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

Jennifer Kao

UCLA Anderson School of Management

February 22, 2021

Abstract

How does basic scientific information shape private sector research investment? I assess the impact of large-scale cancer genome mapping studies, which systematically map the genetic abnormalities in cancer. Using newly-constructed data, I find that publicly available mapping information increases private investment in clinical trials by 50 percent and is disproportionately likely to spur trials evaluating drugs approved or previously tested for one disease in an additional disease. I find evidence that cancer maps improve firms' decision quality: when genetic information is known, firms are more likely to continue investment projects that are most likely to generate promising clinical results.

*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, Juan Alcacer, Megan Bailey, Samantha Burn, Marika Cabral, Caitlin Carroll, Stephen Coussens, Leemore Dafny, Ariella Kahn-Lang, Joshua Krieger, Timothy Layton, Danielle Li, Abhishek Nagaraj, Ramana Nanda, Adrienne Sabety, Rebecca Sachs, Mark Shepard, Gabriel Tourek, Lisa Xu, Heidi Williams, 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 abstracts 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 leading models, private sector research and development (R&D) investments play a key role in driving economic growth (Solow 1957; Romer 1990; Aghion and Howitt 1992). It is widely viewed that basic scientific knowledge is central to the development of these private R&D investments (Nelson 1959; Arrow 1962).[1] Despite widespread interest in the effects of basic scientific knowledge, we have surprisingly little empirical evidence about the context under which basic scientific information shapes the trajectory of private R&D decisions.[2] As Nelson (1959) notes, the effects of scientific knowledge are not homogeneous. Instead, the ability of basic scientific knowledge to help firms navigate the R&D process depends on the information's comprehensiveness, accuracy, and accessibility.

In this paper, I provide evidence of the causal impact of a particular form of basic scientific information: publicly available maps. In recent years, basic scientific maps have received increasing academic and policy attention (Williams 2013; Nagaraj forthcoming).[3] Relative to basic science investigations that may characterize a specific disease in a piecemeal manner, basic science mapping initiatives aim to generate more complete, organized, and standardized representations of the disease. This raises the several questions, for example: what are the effects of placing this detailed blueprint in the public domain on private firms' R&D investments (Nelson 1982; Fleming and Sorenson 2004), and how are these effects shaped by competitive dynamics (Cockburn and Henderson 1994)? Assessing the impact of publicly available maps has important implications for understanding the value of research institutions (Furman and Stern 2011) and policies (e.g., intellectual property policies) that shape knowledge production and dissemination.

To tackle these questions, I leverage the unique features of the pharmaceutical R&D setting to provide causal evidence on the effect of large-scale cancer mapping initiatives on private sector investments in cancer clinical trials. Theoretically, the effects of publicly available maps on private firms' research decisions are ambiguous. On the one hand, private firms with access to mapping information may experience a fundamental shift in their basic scientific knowledge. The comprehensive and detailed information may spur and sharpen private firms' research efforts through improving technological search and lowering entry costs (Rosenberg 1974). Supporting this view, Fleming and Sorenson (2004) use mapping as a metaphor for how science can help private firms navigate the inventive process. On the other hand, the placement of scientific data in the public domain may increase the expected level of competition—e.g., through disproportionately benefitting incumbent firms' existing products or lowering the cost of entry for competitors (Arrow 1962).

[1] In this paper, "basic scientific knowledge" refers to 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] For reviews on studies that examine the linkage between basic science and technical progress, see David, Hall and Toole (2000) and Hall and Van Reenen (2000). For a recent contribution to the area, see Azoulay et al. (2019).

[3] For coverage of recent efforts, see, for example, Benedict Carey, "Mapping the Brain's Genetic Landscape," *New York Times*, December 13, 2018. https://www.nytimes.com/2018/12/13/health/genetics-brain-autism-schizophrenia.html

This expected increase in competition may diminish the incentive of potential entrants to enter and cause the level of subsequent research investments to be, on net, lowered or unchanged. Despite the increasing interest in the economic value of scientific maps, direct empirical evidence on how publicly available maps influence affect research investment incentives and decisions remain scarce (Nagaraj and Stern 2020).[4]

Building on the foundation provided by the Human Genome Project (HGP), cancer genome mapping initiatives systematically catalogue the genetic mutations that might drive the progression and growth of cancer (Lander 2011; Williams 2013; Jayaraj 2018).[5] These large-scale research initiatives aim to facilitate technological search and private firm investment through introducing and validating existing scientific knowledge about the disease (Mardis 2018). Large-scale cancer mapping efforts are believed to play an influential role in the development of novel cancer therapies (Kolata 2013): National Institutes of Health director Francis Collins and former National Cancer Institute deputy director Anna Barker (Barker and Collins 2008, p.50) noted that mapping the cancer genome would "help chart a new course across the complex landscape of human malignancies."

This paper examines how the public introduction of cancer "atlases" shapes the development of new therapies for cancer, a disease whose therapeutic market is the largest in terms of global spending—at $133 billion per year—and as the second leading cause of death in the United States, one in which advances yield tremendous value to society (Heron 2018; IQVIA 2018). Specifically, I focus on changes in clinical trials. Within the pharmaceutical industry, a key requirement for new product entry is the completion of clinical trials mandated by the U.S. Food and Drug Administration (FDA), a risky and costly process. Only about 12 percent 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, Grabowski and Hansen 2016). In light of this, the impact of the cancer mapping efforts on new product development depends in large part on the extent to which the basic science influences private firms' clinical trial investment efforts.

My empirical analysis focuses on how the disclosure of publicly available cancer mapping information shapes two dimensions: (i) the subsequent *quantity* of private clinical trial investments; and (ii) private firms' *decision quality*—the likelihood that firms make decisions that maximize their expected returns based on existing clinical information. I assemble a new dataset of publicly available information produced by 168 large-scale cancer mapping efforts, linked to private sector

[4]Several notable recent contributions include Williams (2013), Jayaraj (2018), and Nagaraj (forthcoming), which provide quasi-experimental evidence documenting that publicly available mapping information leads to an increase in follow-on investments.

[5]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." Furthermore, when presenting the completed draft of the human genome (United States 2000), scientist 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."

clinical trials, over the period 2004–2016, inclusive.[6] I observe many characteristics of the clinical trials, including the cancer types under investigation, the genetic criteria used for patient enrollment, the drug being tested, the sponsoring firm, the trial design, and the clinical outcomes. A key feature of these R&D data is that I am able to follow private firms' R&D investment decisions as they navigate through the multi-phase drug development process.

Using these data, I begin by investigating how publicly available cancer mapping information shapes the subsequent quantity of private sector clinical trials. Publication of results from large-scale cancer 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 framework. To address concerns over selection in the timing of mapping, I control for differences in "research potential" between different gene-cancer pairs with gene-cancer fixed effects and control for secular changes in the pharmaceutical industry over time using year fixed effects and cancer-specific time trends.

I find that mutation-related information disclosures from large-scale cancer genomic efforts increase private sector investment in clinical trials by 50 percent over the study period. If gene-cancer pairs that received mutation-related information had counterfactually experienced the same level of investments as gene-cancer pairs that did not, there would have been up to 46 fewer private sector clinical trials and seven fewer cancer drug approvals. Consistent with the conjectured value of cancer mapping information, I document that the impact on cancer clinical trials is quantitatively and statistically more positive for "driver" mutations, genetic aberrations that are more likely to drive the progression of cancer and therefore more valuable for the drug development process.

To gain insight into how the effects of mutation-related information are distributed among products and firms, I examine how the composition of clinical trials shifts in response to mutation-related information. One mechanism through which cancer mapping increases private investment is by providing evidence of linkages across research opportunities that were previously believed to be distantly related. In particular, genetic mapping can reveal that similar genetic aberrations underlie different cancer types and spur trials testing drugs approved or previously tested for one disease in an additional disease. For example, cancer mapping may show that breast and prostate cancer share similar gene mutations, highlighting the potential for breast cancer drugs to be effective treatments for prostate cancer. While private firms and physicians may have previously identified linkages across diseases through experimentation, large-scale cancer mapping efforts systematically provide information in centralized, publicly accessible genomic databases. Consistent with the above hypothesis, I find that mapping disproportionately increases investment in trials testing

[6] This includes cancer mapping efforts from both government (e.g., National Institutes of Health) and non-government organizations (e.g., Johns Hopkins University) institutions.

drugs that were approved or previously tested.[7] Additional analyses suggest that cancer mapping information disproportionately benefits firms with fewer resources and spurs investment in diseases with relatively low levels of competition.

Finally, I explore whether cancer mapping information increases the quality of private firms' research decisions. Once firms initiate a clinical trial, they must complete a series of additional clinical trials, each with increasing cost and risk. Upon phase completion, firms must decide whether to continue or terminate investment. Clinical trial failures are expected, and one indicator of success is a firm's ability to "fail quickly"—i.e., to maximize their expected returns through minimizing resources allocated to drugs that are unlikely to be successful, and continue investment in drugs that are most likely to generate promising clinical results and successfully come to market (Lendrem et al. 2015). In deciding whether to continue or terminate investment, private firms may exhibit poor decision quality by dismissing or failing to understand existing clinical evidence: indeed, a review of AstraZeneca's drug pipeline revealed that 18 percent of failures occurred because a drug advanced to the next phase of clinical development despite weak evidence from earlier phases (Cook et al. 2014).

Using trial-gene-cancer level data, I investigate whether mapping information is associated with increases in the quality of private firms' research decisions. I find that private firms with access to mutation-related information are more like to terminate drugs that are likely to fail and continue investment in drugs that are most likely to be successful in the long run. Specifically, private firms initiating trials in diseases with mutation-related information are 30 percent less likely to advance drugs with weak clinical evidence to the next phase as compared with trials initiated in diseases without mutation-related information. Examining the outcomes of drugs chosen to advance, I find that cancer mapping information is associated with drugs that lead to greater improvements in patient survival in the next phase, even after controlling for disease and firm characteristics. Taken together, these findings are consistent with the view that firms with increased access to comprehensive and detailed scientific data experience improvements in their decision quality and ultimately increase the presumptive success of their research investments (Nelson 1982; Rosenberg 1990; Sharpe and Keelin 1998; Peck et al. 2015; Spetzler, Winter and Meyer 2016; Bujar et al. 2017).

The remainder of the paper proceeds as follows: Section 2 presents a case study of my results using a single large-scale cancer mapping study. Section 3 introduces the empirical setting and the data. Section 4 analyzes the effect of cancer mapping on the quantity of private firms' clinical trial investments. Section 5 examines the impact of cancer mapping on the quality of private firms' decisions. Finally, Section 6 concludes.

[7] While the prior innovation literature has focused on incentives to develop novel products, this finding focuses on 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 little empirical work on the issue.

2 Case Study

2.1 A Large-Scale Ovarian Cancer Mapping Study

The purpose of the cancer mapping efforts examined in this paper is to create a publicly available "mutational landscape" that serves as a foundation for subsequent cancer research. Large-scale cancer mapping efforts examine hundreds of patients in order to produce novel information about rare gene mutations that were previously overlooked by earlier, small-scale mapping efforts. By having a better understanding of a cancer's biological basis, firms can more easily develop drugs tailored to patient subgroups with specific genetic features (Collins and McKusick 2001). These so-called "targeted" therapy drugs may be more effective for those patients. This, in turn, may have ambiguous effects on private sector research.

Consider, for example, the case of the The Cancer Genome Atlas' (TCGA) serous ovarian cancer study (TCGA, 2011a). Ovarian cancer is diagnosed in roughly 21,000 women each year and is the fifth leading cause of cancer death among women in the United States (American Cancer Society 2021). Of these deaths, 85 percent of deaths are among patients with an aggressive ovarian cancer subtype called "serous" ovarian cancer (TCGA, 2011b). In this mapping study, TCGA researchers systematically catalogued the genetic mutations underlying more than 300 serous ovarian cancer tumors and submitted their findings to the journal *Nature* in 2010.

In addition to other genetic discoveries, the TCGA ovarian study revealed that 21 percent of the tumors contained mutations in the BRCA1 and BRCA2 (collectively referred to here as "BRCA") genes. Previous research had identified BRCA mutations in inherited ovarian cancer. However, the TCGA ovarian cancer study confirmed that mutations also occurred in non-inherited serous ovarian cancers. In light of these findings, TCGA researchers suggested that non-inherited serous ovarian cancer tumors could respond to poly (ADP-ribose) polymerase (or PARP) inhibitors. PARP inhibitors generate an anti-tumor effect: PARP and BRCA genes repair damaged DNA, which makes up genes. In tumors with mutated BRCA genes, PARP inhibitors prevent all potential DNA repair mechanisms, which can ultimately cause cancer cell death.[8] At the time of the TCGA study, PARP inhibitors were already being testing in clinical trials and used to treat other forms of ovarian and breast cancer with mutated BRCA genes.

2.2 Quantity Implications of the Ovarian Cancer Mapping Study

The predicted effects of the ovarian cancer mapping study on the quantity of subsequent private research investments (i.e., clinical trials) are theoretically ambiguous (Nelson 1982).

In terms of increasing subsequent research investments, a sizeable theoretical literature documents the importance of basic scientific knowledge in shaping private firms' research investment decisions (Rosenberg 1974; Mowery and Rosenberg 1979). By providing a genetic blueprint which reveals the structure, organization, and likely function of genetic mutations, a scientific map may

[8] For more details on PARP inhibitors, BRCA mutations, and ovarian cancer, see: Bryant et al. (2005); Farmer et al. (2005); Lijima et al. (2017); Matulonis (2017).

reveal new research opportunities and increase the productivity of technological search, lowering the expected cost of subsequent research investments (David, Mowery and Steinmueller 1992; Arora and Gambardella 1994; Klevorick et al. 1995; Fleming and Sorenson 2003, 2004; Fabrizio 2009; Makri, Hitt and Lane 2010).

In this empirical setting, a large-scale systematic mapping effort like the TCGA's ovarian cancer study may reveal novel information about cancer-causing genes. This information can assist private firms in identifying subgroups of patients that respond most favorably to treatment. Following the concepts outlined in Fleming and Sorenson (2003, 2004), invention is a process of searching for better combinations of components (in this case, drugs and diseases) and facilitates the efficient identification of useful, new combinations. A more efficient drug-disease (or drug-patient) match could enable firms to conduct trials with fewer patients and over a shorter duration, lowering the expected cost of bringing a novel drug to markets and ultimately increase private investment in clinical trials (McShane, Hunsberger and Adjei 2009; Chandra, Garthwaithe and Stern 2018).[9] The increase may be reflected in both clinical trials testing novel drugs and those testing new uses of existing drugs (e.g., PARP inhibitors that were previously tested in a separate disease).

In addition to introducing novel information, mapping efforts like the TCGA's ovarian cancer effort may validate existing information. With respect to that effort, previous sequencing or non-sequencing efforts may have already revealed relationships between BRCA genes and different forms of ovarian cancer before the TCGA revealed its findings in 2010. In particular, private firms with significant resources may have access to private cancer mapping information: they may have their own in-house genomics research efforts, or partner with firms that specialize in genomics research.[10] To illustrate, the drug Olaparib, the first approved PARP inhibitor, had already been clinically tested in several forms of BRCA-mutated ovarian and breast cancer prior to 2010. This suggests that its manufacturer, AstraZeneca—a multinational pharmaceutical firm with $33.3 billion in sales in 2010—was already aware of several different diseases that could be effectively treated by Olaparib (AstraZeneca 2011).[11] In the same vein, prior to the TCGA's release of its findings, non-mapping efforts such as retrospective analyses could also have revealed to AstraZeneca that non-inherited serous ovarian cancer patients with BRCA mutations were most responsive to treatment, suggesting that this particular type of ovarian cancer was driven by BRCA mutations.

Yet, even for private firms with prior knowledge of the linkage between a certain cancer and certain genetic mutations, the information provided by publicly available mapping efforts may still be a useful source of data validation: in describing the impact of TCGA's ovarian cancer study, a leading genomics expert at a pharmaceutical company which manufactures a PARP inhibitor con-

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

[10] For example, FoundationMedicine—a firm that specializes in sequencing tumors and developing genetic tests for evaluating cancer—has a partnership with the pharmaceutical firm Pfizer which allows the company to benefit from access to FoundationMedicine's database of more than 200,000 tumor profiles.

[11] Indeed, a 2017 AstraZeneca Annual Review notes: "In genomics, we have analysed more than 200,000 genomes (including data from *internal* and *external* databases) to inform investment decisions in drug discovery." Emphasis added (AstraZeneca, 2017, p. 90).

firmed that the company may have already known some of the information. However, the TCGA's finding that 21 percent of ovarian cancers exhibit a BRCA mutation was helpful in validating existing hypotheses regarding the share of serous ovarian cancer tumors with BRCA mutations in a large sample.[12,13] Thus, whether the information it provides is new or simply validates earlier research, publicly available mapping efforts can increase the net level of clinical trials testing drugs that are already in the pipeline (e.g., PARP inhibitors) or that have not yet entered clinical development.

While a sizable empirical literature has documented that increased access to basic scientific information leads to an increase in follow-on innovation, the competitive dynamics and regulatory features of the drug development process suggest that the TCGA's ovarian cancer information could also plausibly lower or leave the net level of clinical trials unchanged.[14] The pharmaceutical industry has long been characterized as having first-mover advantages, where firms are able to secure monopoly-like advantages through intellectual property protection and regulatory policies (Bond and Lean 1977; Schmalensee 1982; Comanor 1986; Lieberman and Montgomery 1988; Berndt et al. 1995; Scherer 2000).[15,16] As noted above, placing the ovarian cancer mapping information in the public domain lowers the expected cost of entry (in the TCGA study, the expected cost of developing a drug to treat non-inherited serous ovarian cancer with BRCA mutations). Faced with an information shock that may be quickly appropriated by competitors and a decline in the likelihood of becoming a first-mover, private firms considering whether to invest in clinical trials may expect lower returns to their investments and decline to invest.[17] Relative to investments in early-stage research, such as drug patents (Jayaraj 2018), the clinical development process is long (typically taking 6 to 8 years), costly (typically costing a manufacturer between $800 million to $2.6 billion),[18] and risky (only 12 percent of drugs that begin clinical development ultimately go to market) (DiMasi, Hansen and Grabowski 2003; DiMasi, Grabowski and Hansen 2016). These features suggest that private firms' research decisions to invest in clinical trials may be more responsive to competitors' investment decisions (Krieger forthcoming).

[12] Interviewed by author on April 3, 2018.

[13] To illustrate how pharmaceutical firms use the TCGA's results, a 2017 AstraZeneca study that examined the role of Olaparib in treating non-inherited mutations in serous ovarian cancer cited mutational prevalence estimates from the TCGA ovarian study (Dougherty et al. 2012).

[14] While the focus of this paper is on the economic barriers to investment, other reasons include the scientific challenges of translating basic scientific information into novel drugs (e.g. Hermosilla and Lemus 2017) and the fact that cancer mapping information may confirm the lack of research opportunities (David, Mowery and Steinmueller 1992).

[15] 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.

[16] In recent work, Hill and Stein (2020) examine how first-mover advantages for scientific research can include non-pecuniary benefits, such as researcher credit.

[17] A large but inconclusive literature explores the effects of "winner-take-all situations" on R&D investment decisions. For a review, see Cockburn and Henderson (1994).

[18] These cost estimates reflect the direct cost of research and the opportunity cost of capital. The estimates have been subject to criticism due to small sample size, assumptions about the cost of capital, and the confidential nature of the underlying data. Nevertheless, other efforts have generated similar cost estimates (Avorn 2015).

A unique feature of the drug approval process exacerbates these effects: manufacturers of existing drugs are allowed to progress through the R&D process more quickly relative to new entrants. In particular, manufacturers of existing PARP inhibitors may initiate clinical trials to treat individuals with non-inherited BRCA-mutated ovarian cancers. These firms have an advantage over new entrants (i.e., firms without a PARP inhibitor) because they may be able to skip several stages of the R&D process, such as earlier clinical trials that assess drug safety.[19] Thus, because of first-mover concerns and competition with manufacturers who are not new entrants, potential entrants may choose not to invest in the mapped disease and, in some cases redirect investment efforts to other, non-mapped diseases.

