# Charted Territory: Mapping the Cancer Genome and R&D Decisions in the Pharmaceutical Industry

Jennifer Kao*

*UCLA Anderson School of Management*

December 21, 2023

**Abstract**

How does comprehensive basic scientific information shape private sector research investments among heterogeneous firms? I assess the impact of large-scale public cancer genome mapping studies, which systematically map the genetic abnormalities in cancer. Using newly-constructed data from cancer genome mapping studies and clinical trials, I find that publicly available mapping information increases private investments in clinical trials by 66%. The large-scale public release of such information has nuanced effects: it disproportionately increases research among incumbents with previously tested drugs for related diseases and spurs research activity among firms with limited access to private mapping information. Cancer maps are associated with improvements in firms' decision-making: when genetic data becomes available, firms are more likely to initiate and advance research investments that are likely to yield promising clinical results.

**Keywords: Innovation, Research and Development, Health Care, Information**

---

* Assistant Professor, Anderson School of Management, 110 Westwood Plaza, Los Angeles, CA 90095, USA (e-mail: jennifer.kao@anderson.ucla.edu). I am particularly grateful to my advisors, David Cutler, Pierre Azoulay, and Amitabh Chandra for detailed feedback on this project. I also thank Ariel D. Stern, Jérôme Adda, Ashish Arora, Juan Alcacer, Megan Bailey, Sharon Belenzon, Kevin Bryan, Samantha Burn, Marika Cabral, Caitlin Carroll, Wesley Cohen, Stephen Coussens, Leemore Dafny, Ariella Kahn-Lang, Joshua Krieger, Timothy Layton, Danielle Li, Abhishek Nagaraj, Ramana Nanda, Adrienne Sabety, Rebecca Sachs, Mark Shepard, Olav Sorenson, Gabriel Tourek, Lisa Xu, Heidi Williams, Wes Yin, and numerous seminar participants for helpful comments and suggestions. Mohan Ramanujan provided invaluable help with the data. The American Society of Clinical Oncology Data Library provided conference abstract data. I also thank several pharmaceutical industry experts for their insights on cancer sequencing. This research is supported by the National Institute on Aging, Grant Numbers T32-AG000186 and R24AG048059.

# 1 Introduction

In 2011, the National Institutes of Health's Cancer Genome Atlas (TCGA) program released information from a landmark study. This information was made public amidst an intensive race by firms competing to develop treatments for ovarian cancer—the fifth leading cause of cancer deaths among United States women (American Cancer Society, 2021). In an advance from previous piecemeal scientific investigations, TCGA had systematically mapped genetic mutations in hundreds of ovarian cancer tumors, which uncovered new mutations and revealed unexpected disease connections (Cancer Genome Atlas Research Network, 2011). Notably, TCGA found that ovarian and breast cancers were linked to BRCA gene mutations. This information had the potential to reshape the competitive landscape for ovarian cancer treatment. TCGA's information could level the field for resource-limited firms with little prior mapping information. It could also trigger a significant shift: TCGA indicated that ovarian cancer patients might benefit from poly (ADP-ribose) polymerase inhibitors, which were being tested by some incumbent firms for BRCA-linked breast cancer. This information may have solidified these firms' competitive advantage, enabling swift repurposing of their drugs for ovarian cancer treatment, and prompted rivals developing novel ovarian cancer drugs to consider market withdrawal (Cockburn et al., 2000).

This case illustrates the central question of this paper: How does the public availability of comprehensive basic scientific information shape innovation in competitive settings with heterogeneous firms making high-stakes research investments?[1] Similar questions are emerging in a diverse range of settings such as energy, computing, and transportation. For example, the US Department of Energy's Marine Energy Atlas provides spatially comprehensive marine resource data, aiding marine technology developers in site selection and device design.[2] Through highlighting promising research opportunities and revealing unexpected connections between them, scientific maps are designed to enable researchers to make more informed investment decisions throughout the research and development (R&D) process.[3] Notably, scientific mapping information may incentivize researchers to either initiate or terminate research investments that they otherwise would not.[4] Such information is especially important in the private sector, a key driver of innovation, comprised of heterogenous players making large, high risk investments (Nagaraj and Stern, 2020).[5] Despite the potential impact of basic scientific maps, empirical evidence on its effects on the rate and direction of innovation is limited. This article seeks to address that gap.

I examine the empirical importance of publicly-available scientific maps on private sector R&D investments. Conceptually, scientific maps, unlike more piecemeal forms of basic scientific

[1] In this paper, "basic scientific information" refers to open-ended knowledge about the "fundamental aspects of phenomena and of observable facts without specific applications towards processes or products in mind." (https://grants.nih.gov/grants/glossary.htm).

[2] See "Democratizing the Data", National Renewable Energy Laboratory, August 19, 2022.

[3] See, e.g., Nelson (1959); Arrow (1962); Rosenberg (1974); Mowery and Rosenberg (1979); David et al. (1992); Arora and Gambardella (1994); Klevorick et al. (1995); Fleming and Sorenson (2003, 2004); Fabrizio (2009).

[4] For an excellent discussion of incentives for innovation, see Bryan and Williams (2021).

[5] For reviews on studies that examine the linkage between basic scientific information and technical progress, see, e.g., Hall and Van Reenen (2000).

information, provide standardized information on the *universe* of potential research opportunities for a given domain: all potential genes associated with a certain disease, all tracts of land, etc. In this paper, I investigate how the placement of these detailed blueprints in the public domain shapes the rate and direction of private sector research.

To begin, I present a simple analytical framework of how scientific mapping information impacts private sector research. This framework characterizes firms competing in an R&D race, reflecting the high costs and first-mover advantages associated with product development and commercialization.[6] There is a deterministic relationship between the perceived feasibility of a high-risk research opportunity and how much firms choose to invest in it. When firms anticipate that a research opportunity is unlikely to result in a commercially viable product, often due to limited information, they invest less in it. Against this uncertainty, public scientific maps may provide valuable but imperfect signals about the probable success of a research opportunity. The framework explains that effects of public mapping data depends on its informativeness (Tranchero, 2023) and key firm characteristics, such as the opportunity to quickly utilize the information and its access to private mapping data.

While the analytical framework provides a useful framework, the paper's primary contribution is to empirically demonstrate the causal impact of publicly available scientific maps on private firms' R&D decisions.[7] The influence of public scientific maps on firms' R&D decisions is difficult to isolate empirically for three reasons. First, in practice, public efforts to generate such data are often targeted towards specific research areas (e.g., specific disease subtypes with greater research potential), leading to selection bias. Second, assessing private firms' R&D effort can be challenging. Researchers commonly rely on indicators like successful research *outputs* (e.g., products, discoveries) or investments that may reflect strategic factors rather than innovation-related considerations (e.g., patents). Third, measuring firm characteristics, such a firm's opportunity to leverage public information and its level of prior existing information, is inherently difficult, making it hard for researchers to assess which firms are more likely to benefit from scientific maps.

Three features of the pharmaceutical setting allow me to overcome these challenges. To address selection concerns, I take advantage of the public availability of large-scale cancer genome mapping initiatives. Building on the foundation provided by the Human Genome Project (HGP), cancer genome mapping initiatives systematically catalogue genetic mutations, revealing informative signals about the potential of each gene (e.g., BRCA2) to drive the progression and growth of a given cancer (Williams, 2013; Jayaraj, 2018).[8] Large-scale cancer mapping efforts, which validate

[6] First movers may be able to secure first-mover advantages through intellectual property protection and regulatory policies (Lieberman and Montgomery, 1988; Berndt et al., 1995; Scherer, 2000). For instance, the U.S. Food and Drug Administration (FDA) grants five years of market exclusivity to a drug that contains no "active moiety" that has been previously approved by the FDA. During that time, the FDA is not permitted to review and approve any generic drugs with the same active moiety.

[7] Perhaps most closely related is an important recent contribution by Nagaraj (2022) who examines how the availability of data from satellite maps shapes successful gold discoveries.

[8] In fact, the HGP was largely motivated by a desire to enable future cancer mapping efforts and the development of cancer therapies. In one of the earliest commentaries calling for the HGP, Nobel laureate Renato Dulbecco (1986, p.1055) wrote, "If we wish to learn more about cancer, we must now concentrate on the cellular genome."

existing knowledge about the disease, are believed to significantly influence the development of new cancer therapies (Mardis, 2018). By leveraging large-scale mapping initiatives which often employ whole-genome sequencing techniques that systematically map all genes—not just those with high research potential—to identify potential mutations, I generate causal estimates of the impact of scientific mapping that are less subject to selection concerns.

To overcome the second challenge, measuring firms' research effort, I focus on how the public disclosure of novel cancer mapping information shapes private clinical trial investment. Within the pharmaceutical industry, a requirement for new product entry is the completion of a series of clinical trials mandated by the FDA, a risky and costly process. Only about 12% of drug candidates successfully proceed from the start of clinical testing to approval, and estimated costs for bringing a drug to market are $2.6 billion (DiMasi et al., 2016).[9] I document novel evidence of how pharmaceutical companies adjust their research investments based on publicly available scientific mapping information. Leveraging the fact that they are required to report all late-stage clinical trial activities, I am able to track firms' research efforts, successful and unsuccessful, as they navigate the R&D process.

To address the challenge of measuring firm characteristics, I utilize the rich variation in my data to identify trials associated with firms with greater opportunity to utilize public mapping information and those facing substantial uncertainty due to limited prior mapping information. For example, by accessing firms' drug development portfolios prior to the introduction of public mapping information, I identify clinical trials where firms are testing drugs that were previously tested for a different disease. This enables me to analyze the impact of public mapping information on firms that, due to the regulatory features of my setting, are most likely to gain an advantage in the race to develop a novel treatment.

I assemble a new dataset of publicly available information produced by 168 large-scale cancer mapping efforts, linked to clinical trials, between 2004 and 2016.[10] Journal submission and publication of results from such mapping efforts provides significant variation in the public disclosure that a mutation exists within a particular gene (e.g., BRCA2) in a specific cancer site (e.g., prostate). I isolate quasi-random variation in the timing that the information was submitted to prominent scientific journals using a gene-cancer-year level difference-in-differences (DID) framework. To mitigate selection concerns related to the timing of public mapping, I control for differences in "research potential" with gene-cancer fixed effects and account for cancer-specific secular changes using cancer-year fixed effects.

Using this data, I find that mutation-related information disclosures from large-scale cancer genomic efforts increased net clinical trial investments by 46% over the study period. In my main

Furthermore, when presenting the completed draft of the human genome (United States Office of the Press Secretary, 2000), geneticist and entrepreneur Craig Venter stated, "As a consequence of the genome efforts...and the research that will be catalyzed by this information, there's at least the potential to reduce the number of cancer deaths to zero during our lifetimes."

[9] Here, the term "drug" refers to an active ingredient treating a specific disease.

[10] This includes cancer mapping efforts from both governmental (e.g., the NIH) and nongovernmental institutions (e.g., Johns Hopkins University).

analyses, I focus on phase II clinical trials—the first major test of a drug's safety and efficacy. These effects—which indicate the fact that cancer mapping information is useful in guiding research activities—are largely driven by private sector research investments. In addition, as clarified by my analytical framework, more informative mapping signals have a larger impact on private sector investments. For example, the disclosure of "driver" mutations, genetic aberrations that are more likely to drive cancer progression and therefore more informative for drug development, have a quantitatively and statistically larger impact on clinical trial investment than "passenger" mutations that do not play a significant role in the advancement of cancer.

The question of how large-scale public investments in information shape the direction of subsequent innovation, specifically regarding *who* drives innovation, is a fundamental issue. Prior empirical studies suggest that incumbent firms excel at bringing new products to market due to scale effects and complementarities with non-research related activities (e.g., marketing, navigating regulation).[11] I find that public mapping information spurs a disproportionate increase in investments in trials already approved or tested drugs, supporting the view that such large-scale data releases may favor incumbents.[12] Furthermore, I observe that the effects of public cancer mapping information are particularly pronounced among trials funded by firms with less prior mapping information (as proxied by having having fewer genetic sequencing-related publications before 2004) compared to firms with extensive prior mapping resources.

Firms navigating the drug development process face a sequence of decisions, each carrying distinct costs and risks. For example, after completing a trial phase, firms must evaluate its clinical findings and decide whether to advance or terminate the project. I extend existing research in this area (e.g., Guedj and Scharfstein, 2004; Krieger, 2021) by showing how access to a reliable and organized mapping of the cancer landscape can help decision makers make data-driven continuation-or-termination decisions (Nelson et al., 2015). I find diseases with mapping information are more likely to have clinical trials with promising patient outcomes. Firms with access to mapping information are more likely to continue investment in drugs with higher prospects of gaining approval and to cease investment in less promising drugs. Together, this suggests that publicly available scientific information not only increases the rate of research, but also assists firms' R&D decision-making processes.

The finding that publicly available scientific maps meaningfully shifted private sector research on cancer treatments has important implications for the potential of information dissemination efforts. Cancer is a disease whose therapeutic market is the largest in terms of global spending—at $133 billion per year—and is the second leading cause of death in the United States, one in which advances yield tremendous value to society (Heron, 2018; IQVIA, 2018). Using the estimating effects of large-scale mapping studies on subsequent research, I conduct a series of "back-of-the-envelope"

[11] In addition, a growing number of papers have highlighted that incumbent firms typically prioritize incremental innovations rather than radical breakthroughs. For a detailed survey, see Cohen (2010).

[12] While the prior innovation literature has focused on incentives to develop novel products, this finding highlights incentives for private firms to identify new uses for *existing* products. This topic has been explored by legal scholars (Eisenberg, 2005; Roin, 2013), but there has been relatively little empirical work on the issue.

calculations that support the cost-effectiveness of public mapping information as a policy instrument for stimulating subsequent research and find that $2 billion cancer mapping effort is calculated to translate into four additional cancer drugs and $4 billion in market value. Considering the significant health benefits associated with new cancer therapies (see, e.g., Cutler, 2008; Lichtenberg, 2015; Dubois and Kyle, 2016), the true societal benefits may be even larger.

The paper proceeds as follows: Section 2 presents the analytical framework. Section 3 introduces the empirical setting and the data. Section 4 presents the results. Section 5 provides a cost-benefit analysis. Finally, Section 6 concludes.

## 2 Analytical framework

How does comprehensive basic scientific information shape private sector research among heterogeneous firms? In this section, I outline a framework to motivate my empirical analyses. I relegate the details to online Appendix A.

This framework builds on a large body of prior work on R&D races that describes the level of firms' research investments as the amount of time they spend engaged in an R&D race, where a key decision for them is whether or not to cease research investment in a specific market opportunity (for an excellent survey of research on R&D races, see Reinganum, 1989).[13] Building on Choi (1991) who studies the role of uncertainty in dynamic R&D races, this framework clarifies how informative public signals create exogenous shifts in the likelihood of success within specific market opportunities, influencing firms' research investments. Furthermore, this framework clarifies how competitive dynamics and firm characteristics interact with the effects of public information. While a large theoretical R&D race literature has highlighted the role of competition (Loury, 1979), research uncertainty (Choi, 1991), and firm heterogeneity (Fudenberg et al., 1983), little empirical research exists to either support or refute these views.[14] This paper addresses this gap.

Formally, I conceptualize R&D as the process of searching for better combinations of two components: research investments and market opportunities (Fleming and Sorenson, 2003, 2004). By research investment, I mean an activity that is involved in assessing and developing a product—e.g., initiating a clinical trial to develop a drug for patients with a particular genetic characteristic, collecting data to develop a prototype for a particular consumer segment. By market opportunity, I mean the set of consumers that the product serves—e.g., patients with a specific subtype of disease related to a specific gene mutation.

Consider two firms, $i = 1, 2$, that compete with each other to be the first to successfully develop and commercialize an invention for a market opportunity (e.g., patients with a specific

[13] For simplicity, this follows a large R&D racing literature that focuses on a firm's decision to stay vs. drop out of the R&D race (Reinganum, 1989). An alternative approach would be to focus on firms' entry and reentry decisions which, while important, are not directly related to the goals of the framework (which is to describe the short-term effects of public mapping information within a competitive environment comprised of heterogeneous firms). Though pharmaceutical firms make strategic product launch decisions (Kyle, 2007), wait-and-see approaches for clinical trial investments are less common. That said, the long-term effects of public mapping information on firms' entry decisions is an important topic for future research.

[14] See Cockburn and Henderson (1994) for a survey.

disease subtype) by making research investments (e.g., conducting clinical trials).[15] At each point in time, $t$, given no success to date, firm $i$ must decide whether or not to continue investing in a market opportunity.[16] The longer a pharmaceutical firm is engaged in the R&D race, the greater its likelihood of advancing from preclinical research (that tests a drug candidate in animals and human cells) to clinical research (that clinically tests a drug in patients with a specific disease subtype).[17] If firm $i$ is the first to successfully yield a commercially viable treatment, it obtains a potential payoff of $V$.[18] If the firm ceases investment, it obtains a payoff of 0 with certainty.

Market opportunities vary in their ex ante potential to successfully yield a commercially viable treatment, with success occurring at a stochastic rate. Following the previous literature on innovation races (e.g., Loury, 1979), the probability that a research investment is successful at a given time $t$ takes an exponential distribution. Therefore, the date at which a research investment yields a commercially viable treatment is denoted by $\tau$, where $\mathrm{P}(\tau < t) = 1 - e^{-\lambda^i t}$ and $\lambda^i$ is the hazard rate or the conditional probability of success, given no prior success. $\tau$ can be interpreted as the point when a pharmaceutical firm's clinical trials demonstrate the safety and effectiveness of a drug for a specific patient disease subtype.[19]

For simplicity, I assume that the success rate for a specific market opportunity is either high (i.e., a "high success rate" market opportunity has a hazard rate $\lambda_H^i$) or low (i.e., a "low success rate" market opportunity has a hazard rate $\lambda_L^i$, where $\lambda_H^i > \lambda_L^i > 0$). Reflecting uncertainty in the research process, firms do not know the true success rate of a given market opportunity. Following the framework and notation of previous dynamic R&D race literature, I assume that, all else equal, firms have a prior belief $p_i$ that the market opportunity has a low success rate.

The introduction of public mapping information can be viewed as the provision of positive signals about the success rate of each market opportunity.[20,21] There are three key features of

[15] I focus on two firms for simplicity, but the results can apply to settings with more firms (see Choi, 1991).

[16] Following prior research (e.g., Budish et al., 2016 and Nagaraj, 2022), research decisions in this framework are made within a single stage, though the results apply to a multistage race setting (see Choi, 1991). Although additional features of staged research investment (e.g., real options) are important, they are not the main focus of this framework.

[17] In other words, the firm will conduct a clinical trial if it is still in the R&D race at $t_i \geq \bar{t}$. I treat the threshold $\bar{t}$ as deterministic for simplicity; in practice this parameter would be endogenous. Note that drug development consists of several phases. As a result, the larger the $t_i$, the greater the likelihood that the firm will progress through the drug development pipeline from phase I to phase III clinical trials.

[18] Here, I follow the theoretical literature on R&D races and assume that the race is winner-take-all: only one firm will win $V$ and once a player wins, the game ends (Loury, 1979). Relaxing the winner-take-all assumption and allowing the second place winners to have $V^S$ where $V^S < V$ would lead to similar results. For simplicity, this framework assumes that both firms receive the same value from winning. See, e.g., Harris and Vickers (1985) for an analysis where the potential payoffs vary across firms (i.e., $V_1 \neq V_2$).

[19] For instances where a firm must conduct several phases of clinical trials, $\tau$ can be viewed as the date when the last clinical trial phase demonstrates that the drug is safe and effective.

[20] As noted in Nagaraj (2022), the primary impact of public mapping studies is informational, rather than financial.

[21] This framework characterizes mapping information as positive signals, rather than either positive or negative signals. Anecdotal evidence suggests that cancer mapping efforts are largely viewed as providing positive information about the existence of genetic aberrations as opposed to negative information about their absence (which could be due to a variety of factors, such as rare mutations being more difficult to detect) (Amar et al., 2017). Incorporating negative signals would imply the possibility of a lower success rate ($p^M > p$) for a market opportunity. The framework's main findings would remain the same.

this information. First, consisting of a set of information signals about the success rate of all possible market opportunities, mapping information helps firms navigate the R&D process by distinguishing whether any given market opportunity has a sufficiently high success rate: a mapping signal reveals that a market opportunity has low success rate with probability $p^M$, where $p^M \leq p_i$.[22] As a result, these signals may prompt firms to increase investment in market opportunities, potentially resulting in high quality inventions, which they may have previously not considered making substantial investments in. Second, these mapping signals may have different strengths that reflects how clinically informative they are for drug development. Third, the comprehensive nature of public mapping information allows it to reveal previously unknown connections across market opportunities. This can lead to disproportionate benefits for specific firms that can quickly leverage these newfound connections, especially in competitive environments where there are first mover advantages.

To illustrate how public mapping might shape the rate and direction of R&D efforts, recall TCGA's ovarian cancer study. This large-scale study revealed novel information about rare gene mutations that may have been overlooked by previous, proprietary small-scale mapping efforts. It provided a genetic blueprint, revealing the structure, organization, and likely function of genetic mutations underlying ovarian cancer. By doing so, it offered information signals about which specific patient disease subtypes with certain genetic features might or might not respond favorably to treatment. This, in turn, may have reduced the risks associated with developing "targeted" therapies for those patients (Collins and McKusick, 2001). Importantly, TCGA's study revealed promising connections between disease subtypes. For example, it revealed that BRCA-mutations occurred in ovarian and breast cancer, highlighting new opportunities for incumbent manufacturers of poly (ADP-ribose) polymerase (PARP) inhibitors to repurpose their drugs for ovarian cancer.[23] Drug repurposing is a more cost-effective, lower-risk alternative to developing new treatments compared to de novo drug development (Greenblatt et al., 2023). This is largely due to regulations that enable manufacturers of previously tested drugs to expedite the drug development process by skipping early stages such as preclinical studies and phase I trials, due to the established safety profile of the drug. Following the TCGA study, an incumbent firm that has developed and previously tested a PARP inhibitor for a related disease would have a higher likelihood of success ($\lambda_H^i$ and $\lambda_L^i$) in developing a treatment for ovarian cancer patients with BRCA-mutations than an entrant firm that has not tested any drugs in related diseases.

This empirical work provides support for the idea that given the uncertainty in the R&D process, public mapping information shapes private sector firms' research investments. Under

[22] For simplicity, I treat the mapping parameters as deterministic; in practice, these parameters could be stochastic in the sense that the signals are informative but noisy. This would allow for false positives and Bayesian firms would have to appropriately account for noise. Incorporating the mapping information uncertainty into the framework would not change the implications, but they are excluded for simplicity. I empirically explore the effects of more vs. less noisy mapping signals in Section 4.4.1.

[23] This can apply to drugs that were approved or previously tested. At the time of TCGA's study, no PARP inhibitors had yet been approved. The first approved PARP inhibitor, olaparib, was not approved until 2014. For more details on poly (ADP-ribose) polymerase inhibitors, BRCA mutations, and ovarian cancer, see Matulonis (2017).

plausible assumptions, the framework yields the following implications. Additional details and proofs are presented in Appendix A.

### 2.1 Effect on the quantity of research investments

**Proposition 1.** *The effect of a public mapping signal on the likelihood that it invests in a clinical trial is increasing in the strength of the mapping signal.*

**Corollary 1.** *Competition lowers the impact of public mapping information on the time a firm spends in the R&D race and its likelihood of investing in a clinical trial.*

A mapping signal for a specific market opportunity can increase a firm's expected payoff, potentially leading to an increase in time spent in the R&D race and an increased likelihood of initiating a clinical trial in the market opportunity. However, the impact of public mapping information on firms' clinical trial investments depends on the strength of these signals. Signals that are weak or not clinically relevant are unlikely to affect firms' research investments. Additionally, competition has a negative impact on a given firm's beliefs about its expected payoffs, leading to a decline in the time that the firm spends in the R&D race (Loury, 1979). As a result, even strong mapping signals may not increase firms' research investments if the level of competition is too high. As a result, the net effect on research investment depends on the strength of the public mapping signals and the level of competition. The extent to which public mapping information shapes competing firms' research investments is an empirical question.

### 2.2 Effect on the direction of research investments

My empirical work will provide support for the idea that private sector research increases in response to public mapping information, suggesting that these maps provide strong and valuable signals for firms' R&D efforts. This section investigates the types of firms (and products) that disproportionately benefit from public mapping information, which carries important policy and managerial implications. For simplicity, I separately consider (i) a firm's relative likelihood of success due to its opportunity to leverage the external information and (ii) its level of uncertainty due to limited prior information.

**Proposition 2.** *Assume that after public mapping information is released, firm 1's has a greater opportunity to leverage the information for quickly producing a commercially viable treatment. This increases firm 1's relative likelihood of success (i.e.,* $\lambda_H^1 > \lambda_H^2$ *and* $\lambda_L^1 > \lambda_L^2$*). If the increase in firm 1's expectation of success is sufficiently large, it may cause firm 1 to spend more time in the race, thereby increasing its likelihood of investing in a clinical trial.*

What is the effect on the newly lagging firm 2 and total private research investments from both firm 1 and firm 2? Holding all else equal, firm 2's decision to stay in the R&D race depends on two factors: the size of the resulting gap between the two firms and the stochastic nature of success. If the gap between firm 1 and firm 2 is sufficiently large, firm 2 may drop out of the race, potentially resulting in a decrease in overall private research investments in the market opportunity (Harris and Vickers, 1985). However, firm 2 might still opt to stay in the race due to the uncertain

nature of R&D, which could still enable it to succeed. This could result in a net increase in private research investments. Ultimately, my empirical work will shed light on how shifts in relative firm opportunity to leverage public mapping information affects the payoffs of newly lagging firms.

**Proposition 3.** *Assume that a firm with sufficiently high prior beliefs that a market opportunity has a low success rate. The impact of public mapping information is increasing in the firm's ex ante level of uncertainty.*

This is motivated by the idea public mapping information is most useful for shaping firms' research investments when they have high levels of uncertainty due to limited prior information (e.g., due to limited resources to fund internal mapping efforts).[24]

### 2.3 Empirical implications

Consistent with previous studies on innovation (e.g., Williams, 2013; Azoulay et al., 2019), my analysis takes place at the market opportunity level, exploring how changes in public mapping information at the disease-level influence private sector clinical trial investments. The framework yields the following testable implications for the net effect of public mapping information:

**Prediction 1.** The likelihood of a private sector clinical trial increases (if public mapping signals are sufficiently strong enough to overcome the negative effects of competition).[25]

**Prediction 2.** The increased likelihood of private sector clinical trial investment is driven by trials testing previously tested drugs (if incumbent firms with such drugs have sufficiently higher likelihood of success, due to having a greater opportunity to leverage the public mapping information).

**Prediction 3.** The likelihood of private sector clinical trial investment in novel drugs also increases, though at a lower rate relative to investment in trials testing previously tested drugs (if the R&D process is sufficiently stochastic, allowing laggards to still achieve success, and/or the technological gap between laggards and leaders is relatively narrow).

**Prediction 4.** The impact of public mapping information on private sector clinical trial investment increases in firms' level of uncertainty.

I use the above predictions as a guide for my examination of the data. Overall, the framework suggests that within competitive environments, public mapping information interacts with firm characteristics in complex ways that may ultimately shape firms' research initiation and termination decisions. If the mapping signals are sufficiently strong and there is substantial firm heterogeneity, then the large-scale release of public mapping information can meaningfully increase private sector

[24] This framework closely follows Choi (1991), whose baseline model assumes common priors about the success rates market opportunity and perfect observability in the sense that firms have similar levels of information. The introduction of public mapping information is consistent with this; this information is available to all firms. Choi (1991) also considers an extension where there is imperfect observability. To clarify the role of a firm's ex ante uncertainty, when testing Proposition 3 (in Prediction 4 below), we expand on this extension considering the possibility that some firms may have different levels of private information (e.g., due to internal mapping efforts).

[25] My empirical work examines the effect of public mapping information within a competitive environment. Empirically, I am not able to directly examine the how competition shapes the effects of cancer mapping information as exogenous changes in competition are rare in this setting.

investments and shift the overall trajectory of innovative activity towards both established firms (e.g., those with previously tested drugs in related diseases) as well as firms facing considerably uncertainty due to limited prior mapping information.

# 3 Empirical setting and data

## 3.1 Scientific background

Cancer is caused by changes in the DNA molecule. A gene is a segment of DNA and a gene mutation is a type of DNA change that can modify normal cell behavior, causing excessive growth and tumor development (Stratton et al., 2009).[26] Mutations can cause a cell to produce proteins that can lead cells to grow quickly and cause damage to neighboring areas (TCGA, 2018). The average tumor contains 33 to 66 mutated genes; the number varies across different types of cancers (Vogelstein et al., 2013). For example, the blood cancer acute myeloid leukemia is associated with a median of 8 mutations, while non-small cell lung cancer can have 150 to 200 mutations per tumor.

