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

Jennifer Kao*

January 13, 2019

Please click here for the most recent version.

Abstract

This paper explores how publicly available scientific information shapes the quantity and profitability of private-sector research. I examine the impact of large-scale cancer genome mapping studies, which systematically map the genetic abnormalities in cancer, on research productivity in the pharmaceutical industry. Using a newly-constructed dataset from cancer genome mapping studies and clinical trials, I find that mapping information increases private-sector investment in clinical trials by nearly 50 percent. Considering the types of private-sector research investments, I find that cancer mapping significantly increases trials evaluating drugs previously approved or tested for one disease in an additional disease. Using trial results reported in abstracts submitted to a major cancer conference, I also find that cancer mapping information increases the profitability of firms' research decisions: when genetic information is known, firms are more likely terminate drug investments that are unlikely to be successful in the long run and to continue investment projects that are most likely to generate promising clinical results. This evidence suggests that publicly available, detailed scientific maps can increase and improve private research efforts.

*Harvard Kennedy School. Email: jkao@g.harvard.edu. I am particularly grateful to my advisors, David Cutler, Pierre Azoulay, and Amitabh Chandra, for their guidance and encouragement. I also thank Ariel D. Stern, Megan Bailey, Samantha Burn, Stephen Coussens, Caitlin Carroll, Leemore Dafny, Ariella Kahn-Lang, Joshua Krieger, Timothy Layton, Danielle Li, Ramana Nanda, Adrienne Sabety, Rebecca Sachs, Mark Shepard, Gabriel Tourek, Lisa Xu, Heidi Williams and various seminar participants at Harvard, NBER, and the Wharton Innovation Doctoral Symposium. Mohan Ramanujan provided invaluable help with the data. I also thank several pharmaceutical industry experts for their insights on cancer sequencing. All errors are my own. I am grateful for support from the National Institute on Aging under Award Number R24AG048059 and T32AG000186 to the National Bureau of Economic Research.

1 Introduction

How does publicly available basic scientific information influence the quantity and profitability of private-sector research? On the one hand, fundamental shifts in basic scientific knowledge may spur and sharpen firms' research and development efforts, through improving technological search and lowering entry costs.[1] On the other hand, the public release of detailed scientific data may reinforce the value of existing products or lower the cost of entry for competitors. This may diminish the incentive of potential entrants to enter and cause subsequent private research efforts to be, on net, lowered or unchanged. Thus, the impact of comprehensive advances in publicly available basic scientific knowledge on private-sector research is ambiguous.

Most evaluations of basic science knowledge have focused on whether R&D grants or other subsidies for basic science increase private-sector research.[2] In this paper, I examine the effect of publicly available basic scientific data from large-scale cancer genome mapping initiatives. Building on the foundation provided by the Human Genome Project (Jayaraj, 2018; Williams, 2013), cancer genome mapping initiatives systematically catalogue the genetic mutations that might drive the progression and growth of cancer.[3] These large-scale research initiatives aim to facilitate technological search through introducing and validating existing scientific knowledge about the disease (Fleming and Sorenson, 2004; Mardis, 2018). Large-scale cancer mapping efforts are believed to play an influential role in the development of novel cancer therapies: National Institutes of Health director Francis Collins and former National Cancer Institute deputy director Anna Barker noted that mapping the cancer genome would "help chart a new course across the complex landscape of human malignancies"(Barker and Collins, 2008).

This paper examines how the 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 (CDC, 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 U.S. Food and Drug Administration-required clinical trials, a risky and costly process. Only about 15 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 (Danzon and Keuffel, 2014; DiMasi, 2001; DiMasi, Hansen and Grabowski, 2003; DiMasi

[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 example, Azoulay et al. (2018) examines the impact of National Institutes of Health funding on private sector patenting. For helpful reviews, see David, Hall and Toole (2000) and Hall and Van Reenen (2000). A notable exception to the R&D subsidy literature is Nagaraj (2017), who examines the impact of publicly available satellite maps on discoveries by the gold exploration industry.

[3] In fact, the Human Genome Project (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, the Nobel laureate Renato Dulbecco (1986) wrote "If we wish to learn more about cancer, we must now concentrate on the cellular genome."

et al., 2010; Hay et al., 2014; Thomas et al., 2016). The impact of the cancer mapping efforts on new product development depends in large part on the extent to which the basic science influences firms' clinical trials investment efforts.

In light of this, I focus on two issues: the *quantity* of clinical trials and the *profitability* of firms' research decisions. I assemble a new dataset of publicly available information produced by 168 large-scale cancer mapping efforts, linked to privately-funded clinical trials, over the period 2004-2016.[4] 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.