Let us turn now to the particular case of the TCGA study. Appendix Figure A1 shows the total number of private sector trials enrolling patients with BRCA2-mutated ovarian cancer by year, six years before and after 2010—the year in which the TCGA submitted its findings to *Nature*. For simplicity, this figure excludes any trials testing AstraZeneca's Olaparib which experienced relatively high levels of investment throughout the period, suggesting that AstraZeneca may have relied heavily on its own internal mapping database to guide its research efforts (Appendix Figure A2, Panel A).

The figure reveals a striking relationship between the level of BRCA2-mutated ovarian cancer trials and the disclosure of the TCGA's ovarian study results: the level of trials substantially increases in 2011, the year in which the TCGA's findings were published.[20] Indeed, the increase occurs in trials testing three different PARP inhibitors—all of which had previously been tested in breast cancer. A similar trend occurs among the number of drugs approved to treat patients with BRCA-2 mutated ovarian cancer (Appendix Figure A2, Panel B), though the data are relatively sparse. Though causality has not yet been shown, this figure suggests that mapping information may increase subsequent private investment in clinical trials, particularly among new uses of existing (i.e., previously tested) drugs.

2.3 Decision Quality Implications of the Ovarian Cancer Mapping Study

In the later stages of drug development, firms that have already initiated and completed clinical trials face another type of research investment decision: they must decide whether to terminate or continue investment by advancing their drugs to the next clinical trial phase—a costly decision that involves significant uncertainty. In light of this, a natural follow-up question to ask is whether the TCGA's ovarian study also shaped the quality of firms' termination-or-continuation decisions. Demonstrating the impact of the TCGA ovarian cancer study is difficult in this particular context

[19] This can apply to drugs that were approved or previously tested. At the time of the TCGA study, no PARP inhibitors had yet been approved. The first approved PARP inhibitor, Olaparib, was not approved until 2014.

[20] The figure also shows that one trial was initiated prior to 2010, a fact that may be explained by pre-existing awareness of ovarian cancer-causing genes among some firms.

because of incomplete data.[21] Therefore, my strategy in the remainder of the section is to provide a brief discussion.

Suppose that results of a trial testing a drug on patients with BRCA-mutated non-inherited serous ovarian cancer reveal that the drug is ineffective. For example, the share of treated patients whose tumors shrink may be lower than expected, or patients in the treatment group do not experience any additional gains in months of survival relative to those in the control group. Assume that the negative clinical results accurately reflect the drug's underlying value. In such a case, information from the TCGA ovarian cancer study may either increase or decrease the quality of the trial sponsor's termination-or-continuation decision.

In terms of increasing the trial sponsor's decision quality, the TCGA's mapping information may confirm the negative clinical evidence and thus encourage termination. The firm will be able to save resources by ceasing investment in a drug that is unlikely to be successful. Instead, mapping information may encourage the firm to direct resources towards promising drugs that are more likely to successfully obtain regulatory approval in the long run (Peck et al. 2015). These effects can result from two mechanisms. First, as described above, detailed mapping information can lower the firm's investment uncertainty and help the firm make more informed decisions throughout the multi-phase clinical development process. Qualitative evidence from firm-level case studies suggest that access to basic science can help private firms interpret clinical trial outcomes, and clarify the costs and gains associated with the decision to terminate or continue investment in a particular drug (Sharpe and Keelin 1998; Cook et al. 2014; Morgan et al. 2018).

Second, with access to a reliable, organized view of the ovarian cancer landscape, the firm's decision makers may be less susceptible to biases that can lead to suboptimal outcomes. For example, mapping information may lower the likelihood that the firm's managers compute payoffs incorrectly (e.g., due to confirmation bias, overconfidence, sunk-cost fallacy) (Tversky and Kahneman 1974; Donelan, Walker and Salek 2015; Bujar et al. 2017), fail to consider alternatives (Sharpe and Keelin 1998), follow the decisions of the past or their peers (Bujar et al. 2017), or overemphasize progression-seeking behaviors (Guedj and Scharfstein 2004; Cook et al. 2014). This, in turn, may also lead to cost-saving trial terminations.

In terms of decreasing the trial sponsor's decision quality, the TCGA's mapping information may contradict and subsequently override the negative clinical evidence and lead firms towards continuation. By suggesting that an R&D investment should succeed theoretically, cancer mapping information may encourage the firm to ignore existing clinical evidence that suggests otherwise (Fleming and Sorenson 2004). This may lead to perverse outcomes: by suggesting that a drug-disease pairing should succeed theoretically, the TCGA's mapping information may increase the likelihood that the firm incurs the high development costs associated with late-stage failures (Peck et al. 2015).

[21] Data on trial results is required to understand whether, following the TCGA ovarian cancer study, firms are more likely to terminate trials with ambiguous trial outcomes. However, only two of the trials in Appendix Figure A1 have available data on relevant trial results (see Section 3.3 for a discussion on trial results reporting).

3 Empirical Setting and Data

3.1 Scientific Background

Cancer—the disease I consider—is caused by changes in the DNA molecule.[22] 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, Campbell and Futreal 2009). The average tumor contains 33 to 66 mutated genes; the number varies across different types of mutations (Vogelstein et al. 2013). For example, the blood cancer acute myeloid leukemia is associated with a median number of 8 mutations. In contrast, non-small cell lung cancer is associated with 150 to 200 mutations per tumor. Mutations can cause a cell to produce proteins that can lead cells to grow quickly and cause damage to neighboring areas (TCGA, 2018).

I use gene-cancer pairs as my disease unit of analysis. First, I begin with a list of 80 cancer sites, based on the standard Surveillance, Epidemiology, and End Results (SEER) classification system. Next, I focus on a set of 627 genes listed in the Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer Gene Census, which consists of the set of genes that are believed to be causally associated with cancer.[23] Each gene found in the Cancer Gene Census is listed along with a cancer for which there are at least two independent reports showing that mutations are found in patients with that particular cancer type and are considered to be likely implicated in driving other cancer types. This results in 50,160 possible gene-cancer pairs (627 genes $\times$ 80 cancer sites).

3.2 Large-Scale Cancer Genome Mapping Efforts

The purpose of cancer genome mapping is to identify the specific genes and mutations associated with different types of cancer. This is perfomed by comparing the DNA sequences of cancer cells to those of normal tissue (either from the same individual or a reference DNA). Appendix Figure A3 graphically summarizes this scientific background.

In the past two decades, large-scale systematic cancer genome sequencing initiatives—efforts to catalogue 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 that occur 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

[22] 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, please see https://ghr.nlm.nih.gov/primer.

[23] The original version of the Cancer Gene Census was first published in Futureal et al. (2004). The version used here comes from the Version 82 of the COSMIC database (For more details, see https://cancer.sanger.ac.uk/cosmic/download).

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).[24]

I obtain the information produced through by these large-scale cancer sequencing efforts—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 (Yang et al. 2015). Mapping data include information about a sequenced tumor's cancer type (e.g., ovarian cancer), associated genetic mutations (e.g., BRCA2), and the date in which the associated mapping study was submitted to a scientific journal for publication.[25]

I focus on mapping information from 168 cancer mapping efforts (see Appendix B for a description of how the mapping studies were selected). The cancer mapping studies used in this paper share three important characteristics. First, the studies are cancer-site specific. For example, the TCGA ovarian cancer study described in Section 2 focused only on mapping ovarian cancer tumors. Second, the studies are large-scale and systematic: they typically examine hundreds of tumors, and 91 percent of the studies examine all the protein-coding regions in DNA. Third, following a large literature that uses journal rankings as a proxy for publication impact, I focus on the set of large-scale mapping studies published in highly ranked scientific journals during the study window. Appendix Figure A4 shows the number of cancer mapping studies and cancer tumors mapped between 2004 and 2016. The fall after the original sustained increase likely reflects the finite number of cancer sites (e.g., the marginal value of the fifth large-scale ovarian cancer mapping study may be limited).

I am interested in research activity following the public disclosure that a mutation exists in a gene-cancer pair. Before describing the drug development process, I highlight two features of mutation-related information. 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 may lead to an increase in private sector research activity, relative to gene-cancer pairs that do not have mutation information. However, it is also possible that cancer mapping efforts may provide "negative" mutation information which can lead to a decline in private sector research activity. This can occur, for example, if cancer mapping reveals that a particular gene-cancer harbors a mutation that makes it more difficult to treat patients with the gene mutation and cancer. For example, a TCGA lung cancer study revealed that 3 percent of tumors contained a mutation that allows them to evade the immune system (TCGA, 2012), suggesting that drugs that work through activating the immune system would not be effective treatments for lung cancer patients with that specific gene mutation.

[24] 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.

[25] I focus on mutations that occur in the protein-coding region of the DNA and that are likely to lead to a change in biological structure. See Appendix B for more details.

Second, information produced by large-scale cancer mapping efforts may be known before the cancer mapping study's official publication date: for instance, pharmaceutical firms may first become aware of preliminary mapping results 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.[26]

In a subset of the analyses that follow, I examine how the impact of mapping information varies across information with more (or less) clinical relevance. The scientific literature classifies mutations into two broad categories: those that are likely to drive the growth and progression of cancer (so-called driver mutations) and those that are likely to be harmless (so-called passenger mutations). 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. As a result, I employ two strategies to determine driver mutations (Carr et al. 2016): (1) identifying whether the mutation is highly likely to be a driver mutation based on statistical methods performed by the cancer sequencing researchers and (2) classifying whether the mutation is detected an unusually high number of times ($\geq$ 10 patients) in a particular gene-cancer pair in a given study.[27,28] These probable driver mutations contain the strongest signal of cancer-causing behavior and are typically described in detail in the associated mapping publication.

3.3 Private Research Investments

3.3.1 Drug Development

Drug development typically begins with extensive preclinical laboratory research that involves testing a new candidate on animals and human cells. Once complete, the manufacturer begins the most expensive aspect of drug development: human testing of the drug in a series of clinical trials in which costs increase with each subsequent trial phase. Drugs that successfully demonstrate safety in phase I trials proceed to phase II trials in which their efficacy is tested in a few hundred patients. Phase III is the final stage of clinical development and involves assessing efficacy in thousands of patients and examining them over a longer period of time. Both phase II and phase III trials assess efficacy through measuring changes in overall survival and the objective response rate, the

[26]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).

[27]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)

[28]In particular, a mutation is considered to be occurring at an unusually high number of times if it is discovered in at least 10 patients in the same study year.

share of patients whose tumors reduce by a prespecified amount.[29] Once phase III is complete, manufacturers must submit a new drug application (NDA) for regulatory review.

The development and review process is indication specific—i.e., a drug receives regulatory approval for a specific therapeutic use. However, more than 60% of cancer drugs approved have multiple uses. To expand a drug's label to include a new use, the manufacturer must undertake additional efficacy clinical trials and submit a supplemental new drug application (sNDA) (FDA, 1998b). The amount of resources involved depends on the similarity between the original and new use (FDA, 2004). For example, if the manufacturer of a drug that is approved in one cancer type (e.g., gallbladder) is seeking approval in another tumor type with a common biological origin (e.g., the colon), the manufacturer may skip phase I trials and rely on fewer phase II trials (FDA, 1998a). With less evidence for the FDA to review, average approval times are shorter for sNDAs for new indications and new patient populations relative to NDAs (DiMasi 2013).

New use approvals have high expected social value (Berndt, Cockburn and Grepin 2006; Roin 2013). Francis Collins, the former director of the National Institutes of Health, describes 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 scientific evidence that is useful for clinical decision making, particularly in contexts where off-label use is common. However, despite the relatively lower costs of seeking new use approvals, there is a widespread perception that there is too little investment in new uses of approved drugs. The so-called "problem of new uses" is caused by the limited patent protection for new uses and widespread off-label drug use (Eisenberg 2005).

3.3.2 Clinical Trials Data

Data on private sector clinical trials comes from the Clarivate Cortellis Competitive Intelligence Clinical Trials Database, which includes trial data from public trial registries. Each clinical trial provides detailed information on the cancer being examined (e.g., prostate cancer), the drug being tested (e.g., Olaparib), the sponsoring firm (e.g., AstraZeneca), any collaborating firms, and the trial start date (as measured by the date on which the first patient was enrolled). Importantly, the clinical trials also contain information on protein biomarkers (e.g., the gene EGFR).[30] I restrict the set of clinical trials to those with biomarkers that are used to guide patient selection. Each patient biomarker can then be linked to genes using the Uniprot database to generate a dataset of trials at the gene-cancer level. Since I am interested in private sector investments, I restrict my analysis to the set of clinical trials that are sponsored by a private firm.[31]

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

[30] I am grateful to Ariel D. Stern for sharing the cleaned data from Chandra, Garthwaithe and Stern (2018) for this paper.

[31] 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 firm.

To analyze the impact of mapping information on the quantity of subsequent research, I focus on investments in phase II trials—the first trials that measure efficacy and constitute a major investment for private firms. This results in a sample of 30,137 private sector phase II clinical trials at the gene-cancer level. Appendix Figure A5 shows the growing share of cancer trials that are gene-related, or use gene characteristics to guide patient enrollment, over time. 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. As discussed in the ovarian cancer case study in Section 2, this increase may have been driven by several sources, including manufacturers' retrospective analyses of previous trial results or licensing relationships with genomic firms.[32] This paper aims to examine whether large-scale cancer mapping efforts lead to any additional effect on the level of private sector clinical trials, above and beyond these other factors.

I supplement the clinical trial data in two ways:

(i) *Drug Approval Data:* I link trial data to drug approval data to identify whether a trial is evaluating an approved drug. Data on anticancer drugs originally approved to treat cancer come from the CenterWatch, National Cancer Institute, and Memorial Sloan Kettering Cancer Center websites. This results in 187 drugs originally approved to treat cancer between 1977 and 2015, inclusive. For each drug, I obtain the date of approval and the approved cancer type.

I next classify the drug as being approved for a gene if it is approved with a companion diagnostic, a requirement for drugs aimed at targeting patients with specific genetic types.[33] 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 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 trials into two mutually exclusive categories: trials testing new uses and trials testing novel drugs. A trial-gene-cancer is classified 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. For example, a trial enrolling ovarian cancer patients with BRCA2 gene mutations 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.[34] In contrast, a trial-gene-cancer is

[32] One interpretation is that the pre-2011 increase is driven by trials initiated in gene-cancer pairs that received mutation information before 2011. However, removing these trials does not change the overall trend.

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

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

classified as "testing novel drugs" if *any* of its interventions have never been (i) approved in the focal gene and (ii) previously tested in any gene-cancer pair.[35]

(ii) *Clinical Trial Outcomes:* For a subset of the empirical exercises that follow, I examine the relationship between mapping information on common clinical trial outcomes. Since the Food and Drug Administration Amendments Act (FDAAA) of 2007, most phase II and phase III clinical trials have been required to report results within one year of completion.[36] Despite this requirement, clinical trial results are significantly underreported (it is estimated that just 22 percent of trials meet this reporting requirement) (Prayle, Hurley and Smythe 2012; Anderson et al. 2015).

To obtain additional data on clinical trial outcomes, I turn to abstracts submitted to the ASCO Annual Meeting between 2004 and 2017. ASCO is the primary professional society for medical oncologists and most major research groups submit abstracts describing the findings of their clinical trials to their annual conference. I collect data on two commonly used clinical outcomes in cancer drug development: treatment group gains in overall survival (the time between randomization and death) and objective response rates (the 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 estimate the impact of large-scale cancer mapping on the quantity of private sector trials by randomly assigning mutation information to different gene-cancer pairs. I would then compare the level of subsequently initiated clinical trials in gene-cancer pairs *with mutation information* with all gene-cancer pairs *without mutation information*. This empirical strategy permits within-cancer, across-gene comparisons, in contrast to alternative methods such as comparing mapped gene-cancer pairs that have mutation information with non-mapped gene-cancer pairs that do not have mutation information (see Appendix C for a detailed discussion). Motivated by the ovarian cancer case study in Section 2, I approximate this ideal experiment by using variation in the timing of publicly disclosed information about a mutation in a gene-cancer pair.

This empirical strategy removes cancer-level differences in research potential and payoff by including gene-cancer fixed effects and estimates the impact of mapping information on clinical trials using variation in the timing of the information shock (i.e., when the mutation information is disclosed) between gene-cancer pairs. By comparing gene-cancer pairs that receive an information shock early with those that receive an information shock late (or never receive an information

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

[36] Trials covered by the FDAAA include those that have at least one site in the U.S. and are testing a drug, device, or biological agent (FDA, 2007).

shock), I am able to estimate difference-in-difference regressions with gene-cancer and year fixed effects, and cancer-specific time trends.

4.2 Sample and Descriptive Statistics

I construct a balanced gene-cancer-year panel over the period 2004–2016, inclusive. Since my analysis begins in 2004, the publication year of the first Cancer Gene Census (the source of the cancer genes used in this analysis), and I am interested in quantifying the effect of newly disclosed scientific information (mutation disclosures) on subsequent investment, I drop the 618 gene-cancer pairs (out of 50,160 possible pairs) with known relationships as of 2004. This results in 49,542 gene-cancer pairs and 644,046 gene-cancer-year observations. Appendix Table A1 summarizes how the gene-cancer-year panel is constructed.

Summary statistics at the gene-cancer level are shown in Table 1. Panel A shows that by 2016, a mutation was identified in 58 percent of all 49,542 gene-cancer pairs. Appendix Figure A6 shows the cumulative distribution of the years in which mutations were first identified among the 168 mapping studies. The 2011 increase in the cumulative distribution reflects the disclosure of results from several cancer mapping studies that examined hundreds of tumors (as illustrated in Appendix Figure A4) and were therefore more likely to detect "rare" mutations. Consistent with these trends, Panel B shows that the median year in which mutation information was first disclosed is 2011.

Table 1 also shows that only a minority of mutations are likely cancer-causing: driver mutations are identified in only 9.5 percent of gene-cancer pairs. Panel C shows that from 2004 through 2016, 9 percent of all gene-cancers were targeted in at least one private sector phase II clinical trial. Further, the share of gene-cancer pairs in trials assessing new uses of a drug (8 percent) is higher than the share in trials testing a novel drug (5 percent).

4.3 Estimating Equation and Assumptions

4.3.1 Estimating Equation

Since the clinical trials data are highly skewed, I focus on binary outcomes that measure whether there is any clinical trial investment in a given gene-cancer-year. Thus, my empirical analysis uses variation in the timing of publicly disclosed mapping information to estimate the effect of mapping information on the likelihood of subsequent research investment within a gene-cancer pair. I estimate:

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

where Y_{gct} is an indicator for a private sector clinical trial in gene g, cancer c in year t. $Post \times DisclGeneCancer$ indicate whether gene-cancer gc has been publicly known to be mutated as of that year. $Post \times DisclGeneCancer$ varies within gene-cancers over time, and a transition from 0 to 1 represents the fact that a mutation in a gene-cancer has been publicly disclosed. I include gene-cancer fixed effects, δ_{gc}, to control for

time-invariant differences across gene-cancers, such as a gene-cancer's inherent commercial potential. Year fixed effects τ_t control for year-specific shocks that are common across gene-cancers, such as changes in technology or regulatory standards. Finally, cancer-specific time trends (or cancer-year fixed effects) $\theta_{c,t}$ control for cancer-specific changes that are common across genes within the same cancer. I perform estimates using OLS models and cluster standard errors two-way by (i) gene and (ii) cancer.[37,38]

My coefficient of interest is β. β compares the likelihood of clinical trial investments in gene-cancers for which mapping has provided evidence to those which have not yet been mapped (or will never be mapped).

4.3.2 Assumptions

A key concern is that the research potential of gene-cancer pairs that were sequenced early on in large-scale mapping efforts is significantly different from the potential of those are sequenced at a later date or never sequenced, and that those differences change over time. There are two types of potential selection. The first type of selection is at the cancer-level: large-scale cancer mapping studies (which are typically cancer-site specific) may be more likely to examine tumors that have higher ex ante expected research value. For example, the TCGA prioritized cancers with tumor samples that are more readily available, suggesting that the TCGA was directed towards more widespread cancers (and thus with larger market sizes) and the resulting estimates may be upward biased.[39,40]

I explore whether there is cancer-level selection in Appendix Figure A7 by comparing research proxies for market potential (diagnoses, drugs approvals, and trials) among cancers that were first sequenced before 2011 (the median sequencing year) and cancers that were first sequenced in/after 2011. I examine how the differences in research proxies for these two groups of cancers vary over time. While the difference in diagnoses (Panel A) remains relatively flat, the increasing difference in drug approvals (Panel B) and trials (Panel C) suggest that cancer-level selection is present. However, including cancer-specific time trends (or cancer-year fixed effects) attenuates these concerns by controlling for cancer-level secular changes.

