate information, we find that these patents are prosecuted and maintained by most applicants, are not rejected by examiners, and continue to be cited by some downstream readers after retraction. Insensitivity to inaccurate information may lead to erroneous decisions during examination and has implications for patent quality, disclosure, and knowledge flows.

1 Introduction

The patent system is only as good as the information it produces. If patents contain inaccurate information, examiners and readers may incorrectly rely on these statements. Patents may be erroneously granted to inventors who could not actually make the invention, and downstream patents may be erroneously rejected as preempted by poor-quality prior art.[1]

Consider the example of Theranos. By 2016, it was widely known that Theranos' vaunted technology—the ability to detect molecules in small amounts of blood—did not work. Yet in 2018, the U.S. Patent Office (USPTO) granted a Theranos patent that claims "a method of detecting an analyte in a . . . blood sample having a volume of less than about 500 μL" (U.S. Patent 10,156,579). The examiner did not question Theranos' claim. This reverberated downstream: in 2019, a different examiner cited a Theranos patent as evidence that a University of Arizona patent application claiming methods of detecting analytes in small drops of sweat (WO Patent App. 2018013579) was obvious, without acknowledging the public failure of Theranos' technology.

In this paper, we seek to understand whether the examples above are isolated incidents, or whether participants in the patent system are insensitive to information quality. We make two contributions: a) we develop a new approach to measure inaccurate information in patents and b) we combine empirical and qualitative analyses to find that examiners and inventors are often insensitive to inaccurate information.

Since there is no easy way to identify inaccurate information in patents, we propose a novel method: patent-paper pairs in which the paper has been retracted and the corresponding patent—which we term an "unsupported patent"—makes claims based on the retracted information and is thus unsupported by accurate data. Unlike papers, there is no mechanism to retract a patent,[2] so patents continue through the system even after the corresponding paper has been retracted. We identify the universe of all unsupported

[1] Examiners assess whether patent applications are novel and nonobvious by searching for earlier published disclosures called "prior art."

[2] Patents can be invalidated or found unenforceable in litigation, but this is not the same as retraction because the patent can nonetheless be cited as prior art against later applications, and a loss in litigation does not necessarily mean that information in the patent is wrong. See Appendix D.

patents in the biomedical sciences, and investigate, using matched controls, whether patents are treated differently once material in the patent is publicly acknowledged to be incorrect. Our results from 86 unsupported patent families (589 individual granted patents or patent applications) and 86 control patent families (576 individual patents or applications) suggest that the patent system largely does not react to incorrect information, either during examination or downstream.

During prosecution of the unsupported patents themselves, we find that applicants overwhelmingly (95%) did not disclose the retraction to the USPTO, and, in 63% of families, continued prosecuting or maintaining patents after the corresponding retraction. Examiners in turn almost always (93%) failed to discover that the application contained retracted material and did not reject unsupported patents.

We then turn to the downstream impact of unsupported patents. Confirming other studies of retractions in science (Furman et al. 2012; Azoulay et al. 2015; Azoulay et al. 2017; Jin et al. 2019; Lu et al. 2013), we find that on the paper side, citations to retracted papers dropped significantly after retraction. In contrast, the reaction by the patent system was muted. While unsupported patents received fewer citations from downstream applicants after retraction, this effect was smaller than what was observed on the paper side, and citations from examiners did not change. Further, although some examiners cited unsupported patents as a justification for rejecting downstream patents as obvious or not novel, downstream applicants did not fight back. Only 0.6% challenged the rejection on the basis that the cited prior art contained retracted material.

This paper contributes to a large literature on the prevalence and effects of poor-quality patents (Jaffe and Lerner 2004; Lanjouw and Shankerman 2004; Bessen and Meurer 2008; de Rassenfosse et al. 2016). While the previous literature has predominantly focused on patents that should never have been granted because they are either obvious or not novel, this paper joins a smaller literature that focuses on a different type of poor-quality patents: patents granted incorrectly because they contain problematic information (Ouellette 2017; Freilich 2019; Freilich and Ouellette 2019; Freilich 2020). These patents are of great concern because they may be acquired by patent acquisition entities and asserted in a manner that taxes innovators (Feng and Jaravel 2020), worsens patent thickets (Cohen 2004; Cockburn and MacGarvie 2009), and deters genuine innovators in the area covered by the patent (Government Accountability Office 2016; Tucker 2014).

Although some may argue that it is more efficient to eliminate these patents ex-post through litigation, that patents with some incorrect information may still contain other accurate data, or that not all incorrect information requires the patent's invalidation,[3] patents with incorrect information may still be damaging. Inventors may waste resources relying on or trying to replicate incorrect information. Importantly, inventors may be discouraged altogether from follow-on research if they believe their patents would be rejected as obvious based on incorrect prior art—in which case, there would be no follow-on inventions in the first place and litigation would not occur. This paper does not address all types of informational inputs into patents (some of which may be correct, even for unsupported patents), but the consequences of incorrect information in the disclosure are potentially severe.

While this paper focuses specifically on the biomedical sciences and on the relatively few patents that incorporate retracted material, patents with incorrect information are ubiquitous (Freilich 2020). If the patent system is not sensitive to retracted information—where there is an easily accessible statement that the information is wrong—the patent system is likely also insensitive to other types of incorrect information. Such insensitivity leads to errors in the patent system, spreads poor-quality information to the public, and damages the integrity of the patent system.

2 Institutional Context

Patents contain a substantial amount of information about the invention. Information quality is vital to a functioning patent system at two stages: examination and downstream knowledge flows. Examiners must determine whether an applicant invented something new and useful, and whether the applicant disclosed sufficient information about the invention to teach others how to make the technology. Examiners necessarily rely on the information

[3] Because the standards for scientific retraction and patent invalidity are different, a patent may be valid even if some of the invention's embodiments do not work. See Appendix D.

provided in the application to assess the invention and have access to little evidence beyond the words of the patent (Freilich 2021).

If patent applications provide incorrect instructions on how to make an invention or falsely claim that a technology works, examiners who uncritically rely on information in the application may erroneously grant a patent. Because patent claims are always broader than the underlying data supporting the invention,[4] incorrect data can support claims that cover both a non-functional invention *and* related inventions that (unbeknownst to the applicant) do work. Thus, examiner insensitivity to information quality may result in patentees being granted exclusive rights over useful technologies that the patentees did not invent.

At the downstream stage, the information in a patent is perhaps even more important. First, this information becomes "prior art" to later applications and can be used in rejections for obviousness or lack of novelty (Sampat 2010). Such a rejection is only correct if the information upon which it is based is correct.[5] Erroneous rejections may lead inventors to mistakenly narrow or abandon a meritorious patent application, dampening incentives for innovation. Further, an important purpose of the patent system is to publicly disseminate knowledge that might otherwise be kept private (Ouellette 2012; Sampat 2018). High information quality is key to the disclosure function of patents.

3 Data

1. **Retracted papers:** We began by retrieving all retracted papers from PubMed that were published between 2001 and July 2019. We focused on papers that were indexed in Medline, which exclusively focuses on the life sciences, and matched these papers to data provided by Retraction Watch, which documents the reasons behind the retractions (Oransky and Marcus 2010). We excluded papers that are not original research articles, such as reviews, leaving us with a set of 4,322 retracted papers. Finally, we only kept papers that specified that the retraction occurred because the information in the paper was incorrect; we excluded papers that were retracted due

[4] For example, if a scientist discovers a molecule that reduces tumor in mice, she is likely able to get a patent that claims use of the molecule for any purpose.

[5] More detail is provided in Appendix D.

to reasons that do not cast doubt on the veracity of the retracted information (e.g., plagiarism).

2. **Identifying unsupported patents:** We identified all U.S. patent applications with (1) inventors who share the same name as the first or last author of the focal retracted publication and (2) filing dates within +/- 2 years of paper publication. We then examined which of these patents have a retracted paper pair in two steps. First, we used a word similarity algorithm (described in Appendix A) and identified potential pairs where the patent specification contained at least 90% of the words in the retracted paper abstract. Second, we manually verified that these potential pairs are indeed true patent-paper pairs. Specifically, we read all retraction notices to confirm that the patents' specifications and claims incorporated the retracted material. From our sample of 4,322 retracted papers, we identified 107 patent-paper pairs (101 papers; 96 patents).

   The following example gives a sense of the closeness of the match between retracted material in the paper and the corresponding patent. In 2012, *Cell* published a finding that the compound norspermidine prevented formation of biofilms (communities of bacteria that are resistant to antimicrobials). Shortly before the paper was published, the authors filed a patent application claiming methods of reducing biofilm formation with norspermidine or similar compounds. In 2015, after another group of scientists challenged the study, the authors of the original paper retracted their findings, explaining that after repeating their experiments, "the new results can no longer support our original conclusions." The retracted material was precisely the central finding and was squarely encompassed by claim 1 of the patent.

3. **Controls:** Following papers that have studied the impact of retractions in science (Furman et al. 2012; Azoulay et al. 2015; Jin et al. 2019), we sought matched controls. We found control, non-retracted patent-paper pairs, by matching on both paper and patent characteristics. We first began by identifying all non-retracted, original research papers in Medline that were published in the same year and journal as the retracted papers of our patent-paper pairs (n = 127,271 papers). We then gathered control papers with associated patents to identify patent-paper pairs. We

used the same word similarity algorithm described above and identified potential control pairs with overlap score greater than 90% that share the same primary technology class as the unsupported patents, as defined by the International Patent Classification (IPC) system (n = 4,550 pairs). We manually reviewed potential control pairs to confirm that they are indeed pairs. Specifically, for each retracted patent-paper pair, we sorted the potential control pairs by their word overlap score and reviewed them in descending order of the score until we identified a true control pair. After this procedure, we were able to find control patent-paper pairs for 86 of our retracted patent-paper pairs. More details can be found in Appendix A.

4. **Identifying patent family members:** Because members of a patent family often contain identical or very similar specifications, our unit of analysis is a patent family. For our 86 unsupported and 86 control patents identified above, we sought all family members. Our final sample consists of 86 unsupported patent families (which include 589 individual granted patents and patent applications) and 86 control patent families (576 individual patents and applications).

5. **Patent data:** Filing year and inventor names were obtained from Reed Tech's Bulk Data Downloads. Data on prosecution dates and events were obtained from the USPTO's Patent Application Information Retrieval (PAIR) system. Information on technology class was obtained from PatSnap. Forward citations, maintenance fee payment records, priority dates, and family members were obtained from Google Patents.

   To obtain a more granular understanding of the data, we also manually read all correspondences between the patent applicant and USPTO for both unsupported patents and downstream applications rejected over unsupported patents.

Appendix B provides more details on the data sources.

4 Empirical Strategy

To assess whether the patent system is sensitive to inaccurate information, we studied participants involved in different stages of the patent system and their treatment of unsupported patents as compared to controls.

We first investigate the reaction of applicants and examiners of unsupported patents:

$$\mathbf{Y}_i = \beta_0 \ + \beta_1 \mathit{Treated}_i + \mathbf{X}_i + \epsilon_i \tag{1}$$

$\mathbf{Y}_i$ is an indicator variable for a) whether the applicant prosecuted any application or paid maintenance fees for any granted patent in patent family i after retraction; and b) whether the examiner rejected any application in patent family i for lack of either enablement or written description after retraction. $\mathit{Treated}_i$ is an indicator variable for whether the family is an unsupported or a control patent family. $\mathbf{X}_i$ is a set of controls, including retraction-year fixed effects, age-at-retraction fixed effects, technology class fixed effects, or unsupported-control match fixed effects.[6]

To examine the impact of retraction on downstream applicants and examiners, we employ a staggered difference-in-differences approach to investigate whether citations to unsupported patent families decrease post-retraction relative to control families. Note that we collect information on citations to both applications and granted patents in the family; applications, including those that are eventually abandoned, are still frequently cited (Cotropia and Schwartz 2020). We estimate the following regression equation:

$$\mathit{Cites}_{it} = \beta_0 \ + \beta_1 \mathit{Post\_Retraction}_{it} + \beta_2 \mathit{Post\_Retraction}_{it} \times \mathit{Treated}_i + f(\mathit{age}_{it}) + \delta_t + \gamma_i + \epsilon_{it} \tag{2}$$

[6] Retraction-year fixed effects consist of full set of seventeen indicator variables, (2002-2019). Age-at-retraction fixed effects consist of eight indicator variables, with the age one indicator including all prior age indicators and the age eight indicator variable including all subsequent age indicators. Technology class fixed effects consist of six indicators. Unsupported-control match fixed effects are indicators for every unsupported family and its matched control counterpart; since our sample consists of 1-to-1 matching of unsupported and control families that match on covariates such as technology class and the control family inherits the counterfactual retraction year from its unsupported counterpart, specifications with unsupported-control match fixed effects do not include retraction-year or technology class fixed effects.

$Cites_{it}$ is the number of citations patent family i receives in year t, $Post\_Retraction_{it}$ is an indicator that is zero before retraction and one after retraction, $Post\_Retraction_{it} \times Treated_i$ is an indicator that turns one after retraction for only unsupported patent families. β_1 controls for any leads and lags around the retraction event that are common to both unsupported and control patent families (Jaravel et al. 2018), while β_2 is our coefficient of interest and can be interpreted as the impact of retraction on patent citations (average treatment effect on the treated). Standard errors are clustered at the patent family level.

δ_t are calendar year fixed effects that control for any calendar year shocks that impact citations of all patent families in a given year,[7] while γ_i are patent family fixed effects that control for patent family traits that could affect citations (e.g., technology class). $f(age_{it})$ are indicator variables for patent family age that control for any lifecycle effects (for instance, newer patents may be cited more than older patents).[8] ϵ_{it} is the error term.

To understand the dynamic effects of retraction, we turn to Equation 3:

$$Cites_{it} = \beta_0 \ + \sum_{j=-n}^{N} \beta_1^j \alpha_{it}^j + \sum_{j=-n}^{N} \beta_2^j \alpha_{it}^j \times Treated_i + f(age_{it}) + \delta_t + \gamma_i + \epsilon_{it} \tag{3}$$

Equation 3 is a modified version of Equation 2 and includes separate indicator variables for each year before and after retraction, α_{it}^j, where the subscript j is the window of years we are interested in before and after the retraction year. For our main analyses, we looked at the window of five years around retraction.

Recent work has suggested potential problems with staggered difference-in-differences designs due to treatment heterogeneity (Goodman-Bacon 2021; Sun and Abraham 2021). Sun and Abraham (2021) proposes a new estimator that addresses this problem, which we use in a robustness analysis in Appendix C.

Finally, to conduct our analyses, it is important to understand whether the retraction occurred before or after our outcomes of interest. For instance, to investigate examiner's

[7] Calendar year fixed effects consist of twenty-three indicators (1999-2021), with the 1999 indicator including all prior calendar years.

[8] Patent family age was defined as calendar year minus the patent family's priority year. Due to an imperfect method of determining exact citation dates (see Appendix B), the year of first citation precedes the priority year for some patent families. Patent age fixed effects consist of fourteen age indicators, with the age zero indicator including all prior age indicators and the age thirteen indicator including all subsequent age indicators.

behavior, the retraction must occur before the patent arrives in the examiner's desk; if the retraction occurs after the patent had already been granted or rejected, then this patent should not be included in our analysis on examiner rejection.

Appendix Figure A.1 shows what fraction of our sample of patent families experienced a (real or inherited counterfactual) retraction at each stage of a patent life cycle. We selected the appropriate subsamples for each of our outcomes, such that the timing of the retraction could have impacted the outcome (Appendix Figure A.1). This resulted in the following: 170 families (85 unsupported and 85 control families) were used for our analysis on the impact of retraction on downstream citations, 126 families (68 unsupported and 68 control) were used for our analysis on applicant's decision to prosecute/maintain the patent, and 100 families (50 unsupported and 50 control) were used for our analysis on examiner's decision to reject or grant the patent.

5 Results

Table 2 reports summary statistics for our sample of retracted and control pairs; Panel A shows the patent side of the pairs, while Panel B shows the paper side. Although the retracted and control pairs were only matched on paper publication year, paper journal, and patent technology class, our sample is similar on other covariates, such as priority year, whether the patent family is owned by non-industry institutions, and whether the family is triadic.[9] None of the unsupported patent families were involved in litigation. As shown in Panel B, 60% of the retracted papers were retracted due to error or unreliable results, while 38% were retracted due to fraud or misconduct. The papers in our sample were published in high-ranked journals (on average in the 87th percentile in terms of journal impact factor), with more than 30% of the papers from top journals, such as the *New England Journal of Medicine*, *Nature*, *Nature Medicine*, *Science*, and *Cell.*

[9] A triadic patent family indicates that the applicant filed the application at the USPTO, the European Patent Office, and the Japan Patent Office, indicating that the applicant considers the invention to be potentially of high value.

5.1 How Do Applicants Treat Unsupported Patents?

At the outset, applicants overwhelmingly opt not to tell examiners about the retraction, despite patent law's duty of disclosure.[10] Our small sample size allowed us to manually review all correspondences (prosecution histories) between the examiner and applicant for unsupported patents. Only three applicants disclosed that their application contained retracted material; two examiners promptly rejected the application while one applicant preemptively amended the claims to remove the retracted material (though the material remained in the disclosure), after which the application was granted.

Applicants of unsupported patents treated their patents differently than applicants of control patents, suggesting some sensitivity to inaccurate information, although more than half of applicants continued to invest resources in unsupported patents and continue legal proceedings. For families that were being prosecuted or maintained at the time of the retraction, 63% of applicants of unsupported patents continued to prosecute or maintain the patents,[11] compared to 84% of control applicants. Columns 1 and 2 of Table 3 report estimations from Equation 1 and show that the probability of being maintained or prosecuted after retraction declines by 19% points. This negative reaction by the applicants of unsupported patents is perhaps expected since applications include at least one inventor who is also an author on the corresponding retracted paper and thus applicants must be aware of the retraction. More surprising, however, is that over half of applicants nonetheless continued prosecuting and maintaining unsupported patents—spending money to keep the patents alive despite the retraction.

5.2 How Do Examiners Treat Unsupported Patents?

Examiners appear overwhelmingly unaware that unsupported patent applications contain retracted material. After reviewing all examiner-applicant communications, we found only four examiners who mentioned the retraction, each of whom rejected the application.

However, examiners might reject an application because of the retraction without outright mentioning the retraction. If this was the case, examiners might reject the application either

[10] Applicants have a duty to disclose all material information to the examiner. This arguably includes the retraction, as discussed in Appendix D.

[11] Owners of granted patents must pay maintenance fees to avoid abandoning the patent.

for lack of enablement (on the ground that retracted material cannot teach others how to make and use the invention) or lack of written description (on the ground that retracted material indicates that the inventor was not in possession of the invention). Our results show that examiners are not rejecting applications that contain retracted information. In fact, surprisingly, as shown in Columns 3 and 4 of Table 3, we find that examiners appear to be 16% points *less* likely to reject unsupported patents, relative to controls (although this is imprecisely estimated).

5.3 How Do Downstream Applicants and Examiners Treat Unsupported Patents?

Downstream Citations

To understand the downstream impact of unsupported patents, we ask how citations to applications and granted patents change after the corresponding retraction. Figure 1 shows the mean annual citations to the patent-paper pairs analyzed in this study. Citations to retracted papers drop steeply after retraction, while citations to the corresponding unsupported patent families remain essentially unchanged.

We turn from the descriptive patterns to a difference-in-differences analysis. Table 4 reports the estimations of Equation 2.[12] In Column 1, the outcome is the annual number of total citations, while Columns 2 and 3 decompose the citation counts by whether the citation was added by the examiner or the applicant of the citing patent. Columns 4-6 report Poisson specifications. Interestingly, for examiner citations, the magnitude of the point estimates are positive. As shown in Column 2, unsupported patent families experienced an *increase* of 0.22 annual number of examiner citations after retraction relative to controls, which is a 14% increase from a mean of 1.61.[13] In Poisson specification, as in Column 5, retraction was associated with a 13% increase in examiner citations ($e^{0.119} - 1$). As for applicant citations, we find the expected negative effect due to retraction. Unsupported patent families experienced a decrease of 0.53 annual number of applicant citations (Column

[12] For our main analyses, our sample is unbalanced, as some patent families have fewer pre- or post-periods. Appendix A and Appendix Table A.1 detail a robustness check where we narrow our sample to a smaller but balanced sample.

[13] Appendix Figure A.2 shows the distribution of annual citations.

3) or 18% decrease (Column 6). While our results are imprecisely estimated, the lower bounds of our results suggest that downstream examiners and applicants do not appear to react as negatively to retraction, particularly in contrast to the scientific publication system where citations to retracted papers declined significantly. For instance, as discussed earlier, the annual number of examiner citations to unsupported patent families increased by 0.22, and this point estimate has a 95% confidence interval of [-0.39, 0.84]. On the paper side, Column 1 of Appendix Table A.2 shows that citations to retracted papers declined by -10.6, and this point estimate has a confidence interval of [-17.5, -3.7]. The upper bound of the impact of retraction on papers is more negative than the lower bound of the impact of retraction on corresponding patents. This pattern holds true across all specifications, suggesting that the patent system is less sensitive to retraction than scientific publishing.

Figure 2 plots the event study graphs from estimating Equation 3. There are no noticeable pre-trends before retraction, and downstream examiners and applicants do not appear to react strongly to the retraction event. As a robustness analysis, in Appendix Figure A.3, we use an alternative estimator developed by Sun and Abraham (2021) that accounts for treatment heterogeneity.

Response to Rejections

When examiners reject downstream applications as anticipated or obvious and cite to an unsupported patent as evidence that the invention was previously disclosed, this rejection is arguably incorrect. A retraction suggests that the unsupported patent did not actually disclose the invention, and thus that the downstream patent is novel. Further, a retraction may indicate that the scientific community believed the invention did not work, and thus that the downstream patent is nonobvious. Downstream applicants could therefore argue that a rejection based on an unsupported patent is incorrect.

We read all communications between examiners and downstream applicants where an unsupported patent was cited in rejecting the downstream application. Only three (0.6%) applicants responded to the office action with mention that the prior art contained retracted material. Although downstream applicants are highly incentivized to find and mention the upstream retraction, they do not and appear to be insensitive to the quality of information in cited patents.