I begin by investigating how the disclosure of publicly available cancer mapping information shapes the subsequent quantity of privately-funded 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 differences-in-difference framework. To address concerns over selection in the timing of mapping, I control for differences in "research potential" between different gene-cancers with gene-cancer fixed effects and control for secular changes in the pharmaceutical industry over time using year fixed effects and cancer-year linear 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. 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 97 fewer privately-funded clinical trials and 15 fewer cancer drugs.[5] These results are consistent with recent quasi-experimental work that shows how investments in basic science (Azoulay et al., 2018) and, in particular, how detailed *maps* of basic scientific information (Jayaraj, 2018; Nagaraj, 2017; Williams, 2013), can increase the level of private-sector research.

I next examine the hypothesis that one mechanism through which scientific maps increases private investment is by identifying linkages across research opportunities that were previously believed to be distantly related. In particular, genetic mapping can reveal that similar genetic aberrations underlie different cancers types and spur trials testing drugs approved or previously tested for one disease in an additional disease. To illustrate, 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. Consistent with this hypothesis, I find that mapping significantly increases investment in trials testing drugs that are previously approved (140 percent) or tested (54 percent). In contrast, private investment in trials testing novel drugs remains

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

[5]Here, "drugs" refers to an active ingredient treating a specific disease (See Section 4 for more details).

unchanged.

Finally, I explore whether cancer mapping information increases the profitability of firms' research decisions—in this case, the likelihood that firms make decisions that maximize their expected returns based on existing clinical information. Once firms initiate a clinical trial, they must complete a series of additional clinical trials, each with increasing cost and risk. At each point, 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., maximizing their expected returns through minimizing resources allocated to drugs that are unlikely to be successful, and continuing investment in drugs that are most likely to generate promising clinical results and successfully come to market (Lendrem et al., 2015; Spetzler, Winter and Meyer, 2016). In deciding whether to continue or terminate investment, firms may make low profit decisions by dismissing or failing to understand existing clinical evidence: a review of AstraZeneca's drug pipeline revealed that 18% 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 profitability of firms' research decisions. I find that firms initiating trials in diseases with mutation-related information are 60 percent less likely to advance drugs with weak clinical evidence to the next phase, as compared to 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. These findings are consistent with anecdotal evidence that access to reliable scientific information increases the productivity of firms' research decisions by promoting earlier termination of drugs that are likely to fail and further investment in drugs that are most likely to be successful in the long run (Bujar et al., 2017; Peck et al., 2015; Sharpe and Keelin, 1998).

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 privately-funded clinical trials. I examine the impact of cancer mapping on the profitability of firms' research decisions in Section 5. Finally, Section 6 concludes.

2 Case Study and Conceptual Framework

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 introduce 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 sub-groups with specific genetic features. These so-called "targeted" 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 22,000 women and is the fifth leading cause of cancer death among women in the United States (American Cancer Society, 2018). 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.

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.[6] 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 TCGA's ovarian cancer study has uncertain effects on the quantity of clinical trials enrolling BRCA-mutated non-inherited serous ovarian cancer patients. On the one hand, the TCGA's mapping information may introduce or validate existing information that assists firms in identifying subgroups of patients that respond most favorably to treatment. This is consistent with the theory that invention is a process of searching for better combinations of components (in this case, drugs and diseases) and that science facilitates the efficient identification of useful, new combinations (Fleming and Sorenson, 2004). A more efficient drug-disease (or drug-patient) match could enable firms to conduct trials with fewer patients and over a short duration, ultimately lowering the cost of bringing a novel drug to market (Chandra, Garthwaithe and Stern, 2018).[7] As a result, mapping 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.

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

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

On the other hand, mapping information may lower or not affect the net level of subsequent investment. This may occur through three mechanisms. First, the TCGA's ovarian cancer effort may reveal non-novel information about the relationship between genes and cancers. Previous sequencing or non-sequencing efforts may have already revealed relationships between BRCA genes and different forms of ovarian cancer. For example, firms with significant resources may have their own in-house genomics research efforts, or partner with firms that specialize in genomics research.[8] To illustrate, the drug Olaparib, the first approved PARP-inhibitor, was 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).[9] In the same vein, previous non-mapping efforts such as retrospective analyses may reveal that non-inherited serous ovarian cancer patients with BRCA mutations are most responsive to treatment, suggesting that this particular type of ovarian cancer is driven by BRCA mutations. Thus, publicly available mapping information may have no effect on the net level of private-sector innovation.

However, while the information produced by the TCGA's ovarian cancer effort may be non-novel, the information may still be useful: in describing the impact of TCGA's ovarian cancer study, a leading genomics expert at a pharmaceutical company, which manufacturers a PARP-inhibitor, confirmed that some of the information may have been known already. 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.[10,11]

Second, mapping information may lower or not change the subsequent level of trials because certain firms may be able to take advantage of the information more readily and crowd out potential entrants. Specifically, 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 inhibitors) because they may be able to skip several stages of the research and development process, such as earlier clinical trials that assess drug safety.[12]

Third, in revealing new opportunities and lowering the cost of entry, placement of mapping

[8] 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.