I use gene-cancer pairs as my disease unit of analysis. I begin with a list of 80 cancer sites, based on the standard Surveillance, Epidemiology, and End Results (SEER) classification system.[27] Next, I focus on a set of 627 genes listed in the Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer Gene Census (CGC), which consists of the set of genes believed to be causally associated with cancer.[28] This yields 50,160 possible gene-cancer pairs (627 genes $\times$ 80 cancer sites).

I construct a balanced gene-cancer-year panel from 2004, the publication year of the first CGC, through 2016. Since I am interested in quantifying the effect of newly disclosed scientific information (mutation disclosures) on subsequent clinical research investment, I restrict my analysis to the set of gene-cancer pairs that are likely to have scientific or research potential. My concern with including all 50,160 possible gene-cancer pairs is that the set may include pairs where the chance of a clinical trial is essentially zero, which can artificially inflate the statistical significance and the magnitude of the resulting estimates relative to the sample mean. Therefore, I limit my analysis to the set of gene-cancer pairs that have co-appeared in at least two publications prior to 2004.[29] This results in a final dataset of 30,223 gene-cancer pairs (comprised of 462 genes and 80 cancer sites). Online Appendix Table B1 summarizes how the gene-cancer-year panel is constructed.

Sections 3.2 and 3.3 describes the genetic mapping and research investment data used in this analysis. Online Appendix C describes the data construction in more detail.

[26] The underlying mechanics of genetics is much more complex. However, this is the scientific background needed for the purposes of this paper. For more details, see https://ghr.nlm.nih.gov/primer.

[27] I am grateful to Heidi Williams for sharing the SEER crosswalks used in Budish et al. (2016) for this paper.

[28] Each gene in the CGC is paired with a cancer for which there are at least two independent reports linking the gene to the cancer and which are considered to be likely implicated in driving other cancer types. The original version of the CGC was first published in Futureal et al. (2004). The version used here comes from Version 82 (published on August 3, 2017) of the COSMIC database. For more details, see https://cancer.sanger.ac.uk/cosmic/download.

[29] Online Appendix Table C1 provides more detail on how publications are linked to gene-cancer pairs and shows that the results are robust to different publication count thresholds and different publication data sources.

### 3.2 Large-scale public cancer genome mapping efforts

The purpose of cancer genome mapping is to identify the specific genetic mutations associated with different types of cancer. This is done by comparing the DNA sequences of cancer cells to those of normal tissue, either from the same individual or a reference DNA sequence. Online Appendix Figure B1 graphically summarizes this process.

In the past two decades, large-scale systematic cancer genome sequencing initiatives—efforts to catalog and discover mutations in large numbers of tumors—have been an important source of genomic information. These large-scale efforts include TCGA, the Cancer Genome Project, the International Genome Consortium, the Pediatric Cancer Genome Project, and cancer mapping efforts in universities and other research institutions. Two key factors contributed to the rise of these initiatives (Wheeler and Wang, 2013). The first was the 2003 completion of the HGP, which sequenced the human genome and provided a reference for subsequent cancer mapping efforts. The second factor was improvements in sequencing technology which allowed for more accurate, faster, and cheaper sequencing. It is widely reported that the introduction of so-called next-generation sequencing allowed the cost of sequencing per genome (excluding the cost of data analysis) to fall from $95 million in 2001 to $1,000 in 2017 (Wetterstrand, 2018).[30]

I obtain the information produced by these large-scale cancer sequencing efforts (i.e., mutation data at the gene-cancer-level) from the publicly accessible COSMIC and cBioPortal for Cancer Genomics (cBioPortal) databases (Cerami et al., 2012; Gao et al., 2013; Tate et al., 2018). Similar to biological resource centers which serve as "living libraries" for biological materials (Furman and Stern, 2011), both databases act as centralized repositories of mapping data from hundreds of cancer mapping studies. Further, COSMIC and cBioPortal curate and standardize cancer genome data for subsequent researchers, including information about a sequenced tumor's cancer type (e.g., ovarian cancer), associated genetic mutations (e.g., BRCA2), and the date on which the associated mapping study was submitted to a scientific journal for publication.[31]

I focus on mutation information—hereafter, referred to as mapping information—disclosed from 168 cancer mapping efforts, which share three key characteristics. First, the studies are cancer-site specific—recall that the TCGA ovarian cancer study focused only on mapping ovarian cancer tumors. Second, the studies are large-scale and systematic, involving the examination of hundreds of tumors and with 91% of the studies investigating all genes in the protein-coding regions in DNA. Third, in line with existing research that utilizes journal rankings as a proxy for publication impact, I concentrate on cancer mapping studies published in high-ranking scientific journals. Online Appendix Figure B2 shows the number of cancer mapping studies and cancer tumors mapped from 2004 through 2016. The fall after the original sustained increase likely reflects the finite number of cancer sites and the diminishing marginal value of conducting multiple large-scale mapping studies in the same cancer site.

[30] Technologies have evolved from first-generation Sanger sequencing, a method that sequences a single DNA fragment at a time, to next-generation sequencing, which allows parallel mapping of millions of genes at one time.

[31] I focus on mutations that occur in the protein-coding region of the DNA sequence and that are likely to lead to a change in biological structure.

I am interested in research activity initiated after the public disclosure of a mutation in a gene-cancer pair. Before describing my research investment measures, two features of mutation-related information should be noted. First, I focus on the "positive" impact of mutation information on subsequent research activity—i.e., how disclosure that a mutation occurs in a gene-cancer pair can lead to an increase in private sector research activity relative to gene-cancer pairs without mutation information.[32] Second, information produced by large-scale cancer mapping efforts may be made public before the mapping study's official publication date: for instance, pharmaceutical firms may become aware of preliminary mapping results presented at scientific conferences. To approximate the earliest date that mapping information became publicly known, for each gene-cancer pair in my dataset, I identify the first date that a mapping study containing information about the mutation was submitted to a journal.[33]

In a subset of the analyses that follow, I examine how the impact of public mapping information varies based in on the strength of the information. According to the scientific literature, mutations can be considered driver mutations, which are likely to drive the growth and progression of cancer, or passenger mutations, which are likely to be harmless. To determine driver mutations, I employ two strategies (Carr et al., 2016): (1) identifying mutations that are highly likely to be a drivers based on statistical methods used by the cancer-sequencing researchers and (2) classifying mutations that are detected an unusually high number of times ($\geq$ 10 patients) in a particular gene-cancer pair in a given study.[34] These probable driver mutations contain the strongest signal of cancer-causing behavior and are typically described in detail in the associated cancer mapping publication.

## 3.3 Clinical research investments

### 3.3.1 Drug development

The drug development process involves extensive preclinical research, followed by human testing in a series of clinical trials, in which costs increase with each subsequent phase. These trials progress from phase I, which tests safety, to phase II, which tests safety and efficacy in a larger patient group, to phase III, which assesses safety and efficacy in a larger population over a longer period. For cancer drugs, efficacy is often measured by changes in overall survival and the objective response rate, the share of patients whose tumors reduce by a prespecified amount.[35] In the US, once phase III is complete, manufacturers must submit a new drug application (NDA) for FDA review. The clinical

[32]This is consistent with the analytical framework that characterizes public mapping information as providing positive information shocks (see footnote 21).

[33]The submission date is likely to roughly approximate the time at which final results are presented at a scientific conference. For example, results from a TCGA bladder cancer mapping effort were submitted to the scientific journal *Cell* on March 23, 2017 (Robertson et al., 2017), and presented at the American Society of Clinical Oncology (ASCO) Annual Meeting, a major cancer conference, on June 5, 2017 (https://meetinglibrary.asco.org/record/153648/abstract).

[34]In general, identifying the exact relationship between mutations, patient outcomes, and treatments is difficult, and it is not possible to definitively prove that a mutation is a driver or a passenger. Statistical methods to identify probable driver mutations include the Mutation Significance (MutSig) algorithm (Lawrence et al., 2014) and the Mutational Significance in Cancer (MuSiC) algorithm (Dees et al., 2012).

[35]The objective response rate is commonly measured using Response Evaluation Criteria in Solid Tumors (RECIST) criteria. For more details, see: http://recist.eortc.org/.

development process is long (typically taking 8-12 years), costly (typically costing a manufacturer $800 million - $2.6 billion),[36] and risky (only 9% of drugs that begin clinical development ultimately go to market) (DiMasi, 2001; DiMasi et al., 2003; Danzon and Keuffel, 2014).

The development and regulatory approval process is indication specific—i.e., a drug receives approval for a specific therapeutic use (e.g., bevacizumab is approved for the treatment of colorectal cancer). However, more than 60% of approved cancer drugs have multiple indications. To expand a drug's label to include a new use, the manufacturer must undertake additional clinical efficacy trials and submit a supplemental NDA (sNDA) (FDA, 1998b). The amount of resources required depends on the similarity between the original and new use (FDA, 2004). If a drug is approved for one cancer type (e.g., gallbladder) and the manufacturer seeks approval for another tumor type with a common biological origin (e.g., the colon), they may bypass early research stages (preclinical and phase I trials) and require fewer phase II trials. (FDA, 1998a). With less evidence for FDA review, average approval times may be shorter for sNDAs relative to NDAs (DiMasi, 2013).

New use approvals have high expected social value (Berndt et al., 2006; Roin, 2013; Greenblatt et al., 2023). Former director of the National Institutes of Health, Francis Collins, described the clinical testing of existing drugs for new uses as an opportunity to become "more efficient and effective at delivering therapies and diagnostics to patients" (Collins 2011, p. 397). Further, private firms seeking new use approvals may generate useful scientific evidence for clinical decision making, particularly in contexts where the use of a drug for a non-approved ("off-label") use is common. Despite the potentially lower costs associated with seeking new use approvals, it is believed that there is underinvestment in new uses for approved drugs.[37]

### 3.3.2 Clinical trials data

Data on private sector clinical trials comes from the Clarivate Cortellis Competitive Intelligence Clinical Trials Database ("Cortellis"), which includes data from public trial registries. For each clinical trial, Cortellis provides detailed information on the cancer being examined (e.g., prostate cancer), the drug being tested (e.g., olaparib), and the trial start date (as measured by the date on which the first patient was enrolled). Crucially, the clinical trials also list information on protein biomarkers (e.g., the gene EGFR) that are used to guide patient selection.[38] Each patient biomarker can then be linked to the standardized list of genes in the National Library of Medicine's (NLM) gene database to generate a dataset of trials at the gene-cancer level.

In my main analysis, I focus on phase II trials for several reasons. Phase II trials represent the first major financial investment of a particular drug in a specific disease (costing up to $20 million) (Sertkaya et al., 2016). Additionally, they provide a standardized classification system for evaluating cancer efficacy endpoints, enabling a clear comparison of trial results. Further, unlike for phase

[36] These costs estimates reflect the direct cost of research and the opportunity cost of capital. The estimates have been subject to criticism based on small sample sizes, assumptions about the cost of capital, and the confidential nature of the underlying data. Despite this, other efforts have generated similar cost estimates (Avorn, 2015).

[37] This problem, known as the "problem of new uses," arises due to limited patent protection for new uses and the prevalence of off-label drug use (Eisenberg, 2005).

[38] I am grateful to Ariel D. Stern for sharing the cleaned data from Chandra et al. (2017) for this paper.

I, high-quality data exists for phase II trials due to trial registration requirements. More broadly, categorizing trials by phase allows me to examine the impact of public cancer mapping information on private firms' research investment throughout the drug development process. Restricting my sample to those in phase II trials results in a dataset of 57,683 trial-gene-cancer observations.

Next, I categorize trials based on their funding source into two categories: private sector trials (e.g., funded by AstraZeneca) and public sector trials (e.g., funded by the NIH).[39] Online Appendix Figure B3 shows that the share of private sector phase II cancer trials that are gene related or use gene characteristics to guide patient enrollment has been increasing over time.[40]

**Identifying trials testing new uses and new drugs.** To understand the mechanisms influencing the impact of public mapping information, I would ideally perfectly identify (i) research investments conducted by firms with a higher likelihood of success due to a greater opportunity to leverage the public mapping information and (ii) research investments conducted by firms with high level of uncertainty due to limited prior mapping information.

Due to data constraints, I construct two proxies that capture both aspects (greater opportunity to expedite the research process and uncertainty due to limited prior mapping information), though to differing degrees. In this setting, I use testing new uses of previously tested drugs as a rough proxy for the type of research investments made by firms that are more likely to succeed due to their enhanced opportunity to leverage the public mapping information. Indeed, in 2013, TCGA published the results of a large-scale effort to map nearly 400 endometrial tumors. The results revealed "that the worst endometrial tumors were so similar to the most lethal ovarian and breast cancers, raising the tantalizing possibility that the three deadly cancers might respond to the same drugs" (Kolata, 2013). To construct this proxy, I construct a dataset of FDA-approved drugs to treat cancer, using data from the Drugs@FDA, CenterWatch, National Cancer Institute, and Memorial Sloan Kettering Cancer Center websites.[41] Additionally, I identify drugs approved for specific genes by examining if it was approved with a companion diagnostic, a requirement for drugs aimed at targeting patients with specific genetic types.[42] For each approved drug, I collect information on the approval date(s) and the approved disease(s). For example, in 2014, the PARP inhibitor olaparib was approved to treat ovarian cancer patients with BRCA1 and BRCA2 gene mutations. The drug was approved alongside the companion diagnostic BRACAnalysis CDx, a test

[39] A clinical trial is classified as a private sector trial when it is funded by a private sector firm. When multiple institutions are involved in a clinical trial, I include the clinical trial in my analysis if any of the institutions is a private sector firm. Following Azoulay et al. (2019), all other trials (which are primarily conducted by the US or a foreign government, foundation, university, or hospital are classified as public sector.

[40] There is a notable increase in the share of gene-related trials before 2011, the year in which a large number of mutations were first identified in a given gene-cancer pair. This increase may have been driven by several sources, including early large-scale mapping efforts, firms' earlier small-scale mapping efforts, retrospective analyses of previous trial results, or licensing relationships with genomic firms. One interpretation is that the pre-2011 increase is driven by trials initiated in gene-cancer pairs with known mutation information before 2011. However, removing these trials does not change the overall trend. This paper aims to examine whether large-scale public cancer mapping efforts led to any additional effect on the level of private sector clinical trials, above and beyond these other factors.

[41] This results in a list of 187 cancer drugs originally approved to treat cancer between 1977 and 2015, inclusive.

[42] For more details, see https://www.fda.gov/medicaldevices/productsandmedicalprocedures/invitrodiagnostics/ucm301431.htm.

used to detect mutations in the BRCA genes of ovarian cancer patients. I code this as being an approval in the "BRCA1-Ovarian" and "BRCA2-Ovarian" pairs in 2014. Using the drug approval data, I classify a trial-gene-cancer as "testing new uses" if *all* of its interventions have either been (i) approved in the focal gene or (ii) previously tested in any gene-cancer pair. For example, if a trial enrolls ovarian cancer patients with BRCA2 gene mutations, it is classified as testing new uses if all of its interventions have either been approved to treat patients with BRCA2 gene mutations or previously tested in lung cancer patients with BRCA1 gene mutations.[43] The remaining trial-gene-cancer observations are classified as "testing novel drugs."[44]

Following the disclosure of information from public mapping studies, firms with previously tested drugs in one disease may have a greater likelihood of success in developing novel treatments in a related disease quickly. However, this advantage—primarily driven by regulation that reduces risk and expedites development and approval for new uses—doesn't necessarily imply greater access to useful private genetic mapping information. They may have gained knowledge of gene-cancer pair linkages through non-mapping approaches, such as retrospective analyses of trial results or studying family histories of individuals with cancer, or small-scale mapping efforts that are of limited relevance to other disease areas (Struewing et al., 1997).

**Measuring trial funders' prior mapping information.** To generate a more direct proxy for a trial funder's level of prior mapping information, I collect data firms' mapping-related publications.[45] A firm is classified as having high levels of prior mapping information if it has an above-median number of pre-2004 (i) mapping publications and (ii) phase II trials in the focal cancer (the second condition increases the likelihood that the firms' prior mapping information may be relevant for cancer-related research). Remaining firms are classified as having low levels of prior mapping information. I focus on measures prior to 2004 to mitigate potential confounding effects of disclosures from large-scale cancer mapping studies between 2004 and 2016. A trial-gene-cancer observation is classified as having a funder with high levels of prior mapping information if any of its funders meet the criteria for high levels of prior mapping information.

Before proceeding, I wish to make an additional remark concerning the differences between the two proxies: trials testing previously tested drugs and trials whose funders have higher levels of prior mapping information. These two measures may reflect different types of firms. For example, a trial in a gene-cancer pair may be funded by a large firm that has not yet obtained approval of a drug for the focal gene. This firm may still invest heavily invest in private sequencing efforts or gain private access to genomic information. A widely known example of this can be seen in the rivalry between the pharmaceutical companies, Regeneron and its partner, Sanofi, against Amgen. They competed to be the first to obtain approval of a drug targeting the protein produced by the

[43] Trials can test multiple interventions. This classification scheme considers the novelty of each of the trial's interventions.

[44] Since firms are not required to report phase I trials to public trial registries, this classification may underestimate the number of trials testing previously tested drugs and overestimate the number of trials testing novel drugs.

[45] The NLM maintains Medical Scientific Subject Headings (MeSH), a comprehensive dictionary of scientific terms, and assigns each PubMed publication to the relevant MeSH terms. Publications are considered to be sequencing related publications if they have sequencing-related MeSH terms (e.g., "Genetic Techniques", "Sequence Analysis").

PCSK9 gene. As part of their efforts, each firm invested significant funds in their own sequencing efforts or acquired access to large private sequencing databases (Pollack, 2014).

However, it is important to note that having access to prior mapping information alone does not ensure greater likelihood of success due to faster drug development. As noted above, this is due to the regulatory environment, which lowers the risk and expedites the drug development process for firms seeking to develop and gain approval for new uses of previously tested drugs in related diseases (see, e.g., Cheng et al., 2019 and Pushpakom et al., 2019). These firms, which might have acquired knowledge of gene-cancer relationships through small-scale, targeted (focused on a few, select genes) mapping efforts or non-mapping approaches, may include both large and small entities. The limited overlap between these two proxies is evident in the correlation coefficient, which is 0.21.

**Measuring trial outcomes.** In a subset of the empirical exercises that follow, I examine the relationship between public mapping information on common clinical trial outcomes. The clinical outcome data is comes from the public trial registry, ClinicalTrials.gov, and abstracts submitted to the ASCO Annual Meetings, the world's largest cancer conference.[46] I characterize trials as having a "positive outcome" if it demonstrates improvements in key clinical outcomes including treatment group gains in overall survival (time between randomization and death), progression-free survival (time between randomization and disease progression), or objective response rate (proportion of trial patients who experience a prespecified reduction in tumor size).

# 4 Effects on the quantity of private research investments

## 4.1 Empirical strategy

In an ideal experiment, I would study the impact of large-scale cancer mapping on private sector trials by randomly assigning cancer sites to undergo mapping efforts. I would then compare research investments between cancers assigned to mapping efforts and those that were not. In practice, concerns about cancer-level selection limit this type of cancer-level analysis: large-scale public mapping studies, which are usually cancer-site specific, may prioritize cancers that have higher expected research potential. For example, TCGA prioritized cancers with readily available tumors, suggesting a bias towards cancers with larger market sizes that could potentially lead to upward biased estimates.[47,48] Supporting this view, Online Appendix Figure B4 provides evidence of cancer-level selection by comparing various proxies for market potential (diagnoses, drugs approvals, and trials) between cancers that were first sequenced early (before 2011, the median sequencing year) and those that were first sequenced later (in/after 2011).

[46] ASCO, the primary professional society for medical oncologists, receives abstracts from major research groups describing their clinical trial findings at their annual conference. I rely on ASCO abstracts due to incomplete results reporting on ClinicalTrials.gov. See footnote i for more details.

[47] For more details, see https://cancergenome.nih.gov/cancersselected.

[48] A large literature documents the positive relationship between market size and pharmaceutical research. See e.g., Acemoglu and Linn (2004) and Dubois et al. (2015).

Instead, I focus on the timing of publicly disclosed information across genes within the same cancer, which allows me to address concerns about cancer-level selection and approximate the ideal experiment. I control for secular changes at the cancer level by including cancer-year fixed effects. To address potential gene-level selection bias (i.e., researchers may choose to sequence particular genes with higher ex ante research value), I take advantage of a unique feature of my setting: among the 168 public mapping studies used in this analysis, 89% employ comprehensive mapping techniques (referred to as "whole-genome" or "whole-exome" sequencing) that ensure comprehensive sequencing coverage of all genes.[49] In addition to increasing the likelihood of identifying rare mutations, this minimizes the potential for gene-level selection bias. The remaining 11% of mapping studies employ targeted sequencing, where select genes are targeted. However, given that these studies examine a large number of genes on average (around 3,000), and this analysis focuses on a smaller set of genes (462 "at-risk" cancer genes), the potential for gene-level bias among these studies is relatively low.[50] Moreover, I can demonstrate the robustness of the results by restricting the analysis to studies employing comprehensive mapping techniques, further ensuring that that gene-level selection bias does not influence the findings (see online Appendix Table E1).

With this empirical strategy, I utilize the disclosure of mutation information at the gene-cancer level as the primary source of variation. I then compare the level of clinical trials in gene-cancer pairs with mutation information to the level in gene-cancer pairs without mutation information. Importantly, this permits within-cancer, across-gene comparisons, avoiding the cancer-level biases found in cancer-level analysis and methods that contrast mapped gene-cancer pairs with mutation information against non-mapped pairs without such information (see online Appendix D for a detailed discussion).

### 4.2 Sample and descriptive statistics

Summary statistics at the gene-cancer level are shown in Table 1. Panel A shows that by 2016, at least one mutation was identified in 58% of all 30,223 gene-cancer pairs. Online Appendix Figure B5 shows the cumulative distribution of the years in which mutations were first identified among the 168 mapping studies. The increase in the cumulative distribution at 2011 the disclosure of results from several cancer mapping studies that examined hundreds of tumors (as illustrated in online Appendix Figure B2) increasing the likelihood of detecting "rare" mutations. Consistent with these patterns, Panel B shows that the median year in which mutation information was first disclosed was 2011. Table 1 also shows that only a minority of mutations provide strong signals and are likely cancer-causing: driver mutations were identified in only 9.8% of gene-cancer pairs.

The remaining panels provide characterize clinical trial investments. From 2004 through 2016, 12.7% of all gene-cancer pairs were targeted in at least one phase II clinical trial, and that private sector funding accounted for a lower proportion of phase II trial investments compared to the public sector (9.7% vs. 10.9%, respectively). Among private sector phase II clinical trials, investments

[49] Whole-genome sequencing reads all genes within both protein coding and non-coding regions, while whole-exome sequencing focuses on genes within protein-coding regions.

[50] Indeed, the COSMIC database classifies these specific targeted studies as "large-scale systematic screens".

TABLE 1
Summary statistics: Gene-cancer level data, 2004–2016

| | Mean | Median | Standard Deviation | Minimum | Maximum |
|---|---|---|---|---|---|
| *A. Mapping information* | | | | | |
| Share with mutation: all mutations (%) | 57.95 | 100.00 | 49.36 | 0 | 100 |
| Share with mutation: driver mutations (%) | 9.81 | 0.00 | 29.74 | 0 | 100 |
| *B. Mapping information timing* | | | | | |
| Year first mutation: all mutations | 2011.33 | 2011.00 | 1.33 | 2004 | 2015 |
| Year first mutation: driver mutations | 2012.09 | 2012.00 | 1.25 | 2006 | 2016 |
| *C. Clinical trials: phase II only* | | | | | |
| Any private or public sector trial (%) | 12.73 | 0.00 | 33.33 | 0 | 100 |
| Any private sector trial (%) | 9.68 | 0.00 | 29.58 | 0 | 100 |
| Any public sector trial (%) | 10.88 | 0.00 | 31.13 | 0 | 100 |
| *D. Clinical trials: private sector phase II only* | | | | | |
| Any trial testing new uses (%) | 8.33 | 0.00 | 27.64 | 0 | 100 |
| Any trial testing novel drugs (%) | 4.88 | 0.00 | 21.54 | 0 | 100 |
| Any trial funded by low private information firm (%) | 9.08 | 0.00 | 28.74 | 0 | 100 |
| Any trial funded by high private information firm (%) | 4.70 | 0.00 | 21.15 | 0 | 100 |

*Notes:* This table shows summary statistics at the gene-cancer level. There are 30,223 gene-cancer pairs in this sample. As an example, *Any public or private trial (%)* denotes the share of gene-cancer pairs that had at least one clinical trial funded by just a private sector institution, just a public sector institution, or both types of institutions. See Section 3 and online Appendix C for more detailed data and variable descriptions.

were more likely to focus on testing new uses of drugs rather than on developing novel drugs (8.3% vs. 4.9%, respectively). Furthermore, investments were more likely to be funded by a firm with low rather than with high levels of prior mapping information (9.1% vs. 4.7%, respectively).

## 4.3 Estimating equation and assumptions

### 4.3.1 Estimating equation

Given the skewed clinical trials data, I concentrate on binary outcomes indicating investment in a clinical trial for a specific gene-cancer-year.[51] My empirical analysis relies on the timing of publicly disclosed mapping information to estimate the impact of such information on the probability of subsequent research investment within a gene-cancer pair. I estimate:

$$Y_{gct} = \alpha + \beta Post \times DisclGeneCancer_{gct} + \delta_{gc} + \theta_{ct} + \epsilon_{gct}, \quad (1)$$

[51] I use linear probability models in my baseline specifications for four main reasons. First, since clinical trial investments are rare, I am primarily interested in the extensive margin effects large-scale public cancer mapping efforts. Second, OLS coefficients are straightforward to interpret and compare across different models. Third, linear regressions allow for the inclusion of multiple dimensions of fixed effects and utilize the entire dataset. Fourth, nonlinear methods may not be consistent when dealing with a large number of zeroes in the outcome variable (King and Zeng, 2001). Notably, certain nonlinear methods, such as fixed effects Poisson pseudo maximum likelihood (PPML) estimators, may be less likely to suffer from these concerns (see, e.g., Hausman et al. (1984); Correia et al. (2020)). I therefore report PPML estimates with robust standard errors in online Appendix Table B2. This count-based model serves as both a robustness check on my main findings and demonstrates that public cancer mapping information leads to an increase in subsequent research investments at the intensive margin.

where $Y_{gct}$ is an indicator for a clinical trial in gene $g$, cancer $c$ in year $t$. $Post \times DisclGeneCancer$ indicates whether gene-cancer $gc$ has been publicly known to be mutated as of that year and varies within gene-cancer pairs over time; a transition from 0 to 1 represents the fact that a mutation in a pair has been publicly disclosed. My coefficient of interest is $\beta$ which compares the likelihood of clinical trial investments in gene-cancer pairs for which public mapping has provided evidence to those which have not yet been mapped (or will never be).

I include gene-cancer fixed effects, $\delta_{gc}$, to control for time-invariant differences across gene-cancer pairs, such as a pair's inherent commercial potential. Finally, cancer-year fixed effects, $\theta_{ct}$, flexibly control for cancer-year specific shocks that are common across genes within a cancer, such as changes in technology or political pressure. I perform estimates using ordinary least squares (OLS) models and cluster standard errors two-way by (i) gene and (ii) cancer.[52,53]

### 4.4 Main results

Table 2 documents a positive relationship between public mapping information and subsequent clinical trials. Looking first to the effect on all (private and public sector) trials, column 1 implies a statistically significant 1.1 percentage point increase in the probability of a clinical trial, an increase on the order of 46% relative to the pre-mapping information sample mean. The results suggest that the public release of information from large-scale cancer mapping provided signals about market opportunity success rates. These signals were strong enough to outweigh the inhibitory effects of competition. These results also suggest limited withdrawal of newly lagging firms from the R&D race, which I investigate in more depth in a subsequent analysis.

The remaining columns show that this increase is primarily driven by private sector trials rather than public sector trials: the estimate in column 2 implies a 66% increase in the likelihood of a private sector trial. In contrast, column 3 shows that the effect on public sector trials is economically small and statistically insignificant. One explanation for these different effects is that the public sector being less affected by research uncertainty. Another explanation is that public sector institutions may have private knowledge of gene-cancer pairs from prior smaller-scale mapping studies, which I investigate further in a subsequent analysis (Sampat and Lichtenberg, 2011).[54]

[52] A recent literature in econometrics has demonstrated that staggered difference-in-differences models can lead to average DID estimates that are biased (De Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021; Sun and Abraham, 2021; Athey and Imbens, 2022; Gardner, 2022). I tackle this point in three ways: first, I provide Goodman-Bacon (2021) decompositions that show that the majority of my treatment effect is driven by treated vs. never treated observations. Additionally, I confirm that my main estimates are robust to excluding always-treated observations and to applying an alternative estimator proposed by Gardner (2022). This is discussed further at the end of Section 4.4.