6 Discussion

6.1 Mechanisms

Participants in the patent system are largely insensitive to inaccurate information, leading to dissemination and use of incorrect information. We suggest mechanisms for this insensitivity below.

Applicants: While applicants of unsupported patents are less likely to prosecute and maintain their patents relative to controls, more than half still continued to invest resources into their unsupported patents; several factors may contribute to their continued prosecution and maintenance. First, some authors may believe in their work even after the retraction. For instance, after one inventor's application was rejected by a patent examiner because of the corresponding paper's retraction, the inventor claimed that "[b]y definition, a retraction equates to the data having never been published…It is not a declaration that the data is incorrect" (U.S. Patent App. No. 20150110749).

Second, high rates of continued prosecution may result from a breakdown in institutional communication. Inventors are often not intimately involved in patent prosecution. Rather, they delegate that duty to their attorney and/or to an institutional party such as a technology transfer office (TTO). While the inventor is aware of the retraction, the attorney and TTO making decisions to continue prosecuting or maintaining the patent may remain unaware. Supporting this institutional miscommunication hypothesis, one third of the papers in our sample were retracted after an institutional investigation—yet the corresponding patents were prosecuted and maintained at similar rates to those where an institutional investigation did not occur (68% vs. 61%, t-statistic: 0.56). This suggests that the portion of the institution responsible for the investigation was not communicating with the portion of the institution responsible for prosecuting the patents.

Moreover, applicants' reluctance to abandon unsupported patents may reflect the patents' value. Some retractions may indicate a partially inoperable—but partially operable—technology. Even patents that cover no operable technology can be monetized in nuisance litigation or provide value as part of a large patent portfolio (Hsu and Ziedonis

2008). Perhaps because of these possibilities, Theranos' patent portfolio retained value to investors even after its technology was entirely discredited (McKenna 2018).

Yet, there is little evidence that the technological accuracy of the invention is linked to decisions to continue prosecuting or maintaining. We classified retracted papers based on whether the retraction notice retracted the entire paper or only a portion of the paper. Papers in the latter category might disclose operable discoveries. However, applicants on patents corresponding to partially retracted papers prosecuted and maintained their patents at rates comparable to completely retracted papers (74% vs. 59%, *t*-statistic: 1.09).

Upstream examiners: Why do examiners grant patents containing unsupported material? In fact, our results show that examiners seem less likely to reject unsupported patent families, relative to controls. One possibility behind this counterintuitive result is that inventors who are willing to manipulate data do so thoroughly and produce more comprehensive support for their claims than would otherwise occur, decreasing the likelihood of examiner rejection. While our analysis cannot comment on the exact mechanism, our results broadly suggest that examiners are not rejecting applications that contain retracted material. Furthermore, though we cannot exclude the possibility that examiners were aware of the retraction but did not feel a rejection was merited, we believe this is unlikely. The more plausible explanation is that examiners did not know about the retraction. Indeed, examiners are pressed for time (Merges 1999) and do not have resources to replicate experiments themselves. Although examiners do independently search the field of the invention, they do so in the context of discovering prior art, and therefore truncate searches at the priority date of the application—usually before the retraction. The retraction notice may therefore not come up in an examiner search.

Downstream examiners: Lack of knowledge is also likely why downstream examiners continue to cite unsupported patents. Unlike papers, unsupported patents have no visual notice indicating retraction, providing no warning to citing examiners. Further, while examiners have at least a bachelor's degree in the scientific field in which they work, they may not be sufficiently familiar with the scientific literature to recognize that unsupported patents contain retracted material.

Downstream applicants: Downstream applicants appear to cite unsupported patents less after retraction, while downstream examiners do not appear to react to retraction. We find this result to be consistent with the larger patterns of our results where, in general, applicants are more sensitive to incorrect information than examiners, presumably because applicants are more familiar with academic work.

More strikingly, however, downstream applicants whose patent application has been rejected over unsupported prior art do not raise the presence of retracted material in response to the rejection. Poor communication between parties involved in patent applications may be again a likely explanation. While the applicants themselves are deeply involved in the field of the invention and may (perhaps should) notice that the material in the unsupported patent has been retracted, as a functional matter, attorneys—not inventors—are often the ones answering office actions, and attorneys would be less likely to recognize an unsupported patent.

This insensitivity to incorrect information by the patent system is notable, particularly in contrast to scientific publishing. It is worth noting, however, that standards for scientific retraction and patent invalidity are different, as explained in Appendix D. Moreover, the standards for citing in science and the patent system are also different. In science, the sharp drop in citations after retraction is all the more stark, given that there are still legitimate reasons for downstream papers to cite a retracted paper—for instance, "negative" citations, where downstream papers dispute the retracted paper (Catalini et al. 2015). Even with the possibility of negative citations, it is clear that the scientific community tracks and shuns retracted papers by no longer citing them. On the other hand, in the patent system, downstream readers, especially examiners, appear largely unaware of the incorrect information and continue to cite unsupported patents, when they arguably should not—harming the flow of knowledge, despite the constitutional mandate of the patent system to "promote the progress of science" (U.S. Constitution Art I, Sec. 8, Cl. 8).

6.2 The Scope of Incorrect Information

One may argue that most patents are low value and that important patents that contain inaccurate data will be litigated; thus, battling bad data in the patent system ex-post may be the most efficient way to combat poor information. Furthermore, since the incidence of unsupported patents is low, perhaps one could accept these patents as acceptable costs of having the patent system.

In response, we emphasize that unsupported patents—patents that contain retracted material—represent just a small fraction of the poor-quality data in the patent system. While we focused on unsupported patents because retractions permit clear-cut classification of incorrect information, prior literature suggests that the problem of incorrect information is much more widespread as most experiments reported in patents have major methodological flaws that are linked to irreplicability (Freilich 2020; Ouellette 2012; Ouellette 2017). For instance, Freilich (2019) examined the universe of life science patents and found that half of them contain no experimental data at all and rely only on conjectures, and, of the patents that do disclose experimental data, approximately one quarter contain fictional—speculative—experiments. In addition, we note that 18% of the 4,322 retracted papers in our sample were cited by patents.[14] While a citation linkage does not always mean that retracted information is incorporated into a patent, it does suggest that the problem of retracted information in patents extends far beyond the patent-paper pairs studied in this paper.

Furthermore, there is reason to believe that despite their low incidence, the unsupported patents in our sample may be economically important. In Appendix E, using a measure of patent value developed by Kogan et al. (2017), we conduct a prediction exercise to assess the economic value of the unsupported patents by adopting a method from Hsu et al. (2021). We approximate that the unsupported patents are on average worth $7 million.[15] While this is a back-of-the-envelope exercise, this does suggest that the unsupported patents in our sample may be economically valuable. This is perhaps not surprising given that these patents are life science patents (which tend to be higher value as compared to patents from other industries), and many have a corresponding paper published in a high-impact journal from prestigious institutions. Prior work has also shown that patents that directly cite science have higher economic value than patents that do not (Krieger et al. 2021).

Finally, the patent system's failure to recognize inaccurate data ex-ante—before litigation—may undercut its ability to properly incentivize downstream innovation. Potential innovators may be deterred from innovation in the first place due to existing bad patents; in that case, litigation would not come into play at all. Patents granted on the

[14] Marx and Fuegi (2020) provides data on patents that cite scientific articles.

[15] In 1982 dollars. In Appendix E, we benchmark this number against other references.

basis of poor-quality information to patentees who could not actually develop the technology may therefore tax innovators who could make legitimate progress in the area covered by the patent because it is more difficult both to do downstream research and to get a patent as a subsequent entrant in a field (Roin 2008). Further, even patents that will be found invalid in litigation can be used in nuisance suits, a tactic exploited by so-called "patent trolls" (Cohen et al. 2019; Feng and Jaravel 2020).

7 Conclusion

The patent system is largely insensitive to inaccurate information, and we believe this is likely due to lack of awareness and breakdown in institutional communication. While we do not think patent examiners are well positioned to police information quality for all patents, nor would it be cost-effective for them to do so, we recommend that applicants, particularly universities, conduct internal checks to avoid filing patents with retracted information. Furthermore, downstream applicants should check for the presence of obviously incorrect information in prior art references used to reject their applications.

References

Alcácer, J. and M. Gittelman (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics 88*(4), 774–779.

Azoulay, P., A. Bonatti, and J. Krieger (2017). The career effects of scandal: Evidence from scientific retractions. *Research Policy 46*(9), 1552–1569.

Azoulay, P., J. Furman, J. Krieger, and F. Murray (2015). Retractions. *The Review of Economics and Statistics 97*(5), 1118–1136.

Bacchiocchi, E. and F. Montobbio (2009). Knowledge diffusion from university and public research. A comparison between US, Japan and Europe using patent citations. *The Journal of Technology Transfer 34*(2), 169–181.

Bessen, J. and M. Meurer (2008). *Patent Failure: How Judges, Bureaucrats, and Lawyers Put Innovators at Risk*. Princeton University Press.

Catalini, C., N. Lacetera, and A. Oettl (2015). The incidence and role of negative citations in science. *Proceedings of the National Academy of Sciences 112*(45), 13823–13826.

Cockburn, I. and M. MacGarvie (2009). Patents, thickets and the financing of early-stage firms: Evidence from the software industry. *Journal of Economics & Management Strategy 18*(3), 729–773.

Cohen, L., U. Gurun, and S. D. Kominers (2019). Patent trolls: Evidence from targeted firms. *Management Science 65*(12), 5461–5486.

Cohen, W. (2004). Patent and appropriation: Concerns and evidence. *The Journal of Technology Transfer 30*(1-2), 57–71.

Cotropia, C. and D. Schwartz (2020). The hidden value of abandoned applications to the patent system. *Boston College Law Review 61*, 2809–2867.

de Rassenfosse, G., W. Griffiths, A. Jaffe, and E. Webster (2016). Low-quality patent in the eye of the beholder: Evidence from multiple examiners. *NBER Working Paper No. 22244*.

Feng, J. and X. Jaravel (2020). Crafting intellectual property rights: Implications for patent assertion entities, litigation, and innovation. *American Economic Journal: Applied Economics 12*(1), 140–81.

Freilich, J. (2019). Prophetic patents. *U.C. Davis Law Review 53*, 663–731.

Freilich, J. (2020). The replicability crisis in patent law. *Indiana Law Journal 95*, 431–483.

Freilich, J. (2021). Matching and digging: Evidentiary analysis at the patent office. *Fordham Law Review* (Forthcoming).

Freilich, J. and L. L. Ouellette (2019). Science fiction: Fictitious experiments in patents. *Science 364*(6445), 1036–1037.

Furman, J., K. Jensen, and F. Murray (2012). Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine. *Research Policy 41*(2), 276–290.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics 225*(2), 254–277.

Government Accountability Office (2016). Intellectual property: Patent office should define quality, reassess incentives, and improve clarity. *GAO-16-490*.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2001). The NBER patent citation data file: Lessons, insights and methodological tools. *NBER Working Paper No. 8498*.

Hsu, D., P.-H. Hsu, T. Zhou, and A. Ziedonis (2021). Benchmarking u.s. university patent value and commercialization efforts: A new approach. *Research Policy 50*.

Hsu, D. and R. Ziedonis (2008). Patents as quality signals for entrepreneurial ventures. *Academy of Management Proceedings*.

Jaffe, A. and J. L. Lerner (2004). *Innovation and Its Discontents: How Our Broken Patent System is Endangering Innovation and Progress, and What to do About It.* Princeton University Press.

Jaffe, A. and M. Trajtenberg (1999). International knowledge flows: Evidence from patent citations. *Economics of Innovation and New Technology 8*(1-2), 105–136.

Jaravel, X., N. Petkova, and A. Bell (2018). Team-specific capital and innovation. *American Economic Review 108*(4-5), 1034–73.

Jin, G. Z., B. Jones, S. F. Lu, and B. Uzzi (2019). The reverse Matthew Effect: Consequences of retraciton in scientific teams. *The Review of Economics and Statistics 101*(3), 492–506.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman (2017, 03). Technological Innovation, Resource Allocation, and Growth*. *The Quarterly Journal of Economics 132*(2), 665–712.

Krieger, J. L., M. Watzinger, and M. Schnitzer (2021). Standing on the shoulders of science. *Harvard Business School Working Paper No. 21-128.*

Lanjouw, J. and M. Shankerman (2004). Patent quality indicators and research productivity: Measuring innovation with multiple indicators. *The Economic Journal 114*(495), 441–465.

Lu, S. F., G. Z. Jin, B. Uzzi, and B. Jones (2013). The retraction penalty: Evidence from the web of science. *Nature Scientific Reports 3*, 1–3.

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

McKenna, F. (2018). The last days of Theranos – the financials were as overhyped as the blood tests. *MarketWatch*.

Nicholas, T. (2008). Does innovation cause stock market runups? evidence from the great crash. *American Economic Review 98*(4), 1370–96.

Oransky, I. and A. Marcus (2010). Why write a blog about retractions.

Ouellette, L. L. (2012). Do patents disclose useful information? *Harvard Journal of Law and Technology 25*(2), 545–608.

Ouellette, L. L. (2017). Who reads patents? *Nature Biotechnology 35*(5), 421–424.

Roin, B. (2008). Unpatentable drugs and the standards of patentability. *Texas Law Review 87*, 503–570.

Sampat, B. (2010). When do applicants search for prior art. *The Journal of Law and Economics 53*(2), 399–416.

Sampat, B. (2018). A survey of empirical evidence on patents and innovation. *NBER Working Paper No. 25383*.

Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics 225*(2), 175–199.

Tucker, C. E. (2014). Patent trolls and technology diffusion: The case of medical imaging. *SSRN Working Paper*.

Patents

Gibbons, Ian. 2018. “Methods for the Detection of Analytes in Small-Volume Blood Samples.” US10156579.

Holmes, Elizabeth. 2012. ”Point-of-Care Fluidic Systems and Uses Thereof.” US8283155.

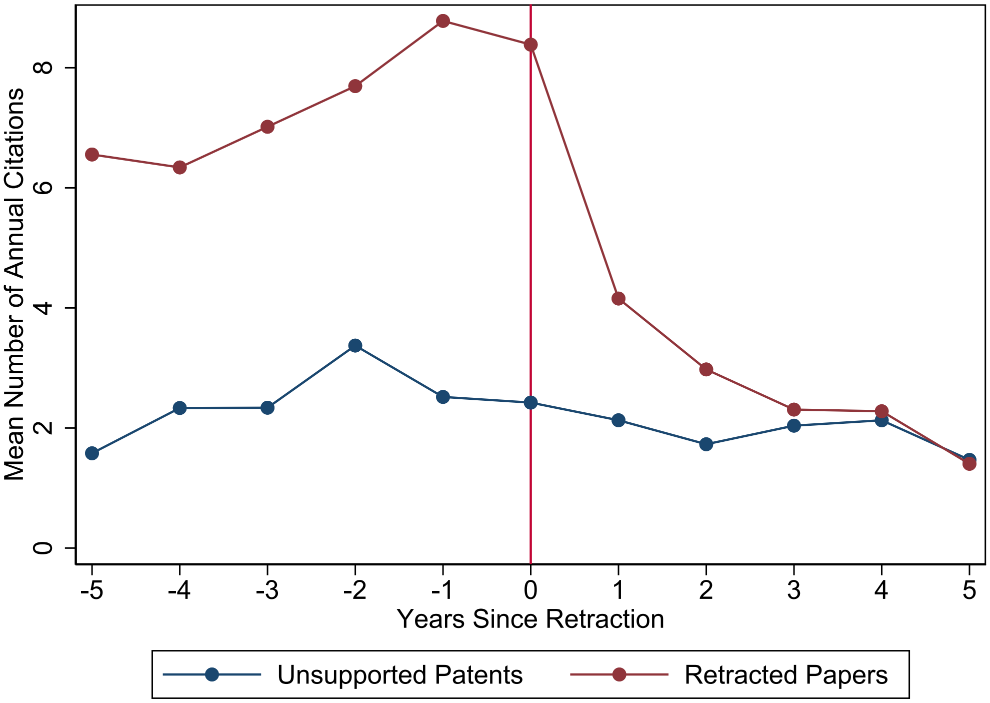
Kaplan, David and Fiorenzo Omenetto. 2017. “Compositions and Methods for Stabilization of Active Agents.” US2017025889.

Katchman, Benjamin. 2017. “Sweat as a Biofluid for Analysis and Disease Identification” WO2018013579A1.

Losick, Richard. 2013. “Methods and Compositions for Treating Biofilms.” US20140056951.

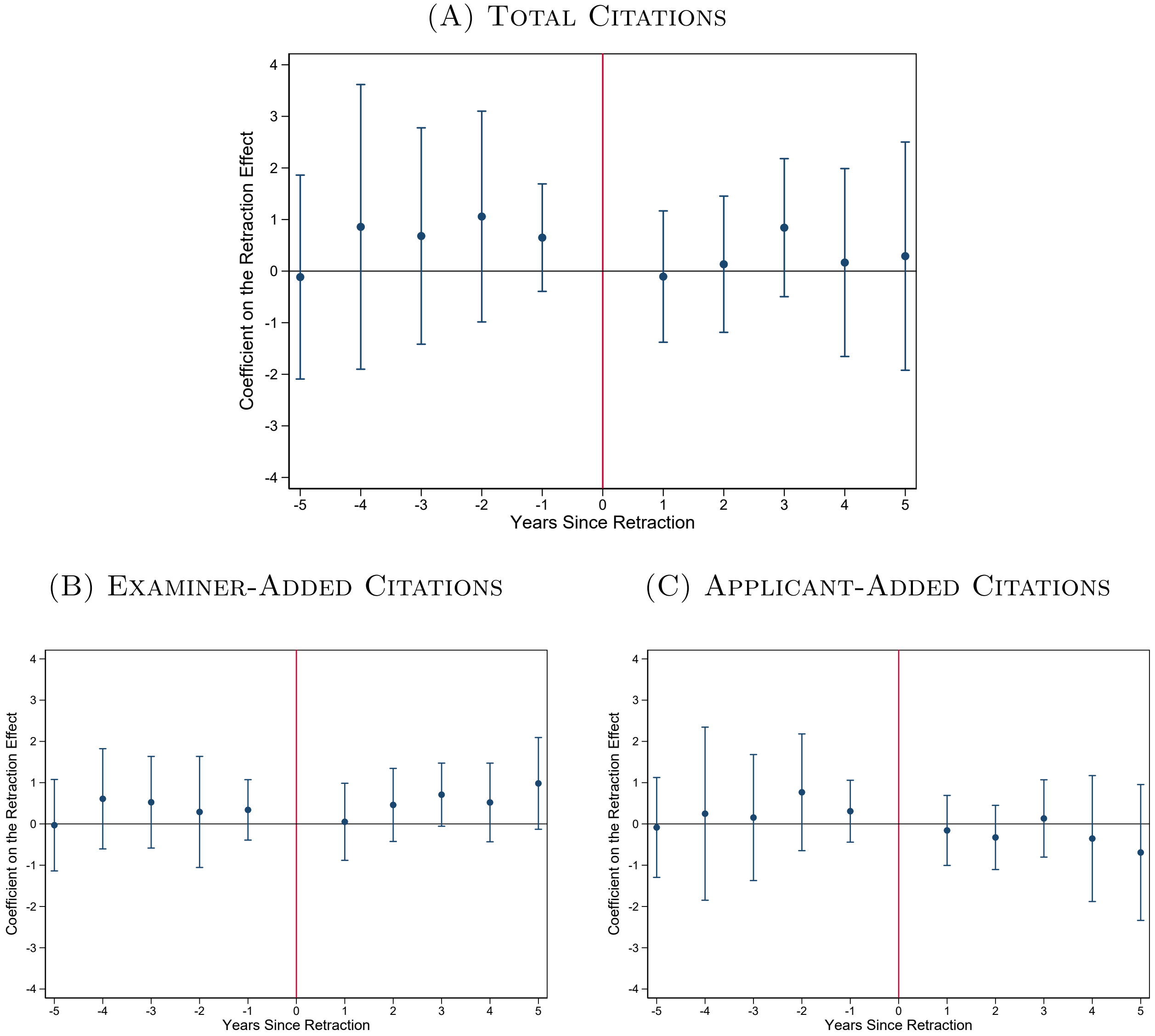
Vacanti, Charles. 2013. “Generating Pluripotent Cells de Novo.” US20150110749.

Figure 1: Unsupported Patents vs. Retracted Papers: Mean Annual Citations Received



Notes: This figure plots the mean number of annual citations received by the 85 unsupported patent families of our main sample and their corresponding retracted papers -/+5 years since retraction.

Figure 2: Event Study: Impact of Retraction on Downstream Citations


(A) Total Citations

(B) Examiner-Added Citations

(C) Applicant-Added Citations

Notes: These figures plot the coefficients from the estimation of Equation 3 and 95% confidence intervals, and show the impact of retraction on downstream citations. The outcome variable is the number of annual citations received by the patent family -/+5 years since retraction; Panel A plots the total citations, while Panels B and C decompose the citation count into examiner-added and applicant-added citations. The unit of analysis is a patent family × year, and the sample includes 170 patent families (85 unsupported and 85 control), which is equivalent to 1,648 patent family-years.