[9] Indeed, an AstraZeneca's 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).

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

[11] 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).

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

information in the public domain may increase the level of competition—i.e., the number of firms with drugs treating non-inherited serous ovarian cancer with BRCA mutations. As a result, firms considering whether to invest may expect lower returns and have lowered incentives to invest.

Figure 1 shows the total number of privately-funded trials enrolling patients with BRCA2-mutated ovarian cancer by year, five years before and after 2010—the year in which the TCGA submitted its findings to the journal *Nature*. For simplicity, this figure excludes AstraZeneca's Olaparib which experiences 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 (see Appendix Figure B.1). The trial in 2008 might be explained by the reasons outlined above (e.g., TCGA's mapping information may have been non-novel for some firms).

Though it may not be causal, the relationship between BRC2-mutated ovarian cancer trials and the disclosure of the TCGA's ovarian study results is striking: the level of trials increases in the same year in which the TCGA submitted its findings to *Nature*. This figure suggests that mapping information may be positively related to the subsequent level of clinical trials.

2.3 Profitability Implications of the Ovarian Cancer Mapping Study

Once firms initiate and complete clinical trials, 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 is to ask whether the TCGA's ovarian study also increases the profitability of firms' termination-or-continuation decisions. Demonstrating the impact of TCGA ovarian cancer study on the profitability of firms' decisions is difficult in this particular context due to incomplete data.[13] Therefore, my strategy in the remainder of the section is to provide a brief conceptual framework.

Suppose that results of a trial testing a drug on patients with BRCA-mutated non-inherited serous ovarian cancer reveals that the drug is ineffective. For example, the share of trial patients whose tumors shrink may be too low, 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. Information from the TCGA ovarian cancer study has uncertain effects on the profitability of the trial sponsor's termination-or-continuation decision.

On the one hand, a detailed scientific map may increase the profitability of the trial sponsor's decision by encouraging the firm to terminate investment in the drug and to save resources by minimizing further investment in a drug that is unlikely to be successful. Instead, mapping information may encourage the firm to direct resources towards drugs that are likely to successfully obtain

[13] For example, 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 three of the trials in Figure 1 have available data on trial results (see Section 3.3.2 for a discussion on trial results reporting).

regulatory approval in the long run (Peck et al., 2015). These effect can result from two mechanisms. First, detailed basic scientific information from the TCGA's ovarian cancer study can lead to more informed-decision making (Arora and Gambardella, 1994; Bujar et al., 2017; Cockburn and Henderson, 1997; Cohen and Levinthal, 1990; Cook et al., 2014; Fleming and Sorenson, 2003, 2004; Morgan et al., 2018; Nelson, 1982; Rosenberg, 1990; Sharpe and Keelin, 1998; Ward and Dranove, 1995). In this case, basic science can help the firm interpret the clinical trial outcomes, and clarify the costs and gains associated with the decision to terminate or continue investment.

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

On the other hand, scientific mapping information may encourage the trial sponsor to *continue* investing in the drug, thus increasing the likelihood that firms incur the high development costs associated with late-stage failures (Peck et al., 2015). Fleming and Sorenson (2004) suggests that science may motivate researchers to continue investing in a particular drug, despite negative clinical feedback. This may lead to perverse outcomes: by suggesting that a drug-disease pairing should succeed theoretically, scientific information may encourage the firm to ignore clinical evidence that indicates otherwise.

3 Empirical Setting and Data

3.1 Scientific Background

Cancer—the disease I consider—is caused by changes in DNA.[14] 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 150-200 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

[14]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`.

sites, based on the standard Surveillance, Epidemiology, and End Results (SEER) classification system. Next, I focus on a set of 627 genes listed in Cancer Gene Census, which are believed to be causally associated with cancer.[15] 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 gene-cancer (627 genes $\times$ 80 cancer sites) pairs possible.

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 executed by comparing the DNA sequences of cancer cells to those of normal tissue (either from the same individual or a reference DNA). Appendix Figure B.2 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 *The Cancer Genome Atlas* (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 Human Genome Project, which sequenced the human genome and provided a reference for subsequent cancer mapping efforts. Williams (2013) finds that intellectual property restrictions that hampered subsequent use of mapping information led to a significant decrease in the level of follow-on innovation. The second factor was improvements in sequencing technology, which allowed for more accurate, faster, and cheaper sequencing. It is widely reported that the introduction of so-called next-generation sequencing allowed the cost of sequencing per genome (excluding the cost of data analysis) to fall from $95 million in 2001 to $1,000 in 2017 (Wetterstrand, 2018).[16]