The second type of selection is at the gene level—that is, conditional on selecting a particular cancer, researchers may choose to sequence particular genes with higher ex ante research value. Due to the mapping technology used, this is unlikely to be a major impediment: of the 168 mapping

[37] Relative to non-linear models, such as probit or logit regressions, ordinary least square regressions generate estimates that are less prone to the incidental parameters problem (Angrist and Pischke 2009).

[38] Clustering by both (i) gene and (ii) cancer allows for within-error correlation across genes and across cancers. However, this two-way clustering structure and a relative large number of fixed effects leads to a rank deficient variance covariance matrix, making an overall F-test of the joint significance of the explanatory variables infeasible. Nonetheless, the individual regression coefficients and standard errors can still be correctly interpreted (Cameron and Miller 2015). As a second best, I cluster at the gene-cancer level and find similar results.

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

[40] 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).

studies used in this analysis, 91 percent employ mapping techniques that are unbiased at the gene level in the sense that they search across all genes in the DNA to identify mutations.[41] A key benefit of this sequencing strategy is that it identifies rare gene mutations. The remaining nine percent of mapping studies use a strategy called targeted sequencing where select genes are targeted ex-ante. While gene-level selection is a concern for these studies, the relatively low number of genes this paper focuses on (627 "at risk" cancer genes) and the large number of genes examined in the targeted sequencing studies included in this paper's analysis (3,000 genes, on average) suggest that the potential bias from gene-level selection is relatively low.

4.4 Results

Table 2 documents a positive relationship between mapping information and subsequent quantity of private sector clinical trials. The first specification in Column 1 includes gene-cancer and year fixed effects. Column 2 adds cancer-specific time trends that control for cancer-specific changes over time. Column 3 controls for cancer-specific changes with cancer-year fixed effects. In all cases, I estimate a strong, positive, and statistically significant effect of cancer mapping on the probability of a private sector clinical trial in the gene-cancer after the cancer mapping information disclosing it is submitted to a scientific journal. Focusing on the set of estimates with cancer-specific time trends and cancer-year fixed effects, I find that cancer mapping information increases the probability of a clinical trial by roughly 50 percent each year relative to the pre-mapping information sample mean.

One interpretation of my findings is that if 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 up to 46 fewer private sector clinical trials at the trial level (as opposed to trial-gene-cancer). This translates into roughly seven fewer cancer drug approvals, or a 3 percent decrease from 2004 through 2016.[42]

[41] More precisely, the specific mapping strategies are whole-genome sequencing and whole-exome sequencing. Whole genome sequencing reads both protein coding and non-coding regions, while whole exome sequencing focuses on protein-coding regions.

[42] I calculate these estimates using the pre-mutation information trial averages as my counterfactual. As of 2016, there were 28,524 gene-cancers that had received mutation information (or 28,524 "mapped" gene-cancers). The likelihood of a gene-cancer pair being targeted in a trial in any given year prior to receiving mutation information is 0.017. This suggests that if the mapped gene-cancers experienced this pre-mutation information likelihood of obtaining a trial, there would be 485 ($\approx$28,524 $\times$ 0.017) trial-gene-cancer observations in each year. Mapping increases the likelihood of a trial by 0.0087 to 0.0257 (= 0.017 + 0.0087). This suggests that if the mapped gene-cancers had this likelihood of experiencing a trial, there would be 734 ($\approx$ 28,524 $\times$ 0.0257) trial-gene-cancer observations in each year. This suggests that mapping leads to a 249 (= 734 - 485) 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 1,496 (= 6 $\times$ 249) trial-gene-cancers. To convert this to the trial level, I note that trials are typically associated with 32 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 (such a trial would appear four times). Converting 1,496 trial-gene-cancer level observations to the trial level gives 46 unique trials. To obtain the estimated number of approved drugs, I take the estimated probability of successfully advancing from phase II to regulatory approval (15.2%) from Thomas et al. (2016), which results in an estimated seven cancer drug approvals.

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

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

where δ_{gc} , τ_t, and θ_{ct} represent gene-cancer fixed effects, year fixed effects, and cancer-specific time trends, 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 for a gene-cancer was disclosed).

Figure 1 presents β_z from this regression and corresponds to a dynamic version of Table 2, Column 2. The vertical lines represent 95 percent confidence intervals and the dashed red line indicates the first year in which a mutation in a gene-cancer is publicly disclosed. The figure illustrates that gene-cancers that received mutation-related information initially exhibit trends in clinical trial research similar to those gene-cancers that did not. However, the probability of a clinical trial rises differentially for gene-cancers that receive mutation-related information and remains elevated afterwards. The timing of the increase ($t = 2$) is consistent with the view that private firms may be initially testing existing drugs that have been approved or previously tested in related diseases. The lack of pre-trends suggests that private firms are not strategically withholding their clinical trial investments in anticipation of the cancer mapping information's public release. Rather, the evidence supports the view that public disclosure of mapping information is an exogenous information shock that spurs firms into increasing their clinical trial investments. Together, these estimates suggest that information from mapping efforts within a particular disease has a positive and significant impact on the subsequent level of clinical trials in the same disease.

In Appendix Table A2, I confirm that the results are robust to concurrent changes in gene-based intellectual property regulation that may influence researchers' and private firms' efforts to identify and conduct clinical trials using gene-based criteria. In the 1990s, the firm Myriad received a patent on the sequenced BRCA1 and BRCA2 genes and associated mutations (Gold and Carbone 2010). Concerned that such patent protection could limit the detection of such mutations, in 2013, the Supreme Court ruled that genes and their mutations could not be patented. Appendix Table A2 shows that the estimated magnitude is similar when excluding gene-cancer pairs with BRCA1 and BRCA2 from the analysis.

4.4.1 Heterogeneity by Clinical Relevance of Mapping Information

The previous analysis rests on the assumption that mapping information contains useful scientific information for drug developers. In this section, I examine this assumption more closely. In particular, I ask: are private firms more likely to respond to mutations that are more clinically relevant—i.e., more likely to contribute to the progression and growth of cancer?

Table 3 shows how the relationship between mapping information and trial quantity varies in response to mutation information with differing levels of clinical relevance. Specifically, using Equation 1, I estimate how investment responds to the first appearance of a driver mutation (column

1) and the first appearance of a passenger mutation (column 2). Column 1 shows that information about a driver mutation leads to a 106 percent increase in the probability of a clinical trial. In contrast, news of a passenger mutation increases the probability of a clinical trial by 31 percent. The difference in percent gains is statistically significant. These estimates support the view that private firms are more responsive to information that is more clinically relevant.

To further examine how the relationship between mapping information and private investment varies by information strength, I investigate whether information about one disease may affect research in a different but closely related disease (Henderson and Cockburn 1996; Sampat 2015). For example, small intestine and large intestine cancer are both in the same cancer site group ("digestive system"). News that the KRAS gene is mutated in small intestine cancer may indicate that KRAS mutations are also likely to occur in large intestine cancer. Appendix Table A4, Column 1 shows that clinical trial investment rises by 27 percent in response to mapping information in the same gene and a different, but closely related cancer. As expected, this effect is smaller than the direct effect of mutation information in the same disease (50 percent). However, Column 2 shows that once the regression controls for mapping information in the same disease, the additional effect of mapping information in a closely related but different disease becomes statistically insignificant.

4.4.2 Composition of Research Investments: New Uses or Novel Drugs?

The increase in the quantity of private sector clinical trials could reflect different types of innovation. First, as in the case of the PARP inhibitors described in Section 2, the increase could represent private investment in trials testing drugs that have been approved or previously tested for one disease in an additional disease (i.e., trials testing new uses).[43] Alternatively, the increase could represent trials testing drugs that have never been approved or previously tested before (i.e., trials testing novel drugs).

In theory, the relative impact of cancer mapping information on these two types of trials is ambiguous. On the one hand, a key benefit of cancer mapping is that it reveals similarities across different cancers. As a result, cancer mapping may reveal that a drug approved to treat or previously tested in one cancer may also be effective for treating other cancers. For example, 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). This, in turn, may lead to a disproportionate increase in trials testing new uses of previously tested drugs.[44] On the other hand, it's plausible that mapping information may not shift the level of investment in trials testing new uses at all. As described in the ovarian

[43] Sampat (2015) notes that research into one disease can often have value for other diseases.

[44] Indeed, this rationale has shifted public investment in clinical trials: the National Cancer Institute's Molecular Analysis for Therapy Choice trial, launched in 2015, aims to identify promising drug-gene mutation matches. This is performed through pairing patients with specific gene mutations with different drugs, regardless of cancer type.

cancer case study, manufacturers of existing drugs may do their own internal mapping. This private information may have already encouraged firms to test their drugs in multiple diseases.[45]

With this motivation, I examine how large-scale cancer genome mapping efforts influences investment in trials testing new uses of approved or previously tested drugs and those testing novel drugs. It should be noted that this comparison is primarily relevant for understanding how the composition of research shifts in the short run. Specifically, it is possible that large-scale cancer mapping spurs additional phase II clinical trials testing novel drugs, but that the effect simply takes more time to observe (relative to investment in trials testing new uses). With this caveats in mind, I estimate regressions similar to Equation 1. In this analysis, the dependent variable is set to 1 if the trial testing new uses or 0 if the trial is testing a novel drug.[46]

Table 4 examines the impact of cancer mapping, separately, for trials testing new uses and trials testing novel drugs. Mapping information increases the probability of a clinical trial that tests new uses by an average of 56 percent in each year following the information disclosure (column 1). In contrast, the probability of a clinical trial that tests novel drugs rises by an average of just 36 percent each year (column 2). The difference in the impact across the two trial types is statistically significant ($p = 0.02$).

Event studies in Appendix Figure A9 echo these findings.[47] The relative timing of phase II clinical trial investment in new uses and novel drugs is consistent with expectations about the timing of clinical investment: private firms investments in clinical trials testing new uses increases soon after the mapping information is disclosed, while investments in trials testing novel drugs take longer to respond. This is consistent with the notion that private firms investing in new use trials may be able to initiate phase II trials sooner than manufacturers of novel drugs because approved or previously tested drugs may be able to skip earlier safety trials or rely on expertise gained from previous experience with the drug.

In sum, I find that the main effect of cancer mapping on private sector trials is largely driven by trials testing new uses of approved or previously tested drugs. These findings suggest that in the short term, firms may prioritize investments in areas where they maintain a competitive advantage.

4.4.3 Heterogeneous Effects Across Firms, Diseases, and Trial Design

I next seek to characterize the types of private firms that respond, the types of diseases that give rise to more clinical trial activity, and the quality of clinical trials undertaken by the private sector in response to cancer mapping information.

[45] Additionally, manufacturers of approved drugs may decide against running an additional trial, and instead use the publicly available information to expand demand for off-label drug use.

[46] This analysis categorizes trials based on the novelty of the drug(s) being tested. As a result, the analysis uses the subset (96%) of private sector phase II trials with a listed drug intervention. 4% of private sector phase II trials have missing drug intervention data. Re-running the previous analysis using the subset of trials with drug intervention leads to similar results. See Appendix Figure A8 and Table A3.

[47] For the event study analysis, I following the main specification: I estimate OLS regressions and cluster standard errors at the gene and cancer level.

(i) *Firm Expertise:* What types of private firms respond most strongly to mutation-related information disclosures from large-scale mapping efforts? Are they firms with extensive prior research experience that have a better ability to act on the information (Nerkar and Roberts 2004; Krieger forthcoming)? Or does the publicly available information disproportionately benefit firms with less research experience (Furman, Nagler, and Watzinger 2018; Nagaraj forthcoming)? Private firms with less experience may also be more likely to be new entrants with less access to mapping technology, and therefore with no related information prior to the public disclosure of cancer mapping results.

Appendix Table A5 parses clinical trials by the experience of the sponsoring firm. I define firm experience based on the total number of research investments by a firm. I split the sample of clinical trials into two mutually exclusive categories based on whether their sponsoring firms are below or above the median of research experience in a specific research type (e.g., gene clinical trials) and time period (e.g., in the prior year).[48,49] Panel A focuses on firm experience in the year prior to the start of the focal trial and examines firm experience across three research domains: gene clinical trials (columns 1 and 2), cancer clinical trials (columns 3 and 4), and patents (columns 5 and 6).[50] Estimates indicate that the firms driving the increase in clinical trial activity following the disclosure of the cancer mapping information have relatively less research experience. In Panel B, I find similar effects when firm experience is calculated over the four years prior to the analysis period (i.e., between 2000 and 2003). Consistent with the findings of Nagaraj (forthcoming), who also examines how the effects of publicly available mapping information vary by firm size, these results support the notion that private firms with less research experience and relatively fewer resources disproportionately benefit from large-scale cancer mapping studies.

(ii) *Market Potential of Disease:* I next investigate how the response in clinical trial activity varies across diseases with varying levels of market potential. I focus on two dimensions of market potential: market size and competition. To assess how the effects vary across diseases with low and high market size, 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 A6, columns 1 and 2 indicate that cancer mapping information has similar effects on clinical trial investment across cancers, regardless of market size. Looking next to how the effect varies across diseases with low and high levels of competition, columns 3 and 4 suggest that relatively low levels of clinical

[48] When I refer to the firm-level distribution of research experience, the relevant universe to compute the median-split is not limited to the firms that sponsor the gene-cancer trials in our data. Rather, the relevant universe includes the entire set of 11,813 firms that conduct either a clinical trial in a cancer or a clinical trial in a gene between 1975 and 2019.

[49] When trials have multiple sponsors, I focus on the experience of the firm with the most experience.

[50] To provide context for the relative number of firms with extensive research experience, this classification suggests that 44 percent of firms had above-median experience in conducting cancer clinical trials in 2004.

trials experience a disproportionate increase in clinical trial investment. This supports the view that private investment responds to competition by investing in areas that provide the greatest opportunities for obtaining first-mover advantages (Rosenberg 1990). Columns 4 and 5 normalize the level of competition by dividing the average number of clinical trials by the average number of diagnoses. While the difference across low and high levels of normalized competition is not statistically significant, the magnitudes of the percentage increase in the likelihood of a clinical trial are similarly ordered.

(iii) *Trial Design Type:* One challenge with interpreting the impact of cancer mapping on total private sector phase II clinical trials investment is that private firms may alter the quality of their clinical trials in response to the information shocks, in which case the value of the clinical trials (e.g., for leading to the development of useful therapies) may vary. For example, one possibility is that in an effort to be the first to develop develop a particular drug for a specific disease, private firms may design clinical trials in a way that increases the likelihood of obtaining favorable trial results, but reduces their generalizability (Hilal, Sonbol and Prasad 2019). 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 forego a control group or may rely on a suboptimal treatment in the control group. Using the recommended trial standards from the scientific literature, I identify trials that are well designed and those that are poorly designed.[51] Appendix Table A7 shows that the effect is similar across the two trial types of trial design, suggesting that cancer mapping has little effect on the quality composition of subsequent clinical trials.

5 Effects on the Private Firms' Decision Quality

5.1 Empirical Strategy

The main results indicate that publicly available, large-scale cancer mapping efforts increases the likelihood that private firms will initiate phase II clinical trials. Importantly, private firms respond strongly to useful (clinically relevant) information and investment disproportionately increases in trials testing new uses. As discussed in the ovarian cancer case study in Section 2, a natural follow-up question is to ask is how cancer mapping information shapes the quality of private firms' termination-or-continuation decisions. To perform this analysis, I first establish patterns in phase II trial outcomes among trials initiated in gene-cancer pairs where mutation is available (hereafter, "trials with information") and those initiated in gene-cancer pairs where genetic information is not yet available (hereafter, "trials without information"). I then consider whether private firms are more likely to terminate phase II trials with weak or ambiguous clinical outcomes when genetic information is available. Finally, to assess whether mapping is associated with an increased likeli-

[51] See Appendix B for more details.

hood that private firms will make choices that meet their objectives, I consider whether drugs that are based on genetic information derived from mapping and that are chosen to advance to phase III ultimately provide better clinical outcomes.

To perform this analysis, I estimate OLS cross-sectional regressions and Cox proportional-hazard models on trial-gene-cancer level data. In this analysis, I focus on phase II and phase III trials because both trial types are relatively well reported and have standardized outcomes relative to phase I clinical trials.[52] Further, using a trial-gene-cancer dataset (as opposed to the gene-cancer-year panel used in my previous analysis), allows me to examine the relationship between mapping information and *any given* trial's likelihood of generating a promising clinical outcome or advancement rate.[53] To isolate the impact of mapping information that is most likely to impact the success of a firm's research decisions, I focus on the impact of mutations that are more likely to be clinically valuable (i.e., driver mutations).

5.2 Sample and Descriptive Statistics

To generate the trial-gene-cancer level dataset used in this analysis, I focus on phase II and phase III trials that satisfy two criteria. First, I restrict the analysis to the set of trials must be completed or terminated status.[54] Appendix Figure A10 shows trends in phase II advancement rates over time. Panel A shows that 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 (Cook et al. 2014; Peck et al. 2015). Panel B indicates that the share of advanced phase II trials that are initiated in gene-cancer pairs with mutation information increases significantly in 2011–2013, which as Appendix Figure A4 shows, is a time period in which mutation information is disclosed for a large share of gene-cancers.

Second, I restrict the analysis to trials that have available data on clinical trial outcomes. I use the most commonly measured clinical trial outcomes for each phase. For phase II trials, I use the objective response rate, or the share of the trial's patients whose tumors respond to treatment according to the trial's pre-set endpoint. For phase III trials, I use gains in overall survival, defined as the average gain in time between randomization and death for the treatment group compared with that of the control group, and widely considered as the traditional gold standard

[52] For example, a common phase II and phase III outcome is objective response rate.

[53] Specifically, I use a trial-gene-cancer dataset to avoid any compositional effects that might arise with a gene-cancer-year panel. For example, suppose that a gene-cancer-year panel is used to examine the relationship between mapping information and the likelihood that a trial demonstrates a statistically significant improvement in overall survival (i.e., is "successful"). Suppose that gene-cancers with mapping information are associated with an increased likelihood of having a successful trial or an increased number of successful trials. This result can be picking up one or two effects: first, mapping increases the likelihood of success, holding the total number of trials constant. Alternatively, mapping increases the total number of trials, holding success constant. With the caveat that the estimates are correlations, using a trial-gene-cancer dataset allows me to examine the relationship between mapping information and trial success, holding the total number of trials constant.

[54] 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."

for demonstrating the clinical benefit of a cancer drug. In total, this results in 2,354 phase II trials and 422 phase III trials, at the trial-gene-cancer level.

Table 5 describes the final trial-gene-cancer level dataset. The table describes trial outcomes, phase II to phase III advancement rates, and trial sponsor characteristics. As a proxy for the trial sponsor's R&D experience, I take the inverse hyperbolic sine of the number of clinical trials that the firm initiated in the same cancer prior to the start of the focal trial. Taking an inverse hyperbolic sine of a variable is similar to a natural logarithm transformation, but the inverse hyperbolic transformation is defined at 0 (Burbidge, Magee and Robb 1988). Table 5 shows that among the trials used in this analysis, phase II trials have advancement rates of 57 percent.[55] Phase II trials with information are significantly less likely to advance to phase III than trials without information (21 vs. 60 percent, respectively). Finally, while phase II trials with information are more likely to be associated with lower objective response rates than trials without information (15 percent vs. 19 percent, respectively), phase III trials with information are significantly more likely to demonstrate a statistically significant improvement in overall survival (69 percent vs. 52 percent, respectively).[56]

5.3 Results

5.3.1 Phase II Trial Outcomes

Before turning to the analysis of private firms' termination-or-continuation decisions, I first establish that private firms with access to genetic information are choosing among drug investments whose clinical quality is similar to those of firms without access to genetic information. Formally, I estimate the following OLS specification:

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

where Y_{igsc} is the inverse hyperbolic sine objective response rate for trial i, gene g, cancer site s, and cancer c. $Post \times DisclGeneCancer_{gc}$ is an indicator for whether information about a clinically relevant mutation is available for gene g and cancer c at least one month prior to the start of trial i. $\boldsymbol{X}_i$ is a vector of trial characteristics including the trial sponsor's R&D experience (as measured by the number of clinical trials initiated in the focal cancer in the prior year, transformed with

[55] This is higher than the most comparable estimates in Wong et al. (2018), which estimates transition rates of 39 percent. This is likely due to selective reporting of trial results: all trial-gene-cancers in my dataset are required to have information on clinical trial results. The phase II to phase III transition rates of all phase II trials (including those without clinical trial results information) is 46 percent. Private firms may be more likely to report positive clinical trial results (and therefore, trials that are more likely to advance to the next phase) to public trial registries or at ASCO. However, it is unlikely that this reporting bias is correlated with the presence of mapping information, suggesting the resulting estimates should be minimally biased.