[53] Clustering by both (i) gene and (ii) cancer allows for within-error correlation across genes and across cancers. The precision of my main results is robust to alternate ways of clustering (i.e., clustering only by gene, only by cancer, or by gene-cancer).

[54] An additional explanation is that the public sector may primarily respond by changing their investments in other areas of basic science (for example, by utilizing newly acquired sequencing techniques), rather than making significant changes to their investments in applied research, such as clinical trials.

TABLE 2
Effect of public cancer mapping information on phase II trials, 2004–2016

| | Dependent variable: Any phase II trial | | |
|---|---|---|---|
| | Any trial | Any private sector trial | Any public sector trial |
| | (1) | (2) | (3) |
| Post × DisclGeneCancer | 0.0113** | 0.00943** | 0.00366 |
| | (0.00386) | (0.00319) | (0.00346) |
| Mean of dep. var. | 0.024 | 0.014 | 0.014 |
| Change in likelihood of trial (%) | 46.27 | 65.73 | 27.00 |
| Gene-cancer FEs | Yes | Yes | Yes |
| Cancer × Year FEs | Yes | Yes | Yes |
| Observations | 392,899 | 392,899 | 392,899 |

*Notes:* This table reports DID estimates of the effect of public cancer mapping information on phase II trials. The level of observation is the gene-cancer-year. Estimates are from OLS models. The outcome variable switches from 0 to 1 if a clinical trial is reported in a gene-cancer-year. *Post* × *DisclGeneCancer* switches from 0 to 1 when a mutation in a gene-cancer pair is publicly disclosed by a mapping study. Robust standard errors, clustered at the gene and cancer level, are shown in parentheses. *Mean of dep. var.* is the mean of the outcome variable in a gene-cancer pair before the first disclosure of a mutation and is used to calculate the percentage change in the likelihood of a clinical trial. See Section 3 and online Appendix C for more detailed data and variable descriptions.
*$p$ <0.10, **$p$ <0.05, ***$p$ <0.01.

To explore the timing of the estimated effects, I estimate:

$$Y_{gct} = \alpha + \sum_{z} \beta_z \times 1(z) + \delta_{gc} + \theta_{ct} + \epsilon_{gct}, \tag{2}$$

where $\delta_{gc}$ and $\theta_{ct}$ represent gene-cancer fixed effects and cancer-year fixed effects, respectively, for gene $g$, cancer $c$, and year $t$. $z$ represents the "lag," or the years relative to a "zero" relative year, which marks the last year a gene-cancer was not known to be mutated (i.e., year 1 marks the first year that a mutation was disclosed).

Figure 1 presents $\beta_z$ from this regression and corresponds to a dynamic version of Table 2. The vertical lines represent 95% confidence intervals and the dashed red line indicates the first year in which a mutation in a gene-cancer was publicly disclosed. Panel A illustrates that gene-cancers that received mutation-related information initially exhibit similar trends in clinical trial research compared to those without such information. However, the probability of any clinical trial rises differentially for gene-cancers that receive mutation-related information and remains elevated afterwards. Supporting the view that private sector research investments are more responsive than public sector research investments after the disclosure of public cancer mapping results, Panel B shows that private sector trials responded relatively quickly, as indicated by the increase at $t = 2$. The timing is consistent with the view that the firms that rapidly responded were those testing products that were "on the shelf" (i.e., approved or previously tested in related diseases), a point that I investigate further in a subsequent analysis.

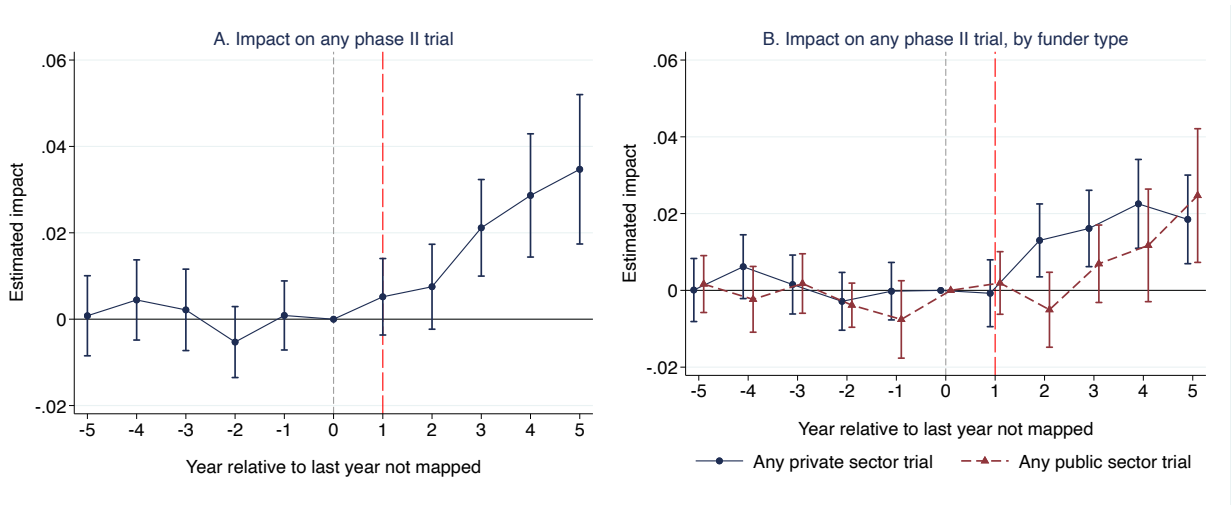


FIGURE 1

Effect of public cancer mapping information on phase II clinical trials, 2004–2016

*Notes:* This figure plots the response of phase II trials following the public release of cancer mapping information. Panel A shows the effect of public cancer mapping information on any phase II trial. Panel B shows the effect of public cancer mapping information on any private sector or public sector phase II trial. Each dot corresponds to coefficients based on estimates of equation (2) and corresponds to a dynamic version of the specifications in Table 2. On the $x$-axis are years $z$ relative to a "zero" relative year that marks the last year the gene-cancer was not known to be mutated based on a cancer mapping study. The dashed red line indicates the first year that a mutation in a gene-cancer pair is publicly disclosed by such a study. Shown are 95% confidence intervals (corresponding to robust standard errors, clustered at the gene and cancer level). This specification is based on gene-cancer-year level observations, the coefficients are estimates from OLS models. Controls include gene-cancer and cancer-year fixed effects.

The lack of pre-trends suggests that firms are not strategically withholding clinical trial investments in anticipation of the public release of cancer mapping information. Instead, the evidence indicates that public disclosure of mapping information is an exogenous information shock that prompts firms to increase their clinical trial investments. Together, these estimates suggest that information from mapping efforts has a positive and significant impact on the subsequent level of clinical trials, particularly in the private sector. Motivated by these findings, I focus on the impact of public cancer mapping on private sector clinical trials in the main analyses that follow.

Recent econometric advances have shown DID models with staggered treatment may lead to biased estimates. Estimated DID estimates are a weighted average of all possible treatment-control pairs, leading to potential issues when a treated unit serves as a control for another unit and when treatment effects vary across treated groups over time. Since the literature has not yet reached consensus on how to address these issues, in Online Appendix Section E.2, I take three approaches (Roth et al., 2023). First, to examine the sources of variation in my DID estimates, I utilize the Goodman-Bacon (2021) decomposition method. This analysis reveals that the main DID estimate primarily relies on comparisons between observations that were never treated (i.e., gene-cancer pairs that were never mapped) and observations that were treated during the sample period (i.e., gene-cancer pairs that were mapped), reducing concerns that time-varying effects are driving my results. Second, I estimate equation (1) after excluding always-treated cancer pairs and find results that

are remarkably similar to the baseline results. Third, I use an alternative estimator proposed by Gardner (2022). This estimator is advantageous due to its simplicity, as it mirrors the identification process in a two-group, two-period setting. It recovers the average difference in outcomes between treated and untreated observations, after accounting for group and period effects. The results from this alternative estimator are similar to the main results, though slightly larger in magnitude.

#### 4.4.1 Heterogeneous effects by strength of public mapping information

The previous analysis suggests that the public release of information from large-scale cancer mapping provided strong signals about market opportunity success rates. The analytical framework predicts that the effect of public mapping information varies by its strength. In this section, I empirically explore the following question: are private firms more likely to respond to information that provides a stronger signal—i.e., mutations that hold more clinically relevance in the progression and growth of cancer?

Table 3 shows that the relationship between public mapping information and trial quantity varies based on the mutation information's strength. Using equation (1), I find that the disclosure of a driver mutation leads to a 154% increase in the probability of a trial (column 1), while news of a passenger mutation leads to a 38% increase (column 2). The difference in point estimates is statistically significant, confirming that firms are more responsive to strong mapping signals.[55]

In addition to clinical relevance, the strength of a mapping signal may also be characterized by its specificity. For example, the impact of research in a focal disease differs depending on whether the information pertains to the same disease or a closely related one (Sampat, 2015). Online Appendix Table B3 confirms that direct public mapping information within the same disease has a significantly larger effect compared to information disclosed in a closely related but different disease.

### 4.5 Understanding mechanisms: effect on the direction of research

In this section, I leverage detailed trial and drug data to examine the mechanisms underlying the increase in private sector research. In addition to examining the types of firms that benefit from such information, I explore how shifts in information shape firms' decision making as they move through the multistage drug development process.

#### 4.5.1 Investigating differences across trials testing new uses and new drugs

The analytical framework presented in Section 2 suggests that public mapping information may increase an incumbent firm's relative likelihood of success since it has a greater opportunity to leverage the information to quickly develop a commercially viable treatment (Prediction 2). This

[55] For this analysis, I estimate a system of seemingly unrelated regressions (SUR). While OLS regressions yield very similar results, SUR permit a direct equality test between coefficients (for driver and passenger mutations). SUR models cannot easily accommodate two-way clustered standard errors. I therefore cluster at the cancer level (Budish et al., 2016) which constitutes larger, more aggregate clusters (relative to clustering at the gene or gene-cancer level) (Cameron and Miller, 2015) and also better reflects the experimental design (Abadie et al., 2023) since the timing of the treatment is likely to be correlated within cancer clusters. In the remainder of the paper, when comparing coefficients across regressions, I always use this method.

TABLE 3
Effect on private sector clinical trials:
Heterogeneity by strength of public mapping signal, 2004–2016

| | Dependent variable: Any private sector phase II trial | |
|---|---|---|
| | Driver mutation (strong signal) (1) | Passenger mutation (weak signal) (2) |
| Post × DisclGeneCancer | 0.0450*** | 0.00514*** |
| | (0.00372) | (0.00123) |
| Mean of dep. var. | 0.029 | 0.014 |
| Change in likelihood of trial (%) | 153.56 | 37.97 |
| Gene-cancer FEs | Yes | Yes |
| Cancer × Year FEs | Yes | Yes |
| Observations | 392,899 | 392,899 |
| Diff. Wald test $p$-value | 0.00 | |

*Notes:* This table reports DID estimates on private sector trials of public mapping information with high and low signal strength. The level of observation is the gene-cancer-year. The outcome variable switches from 0 to 1 if a private sector phase II trial is reported in a gene-cancer-year. *Post* × *DisclGeneCancer* switches from 0 to 1 when a mutation in a gene-cancer pair is publicly disclosed by a mapping study. Column 1 shows the effect of the first driver mutation in a gene-cancer, where driver mutations are identified in one of two ways: 1) the cancer mapping authors list the mutation as a likely driver mutation, or 2) the gene-cancer mutation is detected an unusually high number of times (in $\geq$ 10 patients). All remaining mutations are classified as passenger mutations. Column 2 shows the effect of the first passenger mutation in a gene-cancer. *Mean of dep. var.* is the mean of the outcome variable in a gene-cancer pair before the first disclosure of a mutation and is used to calculate the percentage change in the likelihood of a clinical trial. The estimates in this table are from seemingly unrelated models, which permits a comparison of the coefficient on *Post* × *DisclGeneCancer* across models. Standard errors are clustered at the cancer level (see footnote 55). The $p$-value is from a Wald test that compares the differences in the coefficients on *Post* × *DisclGeneCancer*. See Section 3 and online Appendix C for more detailed data and variable descriptions.
*$p$ <0.10, **$p$ <0.05, ***$p$ <0.01.

advantage is particularly valuable in a racing environment where there are substantial first mover advantages.[56]

I find evidence that phase II trials testing new uses of previously tested drugs show a stronger positive response to public mapping information compared to phase II trials testing novel drugs. Table 4 shows that when estimating equation (1) separately for these two types of trials, the point estimate for trials testing new uses of existing drugs (column 1) is significantly more responsive to public mapping information than trials testing novel drugs (column 2).[57] Appendix Figure B6 shows the quick response among trials testing previously tested drugs, which supports the view that firms conducting such trials can have a greater opportunity to quickly leverage public mapping information, as they can potentially bypass earlier preclinical studies or phase I trials.

[56]A manufacturer of an approved drug might forego trials for additional approval, opting instead to increase promote off-label use. The estimates reflect the effect of public mapping information, beyond these firm responses.

[57]Approximately 95% of clinical trials have a listed drug intervention (see online Appendix C). Rerunning the main analysis shown in Table 2 using the subset of trials with a listed drug intervention leads to similar results. See online Appendix Table B4 and Figure B7.

TABLE 4
Effect on private sector clinical trials testing
by drug type and firm type, 2004–2016

| | Dependent variable: Any private sector phase II trial | | | |
|---|---|---|---|---|
| | Testing new drug uses | Testing novel drugs | Low private mapping information | High private mapping information |
| | (1) | (2) | (3) | (4) |
| Post × DisclGeneCancer | 0.00631*** | 0.00329*** | 0.00634*** | 0.00247*** |
| | (0.00116) | (0.000828) | (0.00112) | (0.000663) |
| Mean of dep. var. | 0.010 | 0.005 | 0.012 | 0.003 |
| Change in likelihood of trial (%) | 63.60 | 63.50 | 51.31 | 88.70 |
| Gene-cancer FEs | Yes | Yes | Yes | Yes |
| Cancer × Year FEs | Yes | Yes | Yes | Yes |
| Observations | 392,899 | 392,899 | 392,899 | 392,899 |
| Diff. Wald test $p$-value | 0.06 | | 0.00 | |

*Notes:* This table reports DID estimates of the effect of public cancer mapping information. The level of observation is the gene-cancer-year. Column 1 estimates the effect on trials testing new uses of drugs approved in the focal gene or previously tested in any gene-cancer pair; column 2, the effect on trials of novel drugs. Column 3 estimates the effect on trials by firms with low levels of private sequencing information; and column 4, the effect on trials by firms with high levels of private sequencing information. *Post* × *DisclGeneCancer* switches from 0 to 1 when a mutation in a gene-cancer pair is publicly disclosed by a mapping study. *Mean of dep. var.* is the mean of the outcome variable in a gene-cancer pair before the first disclosure of a mutation and is used to calculate the percentage change in the likelihood of a clinical trial. The estimates in this table are from seemingly unrelated models, which permits a comparison of the coefficient on *Post* × *DisclGeneCancer* across models. Standard errors are clustered at the cancer level (see footnote 55). The $p$-value is from a Wald test that compares the differences in the coefficients on *Post* × *DisclGeneCancer*. See Section 3 and online Appendix C for more detailed data and variable descriptions. $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

The empirical results, when combined with the analytical framework in Section 2, highlight the significant heterogeneity among firms in their opportunity to effectively utilize public mapping information. Specifically, the results indicate that public mapping information plays a crucial role in reshaping the research landscape, favoring incumbent firms with previously tested drugs. Supporting Prediction 3, the smaller yet positive impact on research investments in trials testing novel drugs suggests either substantial uncertainty in the R&D process, creating opportunities for laggards to still succeed, or that the technological gap between firms with previously tested drugs and those with new drugs, while still significant, may not excessively disadvantage the latter.[58]

Importantly caveats subsist: it is possible that large-scale public mapping spurs additional trials testing novel drugs, but that the effect takes more time to observe relative to investments in trials testing new uses. Motivated by these concerns, I examine the effect of public cancer mapping information on early-stage (phase I) trials in new uses and novel drugs. The findings, presented

[58] Another possibility is that there are smaller, but positive benefits allocated to non-first-movers. See footnote 18.

in online Appendix Table B5, reveal similar patterns, further substantiating the view that public mapping information disproportionately benefits firms with previously tested drugs.[59]

### 4.5.2 Investigating differences in prior mapping information

The analytical framework in Section 2 emphasizes the role of firms' existing information as an important factor in determining the effects of public mapping information (Prediction 4). If conditions outlined in the analytical framework regarding the strength of the public mapping effect and firms' priors about the likelihood of success hold, then the impact of public mapping information will be greater among trials funded by firms with lower levels of prior mapping information compared to firms with higher levels. Consistent with this view, Columns 3 and 4 of Table 4 confirms that clinical trials funded by firms with lower levels of prior mapping information are significantly more responsive to public mapping information.

Combined with the previous results, these findings characterize the nuanced role of publicly available basic scientific information in shaping private research incentives: it provides strong signals about market opportunity success rates and confers advantages to incumbent firms with previously tested drugs in related diseases. Simultaneously, it offers advantages to firms that previously lacked access to basic scientific information.[60] Taken together, this evidence suggests that public mapping information can substitute for prior mapping information, and large-scale efforts may provide an important subsidy for firms lacking the resources to generate their own mapping information.

### 4.5.3 Investigating trial outcomes and continuation decisions

The previous results indicate that publicly available, large-scale cancer mapping efforts increase the likelihood that private firms will initiate clinical trials. To investigate if these research investments are more likely to successfully result in drug approvals, I now explore firms' investments as they navigate through the multistage drug development process.

After initiating a clinical trial, firms face a critical decision: whether to continue investing and advance their drug to the next phase or terminate the drug development process. Access to reliable and organized genetic information can support data-driven decision-making, reducing biases (e.g., overemphasis on progression-seeking behaviors) and improving outcomes in the drug development process (Guedj and Scharfstein, 2004; Krishnamurthy et al., 2022; Minikel et al., 2023).[61] Such information may prompt firms to prioritize promising drugs with "positive" clinical evidence, leading to improved resource allocation and a higher probability of obtaining regulatory approval in the long run.

[59] A caveat to these results is that firms are not required to report phase I trials (as discussed in Section 3.3.2). However, this only raises concerns if there are differential rates of reporting across different types of phase I trials.

[60] Another interpretation of these findings is that large-scale cancer mapping information disproportionately benefits smaller firms facing relatively high costs of research (Nagaraj, 2022).

[61] For example, public mapping information may lower the likelihood that firm managers compute payoffs incorrectly (e.g., due to confirmation bias, overconfidence, or sunk-cost fallacy) (Tversky and Kahneman, 1974; Donelan et al., 2015), fail to consider alternatives (Sharpe and Keelin, 1998), or follow the decisions of the past or their peers (Bujar et al., 2017).

Due to data constraints, I am unable to directly quantify how public mapping information shifts firms' decisions as they weigh the costs of each drug development phase (from phase I to approval) against existing clinical information. Additionally, long development timelines also prevent a clear analysis on how public cancer mapping information shifts later development stages, such as Phase III trials and regulatory approval.[62] As a result, for a subset of trials linked to clinical outcomes, I investigate how access to a reliable, organized view of the cancer information landscape impacts the likelihood that firm investments result in promising clinical outcomes and encourages continued investment in drugs with higher chances of FDA approval. The analytical framework in Section 2 suggests that public mapping information may reveal promising research opportunities, thus increasing the likelihood that clinical investments will generate promising clinical information (that can ultimately lead to drug approval).

To examine the relationship between public mapping information and phase II trial outcomes, I compare outcomes among trials initiated in gene-cancer pairs where mutation information has or has not become available (hereafter, "trials with mapping information" vs. "trials without mapping information"). I then consider whether private firms are more likely to advance trials with positive clinical outcomes when genetic information is available.

For this purpose, I estimate OLS cross-sectional regressions on trial-gene-cancer level data (as opposed to the gene-cancer-year panel used in my previous analyses) which allows me to examine the relationship between public mapping information and *any given* trial's likelihood of generating a promising clinical outcome.[63] I focus on the set of phase II and III trials that have a completed or terminated status[64] and have available data on clinical trial outcomes.[65,66] Within this sample, I estimate the following OLS specification:

$$PositiveOutcome_{igc} = \beta Post \times DisclGeneCancer_{gc} + \boldsymbol{X}_i + \epsilon_{igc}, \tag{3}$$

where $PositiveOutcome_{igc}$ is an indicator for a positive trial outcome for trial $i$ in gene $g$ and cancer $c$. $Post \times DisclGeneCancer_{gc}$ is an indicator for whether information about a clinically

[62] Wong et al. (2019) estimates that the median length of a phase III cancer trial is 5.7 years.

[63] Using the trial-gene-cancer dataset offers an additional advantage as it avoids potential compositional effects that may arise with a gene-cancer-year panel. To illustrate, suppose that the gene-cancer-year panel is used to analyze the relationship between public mapping information and the likelihood of a trial demonstrating a statistically significant improvement in overall survival (i.e., has a positive outcome). If the results show that gene-cancers with public mapping information have a higher likelihood of a trial with a positive outcome, it could be capturing two effects. One effect might be that public mapping increases the likelihood of a positive outcome while keeping the total number of trials constant. Another effect is that public mapping increases the total number of trials while keeping the likelihood of a trial having a positive outcome constant. While these estimates are correlations, using the trial-gene-cancer dataset allows me to examine the relationship between public mapping information and trial success while holding the total number of trials constant.

[64] This refers to the trial's status as of July 14, 2017. This excludes a large share of private sector trials that are "in progress."

[65] Online Appendix Table B6 presents summary statistics on data at the trial-gene-cancer level.

[66] Online Appendix Figure B8 shows the share of phase II trials that successfully advance to phase III is falling over time, a finding consistent with widespread reports about declining productivity in the pharmaceutical industry (Arora et al., 2021).

relevant mutation is available for gene $g$ and cancer $c$ by the end of clinical trial $i$.[67,68] $\boldsymbol{X}_i$ is a vector of trial characteristics including whether the trial is funded by a firm with high levels of prior mapping information, disease (gene and cancer) fixed effects, and year fixed effects.[69] Standard errors are clustered at the gene and cancer level. The results in columns 1 and 2 of Table 5 show that phase II trials with public mapping information are more likely to have a positive clinical outcome, relative to phase II trials without public mapping information—further supporting the idea that public cancer mapping information reveals promising market opportunities.

To analyze firms' continuation-or-termination decisions, I estimate Cox proportional-hazard model regressions to analyze the relationship between public mapping information, phase II outcomes, and phase II continuations:

$$\begin{aligned} h_{ic}(t) = h_{cf0}(t) \times exp[&\beta Post \times DisclGeneCancer_{gc} \\ &+ \gamma Post \times DisclGeneCancer_{gc} \times PositiveOutcome_i \\ &+ \delta PositiveOutcome_i + \boldsymbol{X}_i]. \end{aligned} \tag{4}$$

where $h_{cf0}(t)$ is the baseline hazard rate of trial advancement, stratified by cancer. The coefficient $\gamma$ tells us how the impact of public cancer mapping information on phase II trial continuation rates changes when the trial has a positive outcome. I control for trial characteristics ($\boldsymbol{X}_i$) and cluster standard errors at the gene and cancer level. Columns 3 and 4 of Table 5 show that phase II trials with positive outcome are more likely to advance to phase III. These effects are greater when related public mapping information is available: among trials with positive outcomes, trials with public mapping information are significantly more likely to advance to phase III relative to trials without public mapping information (196% vs. 150%). Online Appendix Table B7 demonstrates that these effects are especially salient in trials funded by firms with large drug portfolios, as they have the flexibility to allocate resources accordingly based on trial results.

Due to the small sample size and the correlational nature of the evidence in Table 5, it is important to interpret these findings with caution. However, in line with the framework (Prediction 1), these estimates are consistent with the idea that public mapping information enhances firms' productivity by increasing the likelihood that a research investment will successfully advance to approval. This is achieved by improving the probability that firms will attain a positive intermediate clinical outcomes, and that they will advance research investments with favorable outcomes and terminate of research investments without such outcomes.

[67] To isolate the impact of public mapping information that is most likely to impact the success of a firm's research decisions, I focus on the impact of mutations with a strong signal (i.e., driver mutations).

[68] This is to take into account that genetic mapping information may be useful for helping firms conduct the clinical trial and influence firm continuation-or-termination decisions even while the clinical trials is on-going.

[69] Due to the small sample size, gene-cancer fixed effects, cancer-year fixed effects, and controls for additional trial characteristics are not included in the analysis.

TABLE 5
Public cancer mapping information and phase II clinical trial outcomes and advancement rates, 2004–2016

| | Dependent variable: Positive trial outcome (mean = 0.33) | | Dependent variable: Advancing to phase III | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Post × DisclGeneCancer | 0.182** | 0.153* | -0.116 | -0.671 |
| | (0.0784) | (0.0788) | (0.291) | (0.454) |
| Post × DisclGeneCancer × PositiveOutcome | | | | 0.839* |
| | | | | (0.479) |
| PositiveOutcome | | | | 1.012*** |
| | | | | (0.291) |
| Change in likelihood of outcome (%) | | | | |
| Mapping info | 55.47 | 46.53 | -10.95 | |
| Positive trial outcome | | | | 175.17 |
| Mapping info, positive trial outcome | | | | 225.34 |
| Mapping info, no positive trial outcome | | | | -48.89 |
| Cancer FEs | Yes | Yes | | |
| Gene FEs | Yes | Yes | | |
| Linear year trend | Yes | Yes | Yes | Yes |
| Trial characteristics | No | Yes | No | Yes |
| No. Trial-gene-cancers | 1,754 | 1,754 | 1,785 | 1,785 |
| No. Trials | 165 | 165 | 177 | 177 |
| No. Genes | 80 | 80 | 80 | 80 |
| No. Cancers | 57 | 57 | 88 | 88 |

*Notes:* This table shows the relationship between public cancer mapping information and phase II outcomes and phase II-to-phase III advancement rates. The level of observation is the trial-gene-cancer. The sample includes all phase II trial-gene-cancer observations associated with phase II clinical trials that began between 2004 and 2016, made clinical outcomes data available, and were completed or terminated as of July 14, 2017. Columns 1 and 2 provide OLS estimates from regressions that examine the relationship between public cancer mapping information and phase II clinical outcomes, as indicated by whether the trial has a "positive outcome." Columns 3 and 4 are from Cox proportional hazard models (stratified by cancer) and examine the relationship between public cancer mapping information and phase II-to-phase III transition rates. Singleton observations are dropped, which accounts for the smaller number of observations in columns 1 and 2 relative to columns 3 and 4. *Post* × *DisclGeneCancer* is an indicator for the disclosure of a driver (clinically relevant) mutation in a gene-cancer by the end of the clinical trial. *PositiveOutcome* is an indicator for whether the trial satisfies any of the following: treatment group has objective response rate above the 75th percentile in the cancer-specific distribution, demonstrates a demonstrates a statistically significant ($p < 0.05$) improvement in overall survival relative to the control group or a historical control, or demonstrates a statistically significant ($p < 0.05$) improvement in progression-free survival relative to the control group or a historical control. *Change in likelihood of outcome (%)* refers to the change in the likelihood of a positive trial outcome (columns 1 and 2) or of advancing to phase III (columns 3 and 4). *Trial characteristics* refers to controls for whether the trial is a private-sector trial and whether the trial is funded by a firm with high levels of prior mapping information. Robust standard errors, clustered at the gene and cancer level, are shown in parentheses. See Sections 3 and 5 and online Appendix C for more detailed data and variable descriptions.
*$p$ <0.10, **$p$ <0.05, ***$p$ <0.01.

### 4.6 Additional results and robustness checks

Online Appendix E presents additional results pertaining to the effects of public cancer mapping information. In the interest of space, I provide a brief discussion of these results. First, I confirm that my main results are robust to accounting for changes in intellectual property regulation that could impact firms' genomics-related research efforts. In the 1990s, the firm Myriad obtained a patent on the sequenced BRCA1 and BRCA2 genes and their associated mutations (Gold and Carbone, 2010). However, in 2013, the Supreme Court ruled that genes and their mutations could no longer be patented. The results are robust to analyzing the period between 2004 to 2012, thus minimizing the impact of these regulatory changes (see online Appendix Table E4).