I obtain the information produced through by these large-scale cancer sequencing efforts —mutation data at the gene-cancer-level—from the Catalogue of Somatic Mutations in Cancer (COSMIC) and the cBioPortal for Cancer Genomics (cBioPortal) databases. Similar to biological resource centers which act as "living libraries" for biological materials, both databases act as repositories of mapping data from hundreds of cancer mapping studies (Furman and Stern, 2011). Further, COSMIC and cBioPortal curate and standardize cancer genome data for subsequent researchers (Yang et al., 2015). Mapping data includes information about a sequenced tumor's cancer type (e.g., ovarian cancer), associated genetic mutations (gene BRCA2), and the

[15] 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 Catalogue of Somatic Mutations in Cancer database (For more details, see `https://cancer.sanger.ac.uk/cosmic/download`).

[16] 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.

date in which the associated mapping study was submitted to a scientific journal for publication (e.g., *Nature*-September 2010).[17]

I focus on mapping information from 168 cancer mapping efforts (see Appendix A1 for a description of how the mapping studies were selected). The cancer mapping studies used in this paper share three important characteristics. First, cancer mapping 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 cancer mapping studies are large-scale and systematic. The cancer mapping studies examined in this paper typically examine hundreds of tumors. 91 percent of the mapping studies examine the entire or all of 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 that are published in highly-ranked scientific journals. Figure 2 shows the number of cancer mapping studies and mapped cancer tumors between 2004 and 2016. The increase and fall 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 possible that cancer mapping efforts may lower subsequent research activity within a disease area. 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 three percent of tumors contained a mutation that allows them to evade the immune system (TCGA, 2012). This suggest that drugs that work through activating the immune system would not be effective treatments for lung cancer patients with that specific gene mutation.

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 conferences. To approximate the earliest date that mapping information was publicly known, I identify, for each gene-cancer pair in my dataset, the first date that a mapping study containing information about a mutation in the gene-cancer is submitted to a journal.[18]

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

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

In a subset of the analysis 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: mutations that are likely to drive the growth and progression of cancer (so-called driver mutations) and mutations that are unlikely to have a deleterious effect (so-called passenger mutations). It is not possible to definitively prove that a mutation is a driver or a passenger—instead, cancer sequencing researchers typically employ a variety of statistical methods to determine whether a given mutations is highly likely to be a driver mutation.[19] 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 drugs 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 objective response rate. Once phase III is complete, manufacturers must submit a new drug application (NDA) for regulatory review. Overall, the average clinical development process is long (typically taking 8-12 years), costly (typically costing a manufacturer $800 million - $2.6 billion), and risky (only 9% of drugs that begin clinical development ultimately go to market) (CSDD, 2014; Danzon and Keuffel, 2014; DiMasi, 2001; DiMasi et al., 2003).[20]

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 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., 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

[19] Common methods include: Mutation Significance (MutSig) algorithm (Lawrence et al., 2014) or the Mutational Significance in Cancer (MuSiC) algorithm (Dees et al., 2012).

[20] These costs 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. Despite this, other efforts have generated similar cost estimates (Avorn, 2015).

indications and new patient populations relative to NDAs (DiMasi, 2013).

New use approvals have high expected social value (Berndt, Cockburn and Grepin, 2006; Roin, 2014). Francis Collins, the former director of the National Institutes of Health (NIH) describes the clinical testing of existing drugs for new uses is an opportunity to become "more efficient and effective at delivering therapies and diagnostics to patients" (Collins, 2011). Further, firms seeking new use approvals may generate scientific evidence that is useful for clinical decision making, particularly in contexts where off-label use is widespread. 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

I collect data on privately-funded clinical trials. Data on clinical trials comes from Clarivate Analytics Cortellis Competitive Intelligence Database, which collects trials 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), and the sponsoring firm (AstraZeneca). The clinical trials also contain information on protein biomarkers (e.g., the gene EGFR).[21] 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 sample of clinical trials to those that are privately-sponsored.

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 firms. This results in 30,137 privately-funded phase II clinical trials at the gene-cancer level. Figure 3 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 share of mutations was 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 retrospective analyses of previous trial results or licensing relationships with genomic firms.[22] This paper aims to examine whether large-scale cancer mapping efforts lead to any additional effect on the level of privately-funded clinical trials, above and beyond these other factors.

I supplement the clinical trial data in two ways:

(i) *Drug Approvals Data:* I link trial data to drug approval data to identify whether a trial is

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

[22] 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.