Table 2: Summary Statistics

(A) Patents

	Unsupported			Controls			
	Mean	Min	Max	Mean	Min	Max	*t*-statistic
Priority year of the family	2007.06	1995.00	2016.00	2007.59	1998.00	2015.00	0.82
Retraction year (real or counterfactual)	2013.26	2002.00	2019.00	2013.26	2002.00	2019.00	0.00
Family age at retraction year	6.20	1.00	19.00	5.66	0.00	14.00	-1.02
Family size at retraction year	6.51	1.00	32.00	5.77	0.00	49.00	-0.75
Triadic	30%	-	-	28%	-	-	-0.33
Non-industry assignee	81%	-	-	87%	-	-	1.05
Technology class							
A61: Medical	56%	-	-	56%	-	-	-
C12: Biochemistry, microbiology, etc.	24%	-	-	24%	-	-	
C07: Organic chemistry	10%	-	-	10%	-	-	
G01: Measuring, testing	5%	-	-	5%	-	-	
A01: Agriculture	3%	-	-	3%	-	-	
G06: Computing, calculating, counting	1%	-	-	1%	-	-	
Maintained or prosecuted by applicant after retraction	63%	-	-	84%	-	-	2.69
Rejected by examiner after retraction	22%	-	-	40%	-	-	1.96
N of citations per year							
5 years before retraction	2.27	0.00	22.83	3.67	0.00	31.60	1.93
5 years after retraction	1.79	0.00	24.80	3.74	0.00	61.60	2.08
Litigated before retraction	0%	-	-	1%	-	-	1.00
Litigated after retraction	0%	-	-	2%	-	-	1.42
N of patent families	86			86			-
N of individual patents (granted or application)	589			576			-

Table 2: Summary Statistics Continued

(B) Papers

	Retracted			Controls			
	Mean	Min	Max	Mean	Min	Max	t-statistic
Publication year	2008.93	2001.00	2017.00	2008.81	2001.00	2017.00	-0.19
Retraction year (real or counterfactual)	2013.35	2002.00	2019.00	2013.26	2002.00	2019.00	-0.14
Age at retraction	4.42	0.00	12.00	4.44	0.00	12.00	0.06
Journal Impact Factor	22.75	3.24	91.25	22.95	3.24	91.25	0.07
Journal Impact Factor Percentile	87.17	41.20	99.70	87.37	41.20	99.70	0.09
Retraction reason							
Error; unreliable results; contaminated materials	60%	-	-	-	-	-	
Fabrication; fraud; misconduct	38%	-	-	-	-	-	-
Unknown	2%	-	-	-	-	-	
Duplication; plagiarism	0%	-	-	-	-	-	
N of citations per year							
5 years before retraction	7.42	0.00	56.00	14.18	0.00	248.67	1.79
5 years after retraction	2.61	0.00	15.60	20.15	0.00	324.20	3.45
N of papers		84			86		-

Notes: This table reports the summary statistics of our main sample: the universe of retracted patent-paper pairs in the biomedical sciences (as indexed in Medline from 2001 to July 2019) and their control, none-retracted patent-paper pairs. The controls were exactly matched on the publication year of the paper, journal of the paper, and the primary technology class of the patent and inherited the counterfactual retraction date from their retracted counterparts. Two of the retracted papers were associated with two patents. The last column reports the t-statistic or the χ^2, comparing the means. All of the patent summary statistics are based on the full sample of 172 patent families (86 unsupported and 86 control), except for the statistics on citations, maintenance/prosecution by the applicant, and examiner rejection, which were based on subsamples that were used for the analyses in Table 3 and Table 4.

Table 3: Impact of Retraction on Prosecution, Maintenance, and Rejection

	(1)	(2)	(3)	(4)
VARIABLES	Applicant Action	Applicant Action	Examiner Rejection	Examiner Rejection
Retracted	-0.192**	-0.192**	-0.162	-0.151
	(0.081)	(0.091)	(0.100)	(0.102)
Retraction-year FE	YES	NO	YES	NO
Age-at-retraction FE	YES	YES	YES	YES
IPC class FE	YES	NO	YES	NO
Unsupported-control match FE	NO	YES	NO	YES
N of patent families	126	126	100	100

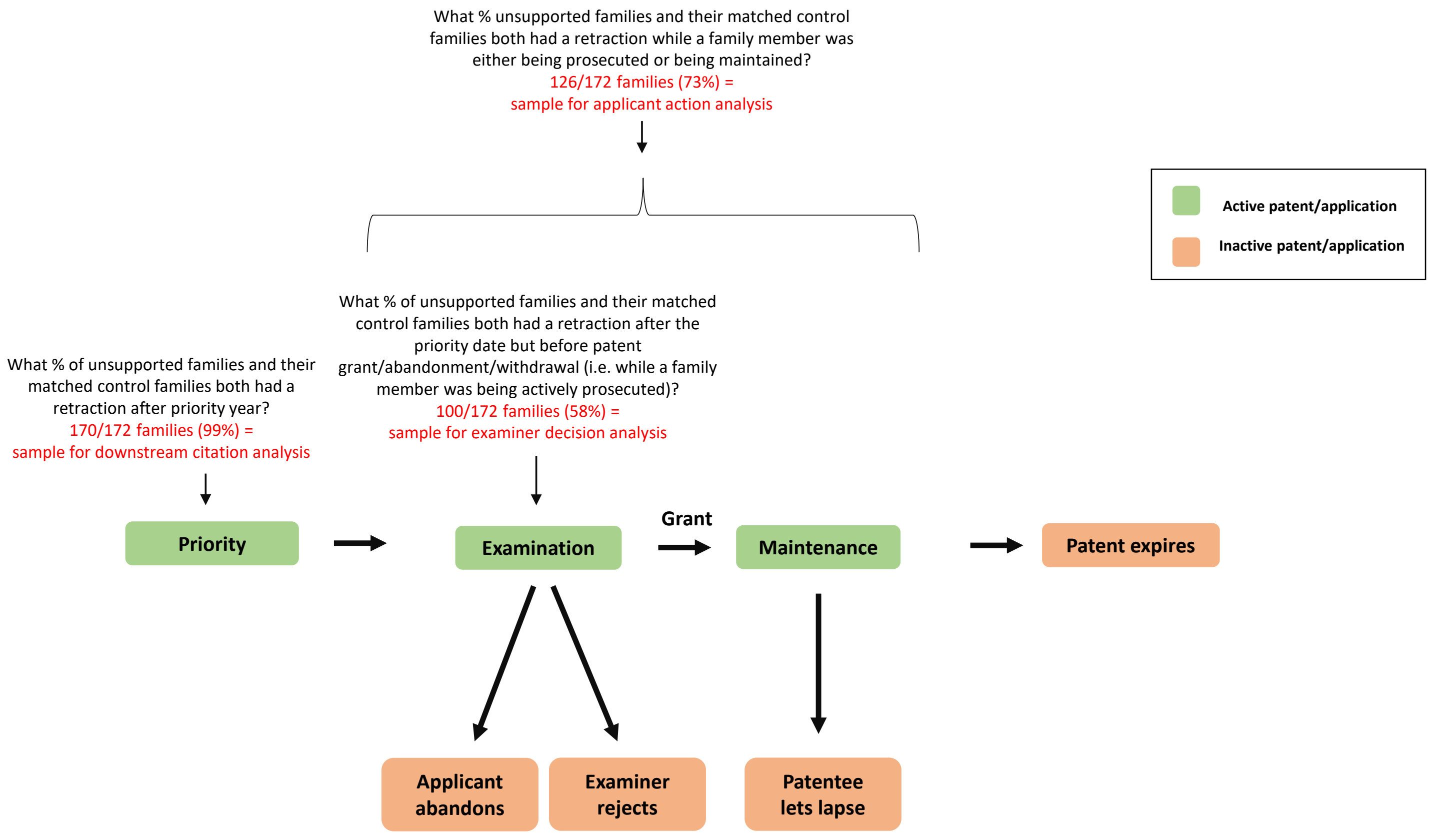
Notes: The table reports estimation from Equation 1. Linear probability model was used, and the unit of analysis is a patent family. Columns 1 and 2 show applicants' reaction to retraction and report whether the applicant continued to pay maintenance fees for any granted patent or prosecute any patent application in the family after retraction. Columns 3 and 4 show examiners' reaction to retraction and whether the examiner rejected any patent in the family for lack of either enablement or written description after retraction. We selected the appropriate subsamples for each of our analyses, such that the timing of the retraction could have impacted the outcome (see Appendix Figure A.1 for more details); 126 patent families (63 unsupported and 63 control) were used for our analysis on applicant action, and 100 patent families (50 unsupported and 50 control) were used for our analysis on examiner rejection. Columns 1 and 3 include retraction-year, age-at-retraction, and technology class fixed effects. Columns 2 and 4 include age-at-retraction and unsupported-control match fixed effects. Unsupported-control match fixed effects are indicators for every unsupported family and its matched control counterpart; since our sample consists of 1-to-1 matching of unsupported and control families that match on covariates such as technology class and the control family inherits the counterfactual retraction year from its unsupported counterpart, specifications with unsupported-control match fixed effects do not include retraction-year or technology class fixed effects. Robust standard errors are in parentheses. Statistical significance is indicated as: *** p<0.01, ** p<0.05, * p<0.1.

Table 4: Impact of Retraction on Downstream Citations

	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	OLS	OLS	Poisson	Poisson	Poisson
VARIABLES	Total	Examiner	Applicant	Total	Examiner	Applicant
Treat × Post-Retraction	-0.302	0.224	-0.526	-0.001	0.119	-0.198
	(0.682)	(0.311)	(0.572)	(0.146)	(0.159)	(0.207)
Post-Retraction indicator	YES	YES	YES	YES	YES	YES
Calendar-year FE	YES	YES	YES	YES	YES	YES
Patent family age FE	YES	YES	YES	YES	YES	YES
Patent family FE	YES	YES	YES	YES	YES	YES
N of patent families	170	170	170	160	160	128
N of patent family-years	1648	1648	1648	1564	1564	1266

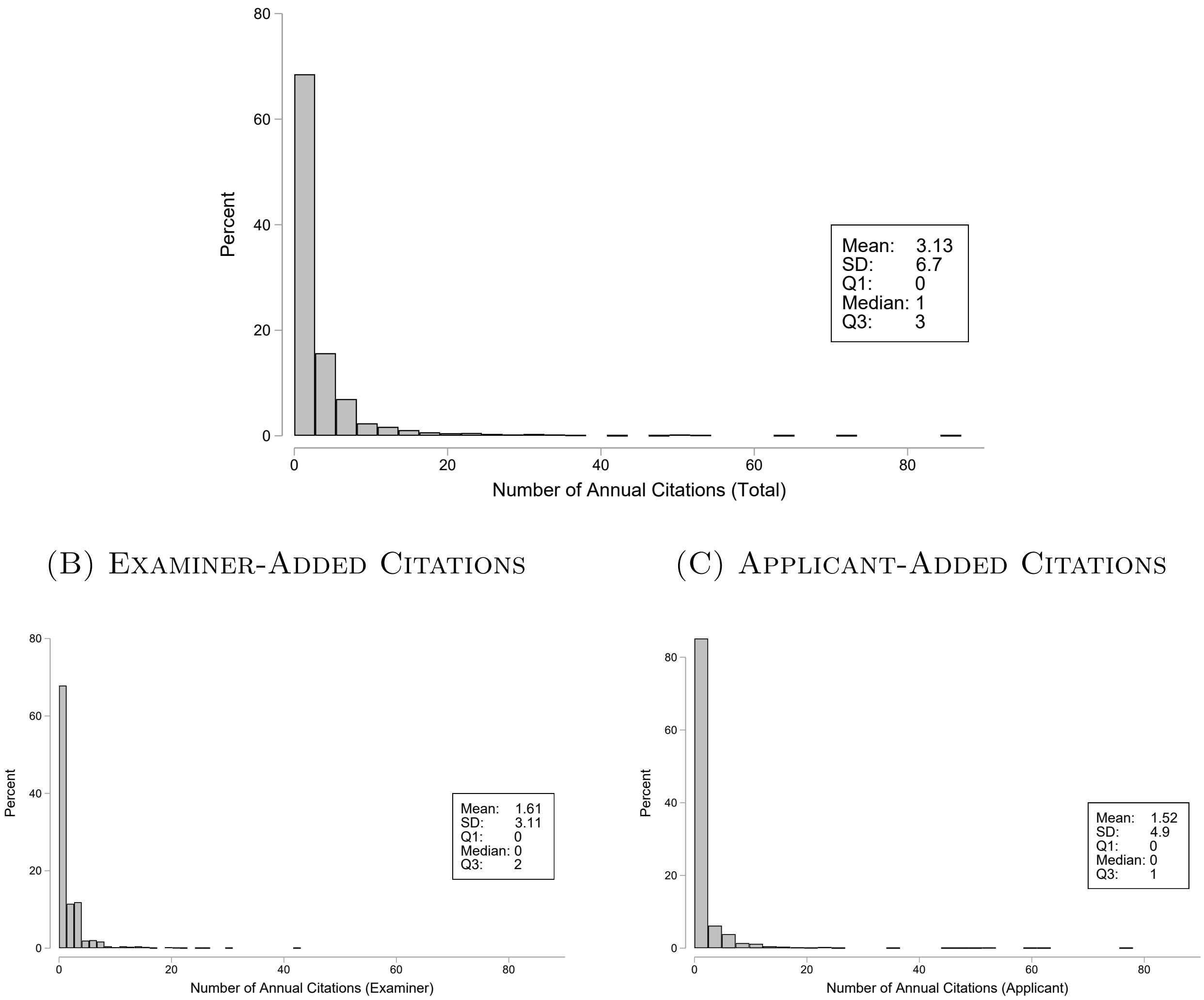
Notes: This table reports results from the estimation of Equation 2 and shows the impact of retraction on downstream citations. The unit of analysis is a patent family × year, and the sample includes 170 families (85 unsupported and 85 control). In Column 1, the outcome variable is the number of total annual citations received by the patent family -/+5 years since retraction, while Columns 2 and 3 decompose the citation counts by whether the citation was added by the examiner or the applicant. Columns 4-6 report results using Poisson specifications; for the Poisson specifications, some patent families never received a citation in our time period and hence dropped out of the regressions. Standard errors are clustered at the patent family level. Statistical significance is indicated as: *** p<0.01, ** p<0.05, * p<0.1.

Appendix Figure A.1: Timing of Retraction and Sample Selection


What % unsupported families and their matched control families both had a retraction while a family member was either being prosecuted or being maintained?
126/172 families (73%) =
sample for applicant action analysis
🟩 Active patent/application
🟧 Inactive patent/application
What % of unsupported families and their matched control families both had a retraction after the priority date but before patent grant/abandonment/withdrawal (i.e. while a family member was being actively prosecuted)?
100/172 families (58%) =
sample for examiner decision analysis
What % of unsupported families and their matched control families both had a retraction after priority year?
170/172 families (99%) =
sample for downstream citation analysis
Priority
Examination
Grant
Maintenance
Patent expires
Applicant abandons
Examiner rejects
Patentee lets lapse

Notes: This figure provides a timeline of a patent application as it progresses through the patent system, as well as what fraction of our sample of 172 patent families (86 unsupported and 86 controls) experienced a (real or counterfactual) retraction at each stage of a patent life cycle. We selected the appropriate subsamples for each of our analyses on applicant actions (Table 3), examiner decisions (Table 3), and downstream citations (Table 4), such that the timing of the retraction could have impacted the outcome. Note that 72/86 (83%) of the unsupported families had a retraction while a family member was either being prosecuted or being maintained, while 61/86 (71%) of the unsupported families had a retraction after the priority date but before patent grant/abandonment/withdrawal. The subsamples in the above figure are smaller since we further restricted the sample to unsupported families whose matched control families also had relevant retraction timing. See Appendix A for details.

Figure A.2: Distribution of Patent Citations

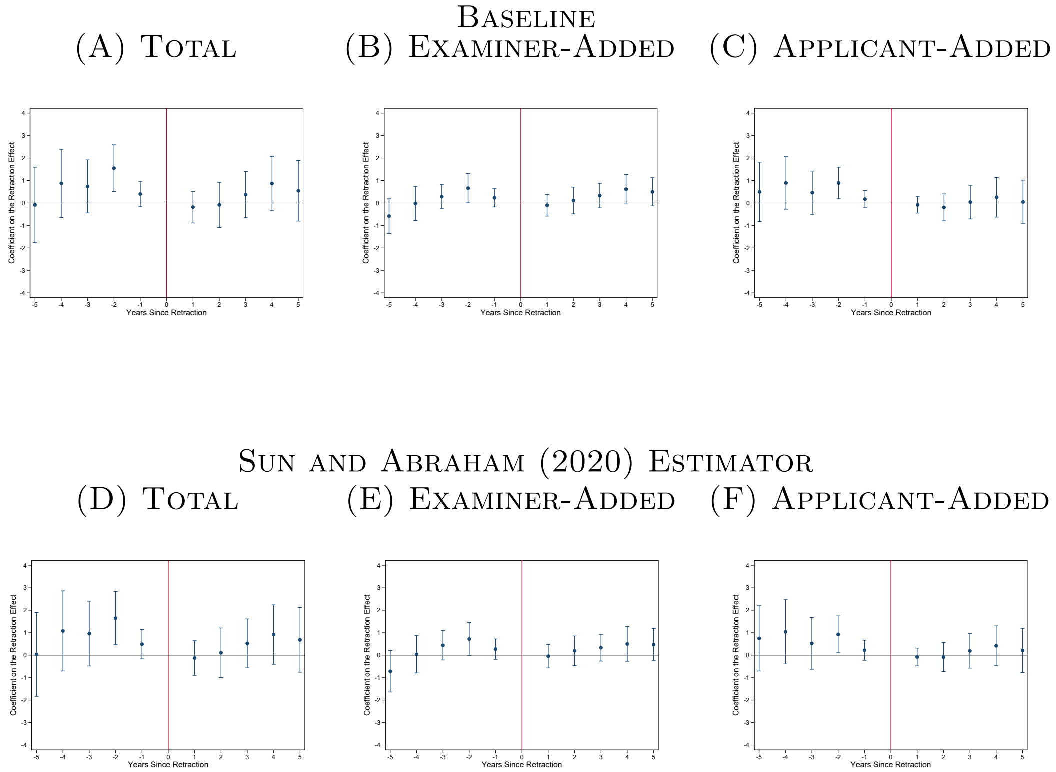

(A) Total Citations

(B) Examiner-Added Citations

(C) Applicant-Added Citations

Notes: These figures plot the distributions of annual citations received by the 170 patent families in our sample on downstream citation analysis (85 unsupported and 85 control families).

Figure A.3: Event Study: Impact of Retraction on Downstream Citations - Treatment Heterogeneity



Baseline

(A) Total (B) Examiner-Added (C) Applicant-Added

Sun and Abraham (2020) Estimator

(D) Total (E) Examiner-Added (F) Applicant-Added

Notes: This figure parallels Figure 2. Panels A-C plot the coefficients from the estimation of Equation 3, modified to drop the indicators for leads and lags around the retraction event that are common to both unsupported and control patent families. Panels D-F plot the coefficients from the same modified version of Equation 3 as Panels A-C but using an alternative estimator developed by Sun and Abraham (2021) that accounts for treatment heterogeneity. Appendix C provides more details.

Appendix Table A.1: Impact of Retraction on Downstream Citations - Balanced Sample

	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	OLS	OLS	Poisson	Poisson	Poisson
VARIABLES	Total	Examiner	Applicant	Total	Examiner	Applicant
Treat × Post-Retraction	-0.176	0.163	-0.339	-0.038	-0.046	-0.098
	(0.715)	(0.392)	(0.606)	(0.142)	(0.181)	(0.174)
Post-Retraction indicator	YES	YES	YES	YES	YES	YES
Calendar-year FE	YES	YES	YES	YES	YES	YES
Patent family age FE	YES	YES	YES	YES	YES	YES
Patent family FE	YES	YES	YES	YES	YES	YES
N of patent families	118	118	118	112	111	90
N of patent family-years	826	826	826	784	777	630

Notes: This table parallels the results from Table 4, but using a balanced sample (in relative time to retraction) and examining -/+3 year window around the retraction event. We selected unsupported patents that have a full citation history of -/+3 years around the retraction event and whose control patents also have a full citation history of -/+3 years around the retraction event. This led to a sample of 118 patent families (59 unsupported and 59 control). The unit of analysis is a patent family × year. Statistical significance is indicated as: *** $p<0.01$, ** $p<0.05$, * $p<0.1$. See Appendix A for more details.

Appendix Table A.2: Impact of Retraction on Downstream Citations - Papers

	(1)	(2)
	OLS	Poisson
VARIABLES	Citations	Citations
Treat × Post-Retraction	-10.599***	-1.470***
	(3.481)	(0.121)
Post-Retraction indicator	YES	YES
Calendar-year FE	YES	YES
Paper age FE	YES	YES
Paper FE	YES	YES
N of papers	166	164
N of paper-years	1505	1488

Notes: This table parallels the results from Table 4, modified for papers. The unit of analysis is a paper × year. Standard errors are clustered at the paper level. Calendar year fixed effects consist of full set of twenty-one indicator variables from 2001 to 2021; age fixed effects consist of twelve indicators, with the age eleven indicator including all subsequent age indicators. The sample includes the corresponding papers of the 170 patents in Table 4 whose publication year occurred after retraction year: 166 papers (82 retracted papers and 84 control papers). For the Poisson specification in Column 2, some papers never received a citation in our time period and hence dropped out of the regression. Statistical significance is indicated as: *** p<0.01, ** p<0.05, * p<0.1.

Appendix Table A.3: Predicting Patent Value

VARIABLES	(1) KPSS value	(2) KPSS value
N of forward citations (5 years after grant)	0.018***	0.023***
	(0.005)	(0.005)
N of backward patent citations	0.002***	-0.001**
	(0.001)	(0.001)
N of backward non-patent citations	0.025***	0.036***
	(0.003)	(0.003)
N of inventors	0.163**	0.055
	(0.065)	(0.063)
N of claims	0.068***	0.071***
	(0.015)	(0.014)
Triadic	4.081***	4.026***
	(0.361)	(0.353)
R&D Intensity		117.794***
		(4.217)
Investment Intensity		-93.529***
		(7.004)
SG&A Intensity		-95.927***
		(1.902)
Ads Intensity		368.342***
		(8.448)
R-squared	0.296	0.330
Issue-year FE	YES	YES
IPC Group FE	YES	YES
N of patents	68712	68712

Notes: This table shows the relationship between patent value and their characteristics. For patent value, we use a measure developed by Kogan et al. (2017)—the "KPSS" measure—which estimates the economic value of a patent by measuring the stock market reaction around the day the patent is issued to the firm. Adopting an approach by Hsu et al. (2021), using an OLS model, we regress the KPSS values on various patent and firm characteristics in the sample of patents owned by public firms provided in Kogan et al. (2017) that were issued from 1997 to 2020, cite at least one scientific article, and can be matched to the CRSP/Compustat Merged Database. Patent characteristics were downloaded from PatSnap. Corporate characteristics were downloaded from CRSP/Compustat Merged Database: R&D intensity (R&D expenditures/total assets), investment intensity (capital expenditures/total assets), SG&A intensity (selling, general, and administrative expenses/total assets), and ads intensity (ad expenses/total assets). Issue year and technology class fixed effects were also included. All of the variables were winsorized at their 1% and 99% percentiles. Statistical significance is indicated as: *** p<0.01, ** p<0.05, * p<0.1.