[56] Specifically, whether the difference in the overall survival between the treatment group and the control (in the trial, or a historical control) is positive ($p < 0.05$).

the inverse hyperbolic sine transformation), disease (gene and cancer) fixed effects, and a trial start-year linear time trend.[57] Standard errors are clustered at the gene and cancer level.

Appendix Table A8 shows that phase II trials with information are not more likely to have higher objective response rates relative to phase II trials without information. The result suggests that the clinical quality of drug investments is similar across gene-cancer pairs with and without mapping information.

5.3.2 Termination Rates for Phase II Trials with Weak Outcomes

Now turning to the analysis of private firms' termination-or-continuation decisions, I examine relationship between mapping information, phase II outcomes, and phase II continuation rates. In particular, I estimate Cox proportional-hazard model regressions of the form:

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

where $h_{cf0}(t)$ is the baseline hazard rate of trial advancement, stratified by cancer and sponsoring firm f's R&D experience.[58] $LowResponseRate_i$ is an indicator for whether trial i's phase II clinical trial objective response rate is "low"—i.e., below the median of the cancer-specific response rate distribution. The coefficient γ tells us how the impact of cancer mapping information on phase II trial continuation rates changes when phase II trial has a low response rate.

I include a set of trial characteristics ($\boldsymbol{X}_i$), including the trial sponsor's R&D experience, the phase II objective response rate (transformed with the inverse hyperbolic sine transformation), and a trial start-year linear time trend. Standard errors are clustered at the gene and cancer level.

Table 6 presents the estimates. Column 1 includes only a mapping information indicator and a linear time trend, and then in columns 2 and 3, I incrementally add baseline controls. Column 3 shows that, holding phase II clinical trial outcomes constant, phase II trials with information are 64 percent less likely to advance to phase III. As expected, phase II trials with higher response rates are more likely to successfully proceed to phase III. Column 4 shows that the negative correlation between cancer mapping information and phase II trial continuation rates is largely driven by phase II clinical trials with low response rates: relative to phase II trials with information and high response rates, phase II trials with information and low response rates are 30 percent less likely to advance to phase III. In sum, the results indicate that on average, private firms with access to

[57] Due to the small sample size, gene-cancer fixed effects are not included in the analysis.

[58] In particular, as in Section 4.4.3, I define firm R&D experience based on the total number of research investments by a firm. I split the sample of clinical trials into two mutually exclusive categories based on whether their sponsoring firms are below or above the median of research experience as measured by the number of clinical trials initiated in the focal cancer in the year prior to the start of the focal trial. Testing the proportional-hazard assumption yielded non-significant results, suggesting that the proportionality assumption holds.

mapping information are more likely to terminate phase II trials with relatively weak or ambiguous trial outcomes.

5.3.3 Outcomes of Drugs that Experience Continued Investment

The previous set of results suggest that upon completing phase II trials, private firms observe where their drug falls in the drug quality distribution. Further, private firms with access to mapping information are more likely to terminate drugs whose quality falls below a certain threshold. This section asks, do drugs that fall above the quality threshold (as of phase II) and are advanced to phase III ultimately demonstrate clinical benefit based on the gold-standard measure of efficacy in cancer, overall survival?

Using a specification similar to that outlined in Equation 3, I examine whether phase III trials with information are more likely to demonstrate improvements in overall survival relative to phase III trials without information. I focus on measuring improvements in overall survival as opposed to assessing whether the drug successfully completes the phase III trial and receives approval because the timing of the mapping initiatives (the median mapping year is 2011) and relatively long length of phase III trials (up to four years) indicate that regulatory approvals are rare in my setting and data. Table 7 shows that even after controlling for disease and firm characteristics, conditional on advancing to phase III, trials with mapping information are 34 percent more likely to demonstrate a statistically significant improvement in overall survival.

Together, this analysis shows that private firms with mapping information are more likely to terminate phase II drug investments with weak clinical outcomes. Drugs advanced by private firms with access to mapping information are more likely to demonstrate improvements in clinical outcomes (and therefore, more likely receive approval). This analysis does not establish causation and is estimated on a relatively small sample size. However, the significant correlations lend a basic level of credence to the idea that when private firms have access to detailed, reliable scientific information, they make higher quality research investment decisions.

6 Conclusion

Basic scientific knowledge is widely viewed as important for stimulating subsequent innovation. This paper shows that publicly available scientific maps have important effects on the subsequent quantity of private sector innovation and private-sector firms' decision quality. I find evidence that such cancer mapping information leads to an estimated 50 percent increase in private sector clinical trials. I estimate that for the period 2004–2016, this translated into up to 46 additional clinical trials and approximately seven additional cancer drug approvals. These results are driven by response to information about mutations most likely to propel cancer, a result consistent with the prediction that mapping information helps private firms address scientific challenges, thus lowering the cost of clinical development. Further, cancer mapping significantly increases investment in trials testing approved or previously tested drugs, suggesting that one way in which large-scale scientific mapping efforts boost private innovation is through identifying clear paths across research

opportunities that were previously believed to be distantly related, and encouraging private firms to direct investment efforts in research areas where they already maintain a competitive advantage. Additionally, I analyze how basic scientific information shapes the presumptive success of firms' investment decisions. I find that private firms with access to mutation-related information are more likely to terminate phase II trials with weak outcomes, and to continue drug investments that are ultimately more likely to demonstrate promising clinical outcomes.

In addition to shaping private sector innovation through directly expanding private firms' scientific knowledge, publicly available scientific information may also shape private firms' R&D investments through influencing consumer demand. In particular, my empirical results on trials testing new uses of existing drugs suggest that cancer mapping may also affect off-label drug use among health care providers, a widespread practice that is poised to continue to grow in importance over the coming years.[59] Given the high cost of seeking drug approval, private firms may choose to invest in expanding demand for off-label drug use (e.g., through disseminating information from large-scale cancer mapping initiatives) rather than conducting trials in new uses to support subsequent regulatory approval. Future work should focus on examining how publicly available basic scientific information shapes consumers' responses and how this, in turn, affects private firms' R&D investment strategies.

More generally, this paper has implications for the role of targeted research policies in stimulating subsequent innovation. As governments consider investments to spur private R&D, understanding the effects of disease-specific investments in basic scientific knowledge is essential for structuring policy that encourages the efficient development of medical technologies. Indeed, similar disease-specific mapping efforts are underway for other conditions, including different types of brain disorders.[60] It would be interesting to know more about the welfare effects of these targeted research efforts (Myers 2020): for example, what is the aggregate effect of cancer mapping efforts on drug development? To what extent are private firms substituting away from developing treatments that are not linked to cancer or specific genes? What are the implications of targeted therapies for pricing and access to novel drugs (Bagley et al. 2019)? In sum, the size of the benefits and costs of publicly available cancer mapping information is a fruitful topic for future research.

Finally, large increases in R&D spending and persistent declines in research productivity have been widely documented across the pharmaceutical industry (Cockburn 2007; Scott Morton and Kyle 2011). This study suggests that the public provision of basic scientific data in the form of scientific maps have the potential to boost medical research productivity. Declining research productivity, however, is ubiquitous across many industries, such as computers and agriculture

[59] Off-label use is estimated to comprise approximately 50% of cancer treatments (Molitor and Agha 2012; Pfister 2012; Conti et al. 2013; Bach 2015).

[60] For example, the Alzheimer's Genome Project (https://curealz.org/the-research/areas-of-focus/alz-genome-project/), the European Human Brain Project (https://www.humanbrainproject.eu/en/), and the US BRAIN Initiative (https://www.braininitiative.nih.gov/).

(Jones 2009; Bloom et al. 2018). Future work should examine the extent to which publicly available scientific maps can help firms in these industries navigate the R&D process.

References

Acemoglu, Daron, and Joshua Linn. 2004. "Market Size in Innovation: Theory and Evidence from the Pharmaceutical Industry." *Quarterly Journal of Economics*, 119(3): 1049–90.

Aghion, Phillipe, and Peter Howitt. 1992. "A Model of Growth through Creative Destruction." *Econometrica*, 60(2): 323–51.

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, Karen, Eric D. Peterson, Asba Tasneen, James Topping, and Robert M. Califf. 2015. "Compliance with Results Reporting at ClinicalTrials.gov." *New England Journal of Medicine*, 372(11): 1031–39.

Angrist, Joshua, and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricists Companion.* Princeton, NJ:Princeton University Press.

Arora, Ashish, and Alfonso Gambardella. 1994. "Evaluating Technological Information and Utilizing It." *Journal of Economic Behavior and Organization*, 24(1): 91–114.

Arrow, Kenneth J. 1962. "Economic Welfare and the Allocation of Resources for Invention." In *The Rate and Direction of Invention Activity: Economic and Social Factors.* , ed. Committee on Economic Growth of the Social Science Research Council Universities-National Bureau of Committee for Economic Research, 609–26. Princeton, NJ:Princeton University Press. https://www.nber.org/system/files/chapters/c2144/c2144.pdf, Accessed on 2021-01-13.

AstraZeneca. 2011. "AstraZeneca Annual Report and Form 20-F Information 2011." AstraZeneca. https://www.astrazeneca.com/content/dam/az/Investor_Relations/annual-reports-homepage/AstraZeneca_AR_2017%20(1).pdf, Assessed on 2021-01-13.

AstraZeneca. 2017. "Delivering the Next Wave of Scientific Innovation: 2017 - a Year in Review." AstraZeneca. https://www.astrazeneca.com/content/dam/az/PDF/2017/IMED_Annual%20Review%202017_FINAL%20APPROVED.pdf, Assessed on 2021-01-13.

Avorn, Jerry. 2015. "The $2.6 Billion Pill—Methodologic and Policy Considerations." *New England Journal of Medicine*, 372(20): 1877–79.

Azoulay, Pierre, Joshua S. Graff Zivin, Danielle Li, and Bhaven N. Sampat. 2019. "Public R&D Investments and Private-Sector Patenting: Evidence from NIH Funding Rules." *The Review of Economic Studies*, 86(1): 117–52.

Bach, Peter. 2015. "Indication-Specific Pricing for Cancer Drugs." *JAMA*, 312(16): 1629–30.

Bagley, Nicholas, Benjamin Berger, Amitabh Chandra, Craig Garthwaite, and Ariel D. Stern. 2019. "The Orphan Drug Act at 35: Observations and an Outlook for the Twenty-First Century. Innovation Policy and the Economy." In *Innovation Policy and the Economy, vol. 19.* , ed. Josh Lerner and Scott Stern, 97–137. Chicago:The University of Chicago Press Journals.

Barker, Anna D., and Francis S. Collins. 2008. "Mapping the Cancer Genome." *Scientific American*, 296(3): 50–7.

Berndt, Ernst R., Iain M. Cockburn, and Karen A. Grepin. 2006. "The Impact of Incremental Innovation in Biopharmaceuticals." *Pharmacoeconomics*, 24(2): 69–86.

Berndt, Ernst R., Linda Bui, David R. Reiley, and Glen L. Urban. 1995. "Information, Marketing, and Pricing in the U.S. Antiulcer Drug Market." *American Economic Review*, 85(2): 100–5.

Bloom, Nicholas, Charles I. Jones, John Van Reenen, and Michael. Webb. 2018. "Are Ideas Getting Harder to Find?" National Bureau of Economic Research Working Paper 23782.

Bond, R.S., and D.F. Lean. 1977. "Sales, Promotion, and Product Differentiation in Two Prescription Drug Markets." Bureau of Economics, Fed-

eral Trade Commission. https://www.ftc.gov/sites/default/files/documents/reports/sales-promotion-and-product-differentiation-two-prescription-drug-markets/197702salespromo.pdf, Accessed on 2021-01-13.

Bryant, Helen E., Niklas Schultz, Huw D. Thomas, Kayan M. Parker, Dan Flower, Elena Lopez, Suzanne Kyle, Mark Meuth, Nicola J. Curtin, and Thomas Helleday. 2005. "Specific Killing of BRCA2-Deficient Tumors with Inhibitors of Poly (ADP-Ribose) Polymerase." *Nature*, 434(7035): 913–17.

Bujar, Magdalena, Neil McAuslane, Stuart R. Walker, and Sam Salek. 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.

Burbidge, John B., Lonnie Magee, and A. Leslie Robb. 1988. "Alternative Transformations to Handle Extreme Values of the Dependent Variable." *Journal of the American Statistical Association*, 83(401): 123–27.

Byar, David P., Richard M. Simon, William T. Friedewald, James J. Schlesselman, David L. DeMets, Jonas H. Ellenberg, Mitchell H. Gail, and James H. Ware. 1976. "Randomized Clinical Trials–Perspectives on Some Recent Ideas." *New England Journal of Medicine*, 295(2): 74–80.

Cameron, A. Colin, and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *The Journal of Human Resources*, 50(2): 317–72.

Campbell, Joshua D., Anton Alexandrov, Jaegil Kim, Jeremiah Wala, Alice H. Berger, Chandra Sekhar Pedamallu, Sachet A Shukla, Guangwu Guo, Angela N. Brooks, and Matthew Meyerson. 2016. "Distinct Patterns of Somatic Genome Alterations in Lung Adenocarcinomas and Squamous Cell Carcinomas." *Nature Genetics*, 48(6): 607–16.

Cancer Genome Atlas Research Network. 2011*a*. "Integrated Genomic Analyses of Ovarian Carcinoma." *Nature*, 474(7353): 609–15.

Cancer Genome Atlas Research Network. 2011*b*. "News Release: The Cancer Genome Atlas Completes Detailed Ovarian Cancer Analysis." https://cancergenome.nih.gov/newsevents/newsannouncements/ovarianpaper, Accessed on 2018-10-01.

Cancer Genome Atlas Research Network. 2012. "Comprehensive Genomic Characterization of Squamous Cell Lung Cancers." *Nature*, 489: 519–25.

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

Carr, T. Hedley, Robert McEwen, Brian Dougherty, Justin H. Johnson, Jonathan R. Dry, Zhongwu Lai, Zara Ghazoui, Darren R. Laing, Naomi M. Hodgson, Francisco Cruzalegui, Simon J. Hollingsworth, and J. Carl Barrett. 2016. "Defining Actionable Mutations For Oncology Therapeutic Development." *Nature Reviews Cancer*, 16(5): 319–29.

Cerami, Ethan, Jianjiong Gao, Benjamin E. Gross, Selcuk Onuer Sumer, Bulent Arman Aksoy, Anders Jacobsen, Caitlin J. Byrne, Michael L. Heuer, and Erik Larsson. 2012. "The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data." *Cancer Discovery*, 2(5): 401–4.

Chandra, Amitabh, Craig Garthwaithe, and Ariel Dora Stern. 2018. "Characterizing the Drug Development Pipeline for Precision Medicines." National Bureau of Economic Research Working Paper 24026.

Cockburn, Iain. 2007. "Is the Pharmaceutical Industry in Productivity Crisis?" In *Innovation Policy and the Economy, Volume 7.* , ed. Josh Lerner and Scott Stern, 1–32. Cambridge, MA:MIT

Press.

Cockburn, Ian, and Rebecca Henderson. 1994. "Racing To Invest? The Dynamics of Competition in Ethical Drug Discovery." *Journal of Economics and Management Strategy*, 3(3): 481–519.

Collins, Francis S. 2011. "Mining for Therapeutic Gold." *Nature Reviews Drug Discovery*, 10(6): 397.

Collins, Francis S., and McKusick. 2001. "Implications of the Human Genome Project for Medical Science." *The Journal of the American Medical Association*, 285(5): 540–44.

Comanor, William S. 1986. "The Political Economy of the Pharmaceutical Industry." *Journal of Economic Literature*, 24(3): 1178–1217.

Conti, Rena M., Arielle C. Bernstein Bernstein, Victoria M. Villaflor, Richard L. Schilsky, Meredith B. Rosenthal, and Peter B. Bach. 2013. "Prevalence of Off-Label Use and Spending in 2010 Among Patent-Protected Chemotherapies in a Population-Based Cohort of Medical Oncologists." *Journal of Clinical Oncology*, 31(9): 1134–39.

Cook, David, Dearg Brown, Robert Alexander, Ruth March, Paul Morgan, Gemma Satterwaite, and Menelas N. Pangalos. 2014. "Lessons Learned from the Fate of AstraZeneca's Drug Pipeline: A Five-Dimensional Framework." *Nature Reviews Drug Discovery*, 13(6): 419–31.

David, Paul A., Bronwyn H. Hall, and Andrew A. Toole. 2000. "Is Public R&D a Complement or Substitute for Private R&D? A Review of the Econometric Evidence." *Research Policy*, 29(4–5): 497–529.

David, Paul A., David C. Mowery, and W. Edward Steinmueller. 1992. "Analyzing the Economic Payoffs from Basic Research." *Economics of Innovation and New Technology*, 2(1): 73–90.

Dees, Nathan D., Qunyuan Zhang, Cyriac Kandoth, Michael C. Wendl, William Schierding, Daniel C. Koboldt, Thomas B. Mooney, Matthew B. Callaway, David Dooling, and Elaine R. Mardis. 2012. "MuSiC: Identifying Mutational Significance in Cancer Genomes." *Genome Research*, 22(8): 1589–98.

DiMasi, Joseph 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, Joseph A., Henry G. Grabowski, and Ronald W. Hansen. 2016. "Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs." *Journal of Health Economics*, 47: 20–33.

DiMasi, Joseph A., Ronald W. Hansen, and Henry G. Grabowski. 2003. "The Price of Innovation: New Estimates of Drug Development Costs." *Journal of Health Economics*, 22(2): 151–85.

Donelan, Ronan, Stuart Walker, and Sam Salek. 2015. "Factors Influencing Quality Decision-Making: Regulatory and Pharmaceutical Industry Perspectives." *Pharmacoepidemiology and Drug Safety*, 24(3): 319–28.

Dougherty, Brian A., Zhongwu Lai, Darren R. Hodgson, Maria C.M. Orr, Matthew Hawryluk, James Sun, Roman Yelensky, Stuart K. Spencer, Jane D. Robertson, and J. Carl Barrett. 2012. "Biological and Clinical Evidence for Somatic Mutations in BRCA1 and BRCA2 as Predictive Markers for Olaparib Response in High-Grade Serous Ovarian Cancers in the Maintenance Setting." *Genome Research*, 22(8): 1589–98.