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 a 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.[23] 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 approvals data, I classify trials into three categories: trials testing approved, pipeline, and novel drugs. A trial-gene-cancer is classified as testing an "approved drug" if its intervention has already been approved in the same gene. For example, a trial enrolling ovarian cancer patients with BRCA2 gene mutations is classified as testing an approved drug if its intervention has been approved to treat patients with BRCA2 gene mutations prior to the start of focal trial. Similarly, a trial-gene-cancer is indicated as testing a "pipeline drug" if its intervention is not approved in the same gene but has been clinically tested previously. Finally, a trial-gene-cancer is classified as testing a "novel drug" if its intervention is not approved in the same gene and has never been clinically tested before.[24]

(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, such as the share of patients who respond to treatment. 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.[25] Despite this requirement, clinical trial results are significantly underreported (it is estimated that just 22 percent of trials meet this reporting requirement) (Anderson et al., 2015; Prayle, Hurley and Smythe, 2012).

To obtain data on clinical trial outcomes, I turn to abstracts submitted to the American Society of Clinical Oncology (ASCO) Annual Meeting. 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. Using abstracts from 2004 to 2017, I collect data on the two commonly used clinical outcomes in cancer drug development:

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

[24] 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.

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

treatment group gains in overall survival (the time between randomization and death) and objective response rates (the proportion of trial patients who experience a reduction in tumor size).

4 Effects on 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 privately-funded 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, to gene-cancer pairs without mutation information. 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. The relative difference in clinical trials—between gene-cancer pairs with mutation information and gene-cancer pairs 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 mutation information. Second, the increase could represent a decrease in gene-cancer pairs without mutation information. I am interested in capturing both effects: the relative difference in clinical trials between gene-cancer pairs with and without mutation information.

This empirical strategy removes cancer-level differences in research potential through including gene-cancer fixed effects and estimates the impact of mapping information on clinical trials using variation in the timing of 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 received an information shock), I am able to estimate difference-in-difference regressions with gene-cancer, year fixed effects, and cancer-year linear 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 year in which the Cancer Gene Census (the source of the cancer genes used in this analysis) was first published—and I am interested in quantifying the effect of newly disclosed scientific information (mutation disclosures) on subsequent investment, I drop all gene-cancer pairs with known relationships as of 2004. This results in 49,542 (=50,160 - 618) gene-cancer pairs and 644,046 gene-cancer-year observations. Table 1 summarizes how the gene-cancer-year panel is constructed.

Table 2 provides summary statistics at the gene-cancer level. Panel A shows that by 2016, a mutation was identified in 58 percent of all 49,542 gene-cancer pairs and the median year in which

mutation information was first disclosed is 2011. Figure 4 shows the cumulative distribution of the years in which mutations were first identified among the 168 mapping studies. Only a minority of mutations are likely cancer-causing: Table 2 shows that driver mutations are identified in only 9.5 percent of gene-cancer pairs. Panel C shows that nine percent of all gene-cancers experience at least one privately-funded phase II clinical trial, between 2004-2016, inclusive. Of this nine-percent, the share of gene-cancer pairs that experience a trial testing a pipeline drug (eight percent) is higher than the share of trials testing an approved drug (less than one percent) and a novel drug (five percent).

4.3 Estimating Equation and Assumptions

4.3.1 Estimating Equation

My empirical analysis uses variation in the timing of publicly disclosed mapping information to estimate the effect of mapping information on the level of subsequent research investment within a gene-cancer pair:

$$Y_{g,c,t} = \alpha + \beta PostDisclGeneCancer_{g,c,t} + \delta_{g,c} + \tau_t + \theta_{c,t} + \epsilon_{g,c,t} \quad (1)$$

where $Y_{g,c,t}$ is an indicator for a clinical trial in gene g, cancer c in year t. The $PostDisclGeneCancer$ variable is an indicator for whether gene-cancer gc has been publicly known to be mutated as of that year. This variable 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, $\delta_{g,c}$, 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. Finally, cancer-linear year trends (or cancer-year fixed effects) $\theta_{c,t}$ control for cancer-specific changes that are common across genes within the same cancer. I cluster standard errors at the gene and cancer level.

My coefficient of interest is β. β compares the average level of clinical trial investments in gene-cancers that received mapping information early to those that received mapping information late (or never received mapping information).

4.3.2 Assumptions

A key concern is that the research potential of gene-cancer pairs that were sequenced early are significantly different those that are sequenced late, and that those differences are changing 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 cancer sites that had more available tumor samples, suggesting that the TCGA was directed towards cancers with large market sizes and the resulting estimates may be upward

biased.[26,27]

I explore whether there is cancer-level selection in Figure 5, by comparing proxies for research potential (diagnoses, drugs approvals, 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-year linear 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—i.e., 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 studies used in this analysis, 91 percent employ mapping techniques that are unbiased at the gene level in the sense that they search across 100 percent of the protein-coding genes in the DNA to identify mutations.[28] 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 3 documents a positive relationship between mapping information and subsequent levels of privately funded clinical trials. The first specification in column 1 includes gene-cancer and year fixed effects, and then in subsequent columns I add cancer-year linear trends (column 2) and cancer-year fixed effects (column 3). In all cases, I estimate a strong, positive, and statistically significant effect of mapping on the relative level of subsequently initiated privately-funded clinical trials. The estimates show that information about a mutation in a gene-cancer is associated with a 0.00874-0.00915 percentage point relative increase on average in clinical trials per year. This translates into an increase in the rate of investment on the order of 50 percent of the pre-mapping information sample mean and 37 percent of the full sample mean.