A Sample Construction

Retracted patent-paper pairs

1. Identifying retracted papers: We retrieved all retracted papers from Medline that were published between 2001 and July 2019. We excluded papers that are not original research articles, such as reviews. We matched the papers to data provided by Retraction Watch in May 2019 in order to obtain data on reason for reaction. We only kept papers that specified that the retraction occurred because information in the paper was incorrect (or did not provide a reason for retraction). We excluded papers retracted due to reasons that cast no doubt on the veracity of the retracted information (e.g. plagiarism, IRB problems).

2. Identifying patent pairs:

   (a) From the set of retracted papers identified in Step 1, we identified all US patent applications with (i) inventors who share the same name (first and last) as the first or last author of the focal retracted publication and (ii) filing dates within +/- 2 years of paper publication. We obtained bibliographic patent data from Reed Tech (`https://patents.reedtech.com/parbbib.php`), which provides bulk files with author information for all applications filed each week.

   (b) We ran a word similarity algorithm that calculated the number of words in common between the paper abstract and the patent specification. The algorithm stems words in the paper abstract and patent specification. The algorithm then takes each word in the abstract and seeks that word in the specification. Finally, the algorithm calculates the percentage of stemmed words in the abstract that are also in the specification.

   (c) For potential pairs that had >90% overlap between words in the paper abstract and the patent specification, we manually reviewed the patent and the paper to verify that the potential pairs identified by the algorithm are indeed true patent-paper pairs. Our manual review incorporated two steps: (i) Review to determine if the potential match is a true patent-paper pair by making sure that the patent and the paper contain the same information in the text or in the figures. (ii) Review to determine if the retracted material (as specified by the retraction notice) from the paper is present in the patent and supports the patent's claims. For retraction notices that did not specify the particular part of the paper that was retracted, we assumed that the entire paper was retracted.

3. Some of our unsupported patents belonged to the same patent family. Since we conducted all of our analyses at the family level, we chose the pair whose paper was retracted first, and if there were still ties, we chose the pair whose paper was published first. After these steps, we identified 107 patent-paper pairs at the family level (101 papers; 96 patent families).

Control patent-paper pairs

1. Identifying control papers: we identified all non-retracted, original research papers in Medline that were published in the same year and journal as the retracted papers of our patent-paper pairs (n = 142,579 papers)

2. Identifying control patent-paper pairs: we determined which control papers had associated patents using the same word similarity algorithm described earlier and identified potential control pairs with word overlap score >0.9 (n = 11,225 pairs). We further narrowed down this pool by focusing on potential control pairs with the same primary technology class as the unsupported patents, using the International Patent Classification (IPC) system at the class level (n = 4,550 pairs).

   Note that while we would have liked to have matched controls on additional covariates of interest, the pool of potential controls for each unsupported patent is highly skewed. Some unsupported patents have very few potential controls left after matching on paper publication year, paper journal, and IPC class and thus we were unable to match on additional covariates.

3. After matching on paper publication year, journal, and IPC class, we then manually reviewed the pool of potential control pairs to confirm that these potential pairs were indeed pairs. Specifically, for each retracted patent-paper pair, we sorted the potential control pairs by the word overlap score measuring correspondence between patent and paper and manually reviewed them in descending order of the score until we identified a true control pair. By sorting the potential control pairs by their word overlap score, we prioritize the manual review of potential control pairs in which the patent closely copies the language from the paper or if the patent and the paper use common, generic

language, but we believe this should not affect the treatment assignment (retraction) or the outcome (e.g., citations that the patent receives).[16]

4. Some unsupported patents have the same journal, publication year, and IPC class (and hence have the same control pool). We randomly assigned one control pair to each such retracted pair, so the control pair can inherit one counterfactual retraction year.

5. Finally, some patents are associated with multiple papers. For these cases, we chose the pair whose paper was retracted first. If there were still patents left associated with multiple papers, then we chose the pair whose paper was published first.

After the 1-to-1 matching procedure, we were able to find control patent-paper pairs for 86 of our retracted patent-paper pairs, leaving us with a sample of 86 retracted pairs (86 unsupported patents and 84 retracted papers; two of the retracted papers were associated with two patents) and 86 control pairs (86 control patents and 86 non-retracted papers).

Identifying patent family members

Because members of a patent family often contain identical or very similar specifications, our unit of analysis is a patent family. For our 86 unsupported and 86 control patents identified above, we sought all of their family members. We defined related applications as both applications filed in other countries (for example, a U.S. patent may have a Japanese counterpart) and parent and/or child applications, including divisionals, continuations, and continuations-in-part. We obtained family information from Google Patents.

Using family members provides a more detailed picture of how applicants are treating these applications. This is particularly true because patent attorneys in this field tend to think about patent strategy on the level of the family (or portfolio) rather than the individual patent, so understanding family-level behavior is a better indicator for applicant behavior than individual applications.

[16] We are facing a measurement error problem since the word similarity algorithm is not perfect. We have a pool of potential control patent-paper pairs that were algorithmically identified—only some of them are true patent-paper pairs, while others are false pairs. We argue that this measurement error is random—being a true vs. false pair in this pool of potential pairs is not a confounder and does not affect the treatment assignment (retraction) or the outcome (e.g., citations that the patent receives). Although patent-paper pairs may receive more citations than non-pairs, all of our potential patent-paper pairs "look like" patent-paper pairs, so even if some of them might not actually be true pairs, they will likely still be cited highly.

Timing of retraction and sample selection

In order to conduct our analyses, it is important to understand whether the retraction occurred before or after our outcomes of interest. For instance, to investigate whether an examiner reacts to retraction, the retraction must have first occurred before the examiner makes a final decision to grant or reject the patent;[17] if the retraction occurred after the patent had already been granted or rejected, then this patent should not be included in our sample for our analysis on examiner rejection.

Appendix Figure A.1 provides a timeline of a patent application as it progresses through the patent system, as well as what fraction of our sample of patent families experienced a (real or inherited counterfactual) retraction at each stage of a patent life cycle. A patent application undergoes the following stages: (i) the applicant files the first patent application in the family (the priority stage); (ii) provided the applicant does not abandon the application beforehand, the application arrives in the examiner's desk, who either rejects or grants the patent; (iii) if the patent is granted, the applicant pays maintenance fees to keep the granted patent active or chooses to let the patent lapse; and (iv) finally, typically after twenty years, the terms of the patent expire.

We selected the appropriate subsample for each of our outcomes, such that the timing of the retraction could have impacted the outcome. Among our sample of 172 patent families, 85 of our unsupported patent families and their matched control families both had a retraction after the priority year, so the sample of 170 families was used for our analysis on the impact of retraction on downstream citations. 63 of our unsupported patent families and their matched control families both had a retraction while a family member was either being prosecuted or maintained, so this sample of 126 families was used for our analysis on applicant's decision to prosecute/maintain the patent. 50 of our unsupported patent families and their matched control families had a retraction after the priority date but before patent grant or abandonment (i.e. during prosecution), so this sample of 100 families was used for our analysis on the examiner's decision to reject or grant the patent.

[17] Note here that we use the term "final" rejection colloquially to mean a rejection after which the applicant stops pursuing the application. "Final rejection" is also a term of art used by PTO examiners to describe certain types of rejections but it is possible under some circumstances for the applicant to continue pursuing the application even after such a rejection.

Unbalanced vs. balanced samples for downstream citation analysis

Our main analysis on downstream citations examines a period of +/- 5 years around the retraction event. This sample is unbalanced, as there are some patent families that experienced retraction early, leaving them with fewer years of pre-retraction citation data, while some patent families were retracted recently, leaving them with fewer years of post-retraction citation data. In order to retain as much of our sample as possible, our main analysis is based on the unbalanced sample.

As a robustness check, we narrow our sample to a smaller but balanced sample (in relative time to retraction). We selected unsupported patents that themselves have full citation history of +/- 3 years around the retraction event and whose control patents also have full history of +/- 3 years around the retraction event.[18] This led to a sample of 59 unsupported and 59 control pairs. As shown in Appendix Table A.1, while the magnitudes slightly change, the results remain similar.

B Data Collection

Prosecution: We obtained data on whether a US patent application was prosecuted after the retraction event. Data was collected manually from the USPTO's Public Patent Application Information Retrieval system (PAIR). We considered an application to have been prosecuted after retraction if the prosecution file history contained a filing that required an affirmative action from the applicant (e.g. responding to an office action or filing an IDS), as opposed to the examiner, and occurred after the retraction date. Our information is current as of August 2020 and applies to US patents only.

Maintenance fee payment: We obtained data on payment and dates of payment from Google Patents, and included that information in our analysis for every country for which the information was available. Our information is current as of August 2020.

Examiner rejection: We obtained data from the USPTO's Office of the Chief Economist.[19]

[18] We chose a smaller window of +/-3 years of the retraction event instead of the +/-5 years window because the +/-5 years window yielded too few patents.

[19] Available here: `https://www.uspto.gov/learning-and-resources/electronic-data-products/office-action-research-dataset-patents`.

Downstream citations: We obtain citation data for all family members from Google Patents, which tracks citations from patents filed in 22 jurisdictions. We included citations to both patent applications and granted patents in the family.

We then sought to determine the citation date for each citing document. Generally, studies that use patent citation dates assume that the citation date is (1) the priority date (Bacchiocchi and Montobbio 2009); (2) the year of patent filing (Alcácer and Gittelman 2006); or, most commonly (3) the patent grant date (Hall et al. 2001; Jaffe and Trajtenberg 1999; Nicholas 2008). Because we seek to understand how citation patterns to patents change when the corresponding paper is retracted, all of these strategies are imperfect measures. Patent grant date is certainly later than both applicant-added and examiner-added citations, and, moreover, we use patent applications, not all of which have been granted. Priority date might be the correct citation date for some applicant-added citations, but is certainly erroneously early for examiner-added citations. We chose to use filing date to approximate citation dates because, although it is also likely early for examiner-added citations, it is more accurate than either the patent grant date or the priority date. Note that due to the imperfect method of determining exact citation dates, the year of first citation can precede the priority year for some of our patent families. Citation data was collected in March 2022.

Partial or full paper retraction: In order to determine whether parts of a retracted paper remained good science even after the retraction, we read the text of each retraction notice and classified it as partially or entirely retracting the paper. Partial retractions suggest that some of the results reported in the paper are valid, despite the retraction. Total retractions suggest that none of the results reported in the paper are valid.

Citation data for papers: Downstream citation data for the papers of our patent-paper pairs was directly exported from PubMed's "Cited By" section in March 2022. The "Cited By" section uses data from publishers and the National Center for Biotechnology Information (which maintains PubMed). Although this citation data may miss citations from, for instance, non-PubMed articles, since our study focuses on the life sciences, we do not expect our papers to receive many citations from non-PubMed articles.

Journal impact factor: Journal Impact Factor (JIF) from 2020 of the papers in our sample were collected from Clarivate Analytics.

Patent owners: We obtained information on assignee from PatSnap and classified the assignee as either industry or non-industry. Individuals were classified as non-industry.

Patent technology class: We obtained the primary International Patent Classification (at the class level) of the patent, as identified by PatSnap.

Litigation: We obtained litigation data from Google Patents. Litigation includes court trials but does not include administrative proceedings.

Triadic patent families: We collected data on whether and when applications were filed in the USPTO, EPO, and JPO from Google patents. All triadic patent families had at least one application filed in each of the USPTO, EPO, and JPO by the retraction date of the corresponding paper (or counterfactual).

C Treatment Heterogeneity

Goodman-Bacon (2021) highlights that when there is heterogeneity in treatment effects over time, DD estimates can be biased due to already-treated units serving as controls in later time periods. A potential solution is an "event studies" framework like Equation 3 that estimates the dynamics of treatment effects over time. However, Sun and Abraham (2021) shows that treatment heterogeneity can also contaminate event studies if the "shape" of the treatment effects changes (a slope change in treatment dynamics for cohorts treated at different times) and proposes a new estimator that addresses this problem.[20]

Our main event studies specification, based on Equation 3 and shown in Figure 2, includes indicators for leads and lags around the retraction event that are common to both unsupported and control patent families (Jaravel et al. (2018)). The alternative estimator proposed by Sun and Abraham (2021) that accounts for treatment heterogeneity does not include these leads and lags, so we first estimated a baseline specification based on a modified version of Equation 3 that dropped these leads and lags (Panels A-C of Appendix Figure A.3). Then in Panels D-F of Appendix Figure A.3, we plotted the coefficients from the same modified version of Equation 3 as in Panels A-C but using the alternative estimator developed by Sun and Abraham (2021). Results remain similar to Figure 2, our main event studies specification.

D Legal Appendix

[20] Note that Sun and Abraham (2021) formerly establishes the validity of their estimator for specifically balanced panels without covariates, and additional assumptions are likely needed to establish validity for unbalanced panels and inclusion of covariates, particularly to deal with problems like attrition in unbalanced panels. Although we have an unbalanced panel with covariates, note that our panel is unbalanced not due to attrition but because patent families are "born" at different calendar times and we do not observe the patents before they are born.

Duty of disclosure

Applicants (including attorneys and assignees) for US patents owe a duty of disclosure, candor, and good faith to the USPTO (37 CFR 1.56) which requires disclosure of "all information known to that individual to be material to patentability." (MPEP 2001.01).

Would knowingly prosecuting a patent application that contains retracted material, and failing to disclose that fact to the PTO, be in violation of the duty of disclosure? The answer depends on the precise facts of the retraction and patent in question, in particular how important the retracted material was to patent grant, but it is likely that such a failing would often violate the duty of disclosure. Although we were unable to find cases directly addressing the question of retracted information, several cases have found that including fabricated data[21] in the patent specification violates the duty of disclosure (e.g., Techno Corp. v. Kenko USA, Inc., 515 F.Supp.2d 1086 (N.D. Cal. 2007) and Hoffmann-La Roche, Inc. v. Promega Corp., 323 F.3d 1354 (Fed. Cir. 2003)).

Ex-post invalidation of patents with inaccurate information

After a patent is granted, it can be challenged in either litigation or a post grant review proceeding at the USPTO. If a patent with inaccurate information is erroneously granted, a challenger can raise the presence of inaccurate information in litigation as part of a validity challenge. However, the presence of inaccurate information cannot be raised in *inter partes* review (IPR) proceedings, the most common type of post grant proceeding.

Circumstances under which a patent with incorrect information is invalid

To be valid, a patent must claim an invention that is useful (35 U.S.C. § 101) and describe the invention in sufficient detail that another skilled in the field of the invention could make and use the invention without undue experimentation (35 U.S.C. § 112).

Incorrect information in the patent may mean that these requirements are not met. If the information is sufficiently problematic that the invention does not work at all, it is invalid under both § 101 and § 112. But a patent with incorrect information could still be valid if it covers many different variations of an invention, and some work but others do not or if another scientist could overcome the incorrect information with a reasonable amount of experimentation.

[21] Note that not all unsupported patents contain *fabricated* data—many contain retracted material that results from errors, not fraud.

The quantum and nature of incorrect information that merits retraction of a paper versus invalidation of a patent differs, and not every paper retraction will require patent invalidation. Moreover, the patent validity requirements described above are not bright line rules and are difficult to interpret in specific cases, making it challenging to ascertain precisely when incorrect information in a patent would require invalidation.

We address this in two ways. First, we removed papers retracted for reasons that would have little bearing on patent validity (e.g. plagiarism). Second, this study includes only patent-paper pairs where the patent claim is supported by the retracted data. This increases the likelihood that the paper retraction indicates a validity problem with the patent.

Circumstances under which an examiner's rejection based on prior art that contains retracted material would be erroneous

Patent applications can be rejected for (1) lack of novelty or (2) obviousness. Examiners issuing either rejection will cite to specific prior art, but the impact of retracted material in that prior art is different for the two rejections.

1. Lack of novelty: Examiners may only make this rejection if the prior art is enabled, meaning that the prior art discloses the invention in sufficient detail that others in the field could make and use the invention. Thus, a rejection for lack of novelty is not correct if the invention described by the prior art does not work.

2. Obviousness: The prior art in obviousness rejections does not need to be enabled. Thus, as a theoretical matter, retracted material could properly be used as prior art. In practice, however, it would often not be proper for an examiner to cite retracted material as part of an obviousness rejection because the fact of retraction demonstrates that the retracting scientist could not actually make the invention, which in turn suggests that it is not obvious to scientists in the field how to make the invention.

E Estimating Economic Value

We use a measure of patent value developed by Kogan et al. (2017)—the "KPSS" measure—which estimates the economic value of a patent by measuring the stock market reaction around the day the patent is issued to the firm. Because the KPSS measure only applies to US patents owned by public firms, and our patents are predominantly owned by non-industry institutions, we adopt an approach taken by Hsu et al. (2021), where the authors estimate the predicted value of academic patents using the KPSS data. We regress the KPSS values on various patent and corporate characteristics in the sample of patents owned by public firms provided by Kogan et al. (2017),[22] and then apply the coefficients estimated from this regression to our unsupported patents to predict their value.

Specifically, we first narrowed the Kogan et al. (2017) data to patents that were issued from 1997 to 2020 and to patents that cite at least one scientific article, using the patent-science linkages provided by Marx and Fuegi (2020), to ensure that these corporate patents have direct links to science like the unsupported patents in our sample. Since the economic value of patents owned by firms in the KPSS data comes from both the technology underlying the patent and the complementary assets of the firms (Hsu et al. 2021), we further narrowed the sample to patents whose firms have non-missing corporate characteristics in the CRSP/Compustat Merged Database. This led to a sample of 68,712 patents. Using an OLS model, we then regressed the KPSS value on various patent characteristics[23] (Column 1) and both patent and corporate characteristics (Column 2), as shown in Appendix Table A.3.

We then applied the coefficients from Column 2 of Appendix Table A.3 on our sample of unsupported family members and estimated that each member is predicted to have on average a value of $42 million.[24] How can we benchmark this number? The average value of patents from public firms in the KPSS data (that cite science at least once and can be matched to the CRSP/Compustat Database) is $27 million, so our unsupported patents are predicted to have higher value than the average patent in the KPSS data. We were also able to directly match one of our unsupported families owned by a corporation to the KPSS data, and this patent had a value of $51 million, which is also higher than the average patent in the KPSS data.

[22] Data can be downloaded at `https://github.com/KPSS2017`.

[23] We are unable to use the datasets used by Hsu et al. (2021) as some of them end before our sample period; we thus use a different data source (PatSnap) and patent characteristics. Note that PatSnap may not define certain types of information exactly the same as Google Patents (the bulk of our data for our main analyses).

[24] In 1982 dollars. Academic patents were assumed to have zero values for corporate complementary assets: R&D intensity (R&D expenditures/total assets), investment intensity (capital expenditures/total assets), SG&A intensity (selling, general, and administrative expenses/total assets), and ads intensity (ad expenses/total assets). The KPSS method cannot be applied to non-US patents or patent applications, so we excluded them from our average (despite the fact that not yet granted patent applications can still be economically valuable and are often licensed or sold before their grant).

A caveat is that we believe that the predicted value reported above is an upper-bound of our unsupported families' economic value. Since Hsu et al. (2021) found that universities can capture approximately 16% of the patent's potential value, we conservatively estimate that the unsupported patents are on average worth \$7 million (\$42 million × 0.16). Despite the limitations of this exercise, the results suggest that the unsupported patents may be economically valuable.

Chapter 4

Insurance Design and Pharmaceutical Innovation

(with Leila Agha and Danielle Li)

Abstract

This paper studies how insurance coverage policies impact pharmaceutical innovation. In the United States, most patients obtain prescription drugs through insurance plans administered by Pharmacy Benefit Managers (PBMs). Beginning in 2012, PBMs began refusing to provide coverage for many newly approved drugs when cheaper alternatives were available. We document a shift in pharmaceutical R&D strategies after this policy took effect: therapeutic classes at greater risk of exclusion experienced a relative reduction in investments. This shift reduced development of drug candidates that appear more incremental: that is, those in drug classes with more pre-existing therapies and less scientifically novel research.

*Published in: Agha, Leila, Soomi Kim, and Danielle Li. 2022. ”Insurance Design and Pharmaceutical Innovation.” *American Economic Review: Insights*, 4 (2): 191-208.

1 Introduction

Technological innovation is a major driver of rising healthcare spending, raising questions as to whether current payment systems appropriately balance incentives to innovate with cost containment. While insurance expansions have been shown to spur R&D investments, critics argue that generous coverage policies create perverse incentives for firms to develop expensive products with little incremental clinical value.[1]

As prescription drug costs rise, politicians and policymakers have increasingly called for the federal government to contain spending by limiting insurance coverage for high-cost, low-value treatments. Despite the importance of this policy debate and the widespread adoption of value-based pricing and coverage decisions outside the US, there is limited empirical evidence on how insurance design shapes incentives for medical innovation.

In this paper, we study the impact of a major change in coverage policies of private sector prescription drug plans on upstream pharmaceutical R&D. Prior to 2012, private prescription drug insurance in the US generally provided coverage for all FDA-approved drugs.[2] To manage costs, plans used a combination of cost-sharing tiers and ordeal mechanisms to direct patients to less expensive drugs. However, these approaches were insufficient to curb prescription drug spending, which grew rapidly during the 1990s and 2000s (Kamal et al. 2018). Beginning in 2012, Pharmacy Benefit Managers (PBMs), the intermediary firms that manage most private prescription drug insurance plans, dramatically shifted their policies and began excluding coverage for some drugs entirely. These exclusions applied to many newly approved drugs without generic equivalents. This practice, known as maintaining a "closed formulary," has since become standard, with 846 branded drugs excluded by at least one of the three largest PBMs as of 2020 (Xcenda 2020).