Second, I consider the role of private-public collaborations. Private firms collaborating with public sector institutions that have extensive research experience (e.g., in cancer) may have a greater absorptive capacity (i.e., ability to assimilate external information) (Cohen and Levinthal, 1990; Cockburn and Henderson, 1998). Supporting this view, online Appendix Table E5 shows that public mapping disproportionately increases investments in such "private-public" trials, particularly when the collaborating public institution has extensive research experience.

Third, I confirm that research activity does not differ significantly across diseases with different market potential, as measured by market size or the number of diagnoses (see online Appendix Table E6). Fourth, one challenge in interpreting the impact of public mapping on private sector phase II trial investments is that firms might alter the design quality of subsequent clinical trials, "cutting corners" in order to expedite the drug development process. Online Appendix Table E7 shows that public cancer mapping has little effect on the design quality of these trials, suggesting that public mapping information does not affect the usefulness of these trials for developing high quality therapies.

## 5 Valuing the potential impact of public scientific maps

In this section, I conduct a "back-of-the-envelope" cost-effectiveness analysis to assess the dollar value of large-scale public cancer mapping information. Due to data limitations, a comprehensive evaluation of both the benefits (e.g., improved patient outcomes) and costs (e.g., costs of each mapping studies and clinical trials) is not feasible. That said, I provide a simple comparison of the estimated cost of public cancer mapping efforts and the implied dollar value of mapping-induced R&D investments. Table 6 summarizes the results of this exercise.

To calculate the cost associated with the large-scale cancer mapping studies, I begin with the estimated cost of TCGA, \$427 million in 2016 dollars.[70] Dividing this amount by 33, the total number of cancers TCGA studied, suggests that the average cost per each TCGA mapping study would be \$12.9 million. This indicates that the combined costs for the 168 cancer mapping studies would be approximately \$2.2 billion ($\approx$ \$12.9 million $\times$ 168 studies).

[70] This figure is derived from the fact that in 2006, TCGA received \$100 million from the NIH, and in 2009, an additional \$100 million from the NIH and \$175 million in American Recovery and Reinvestment Act funding.

TABLE 6
Implied drug valuation of large-scale public cancer mapping efforts

| | |
|---|---|
| *A. Implied Drug Approvals* | |
| Total drug approvals | 4.202 |
| Novel drug approvals | 1.201 |
| New use drug approvals | 3.000 |
| *B. Patent-based market value* | |
| Total market valuation ($ billions) | 2.692 |
| Benefit-cost ratio | 1.238 |
| Rate of return (%) | 23.846 |
| *C. Sales-based market value* | |
| Total market valuation ($ billions) | 4.056 |
| Benefit-cost ratio | 1.865 |
| Rate of return (%) | 86.535 |

*Notes:* This table presents the findings of the implied cost and valuation of large-scale public cancer mapping efforts. Panel A provides calculations of the implied number of drug approvals, which are based on coefficient estimates obtained from Table 2. Panel B presents estimates of the implied market value of drugs based on the market value of associated patents. To calculate this, estimates from Kogan et al. (2017) are used, along with information on the total number of patents per approved drug obtained from the FDA Orange Book. Panel C presents calculations using sales-based measures to estimate the market value of drugs, with drug sales estimates from DiMasi et al. (2004). Market valuations are in 2016 dollars. The benefit-cost ratio and the rate of return are calculated using the implied market valuation estimates and the estimated cost of public cancer mapping $2.2 billion (in 2016 dollars). For additional details, see online Appendix F.

Following Azoulay et al. (2019), I use two methods to quantify the value of large-scale public cancer mapping efforts in dollars: (1) by assessing the market value of patents linked to the mapping-induced drugs, and (2) by evaluating the net present value of lifetime sales associated with these drugs. Because of the long duration of drug development, I am likely underpowered to directly derive meaningful direct estimates of the impact of public cancer mapping information on the number of drug approvals using the current data. To address this limitation, I combine my phase II estimates (as phase II is the latest phase in which a substantial increase is expected) with phase II-to-approval estimates from the literature to derive the implied number of drug approvals. As the dollar value of public mapping information depends on how it shapes *all* subsequent research investments, I focus on the impact of all phase II trials, private and public sector.[71]

I begin with the fact that public cancer mapping information increases investments in phase II trials overall by 46% (Table 2). To determine the additional number of clinical trials due to public mapping, I then ask: If gene-cancer pairs that received mutation-related information had received the same level of investment as pairs that did not, how many fewer trials would have taken place? I take a conservative approach and focus on drug development that would have

[71] As expected, an analysis that focuses only on private sector R&D investments results in a similar (though slightly lower) benefit.

occurred from 2011 (the year in which the effect of public cancer mapping was greatest, as shown in online Appendix Figure B5) to 2016. Panel A of Table 6 shows that there would have been roughly 4.2 fewer drug approvals overall; 3.0 for new uses of approved drugs, and 1.2 for novel drugs (online Appendix F provides a detailed discussion).

Table 6, Panel B presents implied drug valuation estimates using patent-based market values. I set the number of patents per drug approval at 9.74 (corresponding to the average number of patents per drug in the FDA's Orange Book).[72] I use estimates for the average patent dollar value from Kogan et al. (2017) to calculate the implied market value of a drug's associated patents.[73] The average Orange Book patent is worth approximately $65.8 million (2016 dollars). Thus, I calculate a $2.2 billion cancer mapping effort yields $2.7 billion ($\approx$ $65.8 million/patent $\times$ 9.7 patents/drug $\times$ 4.2 drugs) in market value. This implies a benefit-cost ratio of 1.2 and a rate of return of 23.8% based on phase II investments made between 2004 and 2016.

There are limitations to this approach. The Kogan et al. (2017) estimates are based on patents awarded to publicly traded firms (excluding patents owned by privately owned firms and non-profits) suggesting that these figures may be underestimated.[74] Additionally, patent value may vary based on factors such as whether it covers the active ingredient in a novel drug or method-of-use for treating a new indication of an approved drug.[75] Further, patent value may not solely reflect market value and may reflect other strategic concerns–see, e.g., Noel and Schankerman (2013)

Panel C presents an alternative approach to measuring the implied drug valuation of cancer mapping efforts using drug sales. I set the net present discounted value (PDV) of lifetime sales for novel drugs using the average PDV of drugs approved from 1990 to 1994 based on the finding of DiMasi et al. (2004)—$3.38 million in 2016 dollars. These estimates are based on sales data from novel drugs, excluding new indications of existing drugs (because the latter are not commonly reported in existing sales datasets). Multiplying the PDV estimate ($3.4 million) by the implied number of novel drug approvals (1.2) results in $4.1 billion, implying a rate of return of 86.5%.

Taken together, these calculations imply that cancer mapping studies likely yield a positive net return in the form of additional clinical trial investments. A caveat to these calculations is they rely on strong assumptions: in addition to assumptions about patent values across different firm types, these calculations assume that historical phase II-to-approval success rates are similar as those of marginal projects induced by the mapping efforts. If the mapping efforts induced higher quality projects, this would imply that the implied drug valuations could be higher. However, given

[72] The distribution of patents associated with each drug is highly skewed.

[73] In particular, Kogan et al. (2017) estimate the stock market response around the patent grant announcements. I am grateful to the authors for sharing their data.

[74] Bessen (2008) finds that, on average, non-profit organizations have higher patent values compared with publicly traded firms, although the distribution of patent values is heavily skewed.

[75] Specifically, primary patents that cover a drug's active ingredient are generally deemed the most valuable, offering the strongest protection against generic competitors. In contrast, secondary patents, which often encompass other aspects of a drug, may have comparatively lower value. Although conducting a comprehensive analysis of patent value is beyond the scope of this paper, a review of the patent sample supports this perspective. By categorizing Orange Book patents as primary patents when they cover a drug substance, and categorizing all other patents as secondary patents, the findings indicate that the average market value of primary patents is higher than that of secondary patents (i.e., $77.1 million vs. $64.3 million in 2016 dollars).

the limitations of existing data to directly observe changes in drug approvals, I take a conservative approach (consistent with my choice to focus on the effect of public mapping over six years, from 2011 to 2016) and assume that these estimated drug valuations likely represent an upper bound. Reassuringly, the estimated rates of return align closely with the rates of return to R&D reported in the existing literature (Hall et al., 2010). For policymakers, this suggests that the marginal public dollar may be more effectively spent on publicly available scientific maps and similar investments in scientific data, rather than simply directly funding drug development efforts.

# 6 Discussion and conclusion

Existing theory highlights that uncertainty affects innovation investments, leading private firms to underinvest. To tackle this market failure, the public sector has employed two approaches: providing public funding and investing in basic scientific information. While the effects of public funding have been extensively studied, there is relatively less research on how public investments in scientific information influence private firms' research decisions, especially in competitive environments with heterogeneous firms.

In this paper, I investigate the impact of publicly available scientific maps on private firms' R&D decisions. My analytical framework clarifies that that scientific maps—which aim to provide detailed information about the universe of potential research opportunities—can assist private sector firms navigate towards promising opportunities and shape their decisions to initiate or terminate research investments. However, the impact of these maps within competitive settings is heterogeneous and depends on the strength of the public mapping signals and the degree to which public mapping interacts with key firm characteristics.

Empirically, I examine the impact of public scientific maps in the context of the pharmaceutical industry, where large-scale public cancer genome mapping initiatives systematically catalog the genetic mutations associated with different cancers. Using a newly constructed dataset on large-scale cancer mapping studies and clinical trials, I find that public scientific maps significantly increase research investments by private sector firms. These findings support the idea that public maps provide valuable information signals, leading to increased research investments across the drug development process, even in competitive environments. One back-of-the-envelope calculation suggests that $2.2 billion in public cancer mapping investments yields $4.1 billion in market value.

I next exploit the rich heterogeneity in trial and firm characteristics to illuminate the mechanisms underlying the increase in private sector research investments. The large-scale public release of basic scientific information operates in nuanced ways. On one hand, it provides advantages to incumbent firms, enhancing their competitive edge in the race to introduce a new treatment to the market. However, laggards still invest in research, although their increase in investments is comparatively lower compared to the leading firms. Simultaneously, it levels the playing field for firms previously unable to generate to valuable mapping information. These findings indicate that the extensive release of information, even without specifically targeting certain innovators, can have

a profound impact on the direction of future innovation by influencing the types of firms that play a leading role in driving technological change.

My results have at least three important insights that future research could extend. First, my results suggest that large-scale public mapping efforts increased the relative level of private sector research investments in gene-cancer pairs with mutation information relative to gene-cancer pairs without. To quantify the true public and private costs of large-scale public investments, it is necessary to understand whether public mapping information shifted socially valuable innovative effort away or towards gene-cancer pairs without mutation information (Williams, 2013). If publicly available mapping information led to spillovers that also increased socially valuable research in different diseases, then the public and private benefits could be substantial.

Second, my findings primarily speak to the relatively short-term implications of public mapping initiatives, which is important given the substantial economic and mortality burdens associated with cancer. Nevertheless, it is crucial to also consider the lasting effects of these initiatives. The implications of such initiatives depend on various factors in the long run, including the subsequent rate of innovation (e.g., number of trials testing new uses and drug approvals), reallocation of firm resources (e.g., away from private mapping efforts or other on-going drug development projects), the quality and accessibility of new technologies, and the impact of concentrating research activity among incumbent firms (Cohen, 2010). Other key aspects to consider are the influence of intellectual property regulations, the utilization of existing medical technologies (e.g., off-label drug use), and the long-term effects on consumers (e.g., access to targeted therapies, health outcomes). Incorporating these factors is necessary to gain a more comprehensive understanding of the true social value of public mapping initiatives, which may surpass the market values presented in Section 5.

Third, the findings underscore the overall significance of public mapping initiatives and other open-access data efforts (Bryan and Ozcan, 2021; Koffi and Marx, 2023). In recent years, there has been a surge in the large-scale creation and distribution of data (Jones and Tonetti, 2020). Numerous publicly funded mapping projects aiming to collect and distribute basic scientific data, have been launched across a wide range of settings, with important implications for private sector firms motivated to leverage big data to maintain their competitive advantages (Fleming et al., 2019). This includes newer initiatives, such as the NIH's BRAIN Initiative for human brain mapping and the World Bank Group's Global Solar Atlas and Global Wind Atlas to support investments in renewable energy projects. This also includes significant technological enhancements to longstanding efforts, such as the US Census to collect and disseminate real-time socioeconomic information about the US population. Future work should explore the generalizability of these findings to other settings and exploit the unique features of these other data efforts to understand the unique effects of data collection, systematization, and dissemination.

Taken together, this analysis introduces a novel theoretical framework and new data to shed light on a crucial yet understudied aspect of innovation: the role of basic scientific information in guiding firms through the R&D process. As governments explore effective approaches for stimulating innovation, it is essential to understand the impact of public mapping information on the rate

and direction of subsequent private-sector innovation. This understanding is crucial for crafting policies that assist firms in navigating the uncertain technological development landscape, enabling firms to steer their efforts towards the efficient development of valuable new technologies.

## References


Abadie, A., Athey, S., Imbens, G. W. and Wooldridge, J. M. (2023), "When should you adjust standard errors for clustering?", *The Quarterly Journal of Economics* **138**(1), 1–35.

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

Amar, D., Izraeli, S. and Shamir, R. (2017), "Utilizing somatic mutation data from numerous studies for cancer research: proof of concept and applications", *Oncogene* **36**(24), 3375–3383.

American Cancer Society (2021), "Key statistics for ovarian cancer". https://www.cancer.org/cancer/ovarian-cancer/about/key-statistics.html (Accessed on 2021-01-13).

Anderson, Monique L. Chiswell, K., Peterson, E. D., Tasneen, A., Topping, J. and Califf, R. M. (2015), "Compliance with results reporting at clinicaltrials.gov", *New England Journal of Medicine* **372**(11), 1031–39.

Arora, A., Belenzon, S. and Sheer, L. (2021), "Knowledge spillovers and corporate investment in scientific research", *American Economic Review* **111**(3), 871–98.

Arora, A. and Gambardella, A. (1994), "Evaluating technological information and utilizing it", *Journal of Economic Behavior and Organization* **24**(1), 91–114.

Arrow, K. J. (1962), Economic welfare and the allocation of resources for invention, *in* C. o. E. G. o. t. S. S. R. C. Universities-National Bureau of Committee for Economic Research, ed., "The Rate and Direction of Invention Activity: Economic and Social Factors", Princeton University Press, Princeton, NJ, pp. 609–626. https://www.nber.org/system/files/chapters/c2144/c2144.pdf (Accessed on 2021-01-13).

Athey, S. and Imbens, G. W. (2022), "Design-based analysis in difference-in-differences settings with staggered adoption", *Journal of Econometrics* **226**(1), 62–79.

Avorn, J. (2015), "The $2.6 billion pill—methodologic and policy considerations", *New England Journal of Medicine* **372**(20), 1877–1879.

Azoulay, P., Graff Zivin, J. S., Li, D. and Sampat, B. N. (2019), "Public r&d investments and private-sector patenting: Evidence from nih funding rules", *The Review of Economic Studies* **86**(1), 117–52.

Berndt, E. R., Bui, L., Reiley, D. R. and Urban, G. L. (1995), "Information, marketing, and pricing in the u.s. antiulcer drug market", *American Economic Review* **85**(2), 100–5.

Berndt, E. R., Cockburn, I. M. and Grepin, K. A. (2006), "The impact of incremental innovation in biopharmaceuticals", *Pharmacoeconomics* **24**(2), 69–86.

Bessen, J. (2008), "The value of us patents by owner and patent characteristics", *Research Policy* **37**(5), 932–945.

Bloom, N., Schankerman, M. and Van Reenen, J. (2013), "Identifying technology spillovers and product market rivalry", *Econometrica* **81**(4), 1347–1393.

Bryan, K. A. and Ozcan, Y. (2021), "The impact of open access mandates on invention", *The Review of Economics and Statistics* **103**(5), 954–967.

Bryan, K. A. and Williams, H. L. (2021), Innovation: Market failures and public policies, *in* "Handbook of Industrial Organization", Vol. 5, (Elsevier), pp. 281–388.

Budish, E., Roin, B. N. and Williams, H. (2016), "Patents and research investments: Assessing the empirical evidence", *American Economic Review* **106**(5), 183–187.

Bujar, M., McAuslane, N., Walker, S. R. and Salek, S. (2017), "Evaluating quality of decision-making processes in medicines' development, regulatory review, and health technology assessment: A systematic review of the literature", *Frontiers in Pharmacology* **8**, 189.

Cameron, A. C. and Miller, D. L. (2015), "A practitioner's guide to cluster-robust inference", *Journal of human resources* **50**(2), 317–372.

Campbell, J. D., Alexandrov, A., Kim, J., Wala, J., Berger, A. H., Pedamallu, C. S., Shukla, S. A., Guo, G., Brooks, A. N. and Meyerson, M. (2016), "Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas", *Nature Genetics* **48**(6), 607–616.

Cancer Genome Atlas Research Network (2011), "Integrated genomic analyses of ovarian carcinoma", *Nature* **474**(7353), 609–15.

Cancer Genome Atlas Research Network (2018), "What is cancer genomics?". https://cancergenome.nih.gov/cancergenomics/whatisgenomics/whatis (Accessed on 2018-10-01).

Carr, T. H., McEwen, R., Dougherty, B., Johnson, J. H., Dry, J. R., Lai, Z., Ghazoui, Z., Laing, Naomi M. Hodgson, D. R., Cruzalegui, F., Hollingsworth, S. J. and Barrett, J. C. (2016), "Defining actionable mutations for oncology therapeutic development", *Nature Reviews Cancer* **16**(5), 319–329.

Cerami, E., Gao, J., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L. and Larsson, E. (2012), "The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data", *Cancer Discovery* **2**(5), 401–404.

Chandra, A., Garthwaite, C. and Stern, A. (2017), Characterizing the drug development pipeline for precision medicines, *in* E. Berndt, D. Goldman and J. Rowe, eds, "Economic Dimensions of Personalized and Precision Medicine", University of Chicago Press, Chicago, pp. 115–157.

Cheng, F., Lu, W., Liu, C., Fang, J., Hou, Y., Handy, D. E., Wang, R., Zhao, Y., Yang, Y., Huang, J. et al. (2019), "A genome-wide positioning systems network algorithm for in silico drug repurposing", *Nature Communications* **10**(1), 3476.

Choi, J. P. (1991), "Dynamic r&d competition under hazard rate uncertainty", *The RAND Journal of Economics* **22**(4), 596–610.

Cockburn, I. and Henderson, R. (1994), "Racing to invest? the dynamics of competition in ethical drug discovery", *Journal of Economics and Management Strategy* **3**(3), 481–519.

Cockburn, I. M. and Henderson, R. M. (1998), "Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery", *The Journal of Industrial Economics* **46**(2), 157–182.

Cockburn, I. M., Henderson, R. M. and Stern, S. (2000), "Untangling the origins of competitive advantage", *Strategic Management Journal* **21**(10-11), 1123–1145.

Cohen, W. M. (2010), "Fifty years of empirical studies of innovative activity and performance", *Handbook of the Economics of Innovation* **1**, 129–213.

Cohen, W. M. and Levinthal, D. A. (1990), "Absorptive capacity: A new perspective on learning and innovation", *Administration Science Quarterly* **35**(1), 128–52.

Collins, F. S. (2011), "Mining for therapeutic gold", *Nature Reviews Drug Discovery* **10**(6), 397.

Collins, F. S. and McKusick, V. A. (2001), "Implications of the human genome project for medical science", *The Journal of the American Medical Association* **285**(5), 540–544.

Correia, S., Guimarães, P. and Zylkin, T. (2020), "Fast poisson estimation with high-dimensional fixed effects", *The Stata Journal* **20**(1), 95–115.

Cutler, D. M. (2008), "Are we finally winning the war on cancer?", *Journal of Economic Perspectives* **22**(4), 3–26.

Danzon, P. M. and Keuffel, E. L. (2014), Regulation of the pharmaceutical-biotechnology industry, *in* N. L. Rose, ed., "Economic Regulation and Its Reform: What Have We Learned?", University of Chicago Press, Chicago, pp. 407–484.

David, P. A., Mowery, D. C. and Steinmueller, W. E. (1992), "Analyzing the economic payoffs from basic research", *Economics of Innovation and New Technology* **2**(1), 73–90.

De Chaisemartin, C. and d'Haultfoeuille, X. (2020), "Two-way fixed effects estimators with heterogeneous treatment effects", *American Economic Review* **110**(9), 2964–2996.

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., Mooney, T. B., Callaway, M. B., Dooling, D. and Mardis, E. R. (2012), "Music: Identifying mutational significance in cancer genomes", *Genome Research* **22**(8), 1589–1598.

DiMasi, J. A. (2001), "New drug development in the united states from 1963 to 1999", *Clinical Pharmacology and Therapeutics* **69**(5), 286–96.

DiMasi, J. A. (2013), "Innovating by developing new uses of already-approved drugs: Trends in the marketing approval of supplemental indications", *Clinical Therapeutics* **35**(6), 808–18.

DiMasi, J. A., Grabowski, H. G. and Hansen, R. W. (2016), "Innovation in the pharmaceutical industry: New estimates of r&d costs", *Journal of Health Economics* **47**, 20–33.

DiMasi, J. A., Grabowski, H. G. and Vernon, J. (2004), "R&d costs and returns by therapeutic category", *Drug Information Journal* **38**(3), 211–223.

DiMasi, J. A., Hansen, R. W. and Grabowski, H. G. (2003), "The price of innovation: New estimates of drug development costs", *Journal of Health Economics* **22**(2), 151–185.

Donelan, R., Walker, S. and Salek, S. (2015), "Factors influencing quality decision-making: Regulatory and pharmaceutical industry perspectives", *Pharmacoepidemiology and Drug Safety* **24**(3), 319–28.

Dubois, P., de Mouzon, O., Scott-Morton, F. and Seabright, P. (2015), "Market size and pharmaceutical innovation", *The RAND Journal of Economics* **46**(4), 844–871.

Dubois, P. and Kyle, M. (2016), "Are cancer drugs worth the price? the effects of pharmaceutical innovation on cancer mortality rates".

Dulbecco, R. (1986), "A turning point in cancer research: Sequencing the human genome", *Science* **231**(4742), 1055–1056.

Eisenberg, R. S. (2005), "The problem of new uses", *Yale Journal of Health Policy, Law, and Ethics* **5**(2), 717–40.

Fabrizio, K. R. (2009), "Absorptive capacity and the search for innovation", *Research Policy* **38**(2), 255–67.

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

Fleming, L. and Sorenson, O. (2003), "Navigating the technological landscape of innovation", *MIT Sloan Management Review* **44**(2), 15–23.

Fleming, L. and Sorenson, O. (2004), "Science as a map in technological search", *Strategic Management Journal* **25**(8–9), 909–928.

Fudenberg, D., Gilbert, R., Stiglitz, J. and Tirole, J. (1983), "Preemption, leapfrogging and competition in patent races", *European Economic Review* **22**(1), 3–31.

Furman, J. L. and Stern, S. (2011), "Climbing atop the shoulders of giants: The impact of institutions on cumulative research", *American Economic Review* **101**(5), 1933–1963.

Futureal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M. R. (2004), "A census of human cancer genes", *Nature Reviews Cancer* **4**(3), 177–83.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C. and Schultz, N. (2013), "Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal", *Science Signaling* **6**(269), I1.

Gardner, J. (2022), "Two-stage differences in differences", *arXiv preprint arXiv:2207.05943* .

Gold, E. R. and Carbone, J. (2010), "Myriad genetics: In the eye of the policy storm", *Genetics in Medicine* **12**(4 Suppl), S39–S70.

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

Greenblatt, W., Gupta, C. and Kao, J. (2023), "Drug repurposing during the covid-19 pandemic: Lessons for expediting drug development and access", *Health Affairs* **42**(3), 424–432.

Guedj, I. and Scharfstein, D. (2004), Organizational scope and investment: Evidence from drug development strategies and performance of biopharmaceutical firms. http://dx.doi.org/10.2139/ssrn.621322 (Accessed on 2021-01-13).

Hall, B. H., Mairesse, J. and Mohnen, P. (2010), Measuring the returns to r&d, *in* "Handbook of the Economics of Innovation", Vol. 2, Elsevier, pp. 1033–1082.

Hall, B. and Van Reenen, J. (2000), "How effective are fiscal incentives for r&d? a review of the evidence", *Research Policy* **29**(4–5), 449–69.

Harris, C. and Vickers, J. (1985), "Perfect equilibrium in a model of a race", *The Review of Economic Studies* **52**(2), 193–209.

Hausman, J. A., Hall, B. H. and Griliches, Z. (1984), "Econometric models for count data with an application to the patents-r&d relationship".

Heron, M. (2018), "Deaths: Leading causes for 2016", *National Vital Statistics Reports* **67**(6), 1–77.

IQIVIA Institute (2018), "Global oncology trends 2018". https://www.iqvia.com/insights/the-iqvia-institute/reports/global-oncology-trends-2018 (Accessed on 2021-01-13).

Jayaraj, S. (2018), Scientific maps and innovation: Impact of the human genome on drug discovery. PhD diss., Rutgers University.

Jones, C. I. and Tonetti, C. (2020), "Nonrivalry and the economics of data", *American Economic Review* **110**(9), 2819–58.

King, G. and Zeng, L. (2001), "Logistic regression in rare events data", *Political analysis* **9**(2), 137–163.

Klevorick, A., Levin, R., Nelson, R. and Winter, S. (1995), "On the sources and significance of interindustry differences in technological opportunities", *Research Policy* **24**(2), 185–205.

Koffi, M. and Marx, M. (2023), Cassatts in the attic, Technical report, National Bureau of Economic Research.

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

Kolata, G. (2013), "Cancers share gene patterns, studies affirm", *New York Times* . May 1. https://www.nytimes.com/2013/05/02/health/dna-research-points-to-new-insight-into-cancers.html (Accessed 2021-01-13).

Krieger, J. L. (2021), "Trials and terminations: Learning from competitors' r&d failures", *Management Science* **67**(9), 5525–5548.

Krishnamurthy, N., Grimshaw, A. A., Axson, S. A., Choe, S. H. and Miller, J. E. (2022), "Drug repurposing: a systematic review on root causes, barriers and facilitators", *BMC Health Services Research* **22**(1), 1–17.

Kyle, M. K. (2007), "Pharmaceutical price controls and entry strategies", *The Review of Economics and Statistics* **89**(1), 88–99.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivanchenko, A., Carter, S. L., Stewart, C., Mermel, C. H. and Roberts, S. A. (2014), "Mutational heterogeneity in cancer and the search for new cancer genes", *Nature* **499**(7457), 214–18.

Lichtenberg, F. R. (2015), "The impact of pharmaceutical innovation on premature cancer mortality in canada, 2000–2011", *International journal of health economics and management* **15**, 339–359.

Lieberman, M. B. and Montgomery, D. B. (1988), "First-mover advantages", *Strategic Management Journal* **9**, 41–58. S1.

Loury, G. C. (1979), "Market structure and innovation", *The quarterly journal of economics* **93**(3), 395–410.

Mardis, E. R. (2018), "Insights from large-scale cancer genome sequencing", *Annual Reviews of Cancer Biology* **2**, 429–44.

Matulonis, U. A. (2017), "Parp inhibitors in brca-related ovarian cancer–and beyond!".

Minikel, E. V., Painter, J. L., Dong, C. C. and Nelson, M. R. (2023), "Refining the impact of genetic evidence on clinical success", *medRxiv* pp. 2023–06.

Mowery, D. and Rosenberg, N. (1979), "The influence of market demand upon innovation: A critical review of some recent empirical studies", *Research Policy* **8**(2), 102–153.

Nagaraj, A. (2022), "The private impact of public data: Landsat satellite maps increased gold discoveries and encouraged entry", *Management Science* **68**(1), 564–582.

Nagaraj, A. and Stern, S. (2020), "The economics of maps", *Journal of Economic Perspectives* **34**(1), 196–221.

Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P. C., Li, M. J., Wang, J. et al. (2015), "The support of human genetic evidence for approved drug indications", *Nature genetics* **47**(8), 856–860.

Nelson, R. R. (1959), "The simple economics of basic scientific research", *The Journal of Political Economy* **67**(3), 297–306.

Noel, M. and Schankerman, M. (2013), "Strategic patenting and software innovation", *The Journal of Industrial Economics* **61**(3), 481–520.

Pollack, A. (2014), "Aiming to push genomics forward in new study", *New York Times* . Jan 13. https://www.nytimes.com/2014/01/13/business/aiming-to-push-genomics-forward-in-new-study.html (Accessed 2023-07-31).