One interpretation of my findings is that if gene-cancer pairs that received mutation-related information had counterfactually experienced the same level of investments as gene-cancer pairs

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

[27] 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)

[28] 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.

that did not, there would have been up to 97 fewer privately-funded clinical trials at the trial level (as opposed to trial-gene-cancer). This translates into roughly 15 fewer cancer drugs, or a 6 percent decrease between 2004 and 2016.[29]

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

$$Y_{g,c,t} = \alpha + \sum_{z} \beta_z \times 1(z) + \delta_{g,c} + \tau_t + \theta_{c,t} + \epsilon_{g,c,t} \tag{2}$$

where $\delta_{g,c}$, τ_t, and $\theta_{c,t}$ represent gene-cancer fixed effects, year fixed effects, and cancer-year linear 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 6 presents estimates of β_z from this regression and corresponds to a dynamic version of Table 3, Column 2. The light blue colored 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 shows in the second year a gene-cancer is publicly disclosed (t = 2) in the graph, there is a persistent increase in the level of subsequently initiated phase II clinical trials in the same gene and cancer. Given that phase I trials typically last several months, this delay is consistent with the earliest date that we might expect an increase in the number of phase II clinical trials.[30]

Taken 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. Having shown that mapping information increases the likelihood of a privately-funded clinical by 50 percent, I now examine what types of mapping information and clinical trials drive these effects.

[29] I calculate these estimates using the pre-mutation information trial averages as my counterfactual. As of 2016, there are 28,524 gene-cancers that receive mutation information (or 28,524 "mapped" gene-cancers). The likelihood of obtaining experiencing 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 484.908 (28,524 $\times$ 0.017) trial-gene-cancer observations in each year. Mapping increases the likelihood of a trial by 0.0074 to 0.0244 (0.017 + 0.0074). This suggests that if the mapped gene-cancers had this likelihood of experiencing a trial, there would be 695.986 (28,524 $\times$ 0.0244) trial-gene-cancer observations in each year. This suggests that mapping leads to a 211.078 (695.986 - 484.908) yearly increase in the number of trial-gene-cancer observations. Since the majority of gene-cancers are experienced in 2011, to be conservative, I allow mapped gene-cancers to be "mapped" for 6 (2016-2011+1) years, resulting in a total of 1266.466 (6 $\times$ 211.078) trial-gene-cancers. To convert this to the trial level, I note that trials are typically associated with 13 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 patients. This trial would appear 4 times). Converting 1266.466 trial-gene-cancer level observations to the trial level gives 97 unique trials. To obtain the estimated number of approved drugs, I take the estimated probability of successfully advancing from phase 2 to regulatory approval (15.2%) from Thomas et al. (2016), which results in an estimated 15 cancer drugs.

[30] For more details, see `https://www.fda.gov/forpatients/approvals/drugs/ucm405622.htm`

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 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 4 shows how the relationship between mapping information and trial quantity varies 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 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, 2012). 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 likely to occur in large intestine cancer. Appendix Table B.1 provides support for this hypothesis. Column 1 shows that clinical trial investment increases by 35 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 information in the same disease (50 percent). Column 2 shows that once the regression controls for mapping information in the same disease, the additional effect of mapping information in a different disease becomes statistically insignificant.

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

The increase in the quantity of privately-funded clinical trials could reflect several types of innovation. First, the increase could represent trials testing drugs approved for one disease in an additional disease ("approved drugs"). Second, as in the case of the PARP inhibitors described in Section 2, the increase could represent testing drugs that are not approved, but have been previously tested in another disease ("pipeline drugs"). Finally, the increase could represent trials testing drugs that have never been tested in any disease before ("novel drugs").

In theory, the relative impact of cancer mapping information on these three 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 (approved or pipeline) drugs.

However, it's possible that mapping information may not shift the level of investment in new uses of approved or pipeline drugs at all. First, as described in the ovarian cancer case study in Section 2, manufacturers of approved and pipeline drugs with substantial resources may have their internal mapping effort which may have already encouraged firms to test their approved or pipeline drugs in multiple diseases. Second, 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.