Closed formulary policies can substantially reduce the profitability of excluded drugs. When GlaxoSmithKline's blockbuster asthma inhaler, Advair, was excluded by Express

[1] For example, Stanford (2020) and Zycher (2006) have argued that the innovation benefits of generous drug payment policies are large, while Bagley et al. (2015), Frank and Zeckhauser (2018), and Dranove et al. (2020) highlight the risk that generous drug payments may yield excessive incremental innovation.

[2] There are exceptions to this pattern, with some private insurance plans applying restrictive formularies prior to 2012. Importantly, these early formulary restrictions were set by individual plans, unlike the post-2012 restrictions we study in this paper, which were centrally negotiated by Pharmacy Benefit Managers that manage coverage for many different insurance companies with a shared formulary.

Scripts in January 2014, its US sales fell by over 30% within a few months (Pollack 2014). Similarly, exclusions can reduce the expected profitability of drugs that have yet to reach the market. The high blood pressure medication Edarbi received FDA approval in 2011 but was almost immediately excluded by CVS Caremark in 2012, suppressing demand before it could become established. By September 2013, Edarbi's manufacturer, Takeda, had sold off its US distribution rights, despite keeping these rights in other countries.

Declines in the potential profitability of drugs arising from downstream exclusion policies can potentially affect pharmaceutical firms' upstream R&D investments. For instance, since its experience with Edarbi, Takeda has not developed any further drugs for hypertension, choosing instead to focus on oncology and rare diseases, areas that have seen far fewer exclusions.

Studying how PBM policies shape pharmaceutical innovation can inform our understanding of how to design payment policies that balance innovation and cost containment. These lessons, gleaned from the choices of private sector firms, can provide insight into the possible effects of policy proposals governing how public insurers interact with drugmakers.[3] Indeed, the largest PBM, CVS Caremark, manages benefits for 75 million Americans—more than the number of enrollees in either Medicare or Medicaid.

We identify the effect of PBM coverage decisions on upstream innovation by comparing drug development activity across therapeutic classes that vary in their risk of facing exclusions, before and after the introduction of closed formulary policies. We begin by matching hand-collected data on PBM's excluded drugs with information on the characteristics of over 100 therapeutic classes. We show that exclusions were more common in drug markets that, prior to the introduction of closed formularies, had more existing therapies and high prescription volume. Using this information, we create an index that categorizes drug markets on the basis of their ex-ante predicted exclusion risk. We then use data on drug development pipelines to track R&D investments across therapeutic classes that vary in their predicted exclusion risk.

Following the introduction of closed formularies, pharmaceutical investments fell markedly in drug classes at high risk of exclusions relative to trends in low risk classes. For

[3] Congressional Budget Office (2007) predicts that the government will not be able to negotiate lower prices with drug manufacturers unless it adopts a PBM-pioneered model of providing preferential access for specific drugs on publicly-run formularies.

a one standard deviation increase in a drug class's exclusion risk, there was an 11% decline in the number of drugs entering pre-clinical and clinical development. These declines affect drug candidates in all phases of development, but are largest among earlier stage candidates. We find no evidence that drug classes at higher risk of exclusion were on different development trends in the five years prior to the introduction of exclusions. R&D declined the most in high prescription volume markets with a large number of existing therapies, as well as in classes where drug patents were based on older and less disruptive science.

Our analysis identifies a *relative* decline in R&D across drug classes at high vs. low exclusion risk, but cannot distinguish whether this comes from a total decline in innovative activity or a reallocation of R&D investment. As a result, we are limited in our ability to evaluate the full welfare implications of closed formulary policies. Our findings suggest, however, that the policies of downstream drug buyers can influence the economic returns to upstream pharmaceutical R&D. Prior to the introduction of closed formularies, pharmaceutical firms could expect their drugs to be widely covered by insurers if they become FDA approved. In this world, firms have strong incentives to develop incremental drugs aimed at large disease markets—such drugs would be likely to receive FDA approval and to generate a large base of revenues if approved. Yet with closed formularies, these incremental drugs became precisely those at greatest exclusion risk.

We build on a broad literature examining the drivers of innovation across a range of settings. A large body of evidence shows that public health insurance expansions create incentives for firms to develop new technology (Acemoglu et al. 2006; Blume-Kohout and Sood 2013; Clemens and Olsen 2021; Dranove et al. 2020; Finkelstein 2004; Krieger et al. 2017). Kyle and McGahan (2012) and Budish et al. (2015) highlight the role of patent policy in encouraging innovation, while Yin (2008) studies the role of tax credits and Clemens and Rogers (2020) focuses on public procurement incentives. Finally, public research funding has positive spillovers on private patenting (Azoulay et al. 2019; Li et al. 2017), and local agglomeration effects are an important driver of innovation (Jaffe et al. 1993) and technology diffusion (Agha and Molitor 2018; Baicker and Chandra 2010).

Our paper contributes to this literature in two ways. First, to our knowledge, this is the first study of how restricting prescription drug coverage affects pharmaceutical innovation.

Theoretical work in this area highlights the tradeoff between insurance design and innovation (Garber et al. 2006; Lakdawalla and Sood 2009). Although policies that restrict prescription drug coverage and aggressively negotiate prices are widely used in Europe and Asia, there is little empirical evidence of how these policies affect dynamic incentives for innovation. Second, while existing work focuses on the role of public sector policies, ours is the first to show that the decisions of *private* firms can have important effects on pharmaceutical innovation. Our findings suggest that insurance design choices are powerful tools that may shape the direction of pharmaceutical R&D.

2 Institutional Background

2.1 The Role of Pharmacy Benefit Managers (PBMs)

In the US, three key parties are involved in shaping payments and access to prescription drugs: manufacturers who develop and produce new drugs, institutional payers such as insurance companies and large employers, and pharmacy benefit managers (PBMs), who design and administer drug insurance plans.[4]

Historically, PBMs were only responsible for processing insurance claims at the pharmacy: verifying the patient's coverage, obtaining payment from the insurer, and transmitting that payment to the pharmacy. Over time, and in concert with a wave of mergers, PBMs began playing a more active role in designing prescription drug plans on behalf of insurers (Werble 2014). By 2016, the three largest PBMs—CVS Caremark, Express Scripts, and OptumRx—collectively designed and administered 70% of private prescription drug plans (Fein 2017).

Modern PBMs argue that they create value by lowering prescription drug spending for institutional payers. One way that PBMs limit spending is through prescription drug coverage that steers patients toward the lowest cost treatment options. Prior to the use of exclusions, PBMs employed three tools to reduce demand for expensive drugs. First, insurance plans assign expensive drugs to different coverage tiers, with higher patient

[4] There are, of course, other parties involved, such as physicians, wholesalers, and pharmacies. We focus on the parties above because they play the largest role in coverage and R&D decisions. See Government Accountability Office (2019) report for a more complete picture of the supply chain.

cost-sharing. Second, prior authorization requirements imposed on select drugs require physicians to obtain advance approval from the PBM or insurer prior to coverage. Finally, step therapy requirements allow coverage for certain expensive drugs only after the patient has tried and failed cheaper alternatives.

PBMs may also lower costs by pooling demand across multiple payers in order to negotiate bulk discounts. Given the concentration in the industry and their role in shaping patient demand, PBMs have substantial negotiating power with manufacturers. Drugmakers routinely offer large rebates in order to secure more favorable formulary positions. PBMs may return a portion of this savings to institutional payers and keep a portion for themselves.

2.2 The Introduction of Formulary Exclusions

Prior to the introduction of closed formularies, PBMs had limited success in reducing the use of expensive medications because pharmaceutical firms employed a variety of techniques to circumvent their coverage restrictions. For example, to increase the use of drugs placed in more expensive coverage tiers, pharmaceutical firms introduced "co-pay coupons" that reduced patients' out-of-pocket costs.[5] Similarly, drug sales representatives actively helped physicians' offices fill out the paperwork necessary to request a prior-authorization—in some cases by developing specialty software that would auto-fill these forms (Pinsonault 2002).[6]

Beginning with CVS in 2012, major PBMs responded by implementing closed formularies (Pollack 2014). For the first time, PBMs published lists of drugs that their standard plans would not cover at all, directing potential users to recommended alternatives.

Exclusions constituted a much more effective tool for formulary management. In an investor call, Helena Foulkes, the President of CVS Pharmacy at the time, highlighted the efficacy of exclusions:

> "It is only through exclusion where we can prevent manufacturer subversion of a formulary strategy with co-pay coupons. As shown, an exclusion formulary will

[5] Because the average implied co-insurance rate of even the highest tier drugs is roughly 30-40%, subsidizing patient costs still netted pharmaceutical firms substantial revenues via the insurer contribution (Claxton et al. 2011).

[6] One audit study found that 88% of prior authorization requests were approved by health plans (Scott-Levin 2001).

> have more than a 95% preferred drug use versus 55% preferred share in tiered formularies" (Foulkes 2015).

The success of closed formularies at curbing utilization reduces the profitability of targeted drugs. Yet, perhaps more importantly, the *threat* of facing exclusion can also reduce prices even if a drug is never excluded in practice. Stephen Miller, the Chief Medical Officer of Express Scripts, describes using the threat of exclusion in price negotiations with pharmaceutical manufacturers:

> "Who is going to give us the best price? If you give us the best price, we will move the market share to you...We'll exclude the other products" (Miller and Wehrwein 2015).[7]

Consistent with the market dynamics described by Garthwaite and Morton (2017), a credible threat of exclusions reduces the net price that drugmakers can charge, regardless of whether exclusions actually take place.

2.3 Formulary Exclusions and Upstream Innovation

As illustrated above, PBMs may use the threat of exclusions to extract surplus from drug manufacturers. Manufacturers may make price concessions in order to compete for a spot on the restrictive formulary, or the mere threat of exclusion could lead to lower prices even if few drugs are actually excluded in equilibrium. In either scenario, closed formulary policies—which enable the *possibility* of exclusions—may reduce the expected revenues of drug candidates that can be credibly threatened with exclusions.

These changes in expected profitability may in turn influence pharmaceutical firms' upstream R&D decisions. Specifically, concerns about formulary coverage may lead firms to apply a higher "bar" for drugs at greater risk of facing exclusion. After the introduction of formulary exclusions, industry consultants began routinely advising pharmaceutical companies that "[m]arket access strategy should underpin decision-making throughout the entire product lifecycle, including portfolio decision-making" (Siegal and Shah 2019). Rather than simply demonstrating safety and efficacy (the standard for FDA approval),

[7] In line with this description, observers note that within a therapeutic class, PBMs are increasingly selecting a single brand for coverage (Cournoyer and Blandford 2016).

firms were also advised to conduct more ambitious clinical trials to demonstrate superiority in head-to-head comparisons with competitor's drugs (Schafer 2018; Siegal and Shah 2019).[8] Formulary considerations may reduce the number of drug candidates promoted through clinical testing both by weeding out drugs that do not meet this higher standard, and by raising the cost and complexity of clinical trial design.

3 Data

To understand the impact of exclusion policies on innovation, the key economic object we are interested in measuring is pharmaceutical firms' perceptions of exclusion risk associated with developing new drug candidates across different classes. The ideal measure would capture both the risk that the new drug is itself excluded, as well as the risk that the new drug is less profitable because it must offer large price concessions in order to avoid exclusion.

To develop our measure of exclusion risk, we link data on drug market characteristics across classes (from First Data Bank) with the incidence of formulary exclusions (from PBM documents). We then investigate the relationship between exclusion risk and drug development by linking exclusion risk to Cortellis data on R&D activity. The data underlying these analyses is summarized below.

1. **Formulary Exclusions**: We collected data on formulary exclusions, from publicly disclosed standard formulary lists published by CVS Caremark, Express Scripts, and OptumRX through 2017. Together, these firms account for approximately 70% of the PBM market. Our data cover "standard" formulary exclusions: these exclusions apply to most health plans administered by a particular PBM. Insurers may elect to provide more expansive coverage by opting out of the standard formulary, but we do not have information on exclusions within these custom plans.[9]

[8] In a related analysis, Seabright (2013) analyzes how drug procurement may affect trial design, particularly the incentive to investigate treatment effect heterogeneity predictable by biomarkers. Cohen et al. (2021) discuss how timing considerations may impact firms' decisions to seek FDA approval.

[9] Custom plans are less common because they are likely to be substantially more expensive. For example, on its payer-facing website, CVS encourages insurers to choose its standard (closed) formulary, for an estimated 29% savings in per member per month drug costs (Brennan 2017).

2. **First Data Bank**: We collect data on drug markets from First Data Bank (FDB) (2018). FDB is a commercial database that contains information on each approved drug's ATC4 classification, pricing, and generic substitutes. We use this information to construct drug-class level predictors of exclusion risk.

3. **Cortellis Investigational Drugs**: Our main analysis studies the impact of formulary exclusions on drug development. We obtain data on pipeline drugs, including both small molecule and biologic drugs, from Cortellis Investigational Drugs database (Clarivate Analytics 2018). Cortellis tracks drug candidates using data it compiles from public records: company documents, press releases, financial filings, clinical trial registries, and FDA submissions. Drug candidates typically enter the Cortellis database when they enter preclinical development. Because FDA approval is prerequisite for beginning human clinical trials, Cortellis has near complete coverage of drug candidates that advance into human testing.

   Our primary outcome is the total number of drug candidates within a class that entered or advanced to any stage of development each year. Table 1 Panel A reports the summary statistics of development activity across different stages.

Throughout most of the paper, our unit of analysis is a narrowly defined drug class, following the Anatomical Therapeutic Chemical (ATC) classification system. We use an ATC4 (four-digit) level classification, which identifies chemical subgroups that share common therapeutic and pharmacological properties. Appendix Table A.1 lists several examples of ATC4 designations.

We interpret an ATC4 drug class as a "market," where drugs within the class will typically be partial substitutes for one another. We drop ATC4 categories that are not categorized as drugs in FDB, such as medical supplies. We also restrict to ATC4 categories that contain at least one branded drug on the market with no generic equivalent, and to those for which we observe measures of prescription volume and price in 2011. Our primary sample has 127 ATC4 classes. Table 1 Panel B shows the summary statistics of various market characteristics for our sample of ATC4s.

4 Understanding Exclusion Risk

4.1 The Rise of Formulary Exclusions

Figure 1 Panel A illustrates the rise of drug exclusions over time and across PBMs. As described in trade press and national media (Pollack 2014; Fein 2015), CVS began with the exclusion of 38 drugs in 2012. Over the next five years, CVS oversaw a sustained expansion in the number and types of excluded drugs. Express Scripts introduced its exclusion list in 2014, followed by OptumRx in 2016. By 2017, a total of 300 drugs were ever excluded by at least one of the three major PBMs.

Exclusions largely targeted newer branded drugs: 75% of those excluded had no molecularly equivalent generic substitute. Exclusions are concentrated in therapeutic areas with large numbers of patients. Appendix Figure A.1 plots exclusions by disease category at the drug level and shows that diabetes drugs have been the most frequently excluded. Other disease categories with high numbers of exclusions include cardiovascular, endocrine, and respiratory diseases.

PBM formulary choices affect patients' drug use. It has been widely documented that demand for drugs is elastic to out-of-pocket prices, implying that eliminating insurance coverage for excluded drugs will suppress demand (Abaluck et al. 2018; Einav et al. 2017; Choudhry et al. 2011; Tamblyn et al. 2001). In addition, several papers have shown that formulary exclusions specifically reduce utilization of targeted drugs (Chambers et al. 2016; Huskamp et al. 2003; Wang and Pauly 2005).[10] In Appendix Table A.2, we verify this in our own data by tracking how PBM exclusions affect Medicare Part D prescription volume over time. Our findings indicate that a drug's market share of claims (measured as the fraction of the drug's prescription volume relative to other drugs in the ATC4 class) falls by about 25% for each of the 3 major PBMs that exclude it.

[10] While CVS was the first PBM to implement a national closed formulary in 2012, the two older papers cited above provide evidence from smaller scale exclusions by individual insurance plans. These earlier coverage decisions affect many fewer patients than the PBM formularies we study here, but are likely to have similar effects on the drug choices of enrolled patients.

4.2 Predictors of Formulary Exclusion Risk

Using the FDB data, we construct several potential predictors of exclusion risk for ATC4 drug classes. We measure the availability of therapeutic alternatives using the number of existing branded drugs within an ATC4, the number of existing generics within the same class, and the number of finer-grained ATC7 subclasses. To account for the expected size of the patient population, we use the total prescription volume across all drugs in a given ATC4 class; this information is calculated from the Medicare Expenditure Panel Survey. Finally, we collect data on the price of branded and generic drugs, keeping in mind that price data do not reflect the rebates that manufacturers often pay to PBMs. All of these market characteristics are from 2011, before the introduction of exclusions in 2012.

Figure 1 Panel B plots the coefficients from bivariate logit regressions of exclusion on each drug class characteristic. Drug classes with higher prescription volume and more treatment options are more likely to experience exclusions. These patterns are consistent with contemporaneous descriptions of PBMs' exclusion strategies, which indicate that exclusions often target "me-too drugs" with multiple therapeutic substitutes (Reinke 2015), as well as drugs with many prescribed patients: "[T]here's no reason to go after trivial drugs that aren't going to drive savings" (Miller and Wehrwein 2015).[11]

Building on these insights, we estimate a single index of exclusion risk using logistic regression as follows:

$$Pr(\text{Excluded}_c|\mathbf{X}_c) = F(\alpha\mathbf{X}_c) \tag{1}$$

Excluded$_c$ is an indicator for whether drug class c actually experiences exclusions in 2012 or 2013 and X_c is a vector of market characteristics described earlier. We take the resulting fitted values, denoted $\widehat{\text{Pr(Excluded)}}_c$, as our primary measure of exclusion risk for drug class c. Table 2 shows the results of this exercise, and Appendix Figure A.2 plots the resulting distribution of predicted exclusions.

To estimate Equation (1), we use market characteristics from 2011, prior to the introduction of closed formulary policies, in order to avoid confounding our risk measure

[11] We find no statistically significant relationship between drug prices and exclusion risk, but because our data does not measure prices net of rebates, these correlations are difficult to interpret.

with development responses that are endogenous to the exclusion policies we study. We discuss threats to identification further in Section 5.

For $\Pr(\text{Excluded})_c$ to capture firms' perceptions of exclusion risk over the duration of the post-period, it must meet two conditions. First, drug classes predicted to have high exclusion risk in 2012 and 2013 should also be more likely to face exclusions in later years. Second, because exclusion threat can depress profitability even in the absence of actual exclusions (by forcing drugmakers to grant price concessions), our measure should capture the threat of exclusion even in classes where no drugs face early exclusions. Appendix Table A.3 provides support for both predictions. Classes at high risk of early exclusions are also more likely to see later exclusions: a one standard deviation increase in early exclusion risk correlates with a 19 percentage point increase in the likelihood that an ATC4 class experiences exclusions in later periods, from a mean of 39%. Even among drug classes that do not experience any exclusions in 2012-13, those with higher predicted exclusion risk are more likely to see exclusions in later periods: a one standard deviation increase in early exclusion risk generates a 13 percentage point increase in the likelihood of late exclusions, from a base rate of 31%.

5 The Impact of Exclusion Risk on Subsequent Drug Development

5.1 Empirical Strategy

Our main specification compares drug development behavior across ATC4 drug classes that vary in their ex-ante risk of exclusion, before and after the rise of closed formulary policies:

$$\text{Development}_{ct} = \beta_1 \Pr(\text{Excluded})_c \times \mathbb{I}(\text{Year}_t \geq 2012) + \mathbf{X}_{ct}\gamma + \delta_c + \delta_t + \epsilon_{ct} \tag{2}$$

In Equation (2), Development_{ct} measures the number of new drug candidates in drug class c at year t. The index $\Pr(\text{Excluded})_c$ captures a drug class's exposure to exclusions, as defined in the previous section. The regressions control for drug class fixed effects (δ_c), year fixed effects (δ_t), and some specifications include time-varying drug market controls ($\mathbf{X}_{ct}$).

For the coefficient β_1 to represent the causal impact of formulary exclusions on drug development, the exclusion risk index $\Pr(\text{Excluded})_c$ must satisfy a conditional exogeneity assumption. Specifically, market characteristics used to construct this index cannot predict changes in R&D investment that would have occurred even in the absence of exclusive formularies, after conditioning on drug class fixed effects, year fixed effects, and other control variables.

While we cannot directly test this assumption, we can investigate whether these drug classes were on parallel development trends prior to the introduction of PBM formulary exclusions. In Figure 2, we report an event study graph over a 5-year pre-period to assess the plausibility of this assumption. This graph is based on a modified version of Equation (2), which replaces the single indicator variable for the post period ($\mathbb{I}(\text{Year}_t \geq 2012)$) with a vector of indicator variables for each year before and after the introduction of PBM exclusion lists in 2012.

Even with parallel pre-trends, our identification arguments could be threatened if other changes in global drug development incentives coincided with the introduction of PBM formulary exclusions, particularly if these changes disproportionately affected drug classes at high exclusion risk. For example, changes in drug purchasing policies in international markets may have independent effects on innovation, as might changes in industry structure resulting from PBM mergers. We discuss these possibilities and interpret our findings in Section 5.3.

5.2 Main Results

Table 3 presents our main regression results. The outcome is the total number of drug candidates promoted to the next stage of development each year. In Column 1, we estimate that a one standard deviation increase in the risk that the class has formulary exclusions leads to 3.6 fewer advanced drug candidates each year, a 12% reduction from a mean of 30.6 advancing candidates.[12] This estimate reflects declining development in higher-risk classes relative to trends in lower-risk classes. In Column 2, we show that our results are robust to controlling for time-varying market conditions: the number of approved branded

[12] As reported in Appendix Figure A.2, the standard deviation of the probability the class faces exclusions is 0.15. Using the coefficient reported in Table 3, we calculate $-24.04 * 0.15 = -3.6$.

drugs, the number of generic drugs, the mean price of branded drugs minus the mean price of generic drugs, the number of ATC7 subclasses with approved drugs, and prescription volume. Adding these controls lowers our estimated coefficient slightly from 3.6 to 3.3, which translates into an 11% decrease in annual development per standard deviation increase in exclusion risk. In Columns 3 and 4, we consider an alternative functional form: $\log(1 + \text{Development}_{ct})$. The log-transformed outcome suggests that development activity declines by 6% for every 1 standard deviation increase in class exclusion risk. In Appendix Table A.4, we decompose this total effect by drug development stage; across all stages, from preclinical through Phase 3 trials, a one standard deviation increase in exclusion risk predicts a decline in innovation ranging from 8% to 14%. We find no significant effect of exclusions on new drug launches, although our estimate is imprecise relative to the mean frequency of launches.