Dubois, Pierre, Olivier de Mouzon, Fiona Scott-Morton, and Paul Seabright. 2015. "Market Size and Pharmaceutical Innovation." *The RAND Journal of Economics*, 46(4): 844–

71.
Dulbecco, Renato. 1986. "A Turning Point in Cancer Research: Sequencing the Human Genome." *Science*, 231(4742): 1055–56.
Eisenberg, Rebecca S. 2005. "The Problem of New Uses." *Yale Journal of Health Policy, Law, and Ethics*, 5(2): 717–40.
Fabrizio, Kira R. 2009. "Absorptive Capacity and the Search for Innovation." *Research Policy*, 38(2): 255–67.
Farmer, Hannah, Nuala McCabe, Christopher J. Lord, Andrew N.J. Tutt, Damian A. Johnson, Tobias B. Richardson, Manuela Santarosa, Krystyna J. Dillon, Ian Hickson, and Alan Ashworth. 2005. "Targeting the DNA Repair Defect in BRCA Mutant Cells as a Therapeutic Strategy." *Nature*, 434(7035): 917–21.
Fleming, Lee, and Olav Sorenson. 2003. "Navigating the Technological Landscape of Innovation." *MIT Sloan Management Review*, 44(2): 15–23.
Fleming, Lee, and Olav Sorenson. 2004. "Science as a Map in Technological Search." *Strategic Management Journal*, 25(8–9): 909–28.
Furman, Jeffrey L., and Scott Stern. 2011. "Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research." *American Economic Review*, 101(5): 1933–63.
Furman, Jeffrey L., Markus Nagler, and Martin Watzinger. 2018. "Disclosure and Subsequent Innovation: Evidence from the Patent Depository Library Program." National Bureau of Economic Research Working Paper 24660.
Futureal, P. Andrew, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R. Stratton. 2004. "A Census of Human Cancer Genes." *Nature Reviews Cancer*, 4(3): 177–83.
Gao, Jianjion, Bulent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S. Our Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, Ethan Cerami, Chris Sander, and Nikolaus Schultz. 2013. "Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal." *Science Signaling*, 6(269): I1.
Gold, E. Richard, and Julia Carbone. 2010. "Myriad Genetics: In the Eye of the Policy Storm." *Genetics in Medicine*, 12(4 Suppl): S39–S70.
Guedj, Ilan, and David Scharfstein. 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, Bronwyn, and John Van Reenen. 2000. "How Effective Are Fiscal Incentives for R&D? A Review of the Evidence." *Research Policy*, 29(4–5): 449–69.
Henderson, Rebecca, and Iain Cockburn. 1996. "Scale, Scope, and Spillovers: The Determinants of Research Productivity in Drug Discovery." *RAND Journal of Economics*, 27(1): 32–59.
Hermosilla, Manuel, and Jorge Lemus. 2017. "Therapeutic Translation of Genomic Science: Opportunities and Limitations of GWAS." In *Economic Dimensions of Personalized and Precision Medicine.* , ed. Ernst R. Berndt, Dana P. Goldman and John W. Rowe, 21–52. Chicago::University of Chicago Press.
Heron, M. 2018. "Deaths: Leading Causes for 2016." *National Vital Statistics Reports*, 67(6): 1–77.
Hilal, Talal, Mohamad Bassam Sonbol, and Vinay Prasad. 2019. "Analysis of Control Arms Quality in Randomized Clinical Trials Leading to Anticancer Drug Approval by the US Food and Drug Administration." *JAMA Oncology*.
Hill, Ryan, and Carolyn Stein. 2020. "Scooped! Estimating Rewards for Priority in Science."

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, Sebastian. 2018. "Scientific Maps and Innovation: Impact of the Human Genome on Drug Discovery." PhD diss., Rutgers University.

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

Klevorick, Alvin, Richard Levin, Richard Nelson, and Sidney Winter. 1995. "On the Sources and Significance of Interindustry Differences in Technological Opportunities." *Research Policy*, 24(2): 185–205.

Kolata, Gina. 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, Joshua L. Forthcoming. "Trials and Terminations: Learning from Competitors' R&D Failures." *Management Science*.

Lander, Eric S. 2011. "Initial Impact of the Sequencing of the Human Genome." *Nature*, 470: 187–97. 7333.

Lawrence, Michael S., Petar Stojanov, Paz Polak, Gregory V. Kryukov, Kristian Cibulskis, Andrey Sivanchenko, Scott L. Carter, Chip Stewart, Craig H. Mermel, and Steven A. Roberts. 2014. "Mutational Heterogeneity in Cancer and the Search for New Cancer Genes." *Nature*, 499(7457): 214–18.

Lendrem, Dennis W., Clare Lendrem, Richard W. Peck, Stephen J. Senn, Simon Day, and John D. Isaacs. 2015. "Progression-Seeking Bias and Rational Optimism in Research and Development." *Nature Reviews Drug Discovery*, 14: 219–21.

Lieberman, Marvin B., and David B. Montgomery. 1988. "First-Mover Advantages." *Strategic Management Journal*, 9: 41–58. S1.

Lijima, Moito, Kouji Banno, Ryuichiro Okawa, Megumi Yanokura, Miho Lida, Takashi Takeda, Haruko Kunitomi-Irie, Masataka Adachi, Kanako Nakamura, and Daisuke Aoki. 2017. "Genome-wide Analysis of Gynecologic Cancer: The Cancer Genome Atlas in Ovarian and Endometrial Cancer." *Oncology Letters*, 13(3): 1063–70.

Makri, Marianna, Michael A. Hitt, and Peter J. Lane. 2010. "Complementary Technologies, Knowledge Relatedness, and Invention Outcomes in High Technology Mergers and Acquisitions." *Strategic Management Journal*, 31(6): 602–28.

Mardis, Elaine R. 2018. "Insights from Large-Scale Cancer Genome Sequencing." *Annual Reviews of Cancer Biology*, 2: 429–44.

Matulonis, Ursula A. 2017. "PARP Inhibitors in BRCA-Related Ovarian Cancer–and Beyond!"

McShane, Lisa M., Sally Hunsberger, and Alex A. Adjei. 2009. "Effective Incorporation of Biomarkers into Phase II Trials." *Clinical Cancer Research*, 15(6): 1898–1905.

Molitor, David, and Leila Agha. 2012. "Technology Adoption Under Uncertainty: Off-label Prescribing in Cancer Care." PhD diss. chapter, Massachusetts Institute of Technology.

Morgan, Paul, Dean G. Brown, Simon Lennard, Mark J. Anderton, J. Carl Barrett, Ulf Eriksson, Mark Fidock, Bengt Hamren, Anthony Johnson, and Menelas Pangalos. 2018. "Impact of a Five-Dimensional Framework on R&D Productivity at AstraZeneca." *Nature Reviews Drug Discovery*, 17(3): 167–181.

Mowery, David, and Nathan Rosenberg. 1979. "The Influence of Market Demand Upon Innovation: A Critical Review of Some Recent Empirical Studies." *Research Policy*, 8(2): 102–153.

Myers, Kyle. 2020. "The Elasticity of Science." *American Economic Journal: Applied Economics*, 12(4): 103–134.

Nagaraj, Abhishek. Forthcoming. "The Private Impact of Public Maps: Landsat Satellite Imagery and Gold Exploration." *Management Science*.

Nagaraj, Abhishek, and Scott Stern. 2020. "The Economics of Maps." *Journal of Economic Perspectives*, 34(1): 196–221.

Nelson, Richard R. 1959. "The Simple Economics of Basic Scientific Research." *The Journal of Political Economy*, 67(3): 297–306.

Nelson, Richard R. 1982. "The Role of Knowledge in R&D Efficiency." *Quarterly Journal of Economics*, 97(3): 453–70.

Nerkar, Atul, and Peter W. Roberts. 2004. "Technological and Product-Market Experience and the Success of New Product Introductions in the Pharmaceutical Industry." *Strategic Management Journal*, 25(8–9): 779–99.

Peck, Richard W., Dennis W. Lendrem, Iain Grant, B. Clare Lendrem, and John D. Isaacs. 2015. "Why Is It Hard to Terminate Failing Projects in Pharmaceutical R&D?" *Nature Reviews Drug Discovery*, 14(10): 663–64.

Pfister, David G. 2012. "Off-Label Use of Oncology Drugs: The Need for More Data and Then Some." *Journal of Clinical Oncology*, 30(6): 584–86.

Prayle, Andrew P., Matthew N. Hurley, and Alan R. Smythe. 2012. "Compliance with Mandatory Reporting of Clinical Trial Results on ClinicalTrials.gov: Cross Sectional Study." *BMJ*, 344.

Robertson, A. Gordon, Jaegil Kim, Hikmat Al-Ahmadie, Joaquim Bellmunt, Guangwu Guo, Andrew D. Cherniack, Toshinori Hinoue, Peter W. Laird, Katherine A. Hoadley, and Seth P. Lerner. 2017. "Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer." *Cell*, 171(3): 540–56.

Roin, Benjamin. 2013. "Solving the Problem of New Uses." https://ssrn.com/abstract=2337821, Accessed on 2021-01-13.

Romer, Paul M. 1990. "Endogenous Technological Change." *Journal of Political Economy*, 98(5): S71–S102.

Rosenberg, Nathan. 1974. "Science, Invention, and Economic Growth." *The Economic Journal*, 84(333): 90–108.

Rosenberg, Nathan. 1990. "Why Do Firms Do Basic Research (With Their Own Money)?" *Research Policy*, 19(2): 165–74.

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

Scherer, F.M. 2000. "Markets for Pharmaceutical Products." In *Handbook of Health Economics, vol 1.* , ed. Anthony J. Culyer and Joseph P. Newhouse, 1297–336. Amsterdam:Elsevier Science.

Schmalensee, Richard. 1982. "Product Differentiation Advantages of Pioneer Brands." *The American Economic Review*, 72(3): 349–65.

Scott Morton, Fiona, and Margaret Kyle. 2011. "Markets for Pharmaceutical Products." In *Handbook of Health Economics, vol 2.* , ed. Mark V. Pauly, Thomas G. Mcguire and Pedro P. Barros, 763–823. Amsterdam:Elsevier Science.

Sharpe, Paul, and John Keelin. 1998. "How Smithkline Beecham Makes Better Resource-Allocation Decisions." *Harvard Business Review*, 76(2).

Solow, Robert. 1957. "Technical Change and the Aggregate Production Function." *Review of Economics and Statistics*, 39(3): 312–20.

Spetzler, Carl, Hannah Winter, and Jennifer Meyer. 2016. *Decision Quality: Value Creation from Better Business Decisions.* Hoboken, NJ:John Wiley and Sons.

Stratton, Michael R., Peter J. Campbell, and P. Andrew Futreal. 2009. "The Cancer Genome." *Nature*, 458(7239): 719–24.

Tate, John G., Sally Bamford, Harry C. Jubb, Zbyslaw Sandra, David M. Beard, Nidhi Bindal, Harry Boutselakis, Charlotte G. Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C. Jute, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C. Ramshaw, Claire E. Rye, Helen E. Speedy, Ray Stefancsik, Sam L. Thompson, Shicai Wang, Sari Ward, Peter J. Campbell, and Simon A. Forbes. 2018. "COSMIC: The Catalogue of Somatic Mutations." *Nucleic Acids Research*, 47(D1): D941–47.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgement under Uncertainty: Heuristics and Biases." *Science*, 185(4157): 1124–31.

United States, Office of the Press Secretary. 2000. "Remarks Made by the President, Prime Minister Tony Blair of England (via satellite), Dr. Francis Collins, Director of the National Human Genome Research Institute, and Dr. Craig Venter, President and Chief Scientific Officer, Celera Genomics Corporation, 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." 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." 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." https://www.fda.gov/media/71627/download, Accessed on 2021-01-13.

Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz Jr., and Kenneth W. Kinzler. 2013. "Cancer Genome Landscapes." *Science*, 339(6127): 1546–58.

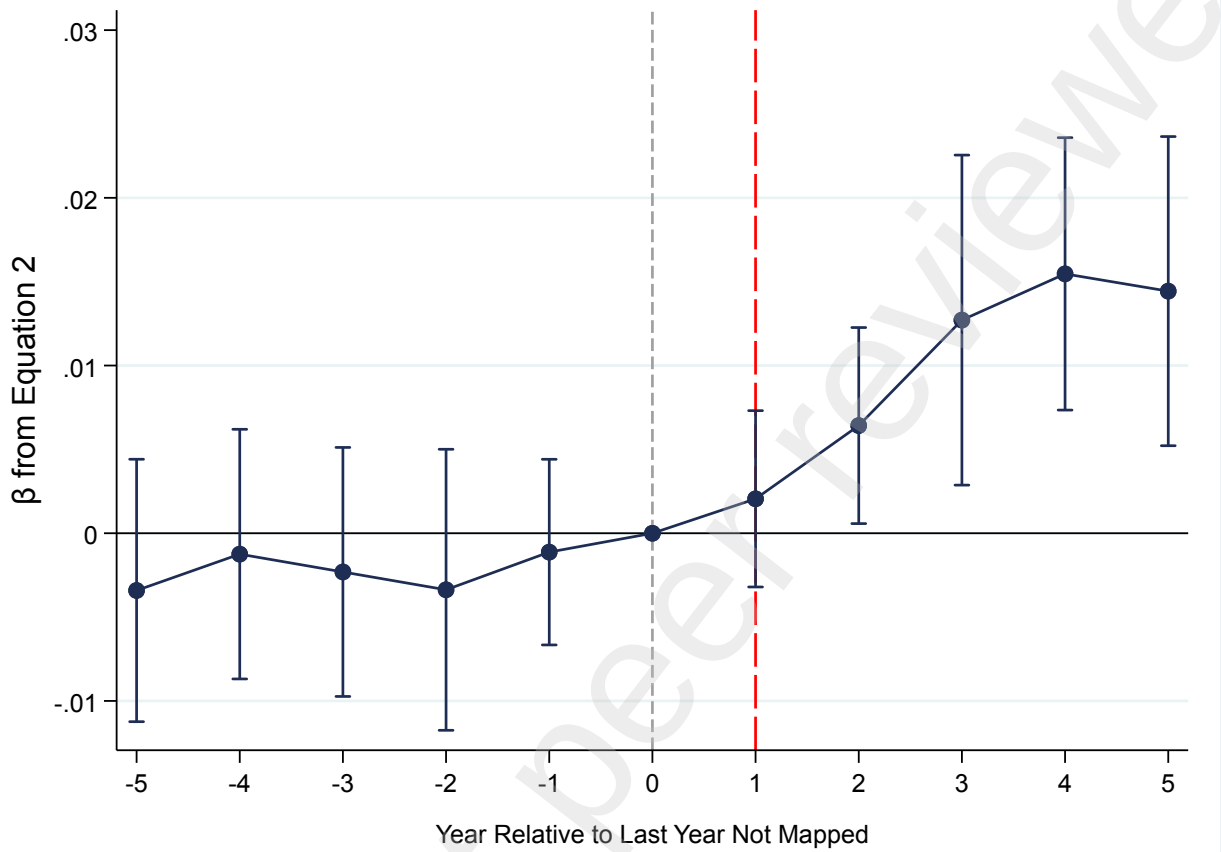
Wetterstrand, Kris. 2018. "DNA Sequencing Costs: Data." National Human Genome Research Institute. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data., Accessed on 2021-01-13.

Wheeler, David A., and Linghua Wang. 2013. "From Human Genome to Cancer Genome: The First Decade." *Genome Research*, 23(7): 1054–62.

Williams, Heidi. 2013. "Intellectual Property Rights and Innovation: Evidence from the Human Genome." *Journal of Political Economy*, 121(1): 1–27.

Yang, Yadong, Xundong Dong, Bingbing Xie, Nan Ding, Juan Chen, Yongjun Li, Qian Zhang, Hongzhu Qu, and Xiangdong Fang. 2015. "Databases and Web Tools for Cancer Genomics Study." *Genomics Proteomics Bioinformatics*, 13(1): 46–50.

Figure 1 – Effect of Cancer Mapping Information on
Private Sector Clinical Trials, 2004–2016



Notes: This figure plots the response of private sector clinical trials following the public release of cancer mapping information. 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 a cancer mapping study. The dashed red line indicates the first year that a mutation in a gene-cancer is publicly disclosed by a cancer mapping study. Shown are 95 percent 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, and the sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004 through 2016. The figure corresponds to a dynamic version of the specification in Table 2, column 2. Controls include gene-cancer fixed effects, year fixed effects, and cancer-specific time trends.

Table 1 – Summary Statistics: Gene-Cancer Level Data, 2004–2016

	Mean	Median	Standard Deviation	Minimum	Maximum
A. Mapping Information					
Share with mutation: all mutations (%)	57.58	100.00	49.42	0	100
Share with mutation: driver mutations (%)	9.48	0.00	29.29	0	100
B. Mapping Information Timing					
Year first mutation: all mutations	2011.36	2011.00	1.26	2004	2016
Year first mutation: driver mutations	2012.10	2012.00	1.23	2008	2016
C. Clinical Trials					
Any trial (%)	8.99	0.00	28.60	0	100
Any trial testing new uses (%)	7.73	0.00	26.71	0	100
Any trial testing novel drugs (%)	5.38	0.00	22.56	0	100

Notes: This table shows summary statistics at the gene-cancer level. There are 49,542 gene-cancer pairs in this sample. All clinical trials are private sector phase II clinical trials. "Any trial" denotes the share of gene-cancer pairs that have at least one clinical trial. See Section 3 and Appendix B for more detailed data and variable descriptions.

Table 2 – Effect of Cancer Mapping Information on Private Sector Clinical Trials, 2004–2016

	Dependent Variable: Any Private Sector Phase II Clinical Trial		
	(1)	(2)	(3)
Post × DisclGeneCancer	0.00585**	0.00874***	0.00915**
	(0.00173)	(0.00255)	(0.00302)
Mean of dep. var.	0.017	0.017	0.017
Percentage gain	34.55%	51.63%	54.02%
Gene-cancer FEs	Yes	Yes	Yes
Year FEs	Yes	Yes	No
Cancer-specific time trends	No	Yes	No
Cancer × Year FEs	No	No	Yes
Observations	644,046	644,046	644,046

Notes: This table reports difference-in-differences estimates of the effect of cancer mapping information on clinical trials by private sector firms. The level of observation is the gene-cancer-year. The sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004 through 2016. Estimates are from OLS models. The outcome variable switches from 0 to 1 if a private sector phase II clinical trial is reported in a gene-cancer-year. "Post × DisclGeneCancer" switches from 0 to 1 when a mutation in a gene-cancer is publicly disclosed by a cancer 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 before the first disclosure of a mutation and is used to calculate "Percentage gain," the percentage change in the likelihood of a clinical trial. See Section 3 and Appendix B for more detailed data and variable descriptions.
*p <0.10, **p <0.05, ***p <0.01.

Table 3 – Effect on Private Sector Clinical Trials:
Heterogeneity by Clinical Relevance of Cancer Mapping Information, 2004–2016

	Dependent Variable: Any Private Sector Phase II Clinical Trial	
	Driver Mutation (High Clinical Relevance)	Passenger Mutation (Low Clinical Relevance)
	(1)	(2)
Post × DisclGeneCancer	0.0392***	0.00511***
	(0.00228)	(0.000853)
Mean of dep. var.	0.037	0.016
Percentage gain	106.06%	30.96%
Gene-cancer FEs	Yes	Yes
Year FEs	Yes	Yes
Cancer-specific time trends	Yes	Yes
Observations	644,046	644,046
Diff. Wald test p-value	0.00	

Notes: This table reports difference-in-differences estimates of the effect of cancer mapping information on clinical trials by private sector firms, separately for mapping information of high and low clinical relevance. The level of observation is the gene-cancer-year. The sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004 through 2016. The outcome variable switches from 0 to 1 if a private sector phase II clinical trial is reported in a gene-cancer-year. "Post × DisclGeneCancer" switches from 0 to 1 when a mutation in a gene-cancer is publicly disclosed by a cancer 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 ($\geq$ 10 patients). All remaining mutations are classified as passenger mutations. Column 2 shows the effect of the first passenger mutation in a gene-cancer. Controls include gene-cancer fixed effects, year fixed effects, and cancer-specific time trends. "Mean of dep. var." is the mean of the outcome variable in a gene-cancer before the first disclosure of a mutation and is used to calculate "Percentage gain," the percentage change in the likelihood of a clinical trial. Estimates are from seemingly unrelated models, which permits a comparison of "Percentage gain" across models. However, the use of seemingly unrelated regressions makes direct two-way clustering by gene and cancer infeasible. As a second best, I cluster at the gene-cancer level. The p-value is from a Wald test that compares the differences in "Percentage gain." See Section 3 and Appendix B for more detailed data and variable descriptions.
*p <0.10, **p <0.05, ***p <0.01.

Table 4 – Effect on Private Sector Clinical Trials Testing New Uses and Novel Drugs, 2004–2016

	Dependent Variable: Any Private Sector Phase II Clinical Trial	
	New Uses	Novel Drugs
	(1)	(2)
Post × DisclGeneCancer	0.00667***	0.00249***
	(0.000796)	(0.000585)
Mean of dep. var.	0.012	0.007
Percentage gain	55.73%	36.41%
Gene-cancer FEs	Yes	Yes
Year FEs	Yes	Yes
Cancer-specific time trends	Yes	Yes
Observations	644,046	644,046
Diff. Wald test p-value	0.02	

Notes: This table reports difference-in-differences estimates of the effect of cancer mapping information on clinical trials by private sector firms, separately for trials new uses and trials testing novel drugs. The level of observation is the gene-cancer-year. The sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004 through 2016. Column 1 examines the effect of cancer mapping information on clinical 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 and tested in any gene-cancer pair. "Post × DisclGeneCancer" switches from 0 to 1 when a mutation in a gene-cancer is publicly disclosed by a cancer mapping study. Controls include gene-cancer fixed effects, year fixed effects, and cancer-specific time trends. "Mean of dep. var." is the mean of the outcome variable in a gene-cancer before the first disclosure of a mutation and is used to calculate "Percentage gain," the percentage change in the likelihood of a clinical trial. Estimates are from seemingly unrelated models, which permits a comparison of "Percentage gain" across models. However, the use of seemingly unrelated regressions makes direct two-way clustering by gene and cancer infeasible. As a second best, I cluster at the gene-cancer level. The p-value is from a Wald test that compares the differences in "Percentage gain." See Section 3 and Appendix B for more detailed data and variable descriptions.
*p <0.10, **p <0.05, ***p <0.01.