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C. et al. (2019), "Drug repurposing: Progress, challenges and recommendations", *Nature Reviews Drug Discovery* **18**(1), 41–58.

Reinganum, J. F. (1989), The timing of innovation: Research, development, and diffusion, *in* R. Schmalensee and R. Willig, eds, "Handbook of Industrial Organization, Vol 1", Elsevier Science, Amsterdam, pp. 849–908.

Robertson, A. G., Kim, J., Al-Ahmadie, H., Bellmunt, J., Guo, G., Cherniack, A. D., Hinoue, T., Laird, P. W., Hoadley, K. A. and Lerner, S. P. (2017), "Comprehensive molecular characterization of muscle-invasive bladder cancer", *Cell* **171**(3), 540–56.

Roin, B. (2013), Solving the problem of new uses. https://ssrn.com/abstract=2337821 (Accessed on 2021-01-13).

Rosenberg, N. (1974), "Science, invention, and economic growth", *The Economic Journal* **84**(333), 90–108.

Roth, J., Sant'Anna, P. H., Bilinski, A. and Poe, J. (2023), "What's trending in difference-in-differences? a synthesis of the recent econometrics literature", *Journal of Econometrics* .

Sampat, B. N. (2015), Serendipity. https://ssrn.com/abstract=2545515 (Accessed on 2021-01-13).

Sampat, B. N. and Lichtenberg, F. R. (2011), "What are the respective roles of the public and private sectors in pharmaceutical innovation?", *Health affairs* **30**(2), 332–339.

Scherer, F. (2000), Markets for pharmaceutical products, *in* A. J. Culyer and J. P. Newhouse, eds, "Handbook of Health Economics, Vol 1", Elsevier Science, Amsterdam, pp. 1297–1336.

Sertkaya, A., Wong, H.-H., Jessup, A. and Beleche, T. (2016), "Key cost drivers of pharmaceutical clinical trials in the united states", *Clinical Trials* **13**(2), 117–126.

Sharpe, P. and Keelin, J. (1998), "How smithkline beecham makes better resource-allocation decisions", *Harvard Business Review* **76**(2), 45–46.

Stratton, M. R., Campbell, P. J. and Futreal, P. A. (2009), "The cancer genome", *Nature* **458**(7239), 719–24.

Struewing, J. P., Hartge, P., Wacholder, S., Baker, S. M., Berlin, M., McAdams, M., Timmerman, M. M., Brody, L. C. and Tucker, M. A. (1997), "The risk of cancer associated with specific mutations of brca1 and brca2 among ashkenazi jews", *New England Journal of Medicine* **336**(20), 1401–1408.

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

Tate, J. G., Bamford, S., Jubb, H. C., Sandra, Z., Beard, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jute, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., Speedy, H. E., Stefancsik, R., Thompson, S. L., Wang, S., Ward, S., Campbell, P. J. and Forbes, S. A. (2018), "Cosmic: The catalogue of somatic mutations", *Nucleic Acids Research* **47**(D1), D941–47.

Tranchero, M. (2023), Finding diamonds in the rough: Data-driven decisions and pharmaceutical innovation. Working Paper.

Tversky, A. and Kahneman, D. (1974), "Judgement under uncertainty: Heuristics and biases", *Science* **185**(4157), 1124–1131.

United States, Office of the Press Secretary (2000), "Text of remarks on the completion of the first survey of the entire human genome project". https://www.genome.gov/10001356/june-2000-white-house-event (Accessed on 2021-01-13).

U.S. Food and Drug Administration (1998*a*), Guidance for industry: Fda approval of new cancer treatment uses for marketed and biological products, Technical report. https://www.fda.gov/media/71396/download (Accessed on 2021-01-13).

U.S. Food and Drug Administration (1998*b*), Guidance for industry: Providing clinical evidence of effectiveness for human drug and biological products, Technical report. https://www.fda.gov/media/71655/download (Accessed on 2021-01-13).

U.S. Food and Drug Administration (2004), Guidance for industry: Ind exemptions for studies of lawfully marketed drug or biological products for the treatment of cancer, Technical report. https://www.fda.gov/media/71627/download (Accessed on 2021-01-13).

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz Jr., L. A. and Kinzler, K. W. (2013), "Cancer genome landscapes", *Science* **339**(6127), 1546–58.

Wetterstrand, K. (2018), Dna sequencing costs: Data, Technical report, National Human Genome Research Institute. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (Accessed on 2021-01-13).

Wheeler, D. A. and Wang, L. (2013), "From human genome to cancer genome: The first decade", *Genome Research* **23**(7), 1054–1062.

Williams, H. (2013), "Intellectual property rights and innovation: Evidence from the human genome", *Journal of Political Economy* **121**(1), 1–27.

Wong, C. H., Siah, K. W. and Lo, A. W. (2019), "Estimation of clinical trial success rates and related parameters", *Biostatistics* **20**(2), 273–286.

# Charted Territory: Evidence from Mapping the Cancer Genome and R&D Decisions in the Pharmaceutical Industry

## Online Appendix

## Appendix A Analytical framework discussion and proofs

This online appendix provides reproduces the discussion in Section 2 in more detail and provides proofs. All notations follow from the definitions in Section 2. This framework closely follows and builds on Choi (1991).

### A.1 Preliminaries

Consider two risk neutral firms, $i = 1, 2$, that compete with each other to be the first to successfully develop and commercialize an invention for a market opportunity (e.g., patients with a specific disease subtype) by making research investments (e.g., conducting clinical trials). A research investment is characterized by the following parameters:

***Duration of Research Investment:*** At each point in time, $t$, given no success to date, firm $i$ must decide whether or not to continue investing in a market opportunity. In the context of the pharmaceutical industry, a research investment refers to developing a particular product (e.g., a new chemical compound). The longer a pharmaceutical firm is engaged in the R&D race, the greater its likelihood of advancing from preclinical research (that tests a drug candidate in animals and human cells) to clinical research (that tests a drug in patients with a specific disease subtype). In other words, the firm will conduct a clinical trial if it is still in the R&D race at $t_i \geq \bar{t}$. I treat the threshold $\bar{t}$ as deterministic for simplicity; in practice this parameter would be endogenous.

***Costs of Research Investment:*** If a firm decides to invest in a market opportunity, it incurs a flow cost of $c$. In the pharmaceutical industry, $c$ can be interpreted as the cost of recruiting study participants, conducting a clinical trial, analyzing clinical trial data, etc.

***Payoffs:*** If firm $i$ is the first to successfully yield a commercially viable treatment, it obtains a potential payoff of $V$. If the firm decides to cease investment, it obtains a payoff of 0 with certainty. Payoffs are discounted at rate $r$.

***Likelihood of Success:*** Market opportunities vary in their ex ante potential to successfully yield a commercially viable treatment, with success occurring at a stochastic rate. Following much of the previous literature on innovation races (e.g., Loury, 1979), the probability that a research investment is successful at a given time $t$ takes an exponential distribution. Therefore, the date at which a research investment yields a commercially viable treatment is denoted by $\tau$, where $\mathrm{P}(\tau < t) = 1 - e^{-\lambda^i t}$ and $\lambda^i$ is the hazard rate or the conditional probability of success, given no prior success. $\tau$ can be interpreted as the point when a pharmaceutical firm's clinical trials demonstrate the safety and effectiveness of a drug for a specific patient disease subtype.

For simplicity, I assume that the success rate for a specific market opportunity is either high (i.e., a "high success rate" market opportunity has a hazard rate $\lambda_H^i$) or low (i.e., a "low success

rate" market opportunity has a hazard rate $\lambda_L^i$, where $\lambda_H^i > \lambda_L^i > 0$). $\lambda^i$ also reflects a firms' likelihood of success. Reflecting uncertainty in the research process, firms do not know the true success rate of a given market opportunity. For simplicity, I follow the framework and notation of previous dynamic R&D race literature (e.g., Choi, 1991) in assuming that firms have a prior belief $p_i$ that a given market opportunity has a low success rate.

In the pharmaceutical industry, firm $i$'s belief about a market opportunity's success rate, $p_i$, is shaped by both "existing" and "novel" information. For simplicity, in this analytical framework, "existing" information refers to data from internal scientific mapping studies that characterize the biological characteristics linked to a particular disease subtype.[i] "Novel" information consists of three elements: (1) newly public information from public data sources (e.g., public maps); (2) newly revealed (common) information on the firm's own research outcomes (as time passes without success, the firm believes that a market opportunity has low success rate); and (3) newly revealed (common) information from competitors' research outcomes.

The costly nature of conducting proprietary mapping experiments can result in low levels of prior existing information for a firm. When a firm has limited access to existing mapping information, leading to increased uncertainty, it will assign greater weight to the novel public signals. Following Krieger (2021), this is weight indicated by $u_i$, where $u_i \in [0, 1]$. For simplicity, we focus on novel information coming from public data sources. A firm $i$'s belief that a market opportunity has a low success rate is represented with the following:

$$p_i = u_i(Public\ Information\ Signal_i) + (1 - u_i)(Existing\ Information\ Signal_i). \tag{A.1}$$

## A.2 Private research investment incentives

First, for clarity, I characterize the key findings of (Choi, 1991) to characterize the optimal level of research investment for a firm without competition. This will serve as a useful benchmark once I introduce decision making within a competitive setting and public mapping information. I begin by defining a firm's expected likelihood of success and then determine its optimal level of research investment.

I define $p(t_i)$ to be the posterior probability at time $t$ that firm $i$ considers the success rate of the market opportunity to be $\lambda_L^i$, given that there has been no success up to time $t_i$.

By Bayes' rule,

$$p(t_i) = Pr(\lambda_L^i | \text{no success until } t_i) = \frac{p_i e^{-\lambda_L^i t_i}}{p_i e^{-\lambda_L^i t_i} + (1 - p_i) e^{-\lambda_H^i t_i}}. \tag{A.2}$$

Let $\lambda(p(t_i))$ denote the expected likelihood of success at time $t$. Then $\lambda(p(t_i)) = p(t_i)\lambda_L^i + (1 - p(t_i))\lambda_H^i$.

Reflecting the fact that firms believe that a market opportunity has a low success rate as more time passes without success, $\lambda^i(p(t_i))$ is a strictly decreasing function of $t$.

[i] In practice, other sources of prior existing information may come from internal clinical trials. However, pharmaceutical firms have strong incentives to closely monitor their rivals' clinical trial activities. Further, since the FDA Amendments Act (FDAAA) of 2007, trial sponsors have been required to report the results of most phase II clinical trials to the public trial registry, ClinicalTrials.gov within one year of completion, though results reporting is far from complete (Anderson et al., 2015).

***Proof that*** $\lambda(p(t_i))$ ***is a strictly decreasing function of*** $t_i$***:***

Recall that $\lambda(p(t_i)) = p(t_i)\lambda_L^i + (1 - p(t_i))\lambda_H^i$. Taking the derivative of $\lambda(p(t_i))$ with respect to $t_i$ yields

$$\begin{aligned}\frac{\lambda(p(t_i))}{\partial t_i} &= \frac{\partial p(t_i)}{\partial t_i}\lambda_L^i - \frac{\partial p(t_i)}{\partial t_i}\lambda_H^i \\ &= \frac{\partial p(t_i)}{\partial t_i}\Big(\lambda_L^i - \lambda_H^i\Big).\end{aligned} \tag{A.3}$$

Since $\lambda_L^i - \lambda_H^i < 0$, if I can show that $\frac{\partial p(t_i)}{\partial t_i} > 0$, then it must be the case that $\frac{\partial \lambda(p(t_i))}{\partial t_i} < 0$.

To sign $\frac{\partial p(t_i)}{\partial t_i}$, consider $p(t_i)$, which is equal to $\frac{p_i e^{-\lambda_L^i t_i}}{p_i e^{-\lambda_L^i t_i} + (1-p_i)e^{-\lambda_H^i t_i}}$ (from equation (A.2)). Taking the derivative $p(t_i)$ with respect to $t_i$ and applying the quotient rules yields

$$\frac{\partial p(t_i)}{\partial t_i} = \frac{-\lambda_L^i p_i e^{-\lambda_L^i t_i}\Big(p_i e^{-\lambda_L^i t_i} + (1-p_i)e^{-\lambda_H^i t_i}\Big) - p_i e^{-\lambda_L^i t_i}\Big(-\lambda_L^i p_i e^{-\lambda_L^i t_i} - (1-p_i)\lambda_H^i e^{-\lambda_H^i t_i}\Big)}{\Big[p_i e^{-\lambda_L^i t_i} + (1-p_i)e^{-\lambda_H^i t_i}\Big]^2}.$$

Rearranging gives

$$\begin{aligned}\frac{\partial p(t_i)}{\partial t_i} &= \frac{-p_i(1-p_i)\lambda_L^i e^{-(\lambda_L^i+\lambda_H^i)t_i} + p_i(1-p_i)\lambda_H^i e^{-(\lambda_L^i+\lambda_H^i)t_i}}{\Big[p_i e^{-\lambda_L^i t_i} + (1-p_i)e^{-\lambda_H^i t_i}\Big]^2} \\ &= \frac{p_i(1-p_i)e^{-(\lambda_L^i+\lambda_H^i)t}(\lambda_H^i - \lambda_L^i)}{\Big[p_i e^{-\lambda_L^i t_i} + (1-p_i)e^{-\lambda_H^i t_i}\Big]^2} > 0.\end{aligned} \tag{A.4}$$

I can sign this expression by noting that $(\lambda_H^i - \lambda_L^i) > 0$. Therefore, $\frac{\partial \lambda(p(t_i))}{\partial t_i} < 0$, as required.

Next, I determine the optimal amount of time that firm $i$ spends in the R&D race (i.e., its optimal amount of research investment). At each point in time $t$, firm $i$ will continue to stay in the R&D race if and only if the expected benefits exceed the costs,

$$\text{Private Investment Occurs} \Leftrightarrow \lambda(p(t_i))V \geq c. \tag{A.5}$$

Firm $i$ will invest until $t_i^*$, where $\lambda(p(t_i^*))V = p(t_i^*)\lambda_L^i + (1 - p(t_i^*))\lambda_H^i = c$. Solving for $t_i^*$ by substituting for $p(t_i)$ in equation (A.2) yields

$$t_i^* = \frac{1}{\lambda_H^i - \lambda_L^i} ln \frac{(1-p_i)\Big(\lambda_H^i - \frac{c}{V}\Big)}{p_i\Big(\frac{c}{V} - \lambda_L^i\Big)}. \tag{A.6}$$

***Proof of equation (A.6):***

I can derive $t_i^*$ by substituting $p(t_i)$ in the equation for $\lambda(p(t_i))$:

$$\begin{aligned}\frac{c}{V} &= \lambda(p(t_i)) \\ &= p(t_i)\lambda_L^i + (1-p(t_i))\lambda_H^i \\ &= \frac{p_i e^{-\lambda_L^i t_i}}{p_i e^{-\lambda_L^i t_i} + (1-p_i)e^{-\lambda_H^i t_i}}\lambda_L^i + \frac{(1-p_i)e^{-\lambda_H^i t_i}}{p_i e^{-\lambda_L^i t_i} + (1-p_i)e^{-\lambda_H^i t_i}}\lambda_H^i \\ &= \lambda_L^i + (\lambda_H^i - \lambda_L^i)\frac{(1-p_i)e^{-\lambda_H^i t_i}}{p_i e^{-\lambda_L^i t_i} + (1-p_i)e^{-\lambda_H^i t_i}}.\end{aligned} \tag{A.7}$$

Rearranging terms and cross-multiplying, yields:

$$\frac{1}{\frac{c}{V} - \lambda_L^i} = \frac{1}{\lambda_H^i - \lambda_L^i}\left[1 + \frac{p_i}{1-p_i}e^{(\lambda_H - \lambda_L^i)t_i}\right]. \tag{A.8}$$

Rearranging terms yields:

$$e^{(\lambda_H^i - \lambda_L^i)t_i} = \frac{(1-p_i)\left(\lambda_H^i - \frac{c}{V}\right)}{p\left(\frac{c}{V} - \lambda_L^i\right)}.$$

Taking the natural logarithm yields the formula for $t_i^*$:

$$\begin{aligned}t_i^* &= \frac{1}{\lambda_H^i - \lambda_L^i} ln \frac{(1-p_i)\left(\lambda_H^i - \frac{c}{V}\right)}{p_i\left(\frac{c}{V} - \lambda_L^i\right)} \\ &= \boldsymbol{I}^{-1} + ln\boldsymbol{B}.\end{aligned} \tag{A.9}$$

where $\boldsymbol{I} = \lambda_H^i - \lambda_L^i$ and $\boldsymbol{B} = \frac{(1-p_i)\left(\lambda_H^i - \frac{c}{V}\right)}{p_i\left(\frac{c}{V} - \lambda_L^i\right)}$.

As described in detail in Choi (1991), the key thing to notice about equation (A.6) is that the optimal level of research investment is determined by an interplay between the firm's own research experience (firm $i$ becomes more pessimistic as time passes without success, at rate $\boldsymbol{I} = \lambda_H^i - \lambda_L^i$) and the firm's ex ante benefit of making a research investment at time zero ($\boldsymbol{B} = (1-p_i)(\lambda_H^i - \frac{c}{V})/p_i(\frac{c}{V} - \lambda_L^i)$).

It is straightforward to extend the analysis to a setting with multiple firms. In this competitive setting, $p(t_i)$ becomes common posterior probability at time $t$ that both firms consider the set of

hazard rates to be $(\lambda_L^1, \lambda_L^2)$, given that there has been no success up to time $t$.[ii] Given no success by either firm, firm $i$ will invest until $t_i = t_i^{**}$ where satisfies $\lambda(p(t_i^{**}))V = p(t_i^{**})\lambda_L^i + (1-p(t_i^{**}))\lambda_H^i = c$. The modified formula for $t_i^{**}$ yields:

$$t_i^{**} = \frac{1}{(\lambda_H^1 - \lambda_L^1) + (\lambda_H^2 - \lambda_L^2)} ln \frac{(1-p_i)\left(\lambda_H^i - \frac{c}{V}\right)}{p_i\left(\frac{c}{V} - \lambda_L^i\right)}. \tag{A.10}$$

As described above in the proof of equation (A.6), $\boldsymbol{I}$ can be interpreted as the rate at which both firms become more pessimistic over time. As information about research investments is perfectly observable, this is a common component shared by both firms. In contrast, $\boldsymbol{B}$ is an idiosyncratic component and reflects firm $i$'s ex ante benefit of making a research investment at time zero.

***Proof of equation (A.10):***

I modify $p(t_i)$ so that it is now the common posterior probability at time $t_i$ that firm $i$ considers the set of hazard rates to be $(\lambda_L^1, \lambda_L^2)$, given that there has been no success up to time $t_i$.

By Bayes' rule,

$$p(t_i) = Pr(\lambda_L^1, \lambda_L^2 | \text{no success until } t_i) = \frac{p_i e^{-(\lambda_L^1 + \lambda_L^2)t_i}}{p_i e^{-(\lambda_L^1 + \lambda_L^2)t_i} + (1-p_i)e^{-(\lambda_H^1 + \lambda_H^2)t_i}}. \tag{A.11}$$

Firm $i$ will invest until $t_i^{**}$, where $\lambda(t_i^{**})V = p(t_i^{**})\lambda_L^i + (1-p(t_i^{**}))\lambda_H^i = c$.

As with the proof of equation (A.6), I can derive $t_i^{**}$ by substituting the modified $p(t_i)$ in the equation for $\lambda(t_i)$:

$$\begin{aligned} \frac{c}{V} &= \lambda(t_i) \\ &= p(t_i)\lambda_L^i + (1-p(t_i))\lambda_H^i \\ &= \frac{p_i e^{-(\lambda_L^1 + \lambda_L^2)t_i}}{p_i e^{-(\lambda_L^1 + \lambda_L^2)t_i} + (1-p_i)e^{-(\lambda_H^1 + \lambda_H^2)t_i}} \lambda_L^i + \frac{(1-p_i)e^{-(\lambda_H^1 + \lambda_H^2)t_i}}{p_i e^{-(\lambda_L^1 + \lambda_L^2)t_i} + (1-p_i)e^{-(\lambda_H^1 + \lambda_H^2)t_i}} \lambda_H^i \\ &= \lambda_L^i + (\lambda_H^i - \lambda_L^i) \frac{(1-p_i)e^{-(\lambda_H^1 + \lambda_H^2)t_i}}{p_i e^{-(\lambda_L^1 + \lambda_L^2)t_i} + (1-p_i)e^{-(\lambda_H^1 + \lambda_H^2)t_i}}. \end{aligned} \tag{A.12}$$

Rearranging terms and cross multiplying, yields

$$\frac{1}{\frac{c}{V} - \lambda_L^i} = \frac{1}{\lambda_H^i - \lambda_L^i}\left[1 + \frac{p_i}{1-p_i} e^{((\lambda_H^1 - \lambda_L^1) + (\lambda_H^2 - \lambda_L^2))t_i}\right]. \tag{A.13}$$

Rearranging terms yields

$$e^{((\lambda_H^1 - \lambda_L^1) + (\lambda_H^2 - \lambda_L^2))t_i} = \frac{(1-p_i)\left(\lambda_H^i - \frac{c}{V}\right)}{p_i\left(\frac{c}{V} - \lambda_L^i\right)}.$$

[ii] The modified formula for $p(t_i)$ becomes $p(t_i) = Pr(\lambda_L^1, \lambda_L^2 | \text{no success until } t_i) = \frac{p_i e^{-(\lambda_L^1 + \lambda_L^2)t_i}}{p_i e^{-(\lambda_L^1 + \lambda_L^2)t_i} + (1-p_i)e^{-(\lambda_H^1 + \lambda_H^2)t_i}}$.

Taking the natural logarithm yields the formula for $t_i^{**}$:

$$t_i^{**} = \frac{1}{(\lambda_H^1 - \lambda_L^1) + (\lambda_H^2 - \lambda_L^2)} ln \frac{(1-p_i)\left(\lambda_H^i - \frac{c}{V}\right)}{p_i\left(\frac{c}{V} - \lambda_L^i\right)} \tag{A.14}$$
$$= \boldsymbol{I}^{-1} + ln\boldsymbol{B}.$$

where $\boldsymbol{I} = (\lambda_H^1 - \lambda_L^1) + (\lambda_H^2 - \lambda_L^2)$ and $\boldsymbol{B} = \frac{(1-p_i)\left(\lambda_H^i - \frac{c}{V}\right)}{p_i\left(\frac{c}{V} - \lambda_L^i\right)}$.

The relative sizes of $t_1^{**}$ and $t_2^{**}$ determine how long each firm stays in the R&D race. Within this dynamic R&D racing framework, if $t_1^{**} > t_2^{**}$, then firm 2 drops out of the R&D race before firm 1. Competition has a negative impact on a given firm's beliefs about its expected payoffs, leading to a decline in the time that the firm spends in the R&D race (Loury, 1979).[iii] A direct comparison of equations (A.6) and (A.10) shows that $t_i^{**} < t_i^*$.[iv]

***Intuition behind the effect of competition:***

When there are multiple firms, as time progresses, firms' beliefs about a market opportunity's likelihood of success (i.e., whether the market opportunity has a high or low success rate) updates in response to (1) newly public information from public data sources (e.g., public maps); (2) newly revealed (common) information on the firm's own research outcomes (as time passes without success, the firm believes that a market opportunity has low success rate); and (3) newly revealed (common) information from competitors' research outcomes.

Regarding (3), at any time $t > 0$, competition has a negative impact on a given firm's beliefs about its expected payoffs, leading to a decline in time spent in the R&D race (Loury, 1979). This effect is exacerbated when $(\lambda_H^1 - \lambda_L^1) + (\lambda_H^2 - \lambda_L^2)$ is large (the gap between the potential success rates of the market opportunity is large). In this setting, competition essentially speeds up the rate at which firms learn about the potential of a market opportunity. As a result, firm $i$ becomes pessimistic as time passes without success at rate $(\lambda_H^1 - \lambda_L^1) + (\lambda_H^2 - \lambda_L^2)$, thus updating (and eroding uncertainty) more quickly than in the case where there is just one firm (i.e., the monopolist case). Recall monopolist firm $i$ updates at rate $\lambda_H^i - \lambda_L^i$. As discussed in Section 2, this framework primarily focuses on the competition-driven learning effects (competitors' experience decreases the probability of winning) rather than the business stealing effects (conditional on winning, payoffs are lower). This is consistent with the dynamic R&D models (Reinganum, 1989) and recent empirical evidence (Bloom et al., 2013; Krieger, 2021). This logic extends to cases with more than one competitor. The addition of a competitor leads to a greater decline in a given firm's time spent in the R&D race (firms learn from their competitors and therefore update and become more pessimistic at a faster rate).

[iii] Reflecting the previous literature on dynamic R&D models (Reinganum, 1989) and recent empirical evidence (Krieger, 2021), in this framework, competition shapes investment through learning spillovers (i.e., a firm becomes more pessimistic that a market opportunity has a high success rate as time passes and it and its rival are not successful) rather than business stealing effects (i.e., conditional on winning, payoffs are lower).

[iv] Theoretically, competition can lower firm 1 and firm 2's combined level of research (Reinganum, 1989).

## A.3 Adding novel public mapping information

I characterize the introduction of public mapping information as the provision of positive signals about the success rate of each market opportunity. There are three key features of this information. First, consisting of a set of information signals about the success rate of all possible market opportunities, mapping information helps firms navigate the R&D process by distinguishing whether any given market opportunity has a sufficiently high success rate: a mapping signal reveals that a market opportunity has low success rate with probability $p^M$, where $p^M \leq p_i$. For simplicity, I treat the mapping parameters as deterministic; in practice, these parameters could be stochastic in the sense that the signals are informative but noisy.[v]

Second, these mapping signals may have different strengths, which reflects how informative they are for drug development. A mapping signal operates through the public information component of equation (A.1). Comparing $p_i$ and $p^M$ indicates that $|p^M - p_i| = u_i \times b$, where $b$ captures the strength of the mapping signal, and $b \in [0, 1]$. For example, $b$ may reflect the clinical relevance of a scientific mapping signal.

Third, the comprehensive nature of public mapping information allows it to reveal previously unknown connections across market opportunities. This can lead to disproportionate benefits for specific firms that can quickly leverage these newfound connections, especially in competitive environments where there are first mover advantages. If firm 1 has a relatively higher likelihood of success due to its increased opportunity to effectively utilize the newly revealed connections relative to firm 2, then $\lambda_H^1 > \lambda_H^2$ and $\lambda_L^1 > \lambda_L^2$.

## A.4 Implications

This empirical work provides support for the idea that given the uncertainty in the R&D process, public mapping information shapes private sector firms' research investments. Before proceeding to the framework's predictions, I simplify my analysis by incorporating the following assumptions, drawn from the previous literature on dynamic R&D races: First, if firm $i$ is certain that a market opportunity has a high success rate ($\lambda_H^i$), then it is optimal for the firm to continue investing in research ($\lambda_H^i V - c > 0$). Second, if firm $i$ is certain that a market opportunity has a low success rate ($\lambda_L^i$), then it is not optimal for the firm to continue investing in research ($\lambda_L^i V - c < 0$). Third, unless otherwise stated, the two firms are identical. Finally, I assume initial entry is always worthwhile: $\lambda_i(p(0)) = p\lambda_L^i + (1 - p_i)\lambda_H^i > \frac{c}{V}$. These admittedly stylized simplifying assumptions allow me to use the duration of a firm's participation in the R&D race (before dropping out) as a direct proxy of its optimal research investment level. Applying this framework to drug development yields three main propositions regarding the effects of public mapping information.

**Proposition 1.** *The effect of a public mapping signal on the likelihood that it invests in a clinical trial is increasing in the strength of the mapping signal.*

**Corollary 1.** *Competition lowers the impact of public mapping information on the time a firm spends in the R&D race and its likelihood of investing in a clinical trial.*

[v] This would allow for false positives and Bayesian firms would have to appropriately account for noise. Incorporating the mapping information uncertainty into the framework would not change the implications, but they are excluded for simplicity. I empirically explore the effects of more vs. less noisy mapping signals in Section 4.4.1.

A firm is likely to spend less time in the race if it believes that the market opportunity as having a low success rate (i.e, $\frac{\partial t_i^{**}}{\partial p_i} < 0$). A mapping signal for a specific market opportunity can increase a firm's expected payoff, potentially leading to an increase in time spent in the R&D race and an increased likelihood of initiating a clinical trial in the market opportunity. In other words, a decrease in $p_i$ indicates a higher expected likelihood of success and increases the investment threshold (higher $\lambda(p(t_i))V$) in equation (A.5).