With this motivation, I examine how large-scale cancer genome mapping efforts influences investment in trials testing new uses of previously-tested (approved or pipeline drugs) and novel drugs. I estimate regressions similar to Equation 1. In this analysis, the dependent variable is set to one if trial is a trial testing an approved drug, pipeline drug, or novel drug.[31]

Estimates are presented in Table 5. Results for trials testing previously-tested drugs are shown in Column 1 (approved drugs) and Column 2 (pipeline drugs). Mapping information increases investment in products that already exist at the time the genetic information is publicly disclosed: investment in trials testing approved drugs increases by 142 percent and in trials testing pipeline drugs by 54 percent.[32] In contrast, Column 3 shows that cancer mapping does not significantly change the rate of investment in trials testing novel drugs. The results are consistent with prior evidence on the relationship between openness and the composition of subsequent R&D: for example, Murray et al. (2016) find that policies which increased access to existing research shifted the composition of follow-on research towards more diverse projects. The findings in this section suggest that publicly available basic scientific data can reinforce the value of existing products and encourage firms to engage in R&D investments that make the most of the products they already have.

4.4.3 Composition of Research Investments: Additional Dimensions

I explore how mapping information shifts the composition of subsequent trials across three additional dimensions.

[31] This analysis categorizes trials based on the novelty of the drug(s) being tested. As a result, the analysis uses the subset (96%) of privately-funded phase II trials with a listed drug intervention. 4% of privately-funded 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 B.3.

[32] I can examine how the effect varies across drug approval novelty, by splitting the trials examined in column 1 into those that test drugs that were approved recently (within 100 days of the clinical trial start date) and drugs that were approved non-recently (more than 100 days before the clinical trial start date). The difference between the effect of information on recently approved and non-recently drugs is not statistically significant.

(i) *Firm Type:* First, I investigate how the impact of mapping information varies across firm types. Smaller (more financially-constrained) firms are less likely to make previous investments in basic science, suggesting that cancer mapping is more likely to disproportionately benefit investments by small firms (Nagaraj, 2017). I divide clinical trials into those conducted by large firms (firms with more than 100 patents prior to 2004) and small firms (all remaining firms). Based on this classification, 68% of trials are conducted by large firms. I find evidence consistent with the view that one mechanism through which public disclosure of basic science information shapes subsequent R&D is through lowering the cost of R&D: Appendix Table B.2 shows that mapping information disproportionately increases investment among smaller (financially-constrained) firms.

(ii) *Disease Type:* I next investigate how private research investment shifts across different disease types. For example, did mapping disproportionately benefit diseases with historically low levels of research investment, or diseases with smaller or larger market sizes? Appendix Table B.3 shows that, in terms of percentage gains, mapping information disproportionately increases investment in diseases with historically low levels of clinical trials, and equally benefits cancers with low and high market sizes.

(iii) *Trial Design Type:* Finally, I explore how cancer mapping shifts private investment in well-designed and non-well-designed trials, where well-designed trials are those that are designed to generate reliable, unbiased scientific evidence. For example, in a randomized controlled trial design, patients are randomly allocated to treatment and control arms. Randomization aims to reduce biases that can be introduced through patient selection. Appendix Table B.4 confirms that mapping leads to a similar increase in both well-designed and non-well-designed trials.

5 Effects on Profitability of Firms' Research Decisions

5.1 Empirical Strategy

In the first set of results, I used a gene-cancer-year panel differences-in-difference research design to show that publicly available, large-scale cancer mapping efforts increases the likelihood that firms initiate phase II clinical trials. As discussed in the ovarian cancer case study in Section 2, a natural follow-up question is to ask how cancer mapping information shapes the profitability of firms' 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 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 likelihood that firms make choices that meet their objectives, I consider whether drugs that are chosen to advance to phase III ultimately result in

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 because, compared to phase I trials, both trial types are relatively well-reported and have standardized outcomes.[33] 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 outcomes or advancement rate.[34] To isolate the impact of mapping information that is most likely to impact the profitability of firm investment, 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.[35] Figure B.4 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, the year in which a large share of gene-cancers first experience mutation information.

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. For phase III, I use gains in overall survival, defined as the gains in time between randomization and death for the treatment group. In total, this results in 2,354 phase II trials and 422 phase III trials, at the trial-gene-cancer level.

Table 6 describes the final trial-gene-cancer level dataset. The table describes trial outcomes, phase II to phase III advancement rates, as well as trial sponsor characteristics. As a proxy for the

[33] For example, a common phase II and phase III outcomes is objective response rate, which is commonly assessed using Response Evaluation Criteria in Solid Tumors (RECIST) criteria (For more details, see: `http://recist.eortc.org/`).

[34] 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. While 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.