Table 5 – Summary Statistics: Trial-Gene-Cancer Level Data, 2004–2016

	Full (1)	Trials With Info (2)	Trials With No Info (3)	Difference (2)-(3) (4)
A. Phase II (N = 2,354)				
Trial outcome: response rate	18.98	15.32	19.25	-3.93**
1(Advance to phase III)	0.57	0.21	0.60	-0.39***
1(Advance to phase III, within 4 years)	0.57	0.20	0.60	-0.39***
Firm Experience (No. of clinical trials)	58.47	94.13	55.83	38.30***
B. Phase III (N = 422)				
1(Improvement in overall survival)	0.54	0.69	0.52	0.17**
Firm experience (No. of clinical trials)	16.99	21.34	16.42	4.92

Notes: This table shows summary statistics at the trial-gene-cancer level. The sample includes all private sector phase II trial-gene-cancer observations associated with phase II clinical trials that began between 2004 and 2016 (excluding gene-cancer pairs known in 2004), made clinical outcomes data available, and were completed or terminated as of July 14, 2017. Column 2 describes trials initiated in gene-cancer pairs where driver (clinically relevant) mutation information was publicly available before the start of the trial. Column 3 describes trials initiated in gene-cancer pairs where driver (clinically relevant) mutation information was not yet available at the start of the trial. Column 4 shows the difference in means. "Trial outcome: response rate" denotes a phase II trial's objective response rate. "**1**(Advance to phase III)" is an indicator variable for a phase II clinical trial's drug that is subsequently tested in a phase III clinical trial. Similarly, "**1**(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 clinical trial start date. "Firm experience" is the number of clinical trials that the focal trial sponsor has conducted in the focal cancer within one year prior to the clinical trial start date. "**1**(Improvement in overall survival) is an indicator variable for a phase III clinical trial whose treatment group demonstrates a statistically significant ($p < 0.05$) improvement in overall survival relative to the clinical trial's control group or a historical control. See Sections 3 and 5 and Appendix B for more detailed data and variable descriptions.
*p <0.10, **p <0.05, ***p <0.01.

Table 6 – Cancer Mapping Information and
Phase II Clinical Trial Transitions, 2004–2016

	Dependent Variable: Advancing from Phase II to Phase III			
	(1)	(2)	(3)	(4)
Post × DisclGeneCancer	-1.039*	-1.033**	-1.143**	-0.664
	(0.532)	(0.523)	(0.542)	(0.590)
Firm experience (No. of clinical trials)		-0.157	-0.0969	-0.0984
		(0.203)	(0.201)	(0.217)
Response rate			0.171*	0.178
			(0.0982)	(0.144)
Post × DisclGeneCancer × Low response rate				-0.891*
				(0.456)
Low response rate				0.0459
				(0.328)
Percentage change				
All trials	-64.62%	-64.41%	-68.10%	
Trials with a low response rate				-78.89%
Trials with a high response rate				-48.52%
Linear time trend	Yes	Yes	Yes	Yes
No. of observations	2,354	2,354	2,354	2,354
No. of trials	164	164	164	164
No. of genes	92	92	92	92
No. of cancers	80	80	80	80
Log Likelihood	-3537	-3528	-3501	-3498

Notes: This table shows the relationship between cancer mapping information and phase II-to-phase III transition rates. The level of observation is the trial-gene-cancer. The sample includes all private sector phase II trial-gene-cancer observations associated with phase II clinical trials that began between 2004 and 2016 (excluding gene-cancer pairs known in 2004), 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 large firm status. "Post × DisclGeneCancer" is an indicator for the disclosure of a driver (clinically relevant) mutation in a gene-cancer by the start of the clinical trial. "Firm experience" is the number of clinical trials that the focal trial sponsor has conduced in the focal cancer within one year prior to the phase II clinical trial start date. "Response rate" is the phase II clinical trial's objective response rate, or the share of patients who respond to treatment. Both "Firm experience" and "Response rate" are transformed with the inverse hyperbolic sine transformation. Controls include a linear time trend. Robust standard errors, clustered at the gene and cancer level, are shown in parentheses. See Sections 3 and 5 and Appendix B for more detailed data and variable descriptions.
*p <0.10, **p <0.05, ***p <0.01.

Table 7 – Cancer Mapping Information and Phase III Clinical Trial Outcomes, 2004–2016

	Dependent Variable: **1**(Improvement in Overall Survival)	
	(1)	(2)
Post × DisclGeneCancer	0.183**	0.185**
	(0.0767)	(0.0749)
Firm Experience (No. of Clinical Trials)		0.0303
		(0.0734)
Mean of dep. var.	0.540	0.540
Percentage gain	33.85%	34.30%
Gene FEs	Yes	Yes
Cancer FEs	Yes	Yes
Linear time trend	Yes	Yes
No. of trial-gene-cancers	394	394
No. of trials	71	71
No. of genes	31	31
No. of cancers	31	31
Adjusted R^2	0.297	0.298

Notes: This table shows the relationship between cancer mapping information and phase III clinical outcomes, as measured by whether the phase III clinical trial's treatment group demonstrates a significant improvement in overall survival over the control group. The level of observation is the trial-gene-cancer. The sample includes all private sector phase III trial-gene-cancers that began between 2004 and 2016 (excluding gene-cancer pairs known in 2004), made clinical outcomes data available, and were completed or terminated as of July 14, 2017. There are fewer than 422 observations because the estimation drops trial-gene-cancer observations with a gene or cancer that just shows up only once. Estimates are from OLS models. The outcome variable is an indicator variable for a phase III clinical trial whose treatment group demonstrates a statistically significant ($p < 0.05$) improvement in overall survival relative to the trial's control group or a historical control. "Post × DisclGeneCancer" is an indicator for the disclosure of a driver (clinically relevant) mutation in a gene-cancer by the start of the clinical trial. "Firm Experience" is the number of clinical trials that the focal trial sponsor has conduced in the focal cancer, within one year prior to the phase III clinical trial start date, and is transformed with the inverse hyperbolic sine transformation. Controls include cancer fixed effects and gene fixed effects. Robust standard errors, clustered at the gene and cancer level, are shown in parentheses. See Sections 3 and 5 and Appendix B for more detailed data and variable descriptions.
*p <0.10, **p <0.05, ***p <0.01.

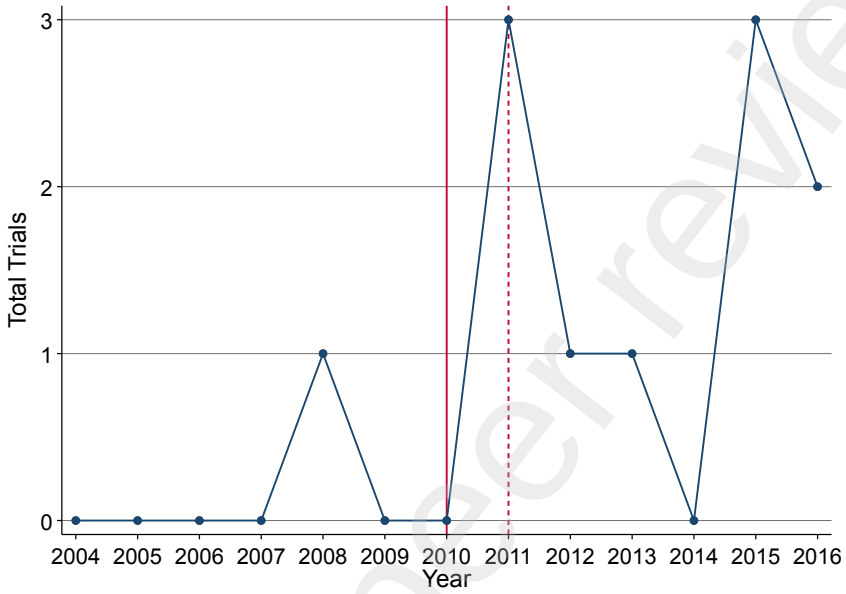
Charted Territory:
Evidence from Mapping the Cancer Genome and R&D Decisions in the Pharmaceutical Industry

Appendix

Jennifer Kao
UCLA
Anderson School of Management
110 Westwood Plaza—D510
Los Angeles, CA 90095

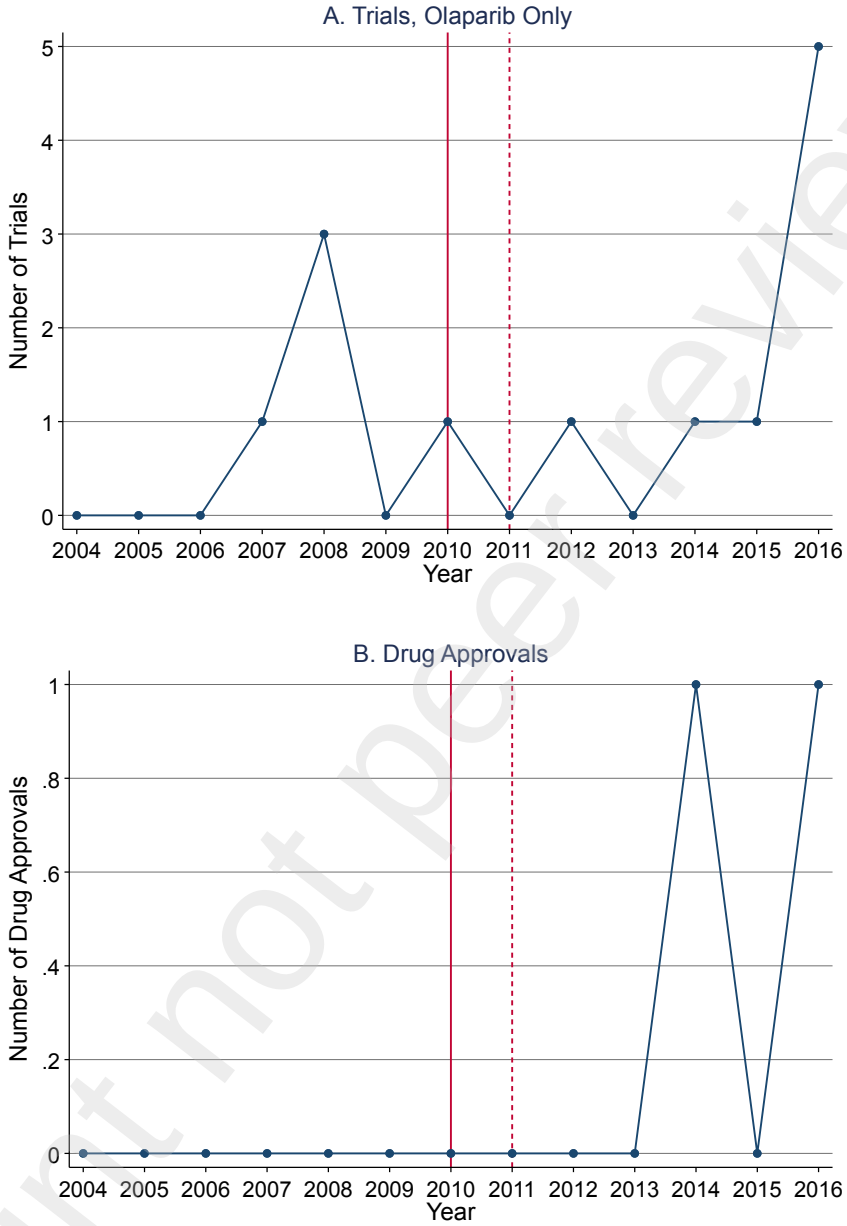
Appendix A.
Additional Figures and Tables

Figure A1 – Private Sector Trials Enrolling Patients with Ovarian Cancer and BRCA2 Gene Mutations, 2004–2016



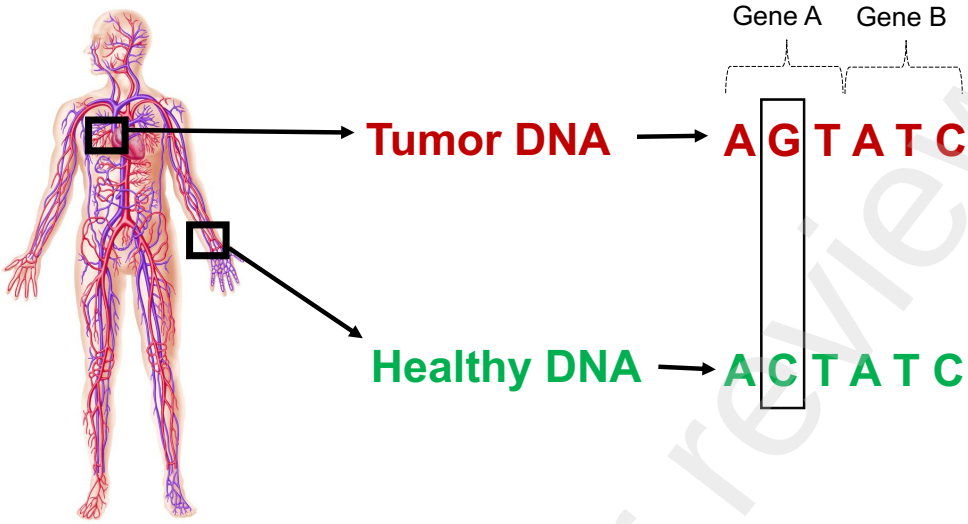
Notes: This figure shows the total number of private sector phase II clinical trials enrolling patients with BRCA2-mutated ovarian cancer from 2004 to 2016. The vertical lines indicated the years in which the TCGA's ovarian cancer study (TCGA 2011) was submitted to (solid line) and published in (dashed line) the journal *Nature*. For simplicity, trials testing the drug Olaparib are omitted (see Appendix Figure A2 for trends in clinical trials testing Olaparib). Since this figure excludes trials testing Olaparib, the figure is generated from the sample of trials with non-missing intervention data.

Figure A2 – Private Sector Trials Testing Olaparib and Drug Approvals for Patients with Ovarian Cancer and BRCA2 Gene Mutations, 2004–2016



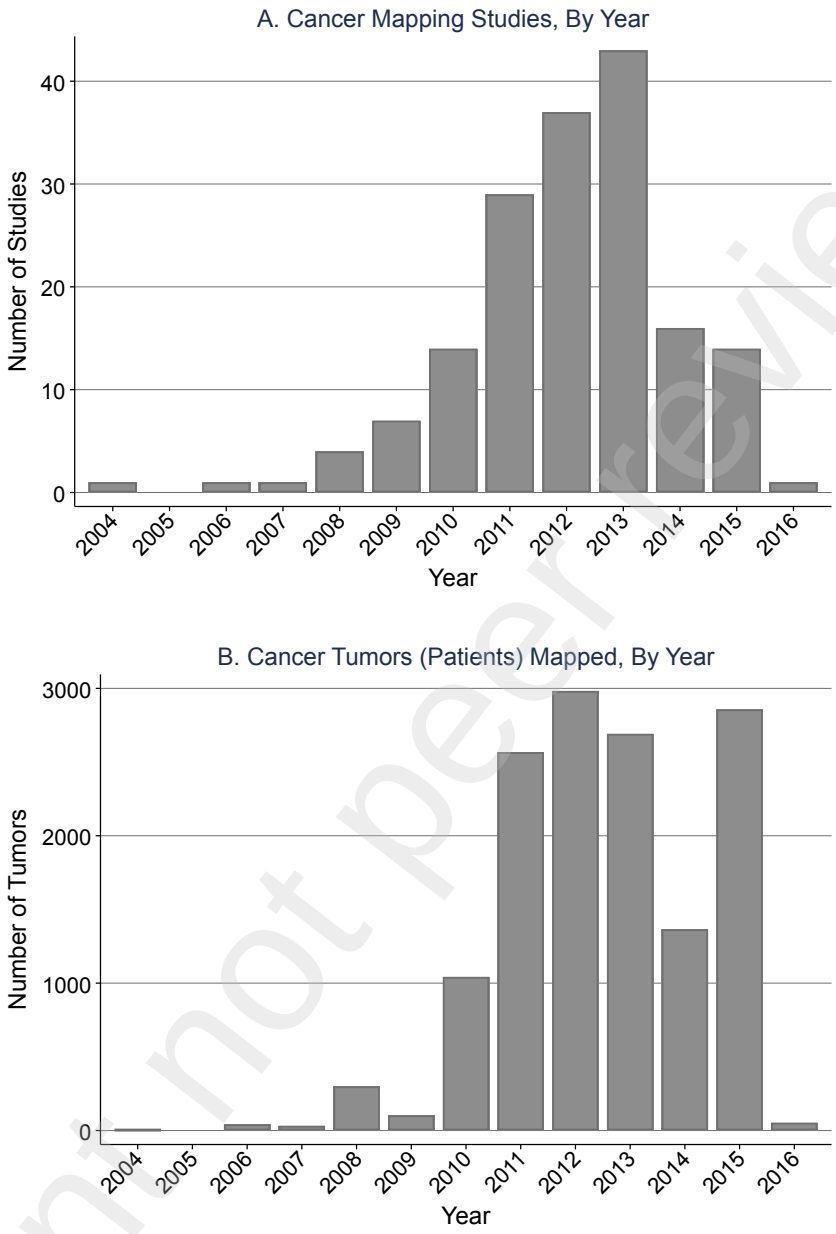
Notes: This figure examines the clinical trials testing Olaparib and drug approvals for the treatment of patients with BRCA2-mutated ovarian cancer in each year from 2004 through 2016. Panel A shows the total number of clinical trials (private sector, phase II only) enrolling patients with BRCA2-mutated ovarian cancer and testing Olaparib. Since it focuses on trials testing Olaparib, the figure is generated from the sample of trials with non-missing intervention data. Panel B shows the total number of drugs approved to treat patients with BRCA2-mutated ovarian cancer. The drug approvals refer to two PARP inhibitors: Olaparib (approved in 2014) and Rucaparib (approved in 2016). The vertical lines indicated the years in which the TCGA's ovarian cancer study (TCGA 2011) was submitted to (solid line) and published in (dashed line) the journal *Nature*.