However, the impact of public mapping information on firms' clinical trial investments depends on the strength of these signals (i.e., represented by $b$). Signals that are weak or not clinically relevant (i.e., $b = 0$) are unlikely to affect firms' research investments. Additionally, competition has a negative impact on a given firm's beliefs about its expected payoffs, leading to a decline in the time that the firm spends in the R&D race (Loury, 1979). As a result, even strong mapping signals may not increase firms' research investments if the level of competition is too high. As a result, the net effect on research investment depends on the strength of the public mapping signals and the level of competition. The extent to which public mapping information shapes competing firms' research investments is an empirical question.

***Proof of Proposition 1:***

For simplicity, I begin with the case where there is one firm. Looking at equation (A.6), if I can show that $ln\frac{(1-p_i)\left(\lambda_H^i-\frac{c}{V}\right)}{p\left(\frac{c}{V}-\lambda_L^i\right)}$ is decreasing in $p_i$, then it must be the case that $\frac{\partial t_i^*}{\partial p_i} < 0$. Taking the derivative of $ln\frac{(1-p_i)\left(\lambda_H^i-\frac{c}{V}\right)}{p\left(\frac{c}{V}-\lambda_L^i\right)}$ with respect to $p_i$ and applying the quotient rule yields

$$\begin{aligned}\frac{\partial ln\frac{(1-p_i)\left(\lambda_H^i-\frac{c}{V}\right)}{p\left(\frac{c}{V}-\lambda_L^i\right)}}{\partial p_i} &= \frac{p\left(\frac{c}{V}-\lambda_L^i\right)}{(1-p_i)\left(\lambda_H^i-\frac{c}{V}\right)}\frac{-p\left(\lambda_H^i-\frac{c}{V}\right)p\left(\frac{c}{V}-\lambda_L^i\right)-(1-p_i)\left(\lambda_H^i-\frac{c}{V}\right)\left(\frac{c}{V}-\lambda_L^i\right)}{\left[p\left(\frac{c}{V}-\lambda_L^i\right)\right]^2} \\ &= \frac{-1}{p_i(1-p_i)} < 0,\end{aligned} \tag{A.15}$$

where I can sign this expression by noting that $1 - p > 0$.

As a result,

$$\frac{\partial t_i^*}{\partial p_i} = \frac{-1}{p_i(1-p_i)(\lambda_H^i-\lambda_L^i)} < 0. \tag{A.16}$$

If $q_i = 1 - p_i$, then,

$$\frac{\partial t_i^*}{\partial q_i} = \frac{1}{q_i(1-q_i)(\lambda_H^i-\lambda_L^i)} > 0. \tag{A.17}$$

Next, I turn to the competitive setting. Based on equation (A.16), it is clear that $\frac{\partial t_i^{**}}{\partial p_i}$ yields:

$$\frac{\partial t_i^{**}}{\partial p_i} = \frac{-1}{p_i(1-p_i)\Big((\lambda_H^1 - \lambda_L^1) + (\lambda_H^2 - \lambda_L^2)\Big)} < 0 \text{ and } \frac{\partial t_i^{**}}{\partial q_i} = \frac{1}{q_i(1-q_i)\Big((\lambda_H^1 - \lambda_L^1) + (\lambda_H^2 - \lambda_L^2)\Big)} > 0. \tag{A.18}$$

Intuitively, this implies that the amount of time firm $i$ spends in the R&D race is increasing in the likelihood that a market opportunity has a high success rate ($p_i$ low, $q_i$ high). This, in turn, increase the likelihood that it invests in a clinical trial (i.e., $t_1^* \geq \bar{t}$) (in a setting with one firm) or $t_1^{**} \geq \bar{t}$ (in a competitive setting).

From equations (A.6) and (A.10), it is clear to see that a decrease in the likelihood that a market opportunity has a low success rate, increases the amount of time a firm spends in the R&D race. The size of this decrease is equal to $|p_M - p_i| = u_i \times b$ for a given market opportunity. This implies that a stronger public mapping signal (high $b$), decreases the likelihood of a low success rate.

***Proof of Corollary 1:***

By comparing equations (A.6) and (A.10), it is clear that $t_i^* > t_i^{**}$. By comparing equations (A.16) and (A.18), it is clear that $\left|\frac{\partial t_i^*}{\partial p_i}\right| > \left|\frac{\partial t_i^{**}}{\partial p_i}\right|$.

**Proposition 2.** *Assume that after public mapping information is released, firm 1's has a greater opportunity to leverage the information for quickly producing a commercially viable treatment. This increases firm 1's relative likelihood of success (i.e., $\lambda_H^1 > \lambda_H^2$ and $\lambda_L^1 > \lambda_L^2$). If the increase in firm 1's expectation of success is sufficiently large, it may cause firm 1 to spend more time in the race, thereby increasing its likelihood of investing in a clinical trial (i.e., $t_1^{**} > t_2^{**}$ and $\frac{\partial t_i^{**}}{\partial \lambda_L^i} > 0$ and $\frac{\partial t_i^{**}}{\partial \lambda_H^i} > 0$).*

While this proposition relies on strong assumptions, its objective is simply to show that public mapping information may have heterogeneous effects on firms in an R&D race, potentially generating a gap between rivals. In particular, if the direct effect of a public mapping signal sufficiently increases $\lambda_L^1$ or $\lambda_H^1$, then firm 1 is expected to successfully commercialize a product first. In this case, firm 1 will stay longer in the R&D race. Empirically, this suggests that firms with a larger opportunity to leverage the public mapping information will constitute a disproportionate share of any positive response to mapping signals.

What is the effect on the newly lagging firm 2 and total private research investments from both firm 1 and firm 2? Holding all else equal, firm 2's decision to stay in the R&D race depends on two factors: the size of the resulting gap between the two firms and the stochastic nature of success. If the gap between firm 1 and firm 2 is sufficiently large, firm 2 may drop out of the race, potentially resulting in a decrease in overall private research investments in the market opportunity (Harris and Vickers, 1985). However, firm 2 might still opt to stay in the race due to the uncertain nature of R&D, which could still enable it to succeed. This could result in a net increase in private research investments. Ultimately, my empirical work will shed light on how shifts in relative firm opportunity to leverage public mapping information affects the payoffs of newly lagging firms.

***Proof of Proposition 2:***

The impact of increasing the relative opportunity of firm 1 leads to two opposing effects on firm 1's research incentives. On the one hand, it is clear from equation (A.6) that an increase in $\lambda_L^1$ and $\lambda_H^1$ directly raises firm 1's post-mapping expectation that it will succeed and win. For example, recall that firm $i$'s likelihood of success is $\lambda(p(t_i)) = p(t_i)\lambda_L^1 + (1 - p(t_i))\lambda_H^1$. As a result, an increase in $\lambda_H^1$ raises firm 1's post-mapping expectation that it will succeed and win (with a rate proportional to the likelihood that it is a high-success market opportunity, $1 - p(t_i)$). This positive effect can encourage firm 1 to stay longer in the R&D race.

On the other hand, recall from the discussion in Section Appendix A that firm $i$ learns from its own research experience and becomes pessimistic as time passes without success (at a rate that is a function of $(\lambda_H^i - \lambda_L^i)$). When the direct effect if $\lambda_H^1$ is sufficiently large (i.e., $1 - p(t)$ is sufficiently high), the direct positive effect dominates the negative learning effect.[vi]

I state this formally as follows. Based on equation (A.10), the following hold:

1. *An increase in $\lambda_L^1$ leads to an increase in the time firm 1 spends in the R&D race. In other words, $\frac{\partial t_i^{**}}{\lambda_L^i} > 0$.*
2. *If the probability of a high-success market opportunity (i.e., $1-p(t_i)$) is sufficiently large, then firm 1 increases the amount of time spent in the R&D race, thus increasing the likelihood that it will invest in a clinical trial (i.e., $t_1^{**} \geq \bar{t}$). Further, firm 1 will drop out after firm 2 (i.e., $t_1^{**} > t_2^{**}$). In other words, $\frac{\partial t_i^{**}}{\lambda_H^i} > 0$.*
3. *If the probability of a high-success market opportunity (i.e., $1-p(t)$) is sufficiently small, then firm 1 reduces the amount of time spent in the R&D race thus decreasing the likelihood that it will invest in a clinical trial (i.e., $t_1^{**} < \bar{t}$). Further, firm 1 will drop out before firm 2 (i.e., $t_1^{**} < t_2^{**}$). In other words, $\frac{\partial t_i^{**}}{\lambda_H^i} < 0$.*

From Equation (A.10), it is straightforward to see that an increase in $\lambda_L^i$ leads to an increase in $t_i^{**}$. As a result, this section focuses on the impact of increasing $\lambda_H^i$. To sign $\frac{\partial t_i^{**}}{\lambda_H^i}$, recall that $t_i^{**}$ is defined by $\lambda(p(t_i^{**}))V = \lambda(t_i^{**})V = p(t_i^{**})\lambda_L^i + (1 - p(t_i^{**}))\lambda_H^i = c$. I totally differentiate this identity with respect to $\lambda_H^i$ and $t_i^*$:

$$\frac{\partial p(t_i^{**})}{\partial \lambda_H^i}\lambda_L^i d\lambda_H^i + \frac{\partial p(t_i^{**})}{\partial t_i^{**}}\lambda_L^i dt_i^{**} + (1 - p(t_i^{**}))d\lambda_H^i - \lambda_H^i\Big(\frac{\partial p(t_i^{**})}{\partial \lambda_H^i}d\lambda_H^i + \frac{\partial p(t_i^{**})}{\partial t_i^{**}}dt_i^{**}\Big) = 0. \quad \text{(A.19)}$$

Rearranging terms, yields

$$(\lambda_H^i - \lambda_L^i)\frac{\partial p(t_i^{**})}{\partial t_i^*}dt_i^{**} = \Big[(1 - p(t_i^{**})) - (\lambda_H^i - \lambda_L^i)\frac{\partial p(t_i^{**})}{\partial \lambda_H^i}\Big]d\lambda_H^i \quad \text{(A.20)}$$

$$\frac{\partial t_i^{**}}{\lambda_H^i} = \frac{(1 - p(t_i^{**})) - (\lambda_H^i - \lambda_L^i)\frac{\partial p(t_i^{**})}{\partial \lambda_H^i}}{(\lambda_H^i - \lambda_L^i)\frac{\partial p(t_i^{**})}{\partial t_i^{**}}}.$$

[vi] Under specific conditions, it is possible that an increase in $\lambda_H^1$ can cause firm 1 to drop out of the R&D race.

Since the denominator is always positive (as shown above, $\frac{\partial p(t_i)}{\partial t_i} > 0$), I focus on determining the sign of the numerator. The numerator can be decomposed into two effects:

- A positive direct effect $(1 - p(t_i))$: The direct effect of $\lambda_H^1$ on firm $i$'s research investment is always positive: an increase in $\lambda_H^1$ is directly associated with a proportional change in $\lambda^1(t_i)$ (at rate that is proportional to the likelihood of a high-success market opportunity, $1 - p(t_i)$). This positive effect can cause firm 1 to stay longer in the R&D race.
- A negative information effect $(\lambda_H^i - \lambda_L^i)$: The information effect of $\lambda_H^1$ on firm $i$'s research investment is always negative: an increase in $\lambda_H^i$ allows firm $i$ to adjust its expectations more quickly such that firm $i$ becomes more pessimistic as time passes without success. This is particularly salient if $\lambda_H^i$ is relatively high (i.e., the gap between $\lambda_H^i$ and $\lambda_L^i$ is relatively large) as firm $i$ will update more quickly. If $\lambda_H^i$ is relatively high, then lack of success early in the race is an event of very low probability, suggesting that the true success rate is $\lambda_L^i$. This negative effect can cause firm 1 to drop out of the R&D race.

The fact that $t_i^{**}$ can be decreasing in $\lambda_H^i$ is counterintuitive. As a result, I draw upon the example provided by Choi (1991). For simplicity, I focus on the case with one firm:

Suppose that in each period, a firm can invest in a market opportunity. In each period, the probability $\pi$ of success is independent and identical across periods—i.e., the probability of developing a commercially viable treatment takes on a Bernoulli distribution with the parameter $\pi$. Firm $i$ has imperfect information about the success rate of a market opportunity, which can take on two values: $\pi_L$ (with probability $p_i$) or $\pi_H > \pi_L$ (with probability $1 - p_i$). As a result, $\pi = p_i\pi_L + (1 - p_i)\pi_H$. Suppose that $c = 8, V = 100$, and $p = 1/2$. I consider two scenarios: a "low" high-success rate scenario and a "high" high-success rate scenario.

Looking first at the "low" high-success rate scenario, suppose that $\pi_L = 0$ and $\pi_H = 1/4$. In the first period, the expected success rate is $\pi_1 = p_i\pi_L + (1 - p_i)\pi_H = (1/2)(0) + (1/2)(1/4) = 1/8$. The expected payoff is then $\pi_1 V - c = (1/8)100 - 8 > 0$. As a result, the firm will invest in the first period.

If the firm is unsuccessful in the first period, the firm will update its prior on the likelihood of success associated with the market opportunity: the updated likelihood that $\pi = \pi_L$ becomes $4/7$. With this posterior probability, the expected success rate in the next period is $\pi_2 = p_i\pi_L + (1 - p_i)\pi_H = (4/7)(0) + (3/7)(1/4) = 3/28$. The expected payoff in the next period is then $\pi_2 V - c = (3/28)100 - 8 > 0$. As a result, the firm will continue in the R&D race and proceed to the next period.

Now, consider the "high" high success rate scenario where $\pi_L = 0$ and $\pi_H = 10/11$. In the first period, the expected success rate is $\pi_1 = p_i\pi_L + (1 - p_i)\pi_H = (1/2)(0) + (1/2)(10/11) = 5/11$. Since the expected payoff $(\pi_1 V - c = (5/11)100 - 8)$ is greater than zero, the firm will invest in the first period.

If the firm is unsuccessful in the first period, the firm will update its prior on the likelihood of success associated with the market opportunity: the updated likelihood that $\pi = \pi_L$ becomes $11/12$. With this posterior probability, the expected success rate in the next period is $\pi_2 = p\pi_L + (1 - p_i)\pi_H = (11/12)(0) + (1/12)(4/5) = 1/15$. The expected payoff in the next period is then $\pi_2 V - c = (1/15)100 - 8 = 100/15 - 8 < 0$. As a result, the firm will not proceed to the next period and instead cease its research investment. This example illustrates how in light

of uncertainty about the potential of a market opportunity, substantial differences in potential success rates (i.e., a relatively high $\pi_H$) can allow firms to learn quickly, ceasing investment quickly following initial failures.

Taken together, the direction of $\frac{\partial t_i^*}{\lambda_H^i}$ depends on the relative size of the positive direct effect and the negative information effect.

**Proposition 3.** *Assume that a firm with sufficiently high prior beliefs that a market opportunity has a low success rate. The impact of public mapping information is increasing in the firm's ex ante level of uncertainty (i.e.,* $\frac{\partial^2 t_i^*}{\partial p_i \partial u_i} < 0$ *and* $\frac{\partial^2 t_i^{**}}{\partial p_i \partial u_i} < 0$*).*

This is motivated by the idea public mapping information is most useful for shaping firms' research investments when they have high levels of uncertainty due to limited existing information (e.g., due limited resources to conduct internal mapping efforts).

***Proof of Proposition 3:***

For simplicity, I focus on the case where there is one firm. I establish the relationship between the impact of public mapping information and uncertainty by determining the sign of $\frac{\partial^2 t_i^*}{\partial p_i \partial u_i}$.

Recall, $p_i = u_i Pub_i + (1 - u_i) Exist_i$, where $Pub_i$ is the novel public information signal and $Exist_i$ is the prior existing information signal. Then:

$$\begin{aligned} \frac{\partial t_i^*}{\partial p_i} &= \frac{-1}{p_i(1-p_i)(\lambda_H^i - \lambda_L^i)} \\ &= \frac{-1}{\big(u_i Pub_i + (1-u_i)Exist_i\big)\big(1 - (u_i Pub_i + (1-u_i)Exist_i)(\lambda_H^i - \lambda_L^i)\big)}. \end{aligned} \tag{A.21}$$

Taking the derivative of $\frac{\partial t_i^*}{\partial p_i}$ with respect to $u_i$, and applying the quotient rule yields:

$$\frac{\partial^2 t_i^*}{\partial p_i \partial u_i} = \frac{-(Pub_i - Exist_i)(-1 + 2 \times Exist_i \times Pub_i - 2(-1+u_i)Exist_i)}{(\lambda_H^i - \lambda_L^i)(u_i Pub_i + (1-u_i)Exist_i)^2 (1 - (u_i Pub_i + (1-u_i)Exist_i))^2}. \tag{A.22}$$

The denominator in equation (A.22) is positive since $\lambda_H^i - \lambda_L^i > 0$. Hence, I can focus on the sign of the numerator. I begin by simplifying the numerator to $-(Pub_i - Exist_i)(-1 + 2 \times Exist_i \times Pub_i - 2(-1+u_i)Exist_i) = -(Pub_i - Exist_i)(-1 + 2p_i)$.

For $\frac{\partial^2 t_i^*}{\partial p_i \partial u_i} < 0$ to hold, then it must be the case that $-(Pub_i - Exist_i)(-1 + 2p_i) > 0$. For this to be the case, one of the following conditions must hold:

(a) $Pub_i > Exist_i$ and $p_i > \frac{1}{2}$, i.e., the public information signal (which includes public mapping information) is sufficiently strong and a firm's prior beliefs that the likelihood of a low success rate market opportunity is sufficiently high.

(b) $Pub_i < Exist_i$ and $(1 - p_i) > \frac{1}{2}$, i.e., the prior existing information signal (which includes internal mapping information) is sufficiently strong and a firm's prior beliefs that the likelihood of a high success rate market opportunity is sufficiently high.

# Appendix B Additional figures and tables

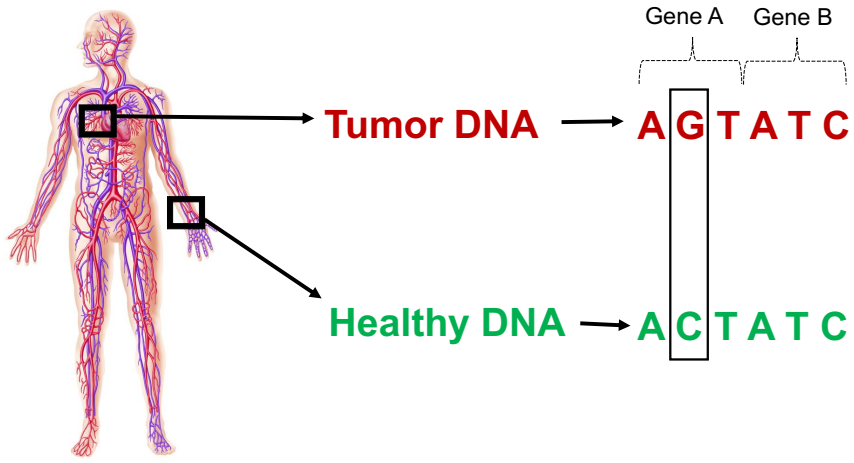


FIGURE B1
Overview of scientific background on cancer genome sequencing

*Notes:* This figure summarizes the scientific background described in Section 3. An individual's genome is the complete set of DNA found in each cell. DNA is comprised of a unique sequence of four bases: adenine (A), cytosine (C), guanine (G), and thymine (T). A gene is a segment of DNA that provides instructions for unique traits. Cancer is caused by a change in the sequence of DNA bases (i.e., a mutation). Cancer genome researchers aim to identify the mutations that drive the development and growth of cancer by comparing the DNA sequences of cancer cells (in red) to those of normal tissue (in green). This figure is a modified version of Figure 1 found in Samuel and Hudson (2013).

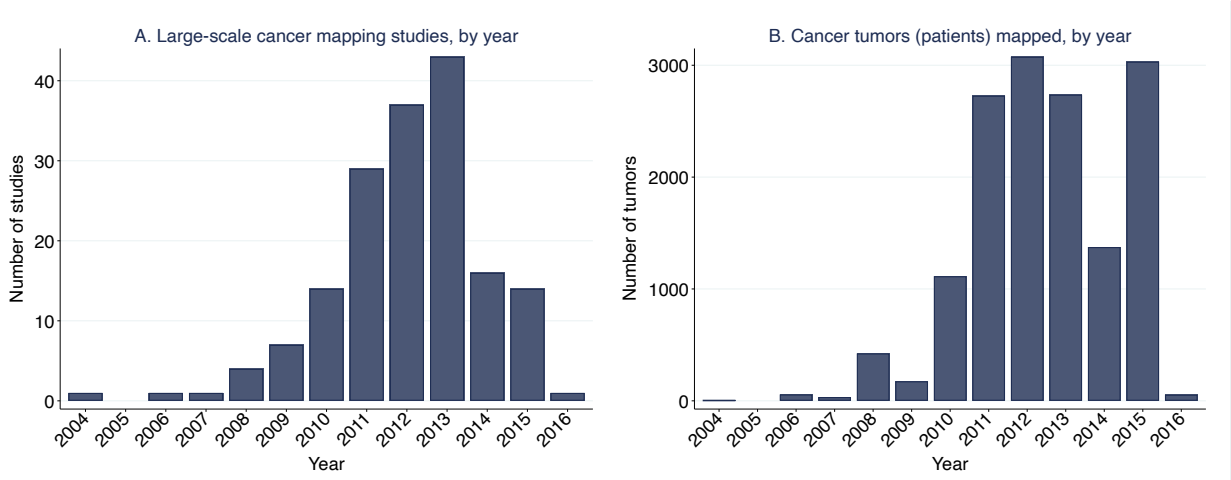


FIGURE B2
Large-scale public cancer mapping studies and mapped tumors by year, 2004–2016

*Notes:* The $x$-axis in Panel A indicates the year in which large-scale public mapping studies were submitted to the publishing journal. All studies were published in a top 25 genetics journal, based on journal rankings between 1999 and 2004. The increase in mapped tumors in 2015 is driven by a single study that sequenced 1,144 lung cancer tumors and was submitted to *Nature Genetics* (Campbell et al., 2016). For details on the construction of the cancer mapping studies sample used in this paper, see online Appendix C.

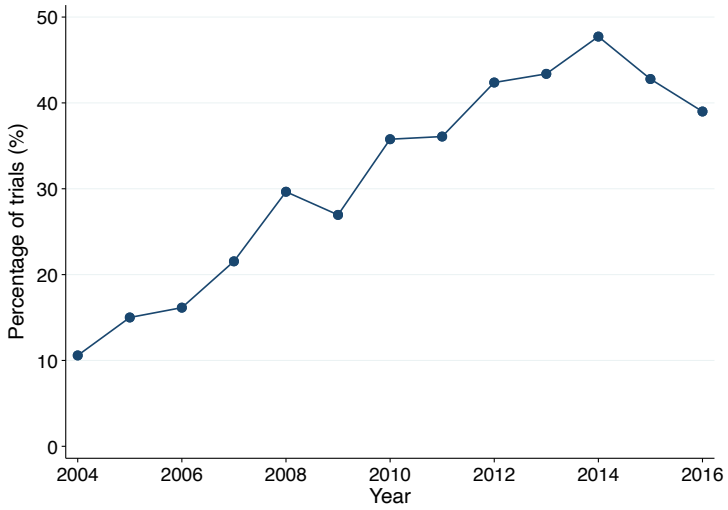


FIGURE B3
Share of private sector phase II cancer trials that enrolled patients based on genes, 2004-2006

*Notes:* This figure plots the percentage of private sector phase II clinical trials in 2004-2016 that were gene-related—i.e., genetic criteria were used to select patients for enrollment. Observations are at the trial-cancer level.

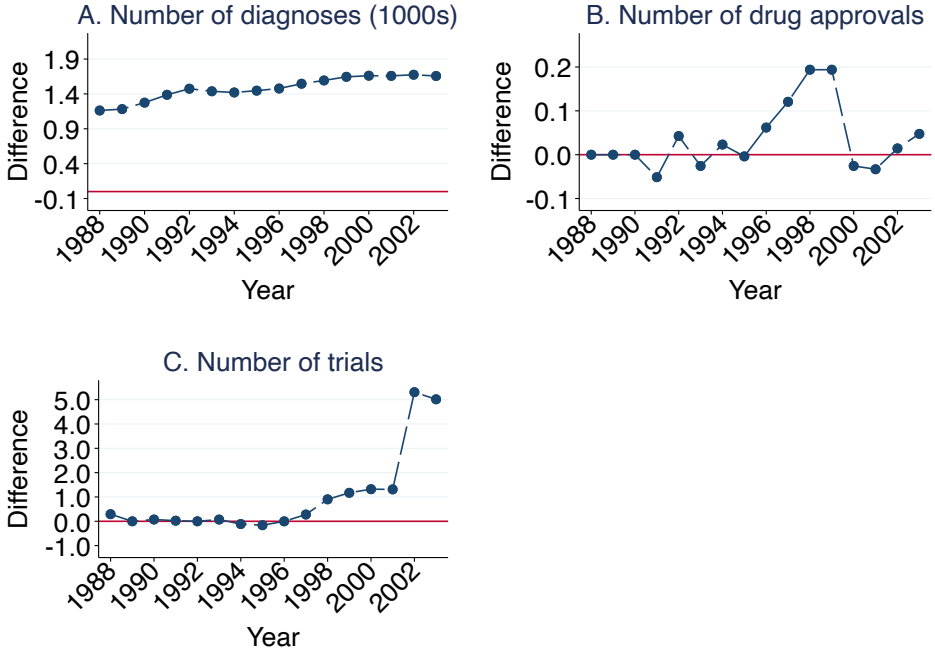


FIGURE B4
Examining cancer-level selection, 1988-2003

*Notes:* This figure examines baseline differences between cancers that were first sequenced relatively early (before 2011) and cancers that were first sequenced relatively late (in/after 2011). For each panel, difference in means of the outcome variable is calculated between the two cancer groups in each year from 1988 (the earliest year in which data for all three outcomes are available) to 2003. For simplicity, Panel C focuses on the number of phase II clinical trials.

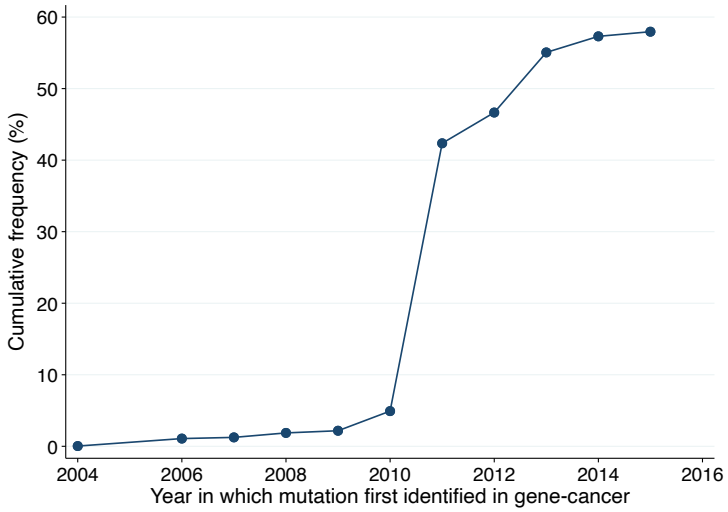


FIGURE B5
Cumulative share of gene-cancer pairs with
mutations identified by large-scale cancer mapping studies, 2004–2016

*Notes:* This figure plots the cumulative share of gene-cancer pairs with mutations identified by large-scale cancer mapping studies. As discussed in Section 3, there are 30,223 gene-cancer pairs possible. See Section 3 and online Appendix C for more detailed data and variable descriptions.

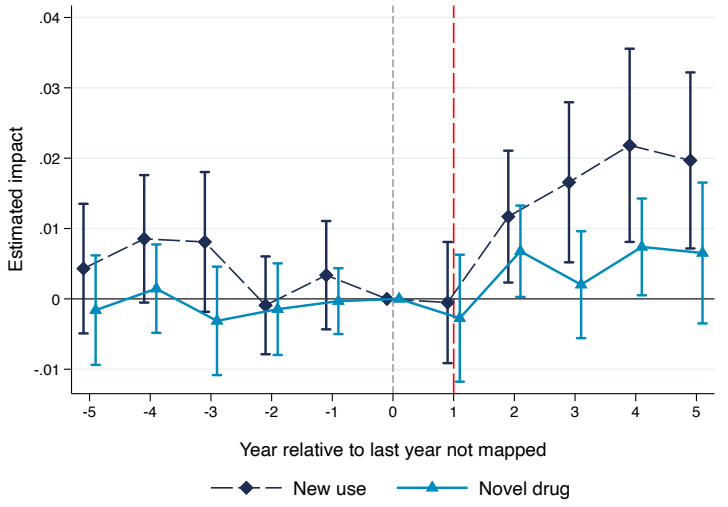


FIGURE B6
Event study estimates: Effect on private sector phase II trials
of new uses vs. novel drugs, 2004-2016

*Notes:* This figure plots the response of private sector phase II trials testing new uses of drugs approved in the focal gene or previously tested in any gene-cancer pair (dark blue line) and trials testing novel drugs (i.e., not previously approved in the focal gene or tested the gene-cancer pair; light blue line). Each dot corresponds to coefficients based on estimates of equation (2). The dashed red line indicates the first year that a mutation in a gene-cancer pair is publicly disclosed by a cancer mapping study. Shown are 95% confidence intervals. This specification is based on gene-cancer-year level observations. The figure corresponds to a dynamic version of the specification in Table 4, though coefficients are estimates from OLS models and standard errors are clustered at the gene and cancer level. See Section 3 and online Appendix C for more detailed data and variable descriptions.