One concern is that innovation in ATC4 classes at high exclusion risk may have been evolving on different trends, for reasons other than the introduction of formulary exclusions. For example, drug classes with many existing treatment options may be both more likely to face exclusions and, independently, also see natural attenuation in innovative activity. Figure 2 plots our results in an event study framework, illustrating that there appears to be little difference in drug development across drug classes at high vs. low risk of exclusions prior to 2011. In Appendix Figure A.3, we report results from various placebo policy tests to provide further evidence that our results are not driven by secular differences in innovative potential across low- and high-exclusion risk classes.

In addition, we conduct a variety of robustness checks. Our results remain statistically significant when applying a wild cluster bootstrap (see Appendix Table A.5), using alternative functional specifications such as Poisson regression or the inverse hyperbolic sine transformed outcome (see Appendix Table A.6), or testing alternative rules for attributing drug candidates to ATC4 classes (see Appendix Table A.7). Our results are also robust to a variety of approaches for assessing exclusion risk: predicting based on the count or share of excluded drugs within an ATC4 class, or simply using an indicator variable for whether a drug class had any realized exclusions in 2012-2013 (see Appendix Table A.8). Finally, we obtain similar estimates when augmenting our predictors of exclusion risk to include 2014 data on copay coupons from Van Nuys et al. (2018) (see Appendix Table A.9).

5.3 Discussion

Our results suggest that the policies of US PBMs have a meaningful impact on the drug development decisions of global firms. To contextualize this result, we consider other possible changes in pharmaceutical markets and quantify the implications of these results for different types of drug classes.

First, the strength of formulary exclusion policies is likely related to the market power of PBMs, which increased over this period through three major mergers: CVS's acquisition of Caremark in 2007 (Harris 2007), Express Scripts' acquisition of Medco Health Solutions in 2012 (Lee 2012), and OptumRx's (owned by UnitedHealth) acquisition of Catamaran in 2015 (Mathews and Walker 2015). In each case, the acquiring PBM introduced its closed formulary 1–5 years after its acquisition. Our results should therefore be interpreted as describing the effect of exclusion policies in a setting where downstream buyers have substantial market power.

Second, while the US drug market plays an outsized role in shaping global development incentives, accounting for 40% of total pharmaceutical spending in 2018 (IQVIA 2019), policy changes in other countries may also contribute to our findings. Any changes to drug purchasing in large markets that occur around 2012 and differentially affect crowded drug classes would be particularly relevant. The European Union does not centrally control prices or coverage of prescription drugs (Rodwin 2019) and the five largest European markets collectively account for only 15% of global spending. As a result, we believe that the ongoing administration of their national formulary policies is unlikely to explain our results. The most relevant policy we have been able to identify is a series of initiatives implemented in Japan beginning in 2006 aimed at encouraging generic substitution of branded drugs. Japan is a large market for branded pharmaceuticals (second after the US[13]), representing 7% of the global spending (IQVIA 2019), and this policy may have depressed incentives for innovation in markets with generic competition. However, the implementation of these policies was gradual and began several years prior to the introduction of closed PBM formularies in the US (Kuribayashi et al. 2015).

[13] The second largest pharmaceutical market in general is China (11% of global spending), but branded drugs comprise a much smaller share of this market than in Japan or the US.

Finally, to better describe the drug markets that experience declines in R&D investment attributable to formulary exclusions, we use our estimates to conduct a quantification exercise considering three dimensions of difference across markets: crowdedness, size, and scientific novelty. Because drug classes with these market characteristics have different predicted exclusion risk (as estimated in Table 2), our findings imply differential impacts of formulary exclusions. In Panel A of Appendix Figure A.4, we predict the largest declines in drug development for drug markets with the most existing therapies; among drug classes in the top tercile of available therapies, exclusions depress development by over 4%. In Panel B, we predict larger R&D declines for drug classes with higher prescription volume, topping out at an 8% fall in the top tercile. In Panels C and D, we apply patent-to-science linkages created by Marx and Fuegi (2020) to assess the scientific novelty of drug classes as measured by citations to recent or "disruptive" science.[14] In both cases, our calculations show that formulary exclusions lead to larger R&D reductions in less scientifically novel drug classes.

These calculations suggest that PBMs wielded the threat of formulary exclusion in a way that disproportionately reduced R&D effort for incremental treatments, with many existing substitutes and older, less novel underlying science. This analysis is suggestive: our finding of differential impact on large, crowded drug classes could reflect the possibility that competition lowered the returns to new investment in these areas. While we see no evidence of this slow-down for more crowded classes in our placebo analysis reported in Appendix Figure A.3, other long-run changes in pharmaceutical markets might affect the nature of these relationships.

6 Conclusion

Amid rising public pressure, government and private payers are looking for ways to contain drug prices while maintaining incentives for innovation. In this paper, we study how the design of insurance policies restricting prescription drug coverage affects upstream investments in pharmaceutical R&D.

[14] Our measure of "disruptiveness" follows Funk and Owen-Smith (2017) and Wu et al. (2019), which captures the idea that a research article representing a paradigm shift will generate forward citations that will not cite the breakthrough article's backward citations.

Drug classes facing a one standard deviation greater risk of exclusions see an 11% decline in drug development activity relative to trends in lower risk classes, following the introduction of closed formulary policies. These declines in development activity occur at each stage of the development process from pre-clinical through Phase 3 trials.

The limitations of our current analysis suggest several important directions for future work. First, our identification strategy allows us to document a relative decline in R&D in high exclusion risk categories. The overall welfare implications of exclusive formularies will depend on their impact on aggregate pharmaceutical R&D, which is not identified by our empirical strategy. Second, it remains challenging to accurately value foregone innovation. While we focus on the availability of existing treatments, prescription volume, and measures of scientific novelty, these are not complete descriptions of the clinical and scientific importance of potentially foregone drugs. Additional research will be needed to quantify the tradeoffs associated with decreased development. Third, because we cannot directly observe drug price rebates, there is more work to be done quantifying the impact of formulary exclusions on pharmaceutical revenue.

Our analysis focuses on the first wave of PBM formulary exclusions, which largely targeted drugs in markets with many available options. In recent years, formularies have begun to exclude therapies for relatively rare and sensitive diseases, including HIV, hemophilia, and certain cancers (The Doctor-Patient Rights Project 2017; Maas 2018). Drug classes that appeared low risk in our analysis based on early exclusion patterns may become higher risk as exclusions expand, possibly leading to declines in R&D in those classes as well.

Viewed from a public policy perspective, this research opens the door for insurance design to be a part of the broader toolkit that policymakers use to encourage and direct investments in innovation. Existing policy efforts to shape innovation have relied almost exclusively on directly influencing the costs and returns to R&D, through patents, tax credits, or research funding. Our results suggest that managers and policymakers can also use targeted coverage limitations and price negotiation—for example, those generated by value-based pricing—to reduce R&D efforts in areas with limited incremental clinical value.

References

Abaluck, J., J. Gruber, and A. Swanson (2018). Prescription drug use under Medicare Part D: A linear model of nonlinear budget sets. *Journal of Public Economics 164*, 106–138.

Acemoglu, D., D. Cutler, A. Finkelstein, and J. Linn (2006). Did Medicare induce pharmaceutical innovation? *American Economic Review 96*(2), 103–107.

Agha, L. and D. Molitor (2018). The local influence of pioneer investigators on technology adoption: evidence from new cancer drugs. *Review of Economics and Statistics 100*(1), 29–44.

Azoulay, P., J. S. Graff Zivin, D. Li, and B. N. Sampat (2019). Public R&D investments and private-sector patenting: evidence from NIH funding rules. *The Review of Economic Studies 86*(1), 117–152.

Bagley, N., A. Chandra, and A. Frakt (2015). *Correcting Signals for Innovation in Health Care*. Brookings Institution.

Baicker, K. and A. Chandra (2010). Understanding agglomerations in health care. In *Agglomeration Economics*, pp. 211–236. University of Chicago Press.

Blume-Kohout, M. E. and N. Sood (2013). Market size and innovation: Effects of Medicare Part D on pharmaceutical research and development. *Journal of Public Economics 97*, 327–336.

Brennan, T. (2017, August). 2018 Formulary strategy. Technical report, CVS Health Payor Solutions. Online at: `https://payorsolutions.cvshealth.com/insights/2018-formulary-strategy`.

Budish, E., B. N. Roin, and H. Williams (2015). Do firms underinvest in long-term research? Evidence from cancer clinical trials. *American Economic Review 105*(7), 2044–85.

Centers for Medicare & Medicaid Services (2012-2018). Cms drug spending: Historical data. `https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Information-on-Prescription-Drugs/Historical_Data`.

Chambers, J. D., P. B. Rane, and P. J. Neumann (2016). The impact of formulary drug exclusion policies on patients and healthcare costs. *Am J Manag Care 22*(8), 524–531.

Choudhry, N. K., J. Avorn, R. J. Glynn, E. M. Antman, S. Schneeweiss, M. Toscano, L. Reisman, J. Fernandes, C. Spettell, J. L. Lee, et al. (2011). Full coverage for preventive medications after myocardial infarction. *New England Journal of Medicine 365*(22), 2088–2097.

Clarivate Analytics (2018). Cotellis investigational drug database. Data licensing information available from `https://clarivate.com/industries/academia/`.

Claxton, G., M. Rae, N. Panchal, J. Lundy, A. Damico, A. Osei-Anto, and J. Pickreign (2011). Employer Health Benefits 2011 Annual Survey. The Kaiser Family Foundation and Health Research and Educational Trust. 1-220.

Clemens, J. and P. Rogers (2020, January). Demand shocks, procurement policies, and the nature of medical innovation: Evidence from wartime prosthetic device patents. Working Paper 26679, National Bureau of Economic Research.

Clemens, J. P. and M. Olsen (2021). Medicare and the rise of American medical patenting: The economics of user-driven innovation.

Cohen, L., U. G. Gurun, and D. Li (2021). Internal deadlines, drug approvals, and safety problems. *American Economic Review: Insights 3*(1), 67–82.

Congressional Budget Office (2007, April). Medicare prescription drug price negotiation act of 2007. Technical report, Congressional Budget Office Cost Estimate. Online at `https://www.cbo.gov/sites/default/files/110th-congress-2007-2008/costestimate/s30.pdf`.

Cournoyer, A. and L. Blandford (2016, October). Formulary exclusion lists create challenges for pharma and payers alike. *Journal of Clinical Pathways*. `https://www.journalofclinicalpathways.com/article/formulary-exclusion-lists-create-challenges-pharma-and-payers-alike`.

Dranove, D., C. Garthwaite, and M. I. Hermosilla (2020, May). Expected profits and the scientific novelty of innovation. Working Paper 27093, National Bureau of Economic Research.

Einav, L., A. Finkelstein, and P. Schrimpf (2017). Bunching at the kink: Implications for spending responses to health insurance contracts. *Journal of Public Economics 146*, 27–40.

Fein, A. J. (2015, August). Here come the 2016 PBM formulary exclusion lists! Technical report, Drug Channels. Online at: https://www.drugchannels.net/2015/08/here-come-2016-pbm-formulary-exclusion.html.

Fein, A. J. (2017, December). The CVS-Aetna deal: Five industry and drug channel implications. Technical report, Drug Channels. Online at: https://www.drugchannels.net/2017/12/the-cvs-aetna-deal-five-industry-and.html.

Filzmoser, P., A. Eisl, and F. Endel (2009). ATC-ICD: Determination of the reliability for predicting the ICD code from the ATC code.

Finkelstein, A. (2004). Static and dynamic effects of health policy: Evidence from the vaccine industry. *The Quarterly Journal of Economics 119*(2), 527–564.

First Data Bank (FDB) (2018). Drug database. Data licensing information available from https://www.fdbhealth.com.

Foulkes, H. (2015, June). CVS Health Corp at Jefferies Consumer Conference.

Frank, R. G. and R. J. Zeckhauser (2018, January). High-priced drugs in Medicare Part D: Diagnosis and potential prescription. Working Paper 24240, National Bureau of Economic Research.

Funk, R. J. and J. Owen-Smith (2017). A dynamic network measure of technological change. *Management Science 63*(3), 791–817.

Garber, A. M., C. I. Jones, and P. Romer (2006). Insurance and incentives for medical innovation. In *Forum for Health Economics & Policy*, Volume 9. De Gruyter.

Garthwaite, C. and F. S. Morton (2017). Perverse market incentives encourage high prescription drug prices. *ProMarket Blog Post*. https://promarket.org/perversemarket-incentives-encourage-high-prescription-drug-prices.

Government Accountability Office (2019, July). Medicare Part D: Use of pharmacy benefit managers and efforts to manage drug expenditures and utilization. GAO-19-498 https://www.gao.gov/assets/710/700259.pdf.

Harris, P. (2007, March). CVS finally wins Caremark for $24 bln. *Reuters*. `https://www.reuters.com/article/us-caremark-cvs/cvs-finally-wins-caremark-for-24-bln-idUSWEN549420070316`.

Hoadley, J., L. Summer, E. Hargrave, J. Cubanski, and T. Neuman (2011). Analysis of medicare prescription drug plans in 2011 and key trends since 2006. *Kaiser Family Foundation Issue Brief. The Henry J. Kaiser Family Foundation.*

Huskamp, H. A., A. M. Epstein, and D. Blumenthal (2003). The impact of a national prescription drug formulary on prices, market share, and spending: Lessons for Medicare? *Health Affairs 22*(3), 149–158.

IQVIA (2019). The global use of medicine in 2019 and outlook to 2023. Technical report, IQVIA Institute for Human Data Science.

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

Kamal, R., C. Cox, and D. McDermott (2018, February). What are the recent and forecasted trends in prescription drug spending? Peterson-Kaiser Health System Tracker. `https://www.kff.org/slideshow/what-are-the-recent-and-forecasted-trends-in-prescription-drug-spending/`.

Krieger, J., D. Li, and D. Papanikolaou (2017). Developing novel drugs. *Available at SSRN 3095246*.

Kuribayashi, R., M. Matsuhama, and K. Mikami (2015). Regulation of generic drugs in Japan: The current situation and future prospects. *The AAPS journal 17*(5), 1312–1316.

Kyle, M. K. and A. M. McGahan (2012). Investments in pharmaceuticals before and after TRIPS. *Review of Economics and Statistics 94*(4), 1157–1172.

Lakdawalla, D. and N. Sood (2009). Innovation and the welfare effects of public drug insurance. *Journal of public economics 93*(3-4), 541–548.

Lee, J. (2012, April). Express Scripts buys Medco for $29 billion. *Modern Healthcare*. `https://www.modernhealthcare.com/article/20120402/NEWS/304029961/express-scripts-buys-medco-for-29-billion`.

Li, D., P. Azoulay, and B. N. Sampat (2017). The applied value of public investments in biomedical research. *Science 356*(6333), 78–81.

Maas, A. (2018, September). As new formulary exclusions are unveiled, drugmakers need to communicate value. *MMIT*. https://www.mmitnetwork.com/member-content/as-new-formulary-exclusions-are-unveiled-drugmakers-need-to-communicate-value.

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

Mathews, A. W. and J. Walker (2015, March). UnitedHealth to buy Catamaran for $12.8 billion in cash. *Wall Street Journal*. https://www.wsj.com/articles/unitedhealth-to-buy-catamaran-for-12-8-billion-in-cash-1427709601.

Miller, S. and P. Wehrwein (2015). A conversation with Steve Miller, MD: Come in and talk with us, pharma. *Managed care 24*(4), 27–8.

Pinsonault, P. (2002, June). When your drug is not on formulary. *PharmExec.com*. http://www.pharmexec.com/when-your-drug-not-formulary.

Pollack, A. (2014, June). Health insurers are pressing down on drug prices. *The New York Times*. https://www.nytimes.com/2014/06/21/business/health-plans-bring-pressure-to-bear-on-drug-prices.html.

Reinke, T. (2015). PBMs just say no to some drugs–but not to others. *Managed Care 24*(4), 24–25.

Rodwin, M. A. (2019, November). What can the United States learn from pharmaceutical spending controls in France? Technical report, Commonwealth Fund Issue Brief.

Schafer, J. (2018, November). Designing clinical trials with the payer in mind. *Clinical Leader*. https://www.clinicalleader.com/doc/designing-clinical-trials-with-the-payer-in-mind-0001.

Scott-Levin (2001, Spring). Managed Care Formulary Drug Audit.

Seabright, P. (2013). Research into biomarkers: How does drug procurement affect the design of clinical trials? *Health Management, Policy and Innovation 1*(3), 1–15.

Siegal, Y. and S. Shah (2019, March). Optimizing market access: How therapeutic area dynamics could influence strategy. *Deloitte Insights*. https://www2.deloitte.com/

us/en/insights/industry/life-sciences/pharmaceutical-pricing-market-access.html.

Stanford, J. (2020, July). Price controls would throttle biomedical innovation. *Wall Street Journal 41*.

Tamblyn, R., R. Laprise, J. A. Hanley, M. Abrahamowicz, S. Scott, N. Mayo, J. Hurley, R. Grad, E. Latimer, R. Perreault, et al. (2001). Adverse events associated with prescription drug cost-sharing among poor and elderly persons. *JAMA 285*(4), 421–429.

The Doctor-Patient Rights Project (2017, December). The de-list: How formulary exclusion lists deny patients access to essential care. Technical report. https://www.healthstrategies.com/sites/default/files/agendas/2015_PBM_Research_Agenda_RA_110714.pdf.

Van Nuys, K., G. Joyce, R. Ribero, and D. Goldman (2018, February). Prescription drug copayment coupon landscape. *University of Southern California Schaeffer Center White Paper*. https://healthpolicy.usc.edu/research/prescription-drug-copayment-coupon-landscape/.

Wang, Y. R. and M. V. Pauly (2005). Spillover effects of restrictive drug formularies on physician prescribing behavior: Evidence from Medicaid. *Journal of Economics & Management Strategy 14*(3), 755–773.

Werble, C. (2014, September). Pharmacy benefit managers: Health policy brief. Technical report, Health Affairs.

WHO Collaborating Centre for Drug Statistics Methodology (2010). Guidelines for ATC classification and DDD assignment. Technical report, World Health Organization. https://www.whocc.no/filearchive/publications/2011guidelines.pdf.

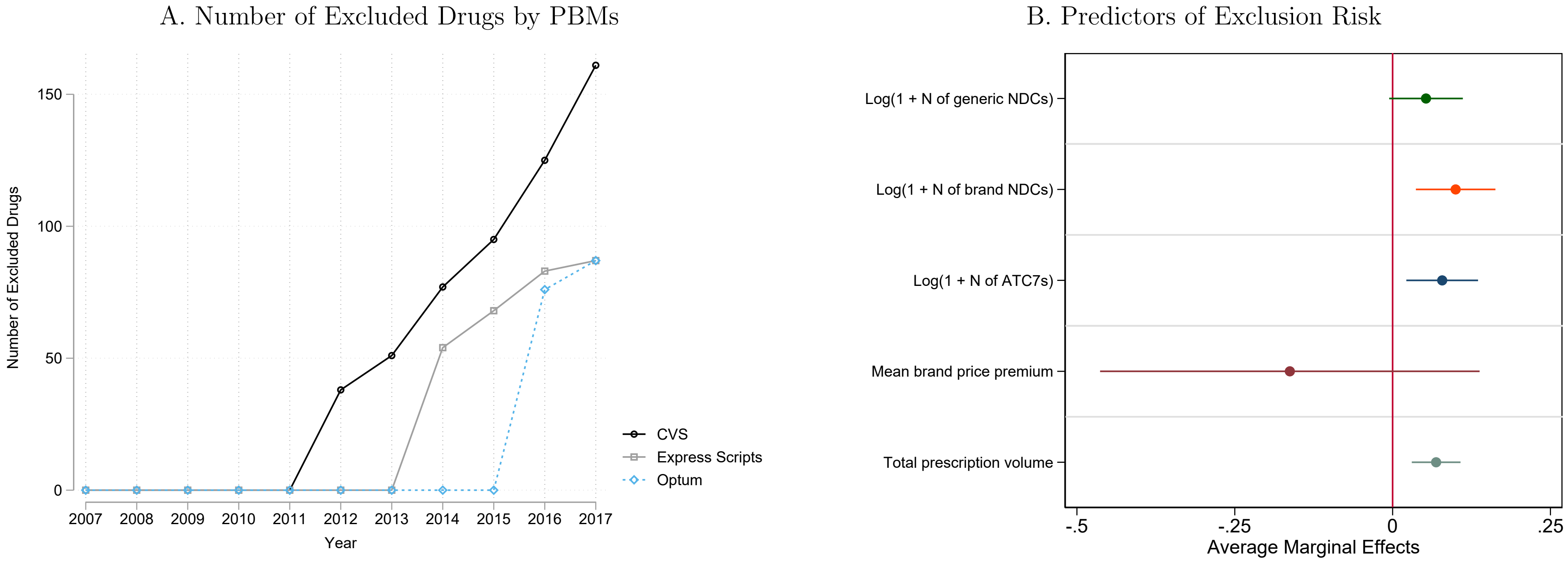
Wu, L., D. Wang, and J. A. Evans (2019). Large teams develop and small teams disrupt science and technology. *Nature 556*, 378–382.

Xcenda (2020, September). Skyrocketing growth in PBM formulary exclusions raises concerns about patient access. Technical report, Amerisource Bergen.

Yin, W. (2008). Market incentives and pharmaceutical innovation. *Journal of Health Economics 27*(4), 1060–1077.

Zycher, B. (2006). *The human cost of federal price negotiations: the Medicare prescription drug benefit and pharmaceutical innovation*. Manhattan Institute, Center for Medical Progress.