Figure A3 – Overview of Scientific Background on Cancer Genome Sequencing



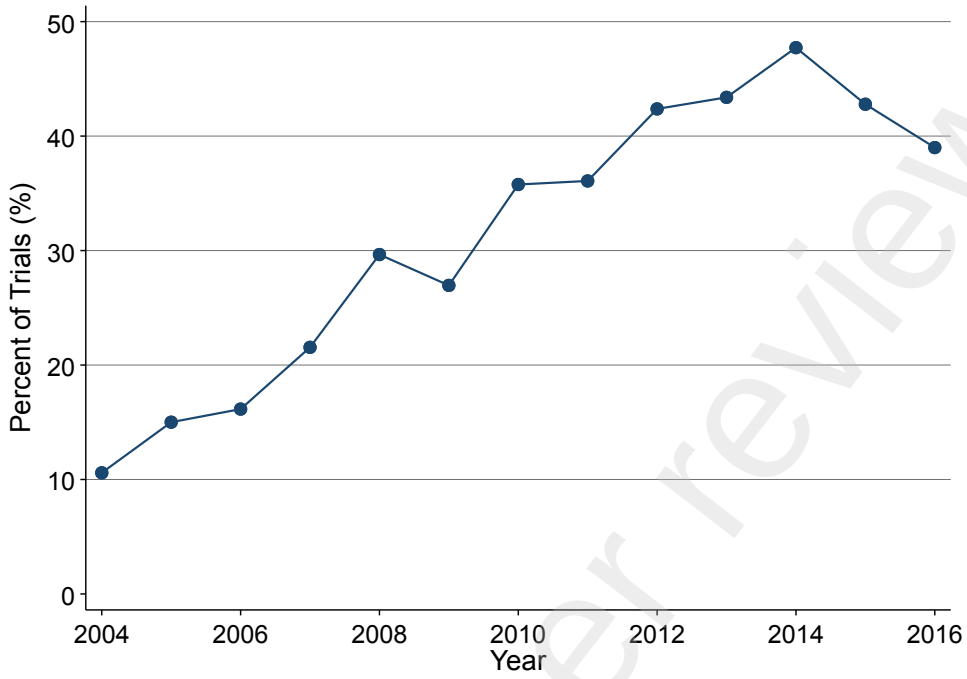
Notes: This figure graphically summarizes the scientific background described in Section 3. An individual's genome is the complete set of DNA found in each cell. DNA is comprised of four bases : adenine (A), cytosine (C), guanine (G), and thymine (T). The unique sequence of these four DNA bases—A, C, G, and T—provides a "blueprint" for the human body (GeneEd: Genetics, Education, Discovery 2018). A gene is a segment of DNA that provides instructions for unique traits. Cancer can be caused by a mutation, or a change in the sequence of DNA bases. 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 A4 – Cancer Mapping Studies and Mapped Tumors by Year, 2004–2016



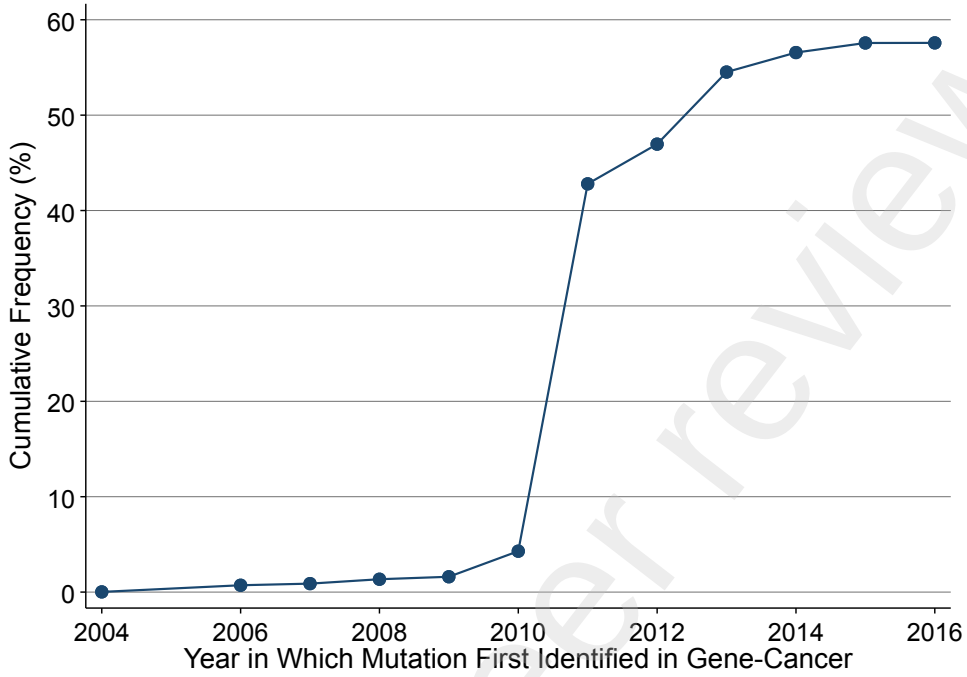
Notes: These figures plot the total number of cancer mapping studies (Panel A) and mapped tumors (Panel B) in each year from 2004 through 2016. The x-axis indicates the year in which the mapping study was submitted to the journal that ultimately published it. Mapping studies are large-scale and published 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* in 2015 (Campbell et al. 2016). For details on the construction of the cancer mapping studies sample used in this paper, see Appendix B.

Figure A5 – Share of Private Sector Cancer Trials that Enroll Patients Based on Genes, 2000-2006



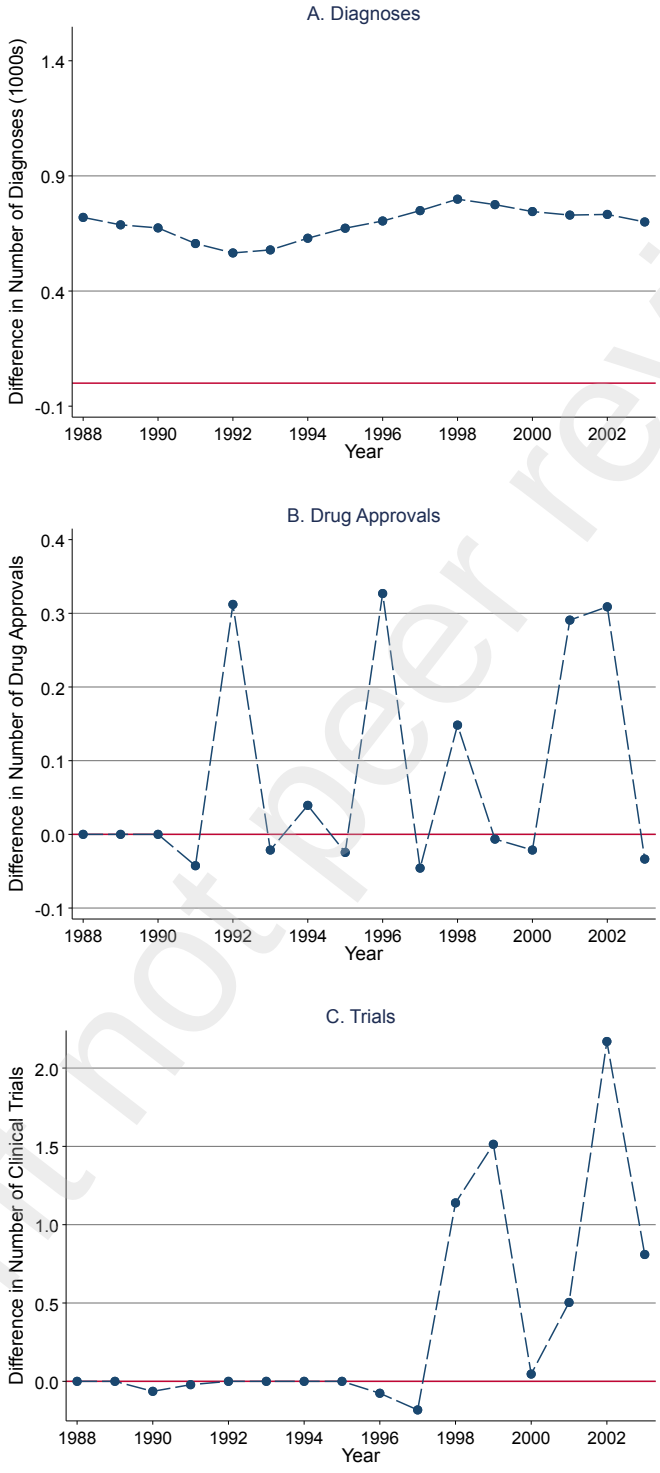
Notes: This figure plots the percentage of private sector phase II clinical trials in 2000-2016 that are gene-related—i.e., genetic criteria was used to select patients for enrollment. Observations are at the trial-cancer level.

Figure A6 – Cumulative Share of Gene-Cancer Pairs with Mutations Identified by Cancer Mapping Studies, 2004–2016



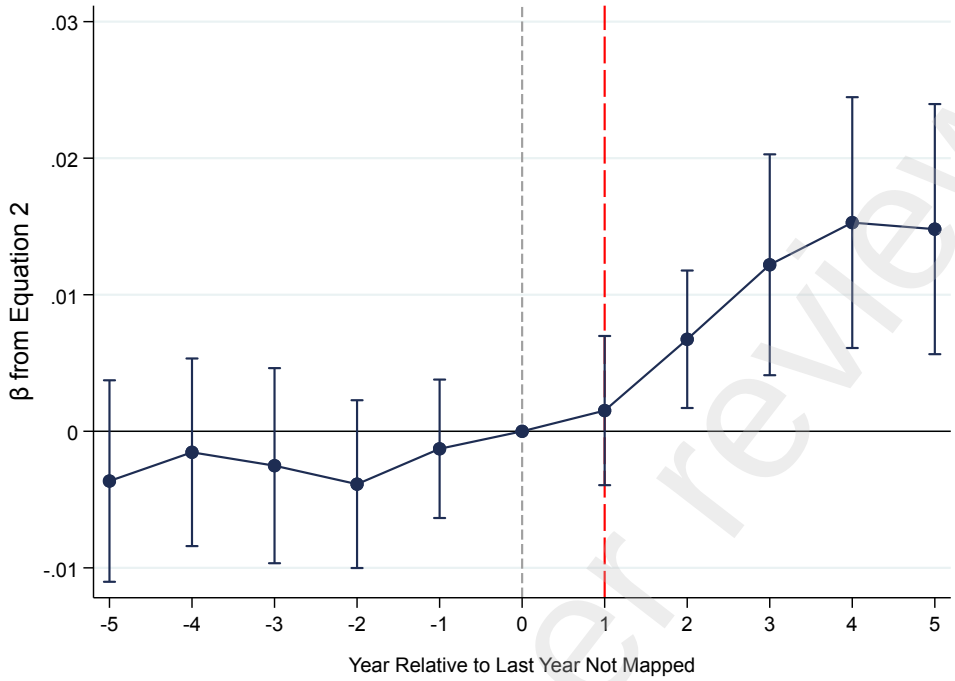
Notes: This figure plots the cumulative share of gene-cancer pairs with mutations identified by cancer mapping studies. As discussed in Section 3, there are 49,542 gene-cancer pairs possible. The period of analysis is 2004–2016. See Section 3 and Appendix B for more detailed data and variable descriptions.

Figure A7 – Examining Cancer-Level Selection, 1988-2003



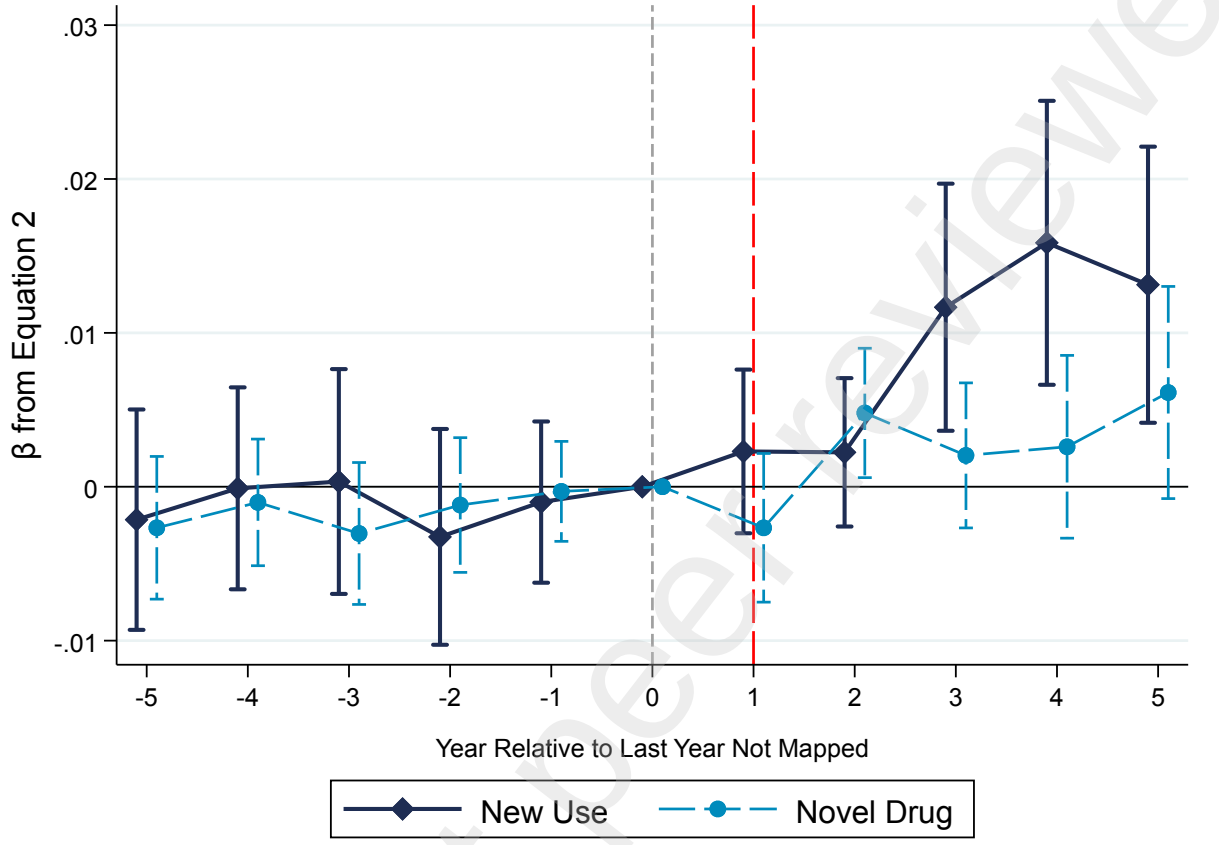
Notes: This figure examines baseline differences between cancers that are first sequenced relatively early (before 2011) and cancers that are first sequenced relatively late (in/after 2011). For each panel, difference in means of the outcome variable is calculated between the two cancer groups in each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

Figure A8 – Effect on Private Sector Clinical Trials with Non-Missing Interventions, 2004–2016



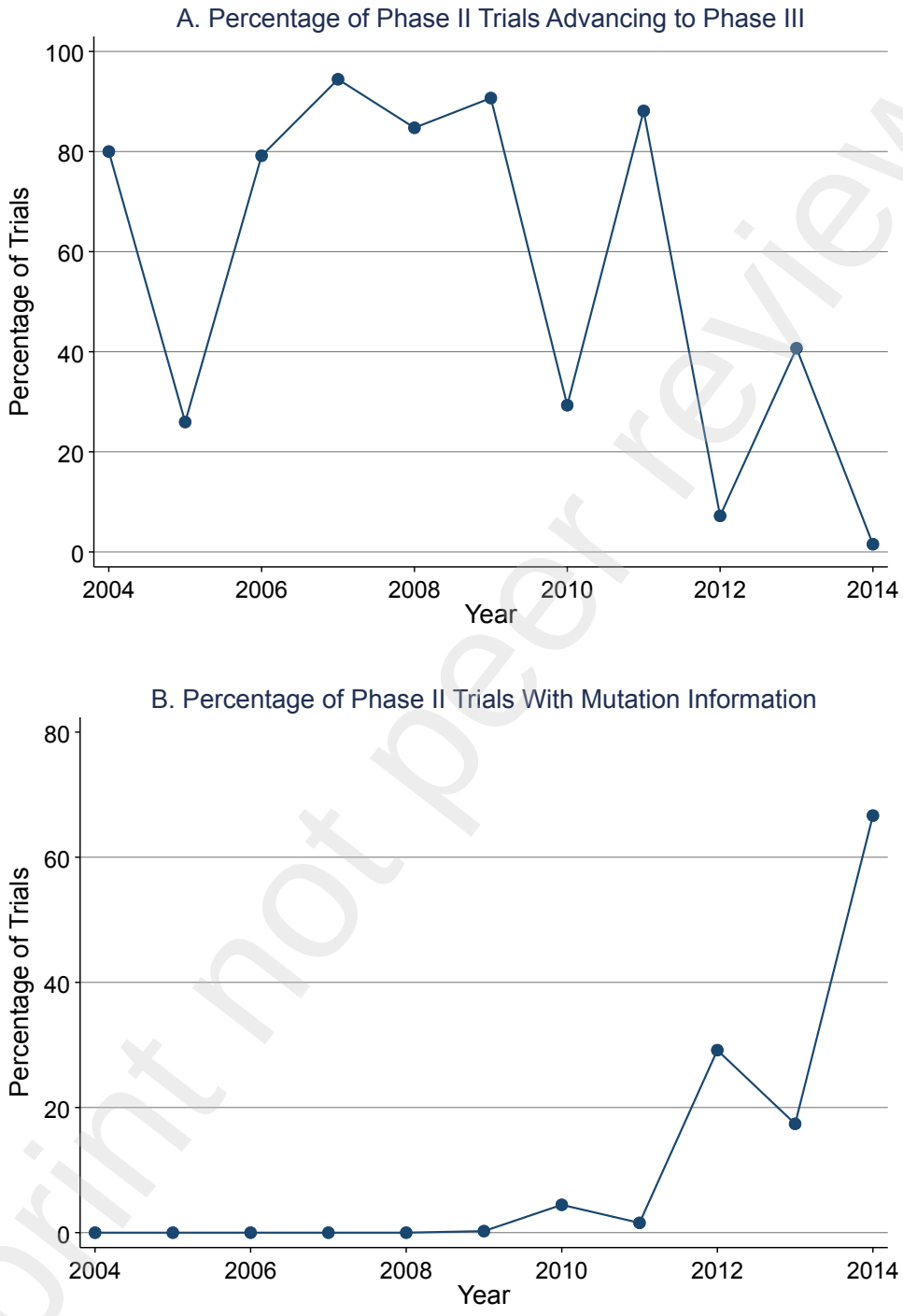
Notes: This figure plots the response of private sector clinical trials following the public release of cancer mapping information using the sample of clinical trials that have non-missing intervention data. Each dot corresponds to coefficients based on estimates of Equation 2. The outcome variable is 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 the cancer mapping studies. The dashed red line indicates the first year that a mutation in a gene-cancer is publicly disclosed by a cancer mapping study. Shown are 95 percent 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, and the sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004 through 2016. The figure corresponds to a dynamic version of the specification in Appendix Table A3, column 2. Controls include gene-cancer fixed effects, year fixed effects, and cancer-specific time trends. See Section 3 and Appendix B for more detailed data and variable descriptions.

Figure A9 – Effect on Private Sector Clinical Trials Testing
New Uses and Novel Drugs, 2004–2016



Notes: This figure separately plots the response of private sector clinical trials following the public release of cancer mapping information, for trials testing new uses and trials testing novel drugs. Each dot corresponds to coefficients based on estimates of Equation 2. The dark blue line plots coefficients from a regression where the outcome is an indicator for any clinical trial whose drugs have been approved in the focal gene or previously tested in any gene-cancer pair. The light blue line plots coefficients from a regression where the outcome is an indicator for any clinical trial whose drugs have not been approved in the focal gene and tested in any gene-cancer pair. 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 the cancer mapping studies. The dashed red line indicates the first year that a mutation in a gene-cancer is publicly disclosed by a cancer mapping study. Shown are 95 percent 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, and the sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004 through 2016. The figure corresponds to a dynamic version of the specification in Table 4. Controls include gene-cancer fixed effects, year fixed effects, and cancer-specific time trends. See Section 3 and Appendix B for more detailed data and variable descriptions.

Figure A10 – Clinical Trial Advancement Rates, by Year, 2004-2014



Notes: Panel A plots the percentage of private sector phase II clinical trials that successfully advanced to phase III. Panel B plots the percentage of private sector phase II clinical trials that are initiated in gene-cancer pairs with mutation information, as a share of the total number of such trials that successfully advanced to phase III. In this figure, 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 sample includes all phase II trials that are completed or terminated as of July 14, 2017. Observations are at the trial-gene-cancer level.

Table A1 – Overview of Gene-Cancer-Year Panel Construction, 2004–2016

	Count
No. of genes (e.g., BRCA1, BRCA2)	627
No. of cancer (e.g., ovarian, small intestine)	80
No. of cancer groups (e.g., digestive)	19
No. of gene-cancer pairs (e.g., BRCA2-prostate)	50,160
No. of gene-cancer pairs, excl. those known in 2004	49,542
No. of years (2004 to 2016)	13
No. of gene-cancer-year (e.g., BRCA2-prostate-2004) observations	652,080
Final Panel: No. of gene-cancer-year observations, excl. gene-cancers known in 2004	**644,046**

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

Table A2 – Effect on Private Sector Clinical Trials
Excluding Genes Affected by Patent Regulation, 2004–2016

	Dependent Variable: Any Private Sector Phase II Clinical Trial		
	(1)	(2)	(3)
Post × DisclGeneCancer	0.00543**	0.00786**	0.00802**
	(0.00170)	(0.00247)	(0.00291)
Mean of dep. var.	0.017	0.017	0.017
Percentage gain	32.82%	47.50%	48.46%
Gene-cancer FEs	Yes	Yes	Yes
Year FEs	Yes	Yes	No
Cancer-specific time trends	No	Yes	No
Cancer × Year FEs	No	No	Yes
Observations	642,018	642,018	642,018

Notes: This table reports difference-in-differences estimates of the effect of cancer mapping information on private sector clinical trials, excluding gene-cancer observations with BRCA1 and BRCA2—genes that are most likely to be affected by changing intellectual property regulation. The level of observation is the gene-cancer-year. The sample includes gene-cancer-years from 2004 through 2016, excluding (i) BRCA1 and BRCA2 genes and (ii) gene-cancer pairs known in 2004. The resulting dataset contains 642,018 gene-cancer-year observations. Estimates are from OLS models. "Post × DisclGeneCancer" switches from 0 to 1 when a mutation in a gene-cancer is publicly disclosed by a cancer 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 before the first disclosure of a mutation and is used to calculate "Percentage gain," the percentage change in the likelihood of a clinical trial. See Section 3 and Appendix B for more detailed data and variable descriptions.
*p <0.10, **p <0.05, ***p <0.01.