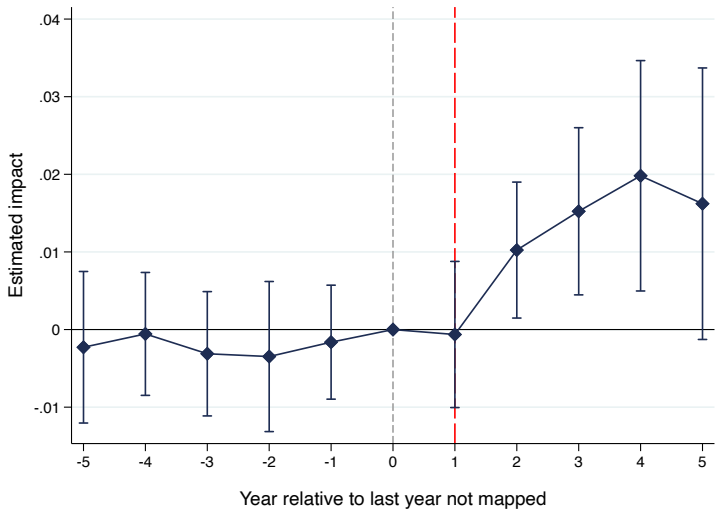


FIGURE B7

Effect on private sector phase II trials with non-missing interventions, 2004–2016

*Notes:* This figure plots the response of private sector phase II trials following the public release of cancer mapping information using the sample of clinical trials with non-missing intervention data. Each dot corresponds to coefficients based on estimates of equation (2). The outcome variable is a binary indicator for whether there is any private sector phase II clinical trial. On the $x$-axis are years $z$ relative to a "zero" relative year that marks the last year the gene-cancer was not known to be mutated based on cancer mapping studies. The dashed red line indicates the first year that a mutation in a gene-cancer pair is publicly disclosed by such a study. Shown are 95% confidence intervals (corresponding to robust standard errors, clustered at the gene and cancer level). This specification is based on gene-cancer-year level observations, and the coefficients are estimates from OLS models. The figure corresponds to a dynamic version of the specification in online Appendix Table B4, column 2. See Section 3 and online Appendix C for more detailed data and variable descriptions.

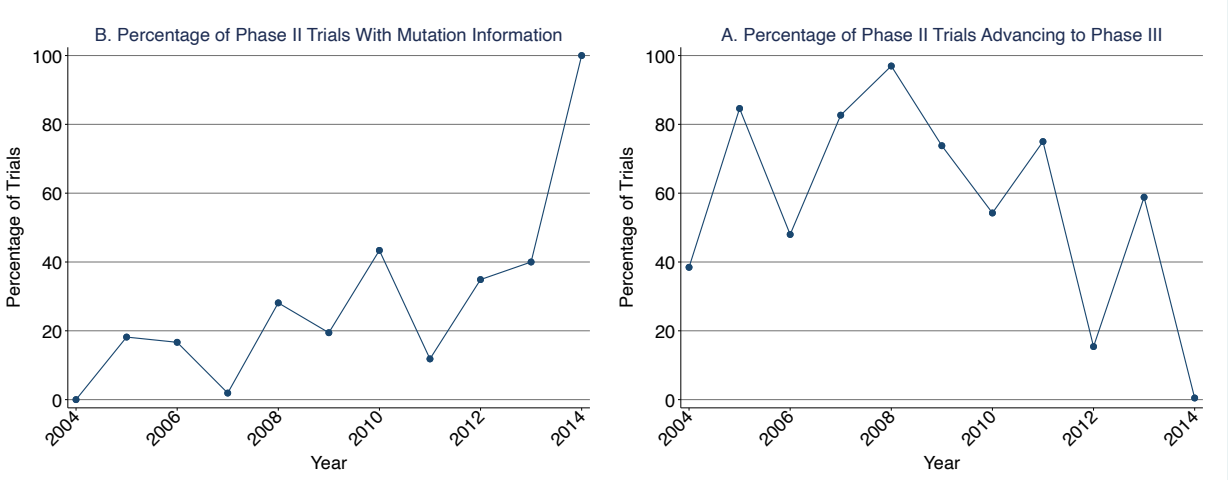


FIGURE B8

Private sector clinical trial advancement rates by year, 2004-2014

*Notes:* Panel A plots the percentage of private sector phase II trials that successfully advanced to phase III. Panel B plots the percentage of private sector phase II clinical trials initiated in gene-cancer pairs with mutation information, as a share of the total number of such trials that successfully advanced to phase III. Here, trials are classified as having successfully advanced to phase III if they transitioned to phase III within 4 years of the phase II trial start date. The share of advanced phase II trials initiated in gene-cancer pairs with mutation information increases significantly in 2011–2013, which as online Appendix Figure B2 shows, is a period in which mutation information was disclosed for a large share of gene-cancers. The sample includes all phase II trials completed or terminated as of July 14, 2017. Observations are at the trial-gene-cancer level.

TABLE B1
Overview of gene-cancer-year panel, 2004–2016

| | Count |
|---|---|
| No. of gene-cancer pairs (e.g., BRCA2-prostate) | 50,160 |
| No. of gene-cancer pairs, coappearing in at least one pub as of 2004 | 30,223 |
| No. of genes (e.g., BRCA1, BRCA2) | 462 |
| No. of cancers (e.g., ovarian, small intestine) | 80 |
| No. of cancer groups (e.g., digestive) | 19 |
| No. of years (2004 to 2016) | 13 |
| **Final Panel: No. of gene-cancer-year observations** | **392,899** |

*Notes:* This table provides an overview of how the gene-cancer-year panel was constructed. See online Appendix C for more details.

TABLE B2
Poisson estimates of public mapping effect on phase II clinical trials, 2004–2016

| | Dependent variable: Number of phase II trials | | |
|---|---|---|---|
| | Any trial | Any private sector trial | Any public sector trial |
| | (1) | (2) | (3) |
| Post × DisclGeneCancer | 0.0678 | 0.283** | -0.215 |
| | (0.110) | (0.117) | (0.148) |
| Change in number of trials (%) | 7.014 | 32.77 | -19.38 |
| Gene-cancer FEs | Yes | Yes | Yes |
| Cancer × Year FEs | Yes | Yes | Yes |
| Observations | 48,311 | 35,936 | 34,071 |

*Notes:* This table reports DID estimates of the effect of public cancer mapping information on phase II trials, but obtained from Poisson pseudo maximum likelihood estimates. The level of observation is the gene-cancer-year. *Post × DisclGeneCancer* switches from 0 to 1 when a mutation in a gene-cancer pair is publicly disclosed by a mapping study. Controls include gene-cancer fixed effects and cancer-year fixed effects. Robust standard errors, clustered at the gene and cancer level, are shown in parentheses. *Change in number of trials (%)* refers to the elasticities, which are calculated by exponentiating the coefficients and differencing one. For example, the estimates in column (2) imply that there is on average a statistically significant 32.77% (=exp[0.283]-1) yearly increase in phase II trials after mapping information is disclosed. Singleton observations and observations that are separated by a fixed effect are dropped, which accounts for the smaller number of observations relative to Table 2. See Section 3 and online Appendix C for more detailed data and variable descriptions.
*$p$ <0.10, **$p$ <0.05, ***$p$ <0.01.

TABLE B3
Effect of public cancer mapping information in the same gene, related cancer, 2004–2016

| | Dependent Variable: Any private sector phase II trial |
|---|---|
| Post × DisclGeneCancerGroup | 0.00343 |
| | (0.00304) |
| Mean of dep. var. | 0.012 |
| Change in likelihood of trial (%) | 29.29 |
| Gene-cancer FEs | Yes |
| Cancer × Year FEs | Yes |
| Observations | 165,204 |

*Notes:* This table reports DID estimates of how private sector phase II trials in a gene-cancer pair respond to public cancer mapping information in the same gene and a different cancer. The level of observation is the gene-cancer-year. In order to limit the control group of gene-cancer pairs to those without any mutation information, this sample excludes pairs where mutation information was directly revealed by a mapping study. Estimates are from OLS models. *Post* × *DisclGeneCancerGroup* switches from 0 to 1 when a mutation in a same gene and different but related cancer is publicly disclosed by a cancer mapping study. Cancers are classified as related if they are in the same cancer site group, based on SEER classification. To illustrate, small intestine and large intestine cancer are both in the same cancer site group ("digestive system"). Robust standard errors, clustered at the gene and cancer level, are shown in parentheses. *Mean of dep. var.* is the mean of the outcome variable in a gene-cancer pair before the first disclosure of a mutation in the same gene and related cancer and is used to calculate the percentage change in the likelihood of a clinical trial that follows the disclosure of a mutation in the same gene and related cancer. See Section 3 and online Appendix C for more detailed data and variable descriptions.
$^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

TABLE B4
Effect on private sector clinical trials with non-missing interventions, 2004–2016

| | Dependent variable: Any phase II trial | | |
|---|---|---|---|
| | Any trial | Any private sector trial | Any public sector trial |
| | (1) | (2) | (3) |
| Post × DisclGeneCancer | 0.0126*** | 0.0104** | 0.00399 |
| | (0.00368) | (0.00315) | (0.00346) |
| Mean of dep. var. | 0.023 | 0.014 | 0.013 |
| Change in likelihood of trial (%) | 54.49 | 75.65 | 31.35 |
| Gene-cancer FEs | Yes | Yes | Yes |
| Cancer × Year FEs | Yes | Yes | Yes |
| Observations | 392,899 | 392,899 | 392,899 |

*Notes:* This table reports DID estimates of the effect of public cancer mapping information on phase II trials using the subset of clinical trials that have non-missing intervention data. The level of observation is the gene-cancer-year. Estimates are from OLS models. The outcome variable switches from 0 to 1 if a phase II trial is reported in a gene-cancer-year. *Post* × *DisclGeneCancer* switches from 0 to 1 when a mutation in a gene-cancer pair is publicly disclosed by a mapping study. Controls include gene-cancer fixed effects and cancer-year fixed effects. Robust standard errors, clustered at the gene and cancer level, are shown in parentheses. *Mean of dep. var.* is the mean of the outcome variable in a gene-cancer pair before the first disclosure of a mutation and is used to calculate the percentage change in the likelihood of a clinical trial. See Section 3 and online Appendix C for more detailed data and variable descriptions.
$^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

TABLE B5
Effect of public mapping information on private sector phase I trials by drug type, 2004–2016

| | Dependent variable: Any private sector phase I trial | |
|---|---|---|
| | Testing new drug uses (1) | Testing novel drugs (2) |
| Post × DisclGeneCancer | 0.00901*** | 0.00576*** |
| | (0.00108) | (0.00114) |
| Mean of dep. var. | 0.016 | 0.023 |
| Change in likelihood of trial (%) | 56.35 | 25.47 |
| Gene-cancer FEs | Yes | Yes |
| Cancer × Year FEs | Yes | Yes |
| Observations | 392,899 | 392,899 |
| Diff. Wald test $p$-value | 0.03 | |

*Notes:* This table reports DID estimates of the effect of public cancer mapping information on private sector phase I trials by drug type. To examine heterogeneity by incumbents and entrants, column 1 estimates the effect on trials testing new uses (i.e., of drugs approved in the focal gene or previously tested in any gene-cancer pair); column 2, the effect on trials of novel drugs. The level of observation is the gene-cancer-year. *Post × DisclGeneCancer* switches from 0 to 1 when a mutation in a gene-cancer pair is publicly disclosed by a mapping study. Controls include gene-cancer fixed effects and cancer-year fixed effects. *Mean of dep. var.* is the mean of the outcome variable in a gene-cancer pair before the first disclosure of a mutation and is used to calculate the percentage change in the likelihood of a clinical trial. The estimates in this table are from seemingly unrelated models, which permits a comparison of the coefficient on *Post × DisclGeneCancer* across models. Standard errors are clustered at the cancer level (see main text footnote 55). The $p$-value is from a Wald test that compares the differences in the coefficients on *Post × DisclGeneCancer*. For sample details, see the text and online Appendix C. *$p$ <0.10, **$p$ <0.05, ***$p$ <0.01.

TABLE B6
Summary statistics: Trial-gene-cancer level data, 2004–2016

| | Full (1) | Trials With Info (2) | Trials With No Info (3) | Difference (2)−(3) (4) |
|---|---|---|---|---|
| Positive clinical outcome | 0.33 | 0.43 | 0.30 | 0.13*** |
| 1(Advance to Phase III) | 0.55 | 0.48 | 0.57 | -0.09*** |
| 1(Advance to Phase III, Within 4 Years) | 0.55 | 0.48 | 0.57 | -0.08*** |
| Private sector trial | 0.81 | 0.91 | 0.79 | 0.12*** |
| Trial funded by firm w/ high private mapping information | 0.09 | 0.14 | 0.08 | 0.06*** |

*Notes:* This table shows summary statistics at the trial-gene-cancer level. The sample includes all 1,785 trial-gene-cancer observations associated with phase II clinical trials that began between 2004 and 2016, made clinical outcomes data available, and were completed or terminated as of July 14, 2017. The table describes trials initiated in gene-cancer pairs where driver (clinically relevant) mutation information was (column 2) and was not (column 3) publicly available by the end of the trial. *Positive trial outcome* is an indicator for whether the phase II clinical trial satisfies has a positive clinical outcome. *Advance to phase III* is an indicator variable for a phase II clinical trial drug that is subsequently tested in a phase III trial. Similarly, *Advance to phase III, within 4 years* is an indicator variable for a phase II clinical trial's drug that is subsequently tested in a phase III clinical trial within four years of the phase II trial start date. *Trial funded by firm w/ high prior mapping information* is an indicator variable for whether the phase II trial is conducted by a firm with high levels of private sequencing information. For sample details, see the text and online Appendix C. *$p$ <0.10, **$p$ <0.05, ***$p$ <0.01.

TABLE B7
Public cancer mapping information and phase II clinical trial outcomes and advancement rates by firm portfolio size, 2004–2016

| | Dependent variable: Advancing to phase III | | | |
|---|---|---|---|---|
| | Firm portfolio of phase II trials | | Firm portfolio of phase IIII trials | |
| | Small | Large | Small | Large |
| | (1) | (2) | (3) | (4) |
| Post × DisclGeneCancer | 0.466 | -0.949* | -0.279 | -0.967** |
| | (0.408) | (0.485) | (0.419) | (0.419) |
| Post × DisclGeneCancer | 0.112 | 0.917 | 0.565 | 1.430* |
| × PositiveOutcome | (0.584) | (0.596) | (0.492) | (0.732) |
| PositiveOutcome | 0.0878 | 1.570*** | 0.794*** | 1.724* |
| | (0.441) | (0.487) | (0.297) | (0.978) |
| Change in likelihood of outcome (%) | | | | |
| Mapping info | | | | |
| Positive trial outcome | 9.18 | 380.59 | 121.27 | 460.8 |
| Mapping info, positive trial outcome | 4.71 | 365.76 | 194.48 | 790.78 |
| Mapping info, negative trial outcome | 59.42 | -61.27 | -24.37 | -62.00 |
| Linear year trend | Yes | Yes | Yes | Yes |
| Trial characteristics | No | Yes | No | Yes |
| Nb. Trial-Gene-Cancers | 752 | 1,033 | 1,262 | 523 |
| Nb. Trials | 96 | 87 | 134 | 49 |
| Nb. Genes | 80 | 74 | 77 | 57 |
| Nb. Cancers | 75 | 54 | 84 | 38 |

*Notes:* This table shows the relationship between public cancer mapping information and phase II-to-phase III advancement rates by firm portfolio size. To examine heterogeneity across firms, the table shows the effect on trials of firms with a small phase II portfolio (column 1) and high a large phase II portfolio (column 2). Column 3 shows the effect on trials of firms with a small phase III portfolio; column 4, the effect on trials of firms with a large phase III portfolio. A firm is considered as having a small portfolio if it has a below median number of yearly trials initiated in the same focal cancer in the past four years. The level of observation is the trial-gene-cancer. The sample includes all phase II trial-gene-cancer observations associated with phase II clinical trials that began between 2004 and 2016, made clinical outcomes data available, and were completed or terminated as of July 14, 2017. Estimates are from Cox proportional hazard models (stratified by cancer) and examine the relationship between public cancer mapping information and phase II-to-phase III transition rates. *Post × DisclGeneCancer* is an indicator for the disclosure of a driver (clinically relevant) mutation in a gene-cancer by the end of the clinical trial. *PositiveOutcome* is an indicator for whether the trial satisfies any of the following: treatment group has objective response rate above the 75th percentile in the cancer-specific distribution, demonstrates a demonstrates a statistically significant ($p < 0.05$) improvement in overall survival relative to the control group or a historical control, or demonstrates a statistically significant ($p < 0.05$) improvement in progression-free survival relative to the control group or a historical control. *Change in likelihood of outcome (%)* refers to the change in the likelihood of a positive trial outcome (columns 1 and 2) or of advancing to phase III (columns 3 and 4). *Trial characteristics* refers to controls for whether the trial is a private-sector trial and whether the trial is funded by a firm with high levels of prior mapping information. Robust standard errors, clustered at the gene and cancer level, are shown in parentheses. For sample details, see the text and online Appendix C. *$p$ <0.10, **$p$ <0.05, ***$p$ <0.01.

# Appendix C Data description

This online appendix provides additional detail on the datasets used in this analysis.

## C.1 Linking publications to gene-cancer pairs

I collect data on the scientific publications related to each gene-cancer pair from two sources: the National Library of Medicine (NLM) National Center for Biotechnology Information Gene database and the Online Mendelian Inheritance in Man (OMIM) database. The NLM gene database is a repository of gene-related information that is presented in individual gene records. As described on the website, the database "integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide."[vii] Similarly, the OMIM database is a comprehensive catalog of human genes and genetic phenotypes.[viii] Each NLM and OMIM gene record cites scientific publications that are relevant to the gene, which I collect.

Next, I link each gene publication to the set of relevant cancer sites. The NLM maintains a comprehensive dictionary of scientific terms called Medical Scientific Subject Headings (MeSH) and assigns each publication to the relevant MeSH terms. To characterize the cancer site associated with each gene publication, I obtain the list of relevant MeSH terms and map them onto the set of 80 cancer sites, based on the standard Surveillance, Epidemiology, and End Results (SEER) classification system.

Table C1
Gene-cancer pair restrictions

| | Dependent variable: any private sector phase II trial | | | | | |
|---|---|---|---|---|---|---|
| Pub. database: | NLM | | | OMIM | | |
| Pub. count threshold: | ≥ 2 pubs | ≥ 3 pubs | ≥ 4 pubs | ≥ 2 pubs | ≥ 3 pubs | ≥ 4 pubs |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Post × DisclGeneCancer | 0.00943** | 0.00780** | 0.00899** | 0.0221*** | 0.0280*** | 0.0323*** |
| | (0.00319) | (0.00353) | (0.00387) | (0.00566) | (0.00720) | (0.00897) |
| Mean of dep. var. | 0.014 | 0.014 | .0145 | 0.035 | 0.045 | 0.0511 |
| Change in likelihood of trial (%) | 65.73 | 55.85 | 61.93 | 63.40 | 61.78 | 63.27 |
| Gene-cancer FEs | Yes | Yes | Yes | Yes | Yes | Yes |
| Cancer × Year FEs | Yes | Yes | Yes | Yes | Yes | Yes |
| No. gene-cancer pairs | 30,223 | 25,632 | 22,737 | 20,736 | 15,257 | 11,831 |
| No. gene-cancer-years | 392,899 | 333,216 | 295,581 | 269,568 | 198,341 | 153,803 |

*Notes:* This table reports DID estimates of the effect of public cancer mapping information on private sector phase II trials, using different gene-cancer pair samples. The level of observation is the gene-cancer-year. Estimates are from OLS models. The table shows the relationship between public cancer mapping information and private sector phase II clinical trials restricting the sample of gene-cancer pairs to those with publications in the NLM National Center for Biotechnology Information Gene database (columns 1–3) and in the OMIM database (columns 4–6). The outcome variable switches from 0 to 1 if a clinical trial is reported in a gene-cancer-year. *Post* × *DisclGeneCancer* switches from 0 to 1 when a mutation in a gene-cancer pair is publicly disclosed by a mapping study. Robust standard errors, clustered at the gene and cancer level, are shown in parentheses. *Mean of dep. var.* is the mean of the outcome variable in a gene-cancer pair before the first disclosure of a mutation and is used to calculate the percentage change in the likelihood of a clinical trial. For sample details, see the text and online Appendix C. *$p$ <0.10, **$p$ <0.05, ***$p$ <0.01.

[vii] For more details, see https://www.ncbi.nlm.nih.gov/gene/.

[viii] For more details, see https://www.omim.org/.

## C.2 Public cancer mapping data

### C.2.1 Mapping studies

Cancer mapping data comes from two publicly available data repositories that contain gene-level mutation data from hundreds of published cancer mapping studies: the Catalogue of Somatic Mutations in Cancer (COSMIC) and the cBioPortal for Cancer Genomes (cBioPortal). COSMIC is considered the primary source of information on somatic mutations relating to human cancers.[ix] (As described below, somatic mutations, the focus of this paper, are non-inherited mutations.) I use data from COSMIC Release v82 and v85 (Tate et al. 2018).[x] The second data repository, cBioPortal, was developed at Memorial Sloan Kettering Cancer Center and provides data from large-scale cancer mapping studies (Cerami et al., 2012; Gao et al., 2013).[xi] I use data downloaded on 7 July 2017 and 7 June 2018. I restrict the downloaded set of cancer mapping studies to those that are (i) large-scale as measured by the number of tumors mapped and (ii) high impact.

- **Large-scale cancer mapping studies:** I define a cancer mapping study as "large-scale" if is published in cBioPortal, which a database that focuses on "large-scale cancer genomics projects" (Cerami et al., 2012) or in COSMIC's "Whole Genome & Large-scale Systematic Screens" sequencing study database.[xii]
- **High impact cancer mapping studies:** To identify "high impact" cancer mapping studies, I isolate the list of cancer mapping studies that were published in highly ranked genetics journals from 2004 through 2016. Journal rankings are based on the Scimago Journal & Country Rank (SJR) system, a yearly ranking scheme that ranks journals using a citation-based algorithm.[xiii] The SJR measures a journal's influence by looking at the number of citations it has received over the past three years (Gonzalez-Pereira et al., 2009). I code a journal as being highly ranked if it is listed among the top 25 journals based on the "Genetics" SJR ranking at least once between 1999 (the earliest year SJR rankings are publicly available) and 2004 (the last year in which a mapping study published in a particular journal cannot influence that same journal's ranking).[xiv]

Using these criteria, the final cancer mapping study sample consists of 168 high-quality and large-scale cancer mapping studies. Nearly all (99%) of the studies receive some form of financial support from the public sector (e.g., the NIH).

### C.2.2 Mutation data

I restrict the gene-level data from the 168 cancer mapping studies in several ways. First, I focus on mutations that occur in the protein-coding region of the DNA sequence. Nearly all cancer mapping studies focus primarily on the mutations in protein-coding regions since, relative to mutations in non-protein-coding regions, the linkages between the mutations, altered proteins, and subsequent diseases are easier to interpret (Vogelstein et al., 2013). In particular, I focus on somatic mutations,

[ix] For more details, see https://cancer.sanger.ac.uk/cosmic.

[x] For more details, see https://cancer.sanger.ac.uk/cosmic/download.

[xi] For more details, see http://www.cbioportal.org/.

[xii] For more details, see https://cancer.sanger.ac.uk/cosmic/papers.

[xiii] For more details, see: https://www.scimagojr.com/.

[xiv] Results using journals ranked in the top 25 using the 2017 "Genetics" SJR ranking, the 1999 to 2004 "Medicine" SJR ranking, or the 2017 "Medicine" SJR ranking produce similar results.

which are DNA aberrations that occur after conception and are not inherited.[xv] According to Stratton et al. (2009, p. 721), "All cancers arise as a result of somatically acquired changes in the DNA of cancer cells."

The focus of this paper is primarily on the impact of relatively localized within-gene changes (e.g., substitutions, deletions, and insertions of DNA bases) or the deletion of whole genes. In addition to these changes, cancer mapping studies may characterize other types of genetic alterations that can also contribute to the progression and growth of cancer. These genetic alterations include DNA rearrangements, where DNA is broken and then fused to a DNA segment from another part of the genome; deletions of large parts of the DNA; and amplifications or excess copies of a gene.[xvi]

The final list of COSMIC mutation types includes Complex, Complex–compound substitution; Complex–deletion frame; Complex–frameshift; Complex–insertion inframe; Deletion–In frame; Insertion–frameshift; Nonstop extension; Substitution–Missense; Substitution–Nonsense; Unknown; Whole gene deletion. Similarly, the cBioPortal mutation types include Frame_Shift_Del, Frame_Shift_Ins, In_Frame_Del, In_Frame_Ins, Missense_Mutation, Nonsense_Mutation, Splice_Site, Splice_Region, Nonstop_Mutation, Translation_Start_Site, De_novo_Start_InFrame, De_novo_Start_-OutOfFrame, and Unknown.

### C.3 Clinical trial data

Clinical trials data comes from the Clarivate Cortellis Competitive Intelligence Clinical Trials Database ("Cortellis"), which is continuously updated with clinical trial data from public trial registries such as ClinicalTrials.gov. In addition to important clinical trial variables, such as the start date and phase (which are both non-missing), key clinical trial variables include:

- **Gene information.** The clinical trials in Cortellis often contain information about the criteria used to enroll patients. Approximately 45% of clinical trials in Cortellis provide information on the biomarkers used to guide patient selection. In my analysis, I focus on the set of clinical trials that use genetic biomarkers (e.g., the gene EGFR). I then link each genetic biomarker to the standardized list of genes using the NLM's gene database. This results in a dataset at the trial-gene level.

- **Cancer information.** In Cortellis, approximately 95% of clinical trials contain information on the disease being examined (e.g., prostate cancer, diabetes). I focus on the subset of clinical trials that enroll patients with cancer, which allows me to generate a trial-gene-cancer dataset. This subset accounts for approximately 31% of all trial-gene observations in the dataset.

- **Trial drug intervention.** Approximately 95% of the clinical trials in Cortellis contain information on the types of drugs being tested. To differentiate between trials that are "testing new uses" and those that are "testing novel drugs," I match the trial's drug intervention information to the drug approval dataset described in Section 3.3.2.

- **Trial funder type.** In Cortellis, approximately 83% of the clinical trials contain information on the types of institutions funding the trials, such as whether they are from the private sector, government, academic, etc. To characterize the types of institutions funding the trials, I define a trial as "private sector" trial if it has any private sector funding. Approximately 52% of

[xv] For more details, see: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/somatic-mutation.

[xvi] For more details, see Stratton et al., (2009), Vogelstein et al. (2013), and https://ghr.nlm.nih.gov/primer/mutationsanddisorders/possiblemutations.

clinical trials are classified as a private sector trial. The remaining 48% of clinical trials are classified as public sector trials.

Approximately 27% of clinical trials are jointly conducted by multiple firms. These trials involve not only the sponsoring firm but also at least one collaborating firm. According to ClinicalTrials.gov, "both sponsors and collaborators are considered funders of the study."[xvii] Both sponsors and collaborators can play important roles in the shaping the study; while sponsors initiate the study, collaborators are defined as providing support that "may include activities related to funding, design, implementation, data analysis, or reporting."

It is important to note that while a given clinical trial (testing a drug for a particular disease) may involve just one or two organizations, a drug's overall development profile may involve many more organizations. For example, throughout a drug's lifecycle there may be participation from multiple firms as the as drug is tested in multiple indications.