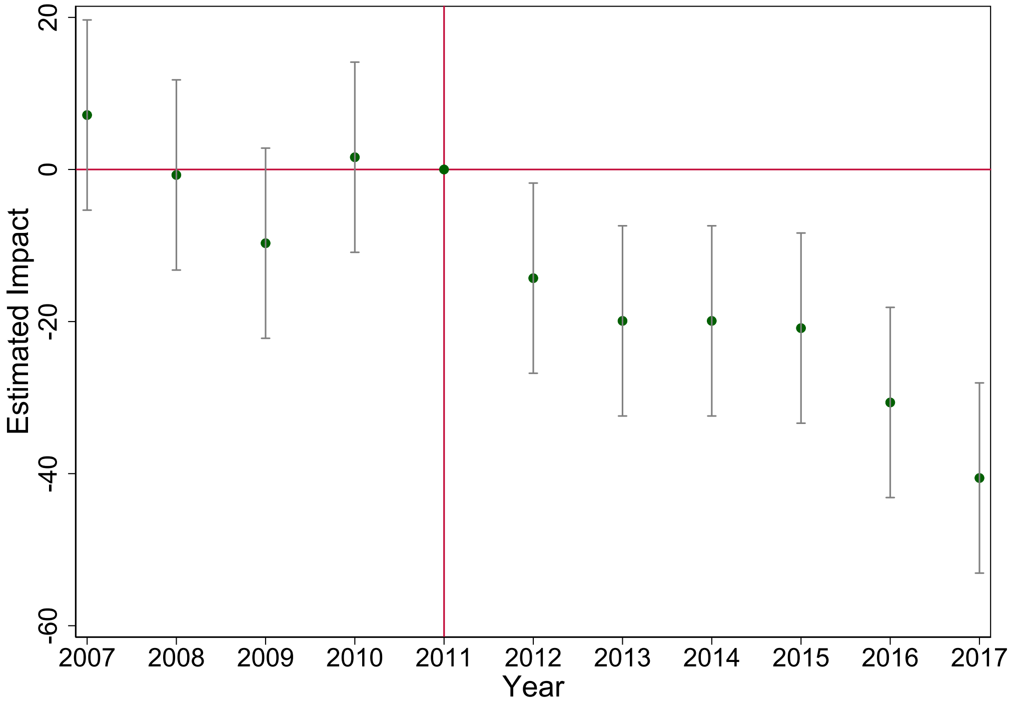
Figure 1: Trends and Predictors of Exclusion


A. Number of Excluded Drugs by PBMs

B. Predictors of Exclusion Risk

Notes: This figure displays the trends and predictors of exclusion. In Panel A, we plot the number of drugs excluded by each of the three largest Pharmacy Benefit Managers. CVS was the first to begin excluding drugs in 2012, followed by Express Scripts in 2014 and OptumRx in 2016. In Panel B, we used the 2011 market characteristics of the ATC4 class to predict exclusion risk. The plotted average marginal effects were generated by conducting bivariate Logit regressions of whether an ATC4 class had at least one drug excluded in 2012 or 2013 on each characteristic of the ATC4 class. Independent variables were standardized (divided by their standard deviation). Data on prices, the number of brand and generic NDCs, and the number of ATC7s are from FDB; data on total prescription volume are from the 2011 Medical Expenditure Panel Survey.

Figure 2: Impact of Predicted Exclusion Risk on New Drug Development: Event Study



Notes: Figure displays coefficient estimates and 90% confidence intervals from a modified version of Equation (2). The outcome variable is the annual count of new development activity (across all stages). To generate the event study graph, we replace the single post-period indicator variable ($\mathbb{I}(\text{Year} \geq 2012)$) with a vector of indicator variables for each year before and after the introduction of PBM exclusion lists in 2012. We plot the coefficients on the interaction of these year indicators and a continuous measure of predicted exclusion risk. (Exclusion risk is predicted using 2011 market characteristics, prior to the introduction of PBM formulary exclusions. Details on the prediction of exclusion risk can be found in Table 2.) The regression controls for ATC4 fixed effects and year fixed effects. The sample includes 1,397 ATC4-year observations.

Table 1: Summary Statistics

(A) New Drug Development

	Mean	Std. Dev.	Median
All	30.61	42.06	13.05
Preclinical	17.39	26.13	6.64
Phase 1	6.54	8.84	3.07
Phase 2	4.57	6.04	2.17
Phase 3	2.11	3.04	1.04
Launch	1.02	1.63	0.31

(B) ATC4 Characteristics

ATC4 market characteristics in 2011	ATC4s with early exclusions	ATC4s without early exclusions
Mean N of generic NDCs	767.9	310.3
Mean N of brand NDCs	268	106.8
Mean N of ATC7s within ATC4	14.60	8.518
Mean brand price - mean generic price	5.822	55.98
Mean total prescription volume (millions)	70.46	17.63
Number of ATC4s	15	112

Notes: Panel A summarizes the annual drug development activity from 2007-2017 in the Cortellis data. The sample includes 1,397 ATC4-year observations. The panel reports the annual number of drug candidates within an ATC4 class that entered different development stages. Panel B summarizes ATC4 market characteristics in 2011. Column 1 reports results for ATC4 classes with at least one excluded drug in 2012-2013; Column 2 reports results for ATC4s with no exclusions in 2012-2013. Data on pricing and the number of available drugs are from First Data Bank; data on total prescription volume are from the 2011 Medical Expenditure Panel Survey.

Table 2: Predicting Exclusion Risk

VARIABLES	(1) Exclusion
Log(1 + N of generic NDCs)	-0.0543**
	(0.0252)
Log(1 + N of brand NDCs)	0.0527
	(0.0415)
Log(1 + N of ATC7s)	0.0861
	(0.0532)
Mean brand price - mean generic price	-0.000695
	(0.000616)
Total prescription volume	1.37e-09**
	(6.17e-10)
Observations	127
Pseudo R2	0.241

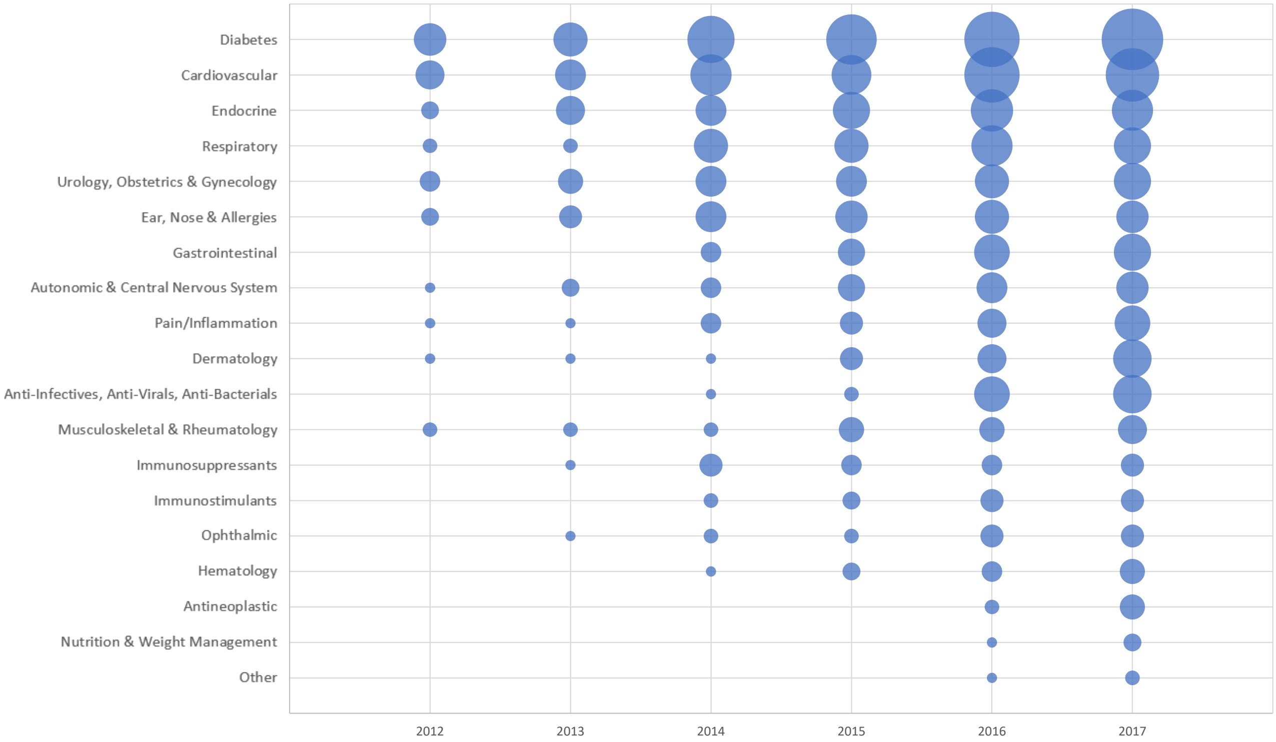
Notes: We used the above 2011 market characteristics of the ATC4 class to predict exclusion risk. Using a logit model, we regressed whether an AT4 class had at least one drug excluded in 2012 or 2013 on all of the characteristics of the ATC4 class listed in the table; average marginal effects are reported. We then used the regression's fitted values to construct predicted exclusion risk of each ATC4. Data on prices, the number of brand and generic NDCs, and the number of ATC7s are from FDB; data on total prescription volume are from the 2011 Medical Expenditure Panel Survey.

Table 3: Impact of Predicted Exclusion Risk on New Drug Development

VARIABLES	(1) New Development	(2) New Development	(3) Log(1+New Dev.)	(4) Log(1+New Dev.)
Post X Pr(Exclusion)	-24.04***	-21.99***	-0.382***	-0.333***
	(5.898)	(6.575)	(0.108)	(0.115)
Observations	1,397	1,397	1,397	1,397
Year FE	YES	YES	YES	YES
ATC FE	YES	YES	YES	YES
Market Controls	NO	YES	NO	YES

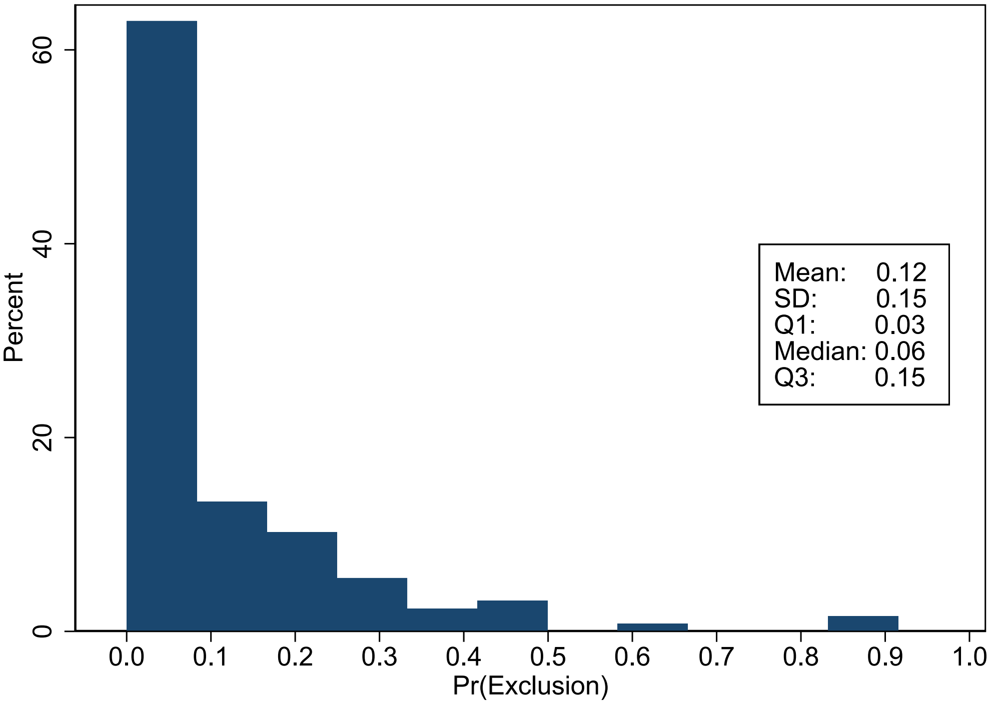
Notes: This table reports results from estimation of equation (2); each column reports a different regression specification. The unit of observation is an ATC4 drug class × year. The outcome variable "New Development" is the annual count of new development activity (across all stages). The treatment variable is a continuous measure of predicted exclusion risk. (Exclusion risk is predicted using 2011 market characteristics, prior to the introduction of PBM formulary exclusions. Details on the prediction of exclusion risk can be found in Table 2.) The "Post" period comprises years 2012 and later, after the introduction of PBM formulary exclusions. All specifications include year fixed effects and ATC4 fixed effects. Columns 2 and 4 include time-varying controls for each of the drug class characteristics listed in Table 1. Standard errors are clustered at the ATC4 level. Statistical significance is indicated as: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Figure A.1: Number of Excluded Drugs by Disease Categories



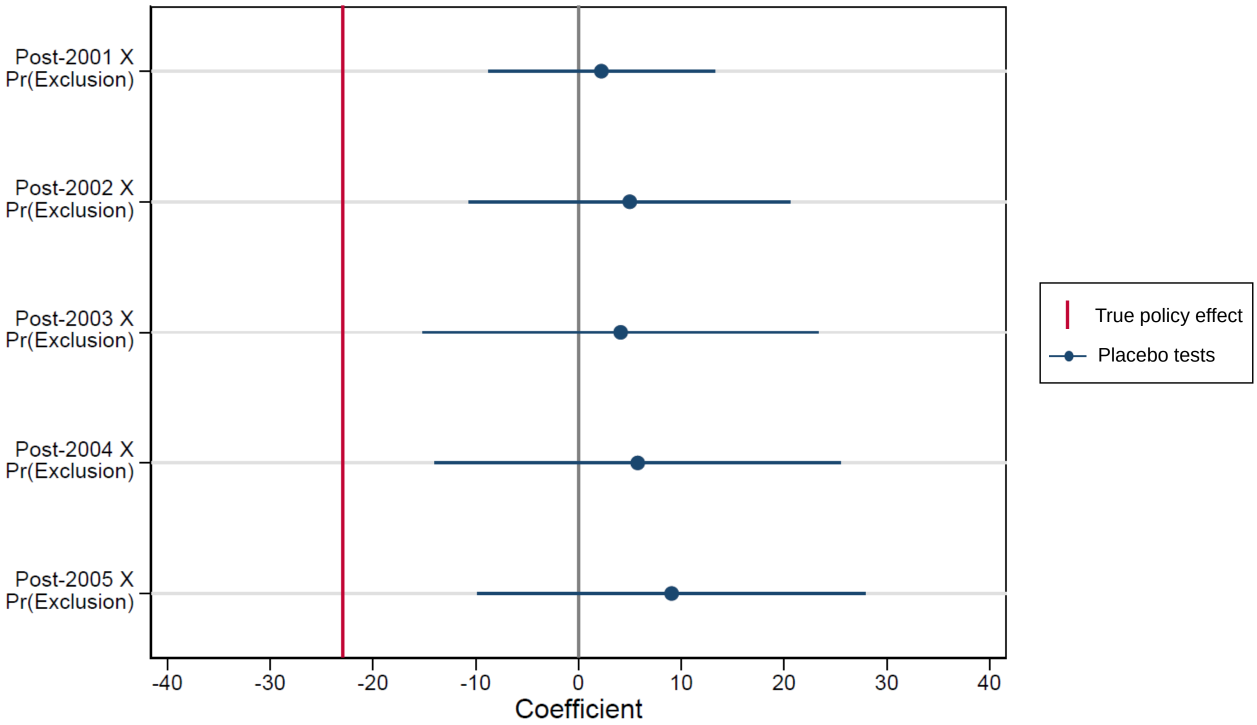
Notes: Each bubble represents a disease category in a year, and the size of the bubble reflects the number of drugs that were excluded by CVS, Express Scripts, or OptumRx in that disease category. There were a total of 300 drugs that were ever excluded from 2012-2017 by at least one of the three PBMs. Of these 300 excluded drugs, we were able to match 260 of them to the First Data Bank data, from which we obtained the ATC4 data and manually matched each ATC4 to a disease category. This disease taxonomy was adapted from the disease categories provided by the PBMs in their exclusion lists and summarized by The Doctor-Patient Rights Project (2017).

Figure A.2: Distribution of Predicted Exclusion Risk



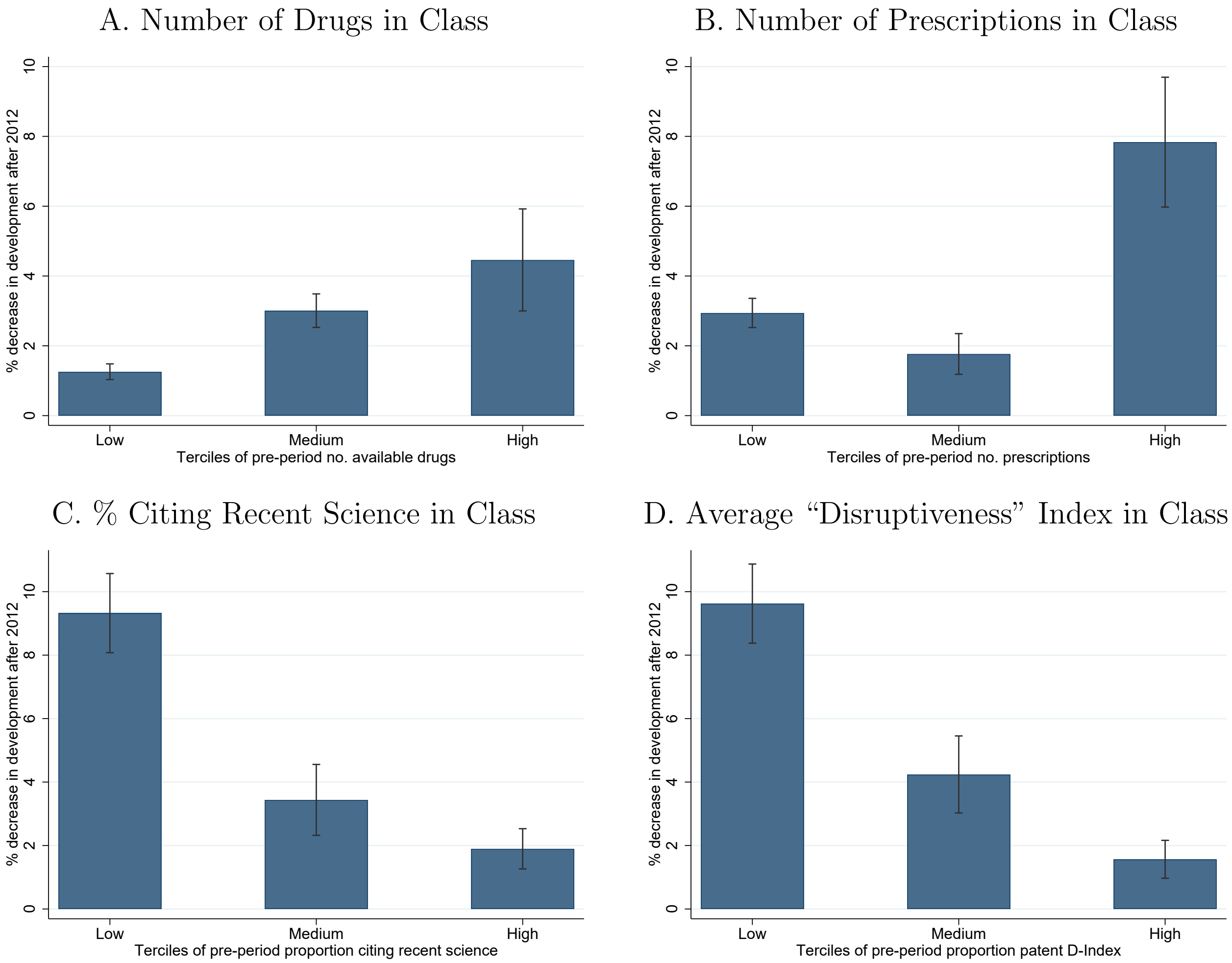
Notes: This histogram plots the distribution of predicted exclusion risk of the 127 ATC4s in our main analyses. Summary statistics are also provided. See notes to Table 2 for details on how the exclusion risk was calculated.

Figure A.3: Placebo Test: Impact of Predicted Exclusion Risk on New Drug Development



Notes: For a more detailed discussion of this placebo analysis, see Appendix B. This coefficient plot shows the "placebo tests" of the results reported in Column 2 of Table 3. The red line indicates the baseline, true policy estimate; it reports β_1, the coefficient on predicted exclusion risk interacted with a post period indicator from Equation 2. This true policy estimate of -22.96 is statistically significant and parallels the specification in Column 2 of Table 3, but the only difference is that when constructing the exclusion risk, we dropped the price variables due to missing historical price data covering the placebo policy periods. The blue coefficients report the "placebo tests" coefficients and 95% confidence intervals, paralleling results reported in Column 2 of Table 3. First, as in the exclusion risk used in Table 3, the model to predict exclusion risk was constructed by using 2011 market characteristics to predict exclusions by 2013, but now we applied the coefficients from this regression to 2001, 2002, 2003, 2004, or 2005 market characteristics to construct new versions of the exclusion risk. Second, the pre-period and post-periods were adjusted depending on the placebo policy year, such that we use the same number of pre- and post-period years as Table 3. For instance, for the 2002 placebo policy, the pre-period was 1997-2001, the post-period was 2002-2007, and we used 2001 market characteristics to construct the exclusion risk. Due to lack of market characteristics data in the earlier period of the data, 3 ATC4s were dropped from the sample for 2006 and 2005 placebo policies, 4 ATC4s for 2004 placebo policy, and 5 ATC4s for 2003 and 2002 placebo policies. None of the placebo estimates were statistically significant.