Table A3 – Effect on Private Sector Clinical Trials with Non-missing Interventions, 2004–2016

	Dependent Variable: Any Private Sector Phase II Clinical Trial		
	(1)	(2)	(3)
Post × DisclGeneCancer	0.00560**	0.00871**	0.00967**
	(0.00174)	(0.00256)	(0.00302)
Mean of dep. var.	0.017	0.017	0.017
Percentage gain	32.95%	51.24%	56.88%
Gene-Cancer FEs	Yes	Yes	Yes
Year FEs	Yes	Yes	No
Cancer-specific time trends	No	Yes	No
Cancer × Year FEs	No	No	Yes
Observations	644,046	644,046	644,046

Notes: This table reports difference-in-differences estimates of the effect of cancer mapping information on private sector clinical trials using the subset of clinical trials that have non-missing intervention data. The level of observation is the gene-cancer-year. The sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004 through 2016. Estimates are from OLS models. "Post × DisclGeneCancer" switches from 0 to 1 when a mutation in a gene-cancer is publicly disclosed by a cancer 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 before the first disclosure of a mutation and is used to calculate "Percentage gain," the percentage change in the likelihood of a clinical trial. See Section 3 and Appendix B for more detailed data and variable descriptions.
*p <0.10, **p <0.05, ***p <0.01.

Table A4 – Effect of Cancer Mapping Information in the Same Gene, Related Cancer, 2004–2016

	Dependent Variable: Any Private Sector Phase II Clinical Trial	
	(1)	(2)
Post × DisclGeneCancerGroup	0.00502**	0.00129
	(0.00233)	(0.00197)
Post × DisclGeneCancer		0.00817***
		(0.00230)
Mean of dep. var.	0.019	0.019
Percentage gain	26.85%	6.909%
Gene-cancer FEs	Yes	Yes
Year FEs	Yes	Yes
Cancer-specific time trends	Yes	Yes
Observations	644,046	644,046

Notes: This table reports difference-in-differences estimates of how private sector clinical trials in a gene-cancer pair respond to cancer mapping information in the same gene and a different cancer. The level of observation is the gene-cancer-year. The sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004 through 2016. 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 the Surveillance, Epidemiology, and End Results (SEER) classification scheme. "Post × DisclGeneCancer" switches from 0 to 1 when a mutation in a same gene and same cancer is publicly disclosed by a cancer mapping study. Controls include gene-cancer fixed effects, year fixed effects, and cancer-specific time trends. 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 before the first disclosure of a mutation in the same gene and related cancer and is used to calculate "Percentage gain," 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 Appendix B for more detailed data and variable descriptions.
*p <0.10, **p <0.05, ***p <0.01.

Table A5 – Effect on Private Sector Clinical Trials:
Heterogeneity by Firm Expertise, 2004–2016

	Dependent Variable: Any Private Sector Phase II Clinical Trial					
	Firm Experience in Gene Trials		Firm Experience in Cancer Trials		Firm Experience in Patents	
	Below Median (1)	Above Median (2)	Below Median (3)	Above Median (4)	Below Median (5)	Above Median (6)
A. Firm Experience in Previous Year						
Post × DisclGeneCancer	0.00757***	0.00250***	0.00549***	0.00498***	0.00509***	0.00596***
	(0.000742)	(0.000571)	(0.000692)	(0.000711)	(0.000610)	(0.000755)
Mean of dep. var.	0.013	0.006	0.006	0.013	0.005	0.014
Percentage gain	21.46%	14.37%	21.35%	17.35%	22.96%	18.13%
Gene-cancer FEs	Yes	Yes	Yes	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes	Yes	Yes	Yes
Cancer-specific time trends	Yes	Yes	Yes	Yes	Yes	Yes
Observations	644,046	644,046	644,046	644,046	644,046	644,046
Diff. Wald test p-value	0.06		0.00		0.00	
B. Firm Experience Between 2000 and 2003						
Post × DisclGeneCancer	0.00814***	0.00187***	0.00849***	0.00363***	0.00590***	0.00619***
	(0.000729)	(0.000524)	(0.000697)	(0.000693)	(0.000564)	(0.000771)
Mean of dep. var.	0.014	0.005	0.006	0.013	0.004	0.015
Percentage gain	23.48	14.16	29.96	13.17	28.26	18.52
Gene-cancer FEs	Yes	Yes	Yes	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes	Yes	Yes	Yes
Cancer-specific time trends	Yes	Yes	Yes	Yes	Yes	Yes
Observations	644,046	644,046	644,046	644,046	644,046	644,046
Diff. Wald test p-value	0.06		0.00		0.00	

Notes: This table reports difference-in-differences estimates of the effect of cancer mapping information on private sector clinical trials, separately for private sector firms with high and low pre-2004 research investment levels. The level of observation is the gene-cancer-year. The sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004 through 2016. The outcome variable reflects the probability of a private sector phase II clinical trial by firms with a particular level of research experience. Focusing on the set of clinical trial sponsors in my sample, I define firm experience based on the total number of research investments by a firm. The outcomes in columns 1, 3, and 5 are indicators for clinical trials whose sponsoring firms have below-median research experience. The outcomes in columns 2, 4, and 6 are indicators for clinical trials whose sponsoring firms have above-median research experience. Panel A focuses on firm experience in the year prior to the start of the focal trial and Panel B focuses on firm experience between 2000 and 2003. "Post × DisclGeneCancer" switches from 0 to 1 when a mutation in a gene-cancer is publicly disclosed by a cancer mapping study. Controls include gene-cancer fixed effects, year fixed effects, and cancer-specific time trends. "Mean of dep. var." is the mean of the outcome variable in a gene-cancer before the first disclosure of a mutation and is used to calculate "Percentage gain," the percentage change in the likelihood of a clinical trial. Estimates are from seemingly unrelated models, which permits a comparison of "Percentage gain" across models. However, the use of seemingly unrelated regressions makes direct two-way clustering by gene and cancer infeasible. As a second best, I cluster at the gene-cancer level. P-values are from Wald tests that compare the differences in "Percentage gain." See Sections 3 and 4 and Appendix B for more detailed data and variable descriptions.
*p <0.10, **p <0.05, ***p <0.01.

Table A6 – Effect on Private Sector Clinical Trials:
Heterogeneity by Market Potential of Disease, 2004–2016

	Dependent Variable: Any Private Sector Phase II Clinical Trial					
	Market Size		Clinical Trials		Clinical Trials/Market Size	
	Below Median	Above Median	Below Median	Above Median	Below Median	Above Median
	(1)	(2)	(3)	(4)	(5)	(6)
Post × DisclGeneCancer	0.00800***	0.00919***	0.00950***	0.00838***	0.00884***	0.00858***
	(0.00119)	(0.00116)	(0.00120)	(0.00118)	(0.00111)	(0.00126)
Mean of dep. var.	0.015	0.018	0.010	0.020	0.017	0.017
Percentage gain	52.67%	50.62%	93.94%	42.48%	52.73%	49.99%
Gene-cancer FEs	Yes	Yes	Yes	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes	Yes	Yes	Yes
Cancer-specific time trends	Yes	Yes	Yes	Yes	Yes	Yes
Observations	323,622	320,424	322,569	321,477	320,021	324,025
Diff. Wald test p-value	0.83		0.00		0.88	

Notes: This table reports difference-in-differences estimates of the effect of cancer mapping information on private sector clinical trials, separately for different diseases with low and high market potential. The level of observation is the gene-cancer-year. The sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004 through 2016. "Post × DisclGeneCancer" switches from 0 to 1 when a mutation in a gene-cancer is publicly disclosed by a cancer mapping study. Each pair of columns splits the sample across the median of market size (as measured by the number of diagnoses) (columns 1 and 2), clinical trials (columns 3 and 4), and clinical trials normalized by market size (columns 5 and 6) for the focal cancer. For example, the first two columns compare the magnitude of the effect of the cancer mapping information on gene-cancer pairs whose focal cancer is below the median annual number of cancer diagnoses between 2000 and 2003. The sum of the number of observations across each pair of columns equals the full sample of gene-cancer pairs (N = 644,046) used in the main analysis. Controls include gene-cancer fixed effects, year fixed effects, and cancer-specific time trends. "Mean of dep. var." is the mean of the outcome variable in a gene-cancer before the first disclosure of a mutation and is used to calculate "Percentage gain," the percentage change in the likelihood of a clinical trial. Estimates are from seemingly unrelated models, which permits a comparison of "Percentage gain" across models. However, the use of seemingly unrelated regressions makes direct two-way clustering by gene and cancer infeasible. As a second best, I cluster at the gene-cancer level. P-values are from Wald tests that compare the differences in "Percentage gain." See Sections 3 and 4 and Appendix B for more detailed data and variable descriptions.
*p <0.10, **p <0.05, ***p <0.01.

Table A7 – Effect on Private Sector Clinical Trials:
Heterogeneity by Trial Design Type, 2004–2016

	Dependent Variable: Any Private Sector Phase II Clinical Trial	
	Well Designed	Poorly Designed
	(1)	(2)
Post × DisclGeneCancer	0.00226***	0.00692***
	(0.000410)	(0.000835)
Mean of dep. var.	0.004	0.015
Percentage gain	63.63%	46.04%
Gene-cancer FEs	Yes	Yes
Year FEs	Yes	Yes
Cancer-specific time trends	Yes	Yes
Observations	644,046	644,046
Diff. Wald test p-value	0.33	

Notes: This table reports difference-in-differences estimates of the effect of cancer mapping information on private sector clinical trials, separately for well designed and poorly designed trials. The level of observation is the gene-cancer-year. The sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004 through 2016. The outcome variable switches from 0 to 1 if a private sector phase II clinical trial is reported in a gene-cancer-year and is a well designed trial (column 1) or poorly designed trial (column 2). (See Appendix B for a description of how I classify trials as well or poorly designed trials.) "Post × DisclGeneCancer" switches from 0 to 1 when a mutation in a gene-cancer is publicly disclosed by a cancer mapping study. Controls include gene-cancer fixed effects, year fixed effects, and cancer-specific time trends. "Mean of dep. var." is the mean of the outcome variable in a gene-cancer before the first disclosure of a mutation and is used to calculate "Percentage gain," the percentage change in the likelihood of a clinical trial. Estimates are from seemingly unrelated models, which permits a comparison of "Percentage gain" across models. However, the use of seemingly unrelated regressions makes direct two-way clustering by gene and cancer infeasible. As a second best, I cluster at the gene-cancer level. The p-value is from a Wald test that compares the differences in "Percentage gain." See Sections 3 and 4 and Appendix B for more detailed data and variable descriptions.
p <0.10, **p <0.05, ***p <0.01.

Table A8 – Cancer Mapping Information and
Phase II Clinical Trial Outcomes, 2004–2016

	Dependent Variable: Response Rate	
	(1)	(2)
Post × DisclGeneCancer	0.343	0.305
	(0.258)	(0.235)
Firm experience (No. of clinical trials)		-0.421***
		(0.122)
Mean of dep. var.	18.98	18.98
Cancer FEs	Yes	Yes
Gene FEs	Yes	Yes
Linear time trend	Yes	Yes
No. of trial-gene-cancers	2,323	2,323
No. of trials	159	159
No. of genes	80	80
No. of cancers	61	61
Adjusted R^2	0.644	0.699

Notes: This table shows the relationship between cancer mapping information and phase II clinical outcomes, as measured by the phase II clinical trial's objective response rate. The level of observation is the trial-gene-cancer. The sample includes all private sector phase II trial-gene-cancer observations associated with phase II clinical trials that began between 2004 and 2016 (excluding gene-cancer pairs known in 2004), made clinical outcomes data available, and were completed or terminated as of July 14, 2017. There are fewer than 2,354 observations because the estimation drops trial-gene-cancer observations with a gene or cancer that just shows up once. Estimates are from OLS models. The outcome variable is the phase II trial's objective response rate, or the share of patients who respond to treatment by a prespecified amount. The outcome variable is transformed with the inverse hyperbolic sine transformation, but the objective response rate mean is reported in levels for ease of interpretation. "Post × DisclGeneCancer" is an indicator for the disclosure of a driver (clinically relevant) mutation in a gene-cancer by the start of the clinical trial. "Firm experience" is the number of clinical trials that the focal trial sponsor has conduced in the focal cancer within one year prior to the phase II clinical trial start date, and is transformed with the inverse hyperbolic sine transformation. Controls include cancer fixed effects and gene fixed effects. Robust standard errors, clustered at the gene and cancer level, are shown in parentheses. See Sections 3 and 5 and Appendix B for more detailed data and variable descriptions.
*p <0.10, **p <0.05, ***p <0.01.

Appendix B.
Data Description

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

B.1 Cancer Mapping Data

B.1.1 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.[i] (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).[ii] 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).[iii] 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.[iv]

- **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.[v] 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 the 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).[vi]

[i] For more details, see https://cancer.sanger.ac.uk/cosmic.
[ii] For more details, see https://cancer.sanger.ac.uk/cosmic/download.
[iii] For more details, see http://www.cbioportal.org/.
[iv] For more details, see https://cancer.sanger.ac.uk/cosmic/papers.
[v] For more details, see: https://www.scimagojr.com/.
[vi] 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.

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 National Institutes of Health).

B.1.2 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. Nearly all cancer mapping studies focus primarily on the mutations in protein-coding regions of the DNA molecule since, relative to mutations in non-protein-coding regions, the linkages between the mutations, altered proteins, and subsequent diseases are easier interpret (Vogelstein et al. 2013). In particular, I focus on somatic mutations, which are DNA aberrations that occur after conception and are not inherited.[vii] According to Stratton, Campbell and Futreal (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.[viii]

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 includes 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.

B.2 Identifying Well Designed and Poorly Designed Trials

This section describes how I classified clinical trials into well designed and poorly designed trials. Using recommended standards outlined in the scientific literature (Adjei et al. 2009; Berger and Alperson 2009; U.S. Food and Drug Administration 2018a, b; Seymour et al. 2010; Prasad et al. 2015; Blumenthal 2017; Dhani et al. 2017; Grossman et al. 2017; Kemp and Prasad 2017; NCI n.d.), I classified 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 surrogated endpoint

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

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

3. Non-randomized, controlled, validated surrogate endpoint

Information on validated surrogate endpoints comes from Prasad et al. (2015). Trials that are not coded as well designed are classified as poorly designed.

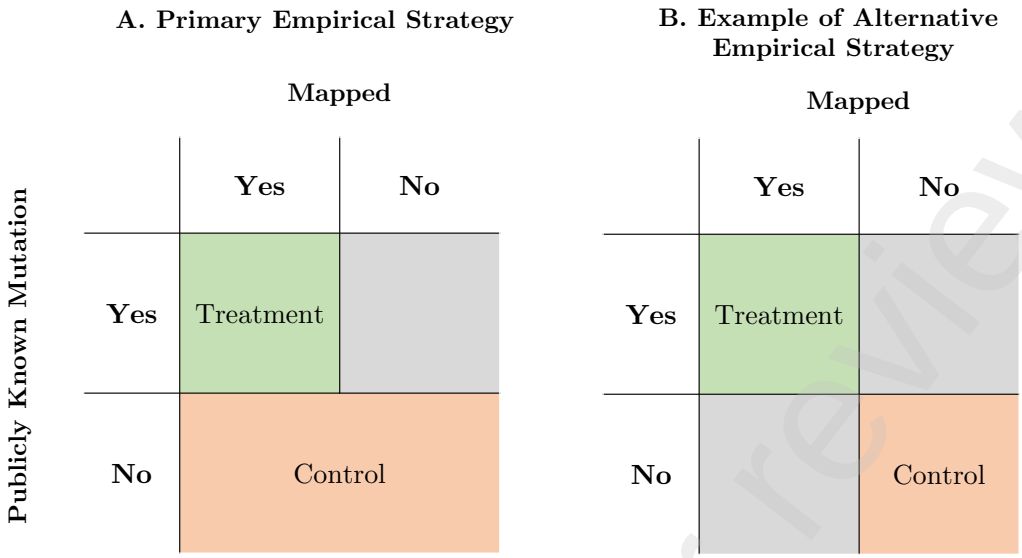
Appendix C.
Treatment and Control Gene-Cancer Pairs

As discussed in Section 4, the empirical strategy employed to measure the impact of large-scale cancer mapping 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 C1 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 to 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 analysis in Section 4.3 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 investment across diseases that are less likely to be substitutable (e.g., ovarian cancer vs. Alzheimer's disease).

Figure C1 – Allocation of Gene-Cancer Pairs to Treatment and Control Groups



Appendix References

Adjei, Alex A., Michaele Christian, and Percy Ivy. 2009. "Novel Designs and End Points for Phase II Clinical Trials." *Clinical Cancer Research,* 15(6): 1866–72.

Berger, Vance W. and Sunny Y. Alperson. 2009. "A General Framework for the Evaluation of Clinical Trial Quality." *Reviews on Recent Clinical Trials,* 4(2): 79–88.

Blumenthal, Gideon. 2017. "Primer on Drug Development." Presentation at Partners in Progress: Cancer Patient Advocates and FDA Public Workshop. https://www.fda.gov/media/109685/download.

Cancer Genome Atlas Research Network. 2011. "Integrated Genomic Analyses of Ovarian Carcinoma." *Nature,* 474(7353): 609–15.

Cerami, Ethan, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Blent Arman Aksoy, Anders Jacobsen, Caitlin J. Byrne, Michael L. Heuer, Erik Larsson, Yevgeniy Antipin, Boris Reva, Arthur P. Goldberg, Chris Sander, and Nikolaus Schultz. 2012. "The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data." *Cancer Discovery,* 2(5): 401–4.

Dhani, Neesha, Dongsheng Tu, Daniel J. Sargent, Lesley Seymour, and Malcom J. Moore. 2017. "Alternate Endpoints for Screening Phase II Studies." *Clinical Cancer Research,* 15(6): 1873–82.

Gao, Jianjiong, Blent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson... 2013. "Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal." *Science Signaling,* 6(269): l1.

GeneEd: Genetics, Education, Discovery. 2018. "DNA, Genes, Chromosomes." https://geneed.nlm.nih.gov/topic_subtopic.php?tid=15.

Gonzalez-Pereiraa, Borja, Vicente P. Guerrero-Boteb, and Flix Moya-Anegn. 2009. "The SJR Indicator: A New Indicator of Journals' Scientific Prestige." https://arxiv.org/ftp/arxiv/papers/0912/0912.4141.pdf.

Grossman, Stuart A., Karisa C. Schreck, Karla Ballman, and Brian Alexander. 2017. "Point/counterpoint: Randomized Versus Single Arm Phase II Trials for Patients with Newly Diagnosed Glioblastoma." *Neuro-Oncology,* 19(4): 469–74.

Kemp, Robert, and Vinay Prasad. 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, Vinay, Chul Kim, Mauricio Burotto, and Andrae Vandross. 2015. "The Strength of Association Between Surrogate Endpoints and Survival in Oncology." *JAMA Internal Medicine,* 175(8): 1389–98.

Samuel, Nardin, and Thomas J. Hudson. 2013. "Translating Genomics to the Clinic: Implications of Cancer Heterogeneity" *Clinical Chemistry,* 59(1): 127–37.

Seymour, Lesley, S. Percy Ivy, Daniel Sargent, David Spriggs, Laurence Baker, Larry Rubinstein, Mark J. Ratain, Michael Le Blanc, David Stewart, and Donald Berry. 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): 1764–69.

Stratton, Michael R., Peter J. Campbell, and P. Andrew Futreal. 2009. "The Cancer Genome." *Nature,* 458(7239): 719–24.

Tate, John G., Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson... 2019. "COSMIC: the Catalogue of Somatic Mutations In Cancer." *Nucleic Acids Research,* 47(D1): D941–47.

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, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz Jr., and Kenneth W. Kinzler. 2013. "Cancer Genome Landscapes." *Science,* 339(6127): 1546–58.