Among the private sector trials, approximately 19% are jointly conducted with public sector institution. The main private sector results are robust to restricting the sample to the set of clinical trials that are only funded by the private sector. For a given firm testing a trial within a gene-cancer pair, I consider the firm to have relatively high levels of prior mapping information if it has an above the median level number of (i) pre-2004 sequencing publications and (ii) pre-2004 clinical trials in the focal cancer. To account for the varying roles of sponsors and collaborators across trials, I consider all clinical trial associated with a firm, regardless of whether the firm is a trial sponsor or collaborator. This information is then used to classify trials as being funded by a firm with low prior mapping information or high prior mapping information. In cases where multiple firms are associated with a clinical trial, I consider a trial as being funded by a firm with high prior mapping information if it is funded by any firm with high prior mapping information.

Overall, approximately 79% of the clinical trials in Cortellis contain information on both the type of trial drug intervention and funder type.

[xvii] For more information, see https://clinicaltrials.gov/ct2/about-studies/glossary.

# Appendix D Treatment and control gene-cancer pairs

As discussed in Section 4.1, the empirical strategy employed to measure the impact of public information from large-scale cancer mapping studies on the quantity of clinical trials compares gene-cancer pairs *with publicly known mutation information* to all gene-cancer pairs *without publicly known mutation information* at any given point in time. Panel A of Appendix Figure D1 shows how gene-cancer pairs are allocated to treatment and control groups under this empirical strategy (hereafter, the "primary empirical strategy").

One alternative strategy is to compare mapped gene-cancer pairs with mutation information to non-mapped gene-cancer pairs (which by definition, do not have publicly known genetic mutation information) as shown in Panel B. A key advantage of the primary empirical strategy over the alternative strategy outlined in Panel B is that within-cancer comparisons are possible. Recall that large-scale cancer mapping efforts are performed at the cancer level. These cancer mapping efforts in turn publicly reveal that a subset of genes have mutations. Under the primary empirical strategy, one gene can be compared with a different gene *in the same cancer*. The alternative strategy outlined in Panel B restricts the comparison to cancers that are mapped and those that are not mapped. However, the discussion in Section 4.1 suggests that estimates generated from across-cancer comparisons may be particularly susceptible to cancer level selection.

One limitation of both empirical strategies is that the relative difference in clinical trials between gene-cancer pairs with mutation information and those without could be picking up one or both of two effects. First, the increase could represent an increase in clinical trials in gene-cancer pairs with publicly known mutation information. Second, the increase could represent a decrease in gene-cancer pairs without publicly known mutation information. Reflecting this limitation, future work will examine how the public disclosure of a mutation within a gene shifts clinical trial investments across diseases that are less likely to be substitutable (e.g., ovarian cancer vs. Alzheimer's disease).

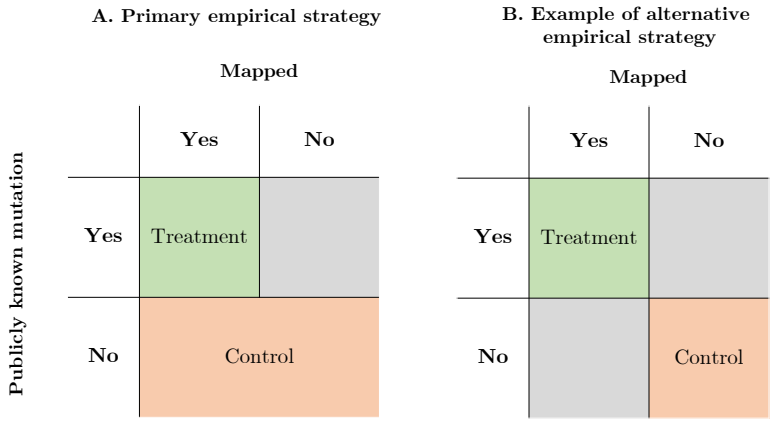


FIGURE D1
Allocation of gene-cancer pairs to treatment and control groups

# Appendix E Extensions and robustness

## E.1 Whole-genome sequencing

This section confirms that the main results are generally robust to restricting the set of large-scale cancer mapping studies to those that utilize whole-genome sequencing.

TABLE E1
Effect of public whole-genome sequencing mapping information on phase II trials, 2004–2016

| | Dependent variable: Any phase II trial | | |
|---|---|---|---|
| | Any trial | Any private sector trial | Any public sector trial |
| | (1) | (2) | (3) |
| Post × DisclGeneCancer | 0.0218*** | 0.0130*** | 0.0113** |
| | (0.00456) | (0.00354) | (0.00437) |
| Mean of dep. var. | 0.027 | 0.017 | 0.014 |
| Change in likelihood of trial (%) | 81.44 | 76.02 | 81.30 |
| Gene-cancer FEs | Yes | Yes | Yes |
| Cancer × Year FEs | Yes | Yes | Yes |
| Observations | 392,899 | 392,899 | 392,899 |

*Notes:* This table reports DID estimates of the effect of public whole-genome cancer mapping information on phase II trials. The level of observation is the gene-cancer-year. Estimates are from OLS models. The outcome variable switches from 0 to 1 if a clinical trial is reported in a gene-cancer-year. *Post* × *DisclGeneCancer* switches from 0 to 1 when a mutation in a gene-cancer pair is publicly disclosed by a mapping study. Robust standard errors, clustered at the gene and cancer level, are shown in parentheses. *Mean of dep. var.* is the mean of the outcome variable in a gene-cancer pair before the first disclosure of a mutation and is used to calculate the percentage change in the likelihood of a clinical trial. See Section 3 and online Appendix C for more detailed data and variable descriptions. *$p$ <0.10, **$p$ <0.05, ***$p$ <0.01.

## E.2 Robustness of difference-in-differences estimates

A recent literature has shown that staggered event studies may produce biased OLS estimates. This section provides evidence that the main results are primarily identified off comparisons between observations that were never treated (i.e., gene-cancer pairs that were never mapped) and those that were treated during the sample period. Furthermore, this section shows that the robust are alternative to excluding always-treated observations and to employing an alternative, heterogeneity-robust estimator.

TABLE E2
Goodman-Bacon decomposition

| | Coef. | P-value |
|---|---|---|
| Post × DisclGeneCancer | 0.00943 | 0.00319 |
| *Decomposition* | | |
| | Beta | Total Weight |
| Timing groups | 0.0023679 | 0.1715446 |
| Always vs. timing | -0.1775148 | 0.0006514 |
| Never vs. timing | 0.0106001 | 0.827804 |

*Notes:* This shows the results of a Bacon-decomposition (Goodman-Bacon, 2021) of the estimate of the effect of public cancer mapping information on private sector phase II trials, with gene-cancer and cancer-year fixed effects. Standard errors are clustered at the gene and cancer level.

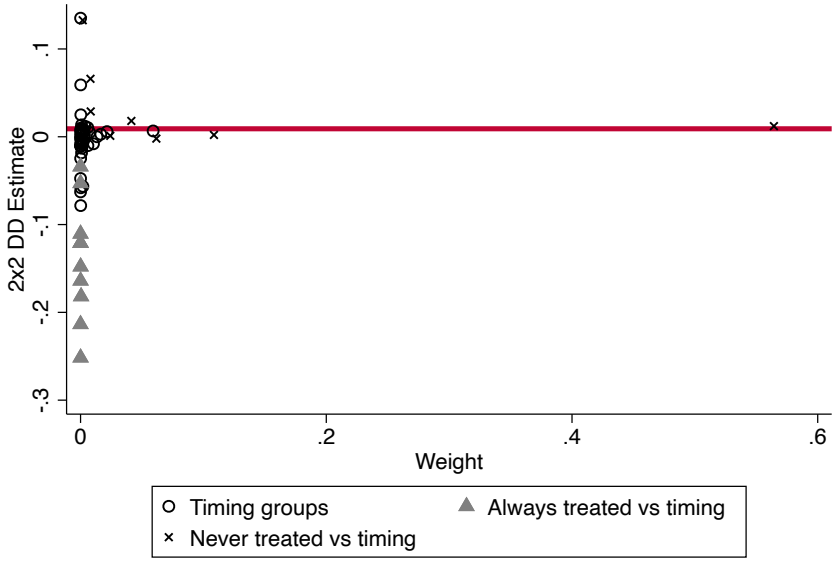


FIGURE E1
Plots of Goodman-Bacon decomposition 2 × 2 DID estimates

*Notes:* The figure reflects the results of a Bacon-decomposition (Goodman-Bacon, 2021) of the estimate of the effect of public cancer mapping information on private sector phase II trials, with gene-cancer and cancer-year fixed effects. Standard errors are clustered at the gene and cancer level. This figure presents a plot of the weight and estimated average effects for each treatment-control pair. The largest weight is associated with the 2011 treated versus never treated 2 × 2 estimate, which coincides with the large increase in genome mapping as shown in the cumulative distribution function (see online Appendix Figure B2).

TABLE E3
Excluding always-treated observations and Gardner (2022) estimation

| | Dependent variable: Any phase II trial | | |
|---|---|---|---|
| | Any trial | Any private sector trial | Any public sector trial |
| | (1) | (2) | (3) |
| Original | 0.0113 | 0.0094 | 0.0037 |
| | (0.0039) | (0.0032) | (0.0035) |
| Excluding always treated | 0.0118 | 0.0098 | 0.0038 |
| | (0.0039) | (0.0032) | (0.0035) |
| Gardner (2022) | 0.022 | 0.0148 | 0.0099 |
| | (0.0015) | (0.0011) | (0.0017) |

*Notes:* This table reports DID estimates of the effect of public cancer mapping information on phase II trials. The level of observation is the gene-cancer-year. Estimates are from OLS models. The outcome variable switches from 0 to 1 if a clinical trial is reported in a gene-cancer-year. Shown are the estimates corresponding to *Post* × *DisclGeneCancer* from equation (1). All regressions include gene-cancer fixed effects and cancer-year fixed effects. *Original* repeats the main results from the baseline regressions in Table 2. *Excluding always treated* reports estimates obtained after excluding always-treated observations. For both sets of results, robust standard errors, clustered at the gene and cancer levels, are shown in parentheses. *Gardner (2022)* reports results obtained by implementing the Gardner (2022) estimator. However, due to limitations in the Stata command used for this estimator, multi-way clustering is not permitted. As a result, the standard errors for this estimator are clustered at the cancer level.

## E.3 2013 Supreme Court gene patent ruling

This section confirms the general robustness of the main results when minimizing the impact of changing intellectual property regulations related to the 2013 Supreme Court ruling on gene patents by restricting the analysis to 2004-2012.

TABLE E4
Effect on phase II trials before 2013 Supreme Court gene patenting ruling

| | Dependent variable: Any phase II trial | | |
|---|---|---|---|
| | Any trial | Any private sector trial | Any public sector trial |
| | (1) | (2) | (3) |
| Post × DisclGeneCancer | 0.00803** | 0.00822** | 0.000686 |
| | (0.00373) | (0.00347) | (0.00273) |
| Mean of dep. var. | 0.023 | 0.013 | 0.013 |
| Change in likelihood of trial (%) | 35.58 | 64.49 | 5.327 |
| Gene-cancer FEs | Yes | Yes | Yes |
| Cancer × Year FEs | Yes | Yes | Yes |
| Observations | 331,092 | 331,092 | 331,092 |

*Notes:* This table reports DID estimates of the effect of public cancer mapping information on phase II trials. The sample includes gene-cancer-years from 2004 through 2012 (331,092 gene-cancer-year observations). The level of observation is the gene-cancer-year. Estimates are from OLS models. *Post* × *DisclGeneCancer* switches from 0 to 1 when a mutation in a gene-cancer pair is publicly disclosed by a mapping study. Robust standard errors, clustered at the gene and cancer level, are shown in parentheses. *Mean of dep. var.* is the mean of the outcome variable in a gene-cancer pair before the first disclosure of a mutation and is used to calculate the percentage change in the likelihood of a clinical trial. For sample details, see the text and online Appendix C. *$p$ <0.10, **$p$ <0.05, ***$p$ <0.01.

## E.4 Public-private phase II trials

This section examines the impact of public cancer mapping information on the set of clinical trials jointly conducted by private and public sector institutions.

TABLE E5
Effect of public cancer mapping information
on private-public phase II trials, 2004-2016

| | Dependent variable: Any private sector phase II trial | |
|---|---|---|
| | Collaboration with public sector institution with low research experience | Collaboration with public sector institution with high research experience |
| | (1) | (2) |
| Post × DisclGeneCancer | -0.000644 | 0.00215* |
| | (0.000805) | (0.000878) |
| Mean of dep. var. | 0.003 | 0.001 |
| Change in likelihood of trial (%) | -4.68 | 26.75 |
| Gene-cancer FEs | Yes | Yes |
| Cancer × Year FEs | Yes | Yes |
| Observations | 392,899 | 392,899 |
| Diff. Wald test $p$-value | 0.01 | |

*Notes:* This table reports DID estimates of the effect of public cancer mapping information on the set of clinical trials jointly conducted by private and public sector institutions. The table shows the impact of cancer mapping on such trials when the public sector institutions have low (column 1) or high (column 2) levels of research . A public sector institution is considered to have high (low) levels of research if it had a level of phase III clinical trial investment above (below) the median in the prior year relative to the focal trial. The level of observation is the gene-cancer-year. Controls include gene-cancer fixed effects and cancer-year fixed effects. *Mean of dep. var.* is the mean of the outcome variable in a gene-cancer pair before the first disclosure of a mutation and is used to calculate the percentage change in the likelihood of a clinical trial. The estimates in this table are from seemingly unrelated models, which permits a comparison of the coefficient on $Post \times DisclGeneCancer$ across models. Standard errors are clustered at the cancer level (see main text footnote 55). The $p$-value is from a Wald test that compares the differences in the coefficients on $Post \times DisclGeneCancer$. For sample details, see the text and online Appendix C. *$p$ <0.10, **$p$ <0.05, ***$p$ <0.01.

## E.5 Market Potential

This section confirms that the impact of public cancer mapping information does not vary across markets with varying levels of market size. For this analysis, I categorize the sample of gene-cancer pairs into two mutually exclusive categories based on whether the focal cancer is below or above the median annual number of cancer diagnoses between 2000 and 2003. In Appendix Table E6, columns 1 and 2 indicate that public cancer mapping information has similar effects on clinical trial investments across cancers, regardless of market size. The difference across low and high levels of market size is not statistically significant.

TABLE E6
Effect on private sector phase II trials:
Heterogeneity by market potential of disease, 2004-2016

| | Dependent variable: Any private sector phase II trial | |
|---|---|---|
| | Small market size (below median number of diagnoses) (1) | Large market size (above median) number of diagnoses) (2) |
| Post × DisclGeneCancer | 0.0110*** | 0.00837*** |
| | (0.000941) | (0.00169) |
| Mean of dep. var. | 0.012 | 0.016 |
| Change in likelihood of trial (%) | 89.60 | 53.31 |
| Gene-cancer FEs | Yes | Yes |
| Cancer × Year FEs | Yes | Yes |
| Observations | 194,012 | 198,887 |
| Diff. Wald test $p$-value | 0.18 | |

*Notes:* This table reports DID estimates of the effect of public cancer mapping information on private sector phase II trials, separately for different diseases with low and high market potential. The level of observation is the gene-cancer-year. *Post* × *DisclGeneCancer* switches from 0 to 1 when a mutation in a gene-cancer pair is publicly disclosed by a mapping study. Columns 1 and 2 split the sample across the median of market size, as measured by the number of diagnoses for the focal cancer between 2000 and 2003. The sum of the number of observations in the two columns equals the full sample of gene-cancer pairs (N = 392,899) used in the main analysis. *Mean of dep. var.* is the mean of the outcome variable in a gene-cancer pair before the first disclosure of a mutation and is used to calculate the percentage change in the likelihood of a clinical trial. The estimates in this table are from seemingly unrelated models, which permits a comparison of the coefficient on *Post* × *DisclGeneCancer* across models. Standard errors are clustered at the cancer level (see main text footnote 55). The $p$-value is from a Wald test that compares the differences in the coefficients on *Post* × *DisclGeneCancer*. For sample details, see the text and online Appendix C. *$p$ <0.10, **$p$ <0.05, ***$p$ <0.01.

## E.6 Design quality

This section confirms that the impact of public cancer mapping information does not vary across trials with varying design quality. The design of a clinical trial shapes the quality of the results that are produced. For example, in a randomized controlled trial design, patients are randomly allocated to treatment and control arms, yielding estimates that are less likely to be biased by patient selection (Byar et al., 1976). Private firms seeking to generate promising results may choose to forgo a control group or may rely on a suboptimal treatment in the control group. Using recommended standards outlined in the scientific literature (e.g., Seymour et al., 2010; Prasad et al., 2015; Dhani et al., 2017; Kemp and Prasad 2017; NCI n.d.; U.S. Food and Drug Administration 2018a, b), I classify phase II trials as well-designed if they satisfied one of the following three criteria: (1) Randomized, controlled, overall survival endpoint; (2) Randomized, controlled, validated surrogate endpoint;[xviii] or (3) Non-randomized, controlled, validated surrogate endpoint. Information on

[xviii] The FDA defines a "surrogate endpoint" as "a clinical trial endpoint used as a substitute for a direct measure of how a patient feels, functions, or survives." For more information, see https://www.fda.gov/drugs/development-resources/surrogate-endpoint-resources-drug-and-biologic-development.

validated surrogate endpoints comes from Prasad et al. (2015). Trials that are not coded as well designed are classified as poorly designed. Online Appendix Table E7 shows that the effect is similar across the two trial types of trial design, suggesting that public cancer mapping information has little effect on the quality composition of subsequent clinical trials.

TABLE E7
Effect on private sector phase II trials: Heterogeneity by trial design type, 2004–2016

| | Dependent variable: Any private sector phase II trial | |
|---|---|---|
| | Well designed (1) | Poorly designed (2) |
| Post × DisclGeneCancer | 0.000851* | 0.00159** |
| | (0.000387) | (0.000557) |
| Mean of dep. var. | 0.001 | 0.004 |
| Change in likelihood of trial (%) | 88.87 | 40.10 |
| Gene-cancer FEs | Yes | Yes |
| Cancer × Year FEs | Yes | Yes |
| Observations | 392,899 | 392,899 |
| Diff. Wald test $p$-value | 0.22 | |

*Notes:* This table reports DID estimates of the effect of public cancer mapping information on private sector phase II trials, separately for well-designed and poorly designed trials. The level of observation is the gene-cancer-year. The outcome variable switches from 0 to 1 if a private sector phase II trial is reported in a gene-cancer-year and is a well designed trial (column 1) or poorly designed trial (column 2). *Post* × *DisclGeneCancer* switches from 0 to 1 when a mutation in a gene-cancer pair is publicly disclosed by a mapping study. *Mean of dep. var.* is the mean of the outcome variable in a gene-cancer pair before the first disclosure of a mutation and is used to calculate the percentage change in the likelihood of a clinical trial. The estimates in this table are from seemingly unrelated models, which permits a comparison of the coefficient on *Post* × *DisclGeneCancer* across models. Standard errors are clustered at the cancer level (see main text footnote 55). The $p$-value is from a Wald test that compares the differences in the coefficients on *Post* × *DisclGeneCancer*. For sample details, see the text and online Appendix C. *$p$ <0.10, **$p$ <0.05, ***$p$ <0.01.

## Appendix F Estimating the implied drug approvals

In this appendix, I provide more information on the "back-of-the-envelope" analysis summarized in Section 5. Specifically, I estimate the implied number of drug approvals that could have resulted from cancer mapping efforts. Using the estimates from Column 1 of Table 2 for phase II trials, I calculate that if the gene-cancer pairs that received mutation-related information had counterfactually experienced the same probability of investments as gene-cancer pairs that did not, there would have been about four fewer drug approvals overall.[xix] After estimating the impact on the total number of potential drug approvals, I then estimate the number of approvals for novel drugs versus new uses.

This exercise requires an estimate of (i) the number of phase II trials that would have been conducted for gene-cancer pairs if they had not received mutation-related information (the "counterfactual number of trials"), (ii) the increase in the number of phase II trials resulting from the mutation-related information ("implied increase in trials"), and (iii) the likelihood of a drug successfully advancing from a phase II trial to approval ("implied increase in drug approvals"):

(i) *Counterfactual number of trials:* I use the phase II estimates from Column 1 of Table 2 to determine the counterfactual number of phase II trials associated with gene-cancer pairs had they not received mutation-related information. As of 2016, there were 17,515 gene-cancers that had received mutation information ("mapped" gene-cancers). Focusing on the pre-mutation information trial averages, I estimate that the likelihood of a gene-cancer pair being targeted in a trial in any given year prior to receiving mutation information is 0.024 overall. This suggests that if the mapped gene-cancers experienced this pre-mutation information likelihood of obtaining a trial, there would be 420 ($\approx$ 17,515 $\times$ 0.024) trial-gene-cancer observations overall trial-gene-cancer observations in each year. Column 1 of Table 2 shows that public mapping information increases the likelihood of a trial by 0.0113 to 0.0353 ($\approx$ 0.024 + 0.0113). This suggests that if the mapped gene-cancers had this likelihood of experiencing a trial, there would be 618 ($\approx$ 17,515 $\times$ 0.0353) trial-gene-cancer observations in each year.

(ii) *Implied increase in trials:* I take a conservative approach for estimating the implied increase in trials. The previous estimates suggest that public mapping information leads to a 198 ($\approx$ 618−420) yearly increase in the number of trial-gene-cancer observations. Since the majority of gene-cancers are mapped in 2011, to be conservative, I allow mapped gene-cancers to be "mapped" for 6 (= 2016−2011+1) years, resulting in a total of 1,187 ($\approx$ 6 $\times$ 198) trial-gene-cancers. To convert this to the trial level, I note that trials are typically associated with approximately 30 trial-gene-cancers. (Trials may enroll patients with a variety of genes or cancers. For example, trials may enroll patients with BRCA1-mutated and BRCA2-mutated breast and ovarian cancer, and such a trial would appear four times.) Converting 212,898 trial-gene-cancer level observations to the trial level gives 7,146 unique trials.

(iii) *Implied increase in drug approvals:* To obtain the estimated number of approved drugs, I take the estimated probability of a cancer drug successfully advancing from phase II to regulatory approval (10.5%) (Hay et al., 2014), which results in an estimated 4.20 cancer drug approvals.

[xix] For simplicity, I use the main linear probability model estimates for this exercise. However, using Poisson estimates as shown in Appendix Table B2 reveals comparable results.

### Additional analyses: new uses, novel drugs

A natural next question is: what share is for novel drugs, and what share is for new uses of previously tested drugs? Appendix Table F1 presents estimates on the impact of public cancer mapping information on all (public and private sector) phase II trials testing new uses and those testing novel drugs. Using these estimates, I apply the same method as outlined in the previous section and calculate that of the 4.20 total drug approvals, 1.20 of them are for novel drugs, and 3.00 for new uses of previously tested drugs.

TABLE F1
Effect on phase II trials of new uses vs. novel drugs, 2004-2016

| | Dependent variable: Any phase II trial | |
|---|---|---|
| | New drug uses | Novel drugs |
| | (1) | (2) |
| Post × DisclGeneCancer | 0.00904*** | 0.00323*** |
| | (0.00145) | (0.000911) |
| Mean of dep. var. | 0.019 | 0.006 |
| Change in likelihood of trial (%) | 47.13 | 55.66 |
| Diff. Wald Test P-value | | 0.00 |
| Gene-cancer FEs | Yes | Yes |
| Cancer × Year FEs | Yes | Yes |
| Observations | 392,899 | 392,899 |
| Diff. Wald test $p$-value | 0.00 | |

*Notes:* This table reports DID estimates of the effect of public cancer mapping information on clinical trials (private and public sector) testing new uses of previously tested drugs and novel drugs. The level of observation is the gene-cancer-year. To examine heterogeneity by drug type, column 1 examines the effect of public mapping information on trials whose drugs have been approved in the focal gene or previously tested in any gene-cancer pair; column 2 estimates the effect on clinical trials whose drugs have not been approved in the focal gene or tested in any gene-cancer pair. *Post × DisclGeneCancer* switches from 0 to 1 when a mutation in a gene-cancer pair is publicly disclosed by a mapping study. *Mean of dep. var.* is the mean of the outcome variable in a gene-cancer pair before the first disclosure of a mutation and is used to calculate the percentage change in the likelihood of a clinical trial. The estimates in this table are from seemingly unrelated models, which permits a comparison of the coefficient on $Post \times DisclGeneCancer$ across models. Standard errors are clustered at the cancer level (see main text footnote 55). The $p$-value is from a Wald test that compares the differences in the coefficients on $Post \times DisclGeneCancer$. For sample details, see the text and online Appendix C. *$p$ <0.10, **$p$ <0.05, ***$p$ <0.01.

## Appendix references


Byar, D. P., Simon, R. M., Friedewald, W. T., Schlesselman, J. J., DeMets, D. L., Ellenberg, J. H., Gail, M. H. and Ware, J. H. (1976), "Randomized clinical trials–perspectives on some recent ideas", *New England Journal of Medicine* 295(2), 74–80. Cancer Genome Atlas Research Network. (2011), "Integrated Genomic Analyses of Ovarian Carcinoma", *Nature* 474(7353), 609–15.

Campbell, J. D., Alexandrov, A., Kim, J., Wala, J., Berger, A. H., Pedamallu, C. S., Shukla, S. A., Guo, G., Brooks, A. N. and Meyerson, M. (2016), "Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas", *Nature Genetics* 48(6), 607–616.

Cerami, E., Gao, J., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. K., Heuer, M. L., Larrson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C. and Schultz, N. (2012), "The cBio

Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data", *Cancer Discovery* 2(5), 401–404.

Choi, J. P. (1991), "Dynamic r&d competition under hazard rate uncertainty", *The RAND Journal of Economics* 22(4), 596—610

Dhani, N., Tu, D., Sargent, D. J., Seymour, L. and Moore, M. J. (2017), "Alternate Endpoints for Screening Phase II Studies", *Clinical Cancer Research* 15(6), 1873–1882.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E... (2013), "Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal", *Science Signaling* 6(269), l1.

Gardner, J. (2022), "Two-stage Differences in Differences", arXiv preprint arXiv:2207.05943.

Gonzalez-Pereiraa, B., Guerrero-Boteb, V. P. and Moya-Anegón, F. (2009), "The SJR Indicator: A New Indicator of Journals' Scientific Prestige", https://arxiv.org/ftp/arxiv/papers/0912/0912.4141.pdf.

Goodman-Bacon, A. (2021), "Difference-in-differences with variation in treatment timing", *Journal of Econometrics* 225(2), 254–27

Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. and Rosenthal, J. (2014), "Clinical development success rates for investigational drugs", *Nature Biotechnology* 32(1), 40–51.

Kemp, R. and Prasad, V. (2017), "Surrogate Endpoints in Oncology: When Are They Acceptable For Regulatory and Clinical Decisions, and Are They Currently Overused?" *BMC Medicine* 15(1),134.

NCI Center for Cancer Research. (n.d.), "Clinical Trial Design", https://docplayer.net/15224109-Clinical-trial-design-sponsored-by-center-for-cancer-research-national-cancer-institute.html.

Prasad, V., Kim, C., Burotto, M. and Vandross, A. (2015), "The Strength of Association Between Surrogate Endpoints and Survival in Oncology", *JAMA Internal Medicine* 175(8), 1389–1398.

Samuel, N. and Hudson, T. J. (2013), "Translating Genomics to the Clinic: Implications of Cancer Heterogeneity" *Clinical Chemistry* 59(1), 127–137.

Seymour, L, Ivy, S. P., Sargent, D., Spriggs, D., Baker, L., Rubinstein, L., Ratain, M. J., Le Blanc, M., Stewart, D., and Berry, D. (2010), "The Design of Phase II Clinical Trials Testing Cancer Therapeutics: Consensus Recommendations from the Clinical Trial Design Task Force of the National Cancer Institute Investigational Drug Steering Committee", *Clinical Cancer Research* 16(6), 1764–1769.

Stratton, M.R., Campbell, P. J., and Futreal, P. A. (2009) "The Cancer Genome", *Nature* 458(7239), 719–724.

Tate, J.G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E... (2019), "COSMIC: the Catalogue of Somatic Mutations In Cancer", *Nucleic Acids Research* 47(D1), D941–D47.

U.S. Food and Drug Administration. (2018a), "Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics", https://www.fda.gov/media/71195/download (Accessed on 2021-01-13).

U.S. Food and Drug Administration. (2018b), "Master Protocols: Efficient Clinical Trial Design Strategies to Expedite Development of Oncology Drugs and Biologics Guidance for Industry", https://www.fda.gov/media/120721/download (Accessed on 2021-01-13).

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz Jr., L. A., and Kinzler, K. W. (2013), "Cancer Genome Landscapes", *Science* 339(6127), 1546–1558.