Figure A.4: Counterfactual Development Activity by Pre-Period Attributes of Drug Class: Existing Therapies, Prescriptions, and Scientific Novelty



Notes: This figure displays the percent decrease in annual development attributable to exclusions. Predictions are based on our estimation of equation (2), matching the specification reported in Table 3 Column 2. To measure predicted new drug candidates in the presence of exclusions, we calculate the fitted value of drug development activity for every year of the post-period. To recover the predicted new drug candidates absent exclusions, we repeat this exercise after setting the treatment variable $\Pr(\text{Excluded})_c \times \mathbb{I}(\text{Year}_t \geq 2012)$ equal to zero for all observations. The figure shows the percent difference between predictions at the ATC4 × year with and without exclusions, averaged over the post-period (2012-2017). In Panel A, we group ATC4 drug classes by terciles of the number of existing drugs in the class (in 2011); data on existing drugs is from First Data Bank. In Panel B, we group ATC4 drug classes by the number of prescriptions written in the class (in 2011); data on prescriptions is from the 2011 Medical Expenditure Panel Survey. Drug classes are weighted by the number of drugs with advancing development over the pre-period. In Panels C and D, drug classes are divided into terciles according to attributes of patents associated with drug development activity over the pre-period, averaged from 2007-2011. Panel C groups drug classes by the share of pre-period patents in a drug class citing recent science as of 2011 (recent is defined as publications since 2006). Panel D groups drug classes by the average "disruptiveness" index of patents in the drug class over the pre-period, which is a measure that captures how disruptive the scientific articles associated with the patent are; the index ranges from -1 (least disruptive) to 1 (most disruptive) and was originally developed by Funk and Owen-Smith (2017).

Table A.1: Examples of ATC4 Codes Defining Drug Markets

A10 Diabetes drugs
A10A Insulins and analogues
A10B Blood glucose lowering drugs, excluding insulins
A10X Other drugs used in diabetes
C07 Beta blocking drugs
C07A Beta blocking agents
C07B Beta blocking agents and thiazides
C07C Beta blocking agents and other diuretics
C07D Beta blocking agents, thiazides and other diuretics
C07E Beta blocking agents and vasodilators
C07F Beta blocking agents, other combinations

Notes: This table provides examples of ATC4 classes for illustrative purposes. Our sample includes 127 distinct ATC4 classes. A complete listing of the ATC4 class definitions that guided this analysis can be found in WHO Collaborating Centre for Drug Statistics Methodology (2010).

Table A.2: Prescription Volume

A. Summary Statistics, Part D Claims per Drug

	Mean	Std. Dev.	Median	Count
Claims for non-excluded drugs	178,503	932,026	3,841	3,046
Claims for excluded drugs	477,332	1,220,225	52,929	791
Market share, non-excluded drugs	0.225	0.328	0.042	3,046
Market share, excluded drugs	0.116	0.213	0.029	791

B. Impact of Exclusions on Prescription Volume

	(1)	(2)
VARIABLES	Log(Market Share)	Log(Market Share)
Number of Excluding PBMs	-0.206**	-0.293***
	(0.0823)	(0.0756)
Observations	3,699	3,475
Drug FE	YES	YES
Cohort X Year FE	YES	YES
Market Controls	NO	YES

Notes: For a more detailed discussion of this analysis, see Appendix A. Panel A reports summary statistics from the Medicare Part D public use file. Data tracks annual claims per drug in 2012-2017; the unit of observation is the drug-year pair. Market share is calculated as the fraction of prescription drug claims in the ATC4 class that are for the index drug. The table compares drugs that were ever excluded to those that were never excluded during the sample period. Panel B estimates the impact of PBM formulary exclusion on the volume of Medicare Part D insurance claims. The unit of observation is a drug × year. The outcome variable is the annual market share of the index drug relative to all other drugs in the ATC4 class, described in Panel A. The key independent variable of interest is the number of PBMs excluding the drug that year. All regressions include drug fixed effects and drug age × calendar year fixed effects. (Drug age is measured as number of years elapsed since market entry.) Specification (2) includes additional controls for ATC4 class × calendar year fixed effects to account for trends in demand for different drug classes. We analyze exclusions on 161 excluded drugs that are prescribed to Medicare Part D enrollees and are not in a protected class. Standard errors are clustered at the drug level. Statistical significance is indicated as: *** p<0.01, ** p<0.05, * p<0.1.

Table A.3: Early Exclusion Risk and Later Exclusions

	(1)	(2)
VARIABLES	Late Exclusion	Late Exclusion
Standardized exclusion risk	0.189***	0.134**
	(0.0468)	(0.0543)
Observations	127	112
Sample	All ATC4s	ATC4s without early exclusions
Fraction with Late Exclusions	0.39	0.31

Notes: Using a logit regression, we investigate whether ATC4 classes that were highly predicted to be excluded by 2013 were more likely to be actually excluded later after 2013. Early exclusion risk is a continuous measure defined using the same specification underlying Table 3; we used 2011 market characteristics of the ATC4 class to predict whether the ATC4 class was at risk of exclusion by 2013. We then standardized this early exclusion risk variable, dividing by its standard deviation. The outcome variable, late exclusion, is a binary variable that indicates whether the ATC4 was on any of the PBM's exclusion list at least once in 2014-2017. Column 1 includes all ATC4s, while Column 2 drops ATC4s that were actually excluded by 2013. Average marginal effects are reported. Statistical significance is indicated as: *** p<0.01, ** p<0.05, * p<0.1.

Table A.4: Impact of Predicted Exclusion Risk on New Drug Development By Stages

VARIABLES	(1) All	(2) Preclinical	(3) Phase 1	(4) Phase 2	(5) Phase 3	(6) Launch
Post X Pr(Exclusion)	-21.99***	-11.05***	-6.010***	-3.831***	-1.100**	0.220
	(6.575)	(3.405)	(2.078)	(1.350)	(0.422)	(0.496)
Observations	1,397	1,397	1,397	1,397	1,397	1,397
Year FE	YES	YES	YES	YES	YES	YES
ATC FE	YES	YES	YES	YES	YES	YES
Market Controls	YES	YES	YES	YES	YES	YES
N of Drug Candidates Mean	30.61	17.39	6.54	4.57	2.11	1.02

Notes: See notes to Table 3. Each column reports a regression with a different outcome variable. Column 1 replicates the result reported in Table 3 Column 2 on total development activity. The additional columns decompose this affect to explore how drug development changes at each phase, moving from the earliest observed preclinical activity in Column 2 through the each phase of clinical trials and eventual launch on the market. Standard errors are clustered at the ATC4 level. Statistical significance is indicated as: *** p<0.01, ** p<0.05, * p<0.1.

Table A.5: Impact of Predicted Exclusion Risk on New Drug Development: Wild Cluster Bootstrap

VARIABLES	(1) New Development	(2) Log(1+New Dev.)
Post X Pr(Exclusion)	-21.99*** [-37.79, -5.854]	-0.333** [-.5375, -.03391]
Observations	1,397	1,397
Year FE	YES	YES
ATC FE	YES	YES
Market Controls	YES	YES

Notes: Columns 1 and 2 of this table repeat the specifications reported in Table 3 Columns 2 and 4, but now using wild cluster bootstrap to calculate the 95% confidence interval (rather than using conventional inference). Clustering is performed at the ATC4 level. Statistical significance is indicated as: *** p<0.01, ** p<0.05, * p<0.1.

Table A.6: Impact of Predicted Exclusion Risk on New Drug Development: Alternative Functional Forms

	(1)	(2)	(3)	(4)
VARIABLES	IHS New Dev	IHS New Dev	Poisson New Dev	Poisson New Dev
Post X Pr(Exclusion)	-0.368***	-0.317**	-0.524***	-0.455***
	(0.123)	(0.131)	(0.0834)	(0.0999)
Observations	1,397	1,397	1,397	1,397
Year FE	YES	YES	YES	YES
ATC FE	YES	YES	YES	YES
Market Controls	NO	YES	NO	YES

Notes: These results parallel the results in Table 3, but with alternative functional forms. Columns 1-2 report regressions using the inverse hyperbolic sine transformation of development activity as the outcome, while Columns 3-4 report results using Poisson regressions. Standard errors are clustered at the ATC4 level for the regressions with inverse hyperbolic sine transformation, and robust standard errors are reported for the Poisson regressions. Statistical significance is indicated as: *** p<0.01, ** p<0.05, * p<0.1.

Table A.7: Impact of Predicted Exclusion Risk on New Drug Development: Alternative ATC4 Linking

	Direct Linking Approach		*Indirect Linking Approach*	
	(1)	(2)	(3)	(4)
VARIABLES	New Development	New Development	New Development	New Development
Post X Pr(Exclusion)	-20.98***	-18.60***	-4.308***	-4.460***
	(6.053)	(6.749)	(1.331)	(1.474)
Observations	1,397	1,397	1,397	1,397
Year FE	YES	YES	YES	YES
ATC FE	YES	YES	YES	YES
Market Controls	NO	YES	NO	YES

Notes: For a more detailed discussion of ATC4 linking, see Appendix C. These results parallel the specification underlying Table 3, but with alternative methods for linking drug candidates to ATC4 classes. We have replaced our baseline outcome measure of development activity with two alternative outcomes that take different approaches to matching. In Columns 1-2, we only count track development activity among the subset of drug candidates for which Cortellis directly reports the drug class. In Columns 3-4, we impute ATC4s from ICD9 codes for all drug candidates, rather than relying on Cortellis' reporting of drug class. Standard errors are clustered at the ATC4 level. Statistical significance is indicated as: *** p<0.01, ** p<0.05, * p<0.1..

Table A.8: Impact of Exclusion Risk on New Drug Development: Alternative Definitions of Exclusion Risk

	Predicted Count Exclusion		*Predicted Share Exclusion*		*Realized Exclusion*	
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	New Dev.	New Dev.	New Dev.	New Dev.	New Dev.	New Dev.
Post X Exclusion Risk	-7.867***	-7.136**	-59.12*	-56.76*	-5.824**	-4.534**
	(2.578)	(2.748)	(33.77)	(31.22)	(2.568)	(2.290)
Observations	1,397	1,397	1,397	1,397	1,397	1,397
Year FE	YES	YES	YES	YES	YES	YES
ATC FE	YES	YES	YES	YES	YES	YES
Market Controls	NO	YES	NO	YES	NO	YES

Notes: For a more detailed discussion of alternative measures of exclusion risk, see Appendix D. This table reports results from estimating a modified version of Equation (2), applying alternative definitions of exclusion risk. Instead of defining exclusion risk as whether an ATC4 class is predicted to have at least one drug with an exclusion as in Table 3, the exclusion risk here is defined as how many drugs are predicted to be excluded in an ATC4 class in Columns 1-2 and what share of drugs are predicted to be excluded in an ATC4 class in Columns 3-4. In Columns 5-6, rather than using continuous measures of predicted exclusion risk as our measure of treatment, we use a binary definition of treatment by looking at realized exclusions: whether at least one drug in an ATC4 class was actually on a PBM exclusion list. For further details on the regression specifications, see notes to Table 3. Standard errors are clustered at the ATC4 level. Statistical significance is indicated as: *** p<0.01, ** p<0.05, * p<0.1.

Table A.9: Impact of Predicted Exclusion Risk on New Drug Development: Incorporating Coupon Data

A. Predicting Exclusion Risk with Coupon Data

VARIABLES	(1) Exclusion
ATC4 class with copay coupons	0.153***
	(0.0495)
Log(1 + N of generic NDCs)	-0.0412*
	(0.0246)
Log(1 + N of brand NDCs)	0.0304
	(0.0383)
Log(1 + N of ATC7s)	0.0519
	(0.0471)
Mean brand price - mean generic price	-0.000580
	(0.000553)
Total prescription volume	1.03e-09*
	(5.94e-10)
Observations	127

B. Impact of Predicted Exclusion Risk on New Drug Development

VARIABLES	(1) New Development	(2) New Development	(3) Log(1+New Dev.)	(4) Log(1+New Dev.)
Post X Pr(Exclusion)	-18.18***	-16.59***	-0.404***	-0.383***
	(4.093)	(3.992)	(0.102)	(0.112)
Observations	1,397	1,397	1,397	1,397
Year FE	YES	YES	YES	YES
ATC FE	YES	YES	YES	YES
Market Controls	NO	YES	NO	YES

Notes: For more details on the measurement of copay coupons see Appendix D. Panel A parallels Table 2 and Panel B parallels Table 3, but now with a measure of drug copay coupons as an additional predictor of exclusion risk. Statistical significance is indicated as: *** p<0.01, ** p<0.05, * p<0.1. .

A Impact of Exclusions on Drug Utilization in Medicare Part D

As discussed in Section 4.1, a PBM's formulary choices (coverage and prices) have been shown to have an impact on patients' drug use. To test whether these patterns hold in our setting, we investigate the link between PBM formulary exclusions and drug sales. Because sales volume is not measured by FDB, we turn to publicly available data on annual Medicare Part D claims volume by drug.[1] Most Medicare Part D plan sponsors contract with PBMs for rebate negotiation and benefit management (Government Accountability Office 2019), and many Part D plans feature closed formularies (Hoadley et al. 2011), making Medicare Part D a suitable context to study the impact of exclusions. This data is available from 2012-2017 and reports the annual number of claims for all drugs with at least 11 claims.

We estimate the following regression equation:

$$\text{Log(Claims)}_{dt} = \beta_1 \text{Excluded}_{dt} + \mathbf{X}_{dt} + \delta_d + \delta_t + \epsilon_{dt} \tag{3}$$

Here, Claims_{dt} refers to the fraction of Medicare Part D claims made on drug d in year t, relative to all other drugs in the ATC4 class (i.e., the drug d's market share in year t). Because the distribution of Part D claims per drug is highly right-skewed (see Appendix Table A.2), we report our results in terms of the natural log of the drug's market share. The key variable of interest is Excluded_{dt}, how many of the three main PBMs were excluding the drug in a given year. We include drug fixed effects in all specifications so that our effect is identified from within-drug changes in formulary exclusion status. We also include drug age × calendar year fixed effects to capture time trends and drug lifecycle patterns.

Our sample consists of branded drugs that were on the market prior to the introduction of exclusions, had no generic substitutes, and have at least 11 annual Part D claims. Because Medicare Part D regulation over this period disallowed formulary exclusions from six protected drug classes, this analysis studies the 161 excluded drugs that are not in a

[1] This data is published annually by the Centers for Medicare & Medicaid Services (2018). We accessed it online at `https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Information-on-Prescription-Drugs/Historical_Data`, in November 2019.

protected class.[2] Further note that in some cases different formulations or packaging of the same drug are listed with separate drug names on formulary exclusion lists, but are reported as a single drug in the Medicare Part D data; we use the more aggregate definition of a drug for this analysis in keeping with the unit of observation in Part D.

In Appendix Table A.2, we show that each excluding PBM decreases a drug's market share by 25% ($e^{-0.293}-1$), relative to comparable drugs that did not experience an exclusion. Column 2 shows that our results are robust to including additional controls for time-varying demand for the drug class, captured with ATC4 X calendar year fixed effects. We note that this analysis does not allow us to measure prescription drug sales that are not claimed in Medicare Part D; if formulary exclusions lead patients to pay fully out-of-pocket for the drugs without requesting insurance coverage, we will not have a record of it in our data.

The effects we measure capture the combined effect of reduced prescriptions for the focal drug, as well as possible reallocation toward non-excluded drugs in its category. These findings show that exclusions had a major impact on shifting sales and market share across competitor drugs, beyond what PBMs previously accomplished for these drugs with traditional demand management tools such as tiering, prior authorization, or step therapy. Moreover, our magnitudes are consistent with anecdotal case by case reporting: for example, after its exclusion by Express Scripts, sales of the asthma inhaler Advair fell 30% while sales for its non-excluded competitor Symbicort increased 20% over the same period (Pollack 2014).

[2] The protected classes are antidepressants, antipsychotics, anticonvulsants, antineoplastic agents, antiretroviral agents, and immunosupressants. Of the 181 excluded drugs prescribed in Part D, only 20 fall into these classes.

B Placebo Policy Analysis

We conduct a series of placebo tests of the introduction of closed formularies. If our measure of exclusion risk captures aspects of a drug class—crowdedness, for instance—that are predictive of declining R&D independent of formulary exclusions, then we would expect drug classes with high exclusion risk (measured in earlier pre-period years) to see innovation fall in response to pre-period placebo exclusion policies. To test this, we use our coefficient estimates reported in Table 2 to identify drug classes that appear at risk of exclusion based on their market characteristics as of each year in 2001-2005. That is, we look for drug classes that, in earlier years, shared the same mix of treatment options and prescription volumes that would have put them at high risk of exclusions in 2011. These are drug classes that, at a given point in time, have a relatively large number of treatment options, as well as high prescription volume. If our results were driven by trends unrelated to exclusions, we should see R&D in these classes fall in the years following our assessment of their exclusion risk. It is worth noting that there were other changes in prescription drug markets over this early pre-period, such as the introduction of Medicare Part D in 2006. While Medicare Part D did affect drug development investments, there is no evidence to suggest that it differentially impacted drug classes based on their exclusion risk. To make sure that our results are not driven by this change, we study a variety of placebo test timing.

Appendix Figure A.3 plots out results for five different tests, corresponding to a placebo policy change in each of the years 2002 through 2006. The blue horizontal lines plot the placebo policy estimates and 95% confidence interval, while the vertical red line highlights the true estimated policy effect. These estimates mirror the specification in Column 2 of Table 3, except that we drop price when constructing the exclusion risk due to missing historical price data covering the placebo policy periods.[3] For example, the 2002 placebo policy estimates a positive $\hat{\beta}$ coefficient of 2.2 on predicted exclusion risk interacted with a post period indicator from Equation 2. For this placebo policy, the post period begins in 2002; exclusion risk is measured using 2001 market characteristics; and we use a corresponding 11-year sample

[3] The true estimated policy effect of -22.96 is statistically significant and very similar to the estimate of -21.99 reported in Table 3.

period from 1997-2007. We end the placebo tests with the 2006 placebo policy change, because its 5-year post-period ends in 2011, the last year of our true policy pre-period.

Appendix Figure A.3 suggests drug classes with similar features to those eventually targeted with exclusions did not experience declining investment over the pre-period; compared to the statistically significant true policy estimate of -22.96, the placebo estimates range from 2.2 to 9.1, and none are statistically significant.

C Linking Drug Candidates to ATC4 Classes

We matched the pipeline drug candidates in Cortellis to ATC4 codes in two ways: directly via EphMRA codes and indirectly via ICD9 codes if the EphMRA codes were missing.

Direct method: matching via EphMRA codes. Cortellis links drug candidates to chemical drug classes (specifically the EphMRA code, which is a close derivative of the ATC classification). Using a manually created crosswalk of EphMRA codes to ATC4 codes, we used the EphMRA codes of the drug candidates to link the drugs to ATC4 classes. A drug can be linked to many ATC4 classes, and we assigned equal weights of 1 to all ATC4 classes that directly matched to a given drug through their EphMRA codes.

Indirect method: matching via ICD9 codes. An alternative way to link the drug candidates to ATC4 classes is through the drugs' areas of therapeutic use (ICD9) provided by Cortellis. Using the drug to ICD9 crosswalk from Cortellis, we linked to a crosswalk of ICD9 to ATC4 codes created by Filzmoser et al. (2009), in which the authors assigned a probabilistic match score of ICD9-ATC4 pairs.[4] Since this results in a drug being matched (indirectly via ICD9) to many ATC4s, we assigned the likelihood of an ATC4 matching to a drug based on the probabilistic match scores from Filzmoser et al. (2009), such that the assigned weights sum to 1 for each drug.

For our main analyses, we matched the drug candidates to ATC4 codes using the direct method via EphMRA codes and used the indirect method via ICD9 codes for drugs with missing EphMRA codes. As shown in Appendix Table A.7, our results are similar regardless of the linking method used.

[4] Filzmoser et al. (2009) merged a dataset of prescriptions (with ATC4 codes) and a dataset of hospital admissions (with ICD9 codes) at the patient-level. Since the ATC4 code of a patient's drug matches to many diagnosis codes of the patient, the authors use a frequency-based measure to calculate a probabilistic match score of an ICD9-ATC4 pair. They conduct this match specific to gender/age group of the patients. For our analysis, we take the average match probability across the gender/age groups for a given ICD9-ATC4 pair.

D Alternative Measures of Exposure to Exclusion Risk

Our analysis is based on differentiating drug classes at varying risk of formulary exclusion. In our primary analysis, we use 2011 ATC4 market level characteristics to predict exclusion risk, defined as whether an ATC4 class is predicted to have at least one drug with an exclusion by 2013. In this section, we describe several alternative approaches.

Alternative functional forms

Appendix Table A.8 tests alternative functional forms for predicting exclusion risk. Columns 1-2 use 2011 ATC4 market characteristics to predict the *count* of excluded drugs in a class by 2013, while columns 3-4 use 2011 ATC4 market characteristics to predict the *share* of excluded drugs in a class by 2013. Like our main measure of exclusion risk, both of these alternatives provide continuous measures of predicted exclusion risk, and thus have the benefit of capturing variation in the *threat* of exclusions—in drug classes that are similar to the initially targeted set but that did not experience early exclusions. Columns 5-6 present results using a binary definition of *realized* exclusions (whether at least one drug in an ATC4 class was on a PBM exclusion list by 2013) and show a similar pattern of results as our main analysis. All of these approaches find that new drug development is declining in exclusion risk. Scaling each of the coefficients in Appendix Table A.8 by the standard deviation of the relevant exclusion risk measure, we predict a similar magnitude reduction in drug development in each specification: 2.7 (column 2), 1.7 (column 4), and 1.5 (column 6).

Copay coupons

Contemporaneous industry reports describe drugs with copay coupons as a major target of PBM formulary exclusions (Foulkes 2015). This motivates an additional analysis using copay coupons as a predictor of exclusion risk. We use copay data from Van Nuys et al. (2018), which are available in the year 2014 and for the top 200 drugs (by sales volume). Because this coupon data comes from the post-period, after the introduction of PBMs' closed formularies, we do not include it in our baseline measure of exclusion risk. We incorporate copay coupons into our prediction of exclusion risk as an additional robustness check. As reported in the logit regression in Panel A of Appendix Table A.9, drug classes targeted with copay coupons have a large and statistically significant increase in exclusion risk, even after conditioning on the other measured market characteristics. Using this augmented measure of exclusion risk, we repeat our analysis testing how exclusion risk predicts changes in development activity after 2012. Results reported in Panel B of Appendix Table A.9 continue to find that drug classes at higher risk of exclusion experience a relative reduction in exclusion risk after 2012; a one standard deviation increase in exclusion risk predicts 3.0 fewer promoted drugs per ATC4 class-year.