[35] This refers to the trial's status as of July 14, 2017. This excludes a large share of firms that are "in-progress."

trial sponsor's R&D experience, I use the log number of clinical trials that the firm initiated in the same cancer, prior to the start of the focal trial. Table 6 shows that, among the trials used in this analysis, phase II trials have advancement rates of 57 percent.[36] Phase II trials with information are significantly less likely to advance to phase III and phase III trials with information are significantly more likely to demonstrate a statistically significant improvement in overall survival.[37]

5.3 Results

5.3.1 Phase II Outcomes

Before turning to analysis of firms' termination-or-continuation decisions, I first establish that 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:

$$Y_{i,g,s,c} = \beta PostDisclGeneCancer_{g,c} + \boldsymbol{X}_i + \epsilon_{i,g,c} \tag{3}$$

where $Y_{i,g,s,c}$ is the log objective response rate for trial i, gene g, cancer site s, and cancer c. The main coefficient of interest, $PostDisclGeneCancer_{g,c}$, is an indicator for whether information about a clinically relevant mutation is available for gene g, 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, disease (gene and cancer) fixed effects, and trial start year linear trends.[38] Standard errors are clustered at the gene and cancer level.

Table 7 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 results suggests that the quality of drug investments are similar across gene-cancer pairs with and without mapping information.

[36] 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. 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.

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

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

5.3.2 Termination Rates for Trials with Weak Outcomes

To understand the relationship between mapping information, phase II outcomes, and phase II advancement rates, I estimate Cox proportional hazard model regressions of the form:

$$h_{i,c,f}(t) = h_{c,f,0}(t) \times exp[\beta PostDisclGeneCancer_{g,c} + \lambda ResponseRate_i + \boldsymbol{X}_i] \quad (4)$$

where $h_{c,f,0}(t)$ is the baseline hazard rate of trial advancement, stratified by cancer and sponsoring firm f's "high R&D experience status."[39] I consider a trial sponsor as having "high R&D experience" if its R&D experience is above the median of the firm experience distribution. $ResponseRate_i$ reflects trial i's phase II clinical outcomes and is the log of trial i's response rate. As before, $\boldsymbol{X}_i$ is a set of trial characteristics, including the trial sponsor's R&D experience and a trial start year linear trend. Standard errors are clustered at the gene and cancer level.

Table 8 presents the estimates. Before examining how the relationship between mapping information and phase II advancement rates varies by phase II outcomes, I examine the relationship between mapping information and phase II advancement rates across the full phase II trial-gene-cancer sample. Column 1 includes only a mapping information indicator and a linear year trend, and then in Column 2 and Column 3, I incrementally add baseline controls. Column 3 shows that, holding phase II clinical trial outcomes constant, phase II trials with information are 49 percent less likely to advance to phase III. As expected, phase II trials with higher response rates are more likely to successfully proceed. Appendix Table B.5 provides additional support for the positive relationship between promising phase II outcomes and phase II to phase III transition rates.

To examine how the relationship between mapping information and phase II advancement rates vary by phase II trial outcomes, I split the sample of phase II trials into those whose response rates were below or equal to median of the cancer-specific response rate distribution (Column 4) and those with response rates above the median (Column 5). Columns 4 suggests that conditional on having weak phase II clinical results, phase II trials with information are significantly less (60 percent) likely to advance to phase III. In contrast, column 5 shows that there is no statistically significant relationship between mapping information and advancement rates among phase II trials with positive clinical results. The results indicate that on average, firms with access to 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

Given the evidence that firms with access to genetic information are more likely to terminate phase II trials with weak clinical evidence, I next examine whether firms with access to mapping information continue to invest in drugs that ultimately experience better clinical outcomes.

[39] Testing the proportional-hazards assumption yielded non-significant results, suggesting that the proportionality assumption holds.

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 my setting and data. Table 9 shows that, even after controlling for disease and firm characteristics, conditional on advancing to phase III, trials with mapping information are 40 percent more likely to demonstrate a statistically significant improvement in overall survival.

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

6 Conclusion

This paper shows that large-scale publicly available scientific maps have important effects on the quantity and profitability of private-sector innovation. Data from large-scale cancer sequencing efforts and privately-funded clinical trials reveal that cancer mapping information leads to an estimated 50 percent increase in privately-funded clinical trials. I estimate that this translates into up to 97 additional trials and approximately 15 additional drugs. These results are driven by response to information about mutations most likely to propel cancer, a result consistent with the prediction that mapping information produces information through helping firms address scientific challenges, thus lowering the cost of clinical development. Further, cancer mapping significantly increases investment in 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 through encouraging firms to make the most of the products they already have.

I analyze the relationship between cancer mapping information and the profitability of firms' research investments by looking at whether cancer mapping information leads firms to make choices that are more informed and consistent with their objectives: I find that firms are more likely to terminate phase II trials with weak outcomes, and to continue drug investments that ultimately more likely to demonstrate promising clinical outcomes. These results complement other research into the importance of understanding what guides productivity in research and development: for example, case studies suggest that access to detailed scientific information can increase the likelihood that any given investment is successful (Cook et al., 2014; Morgan et al., 2018). Guedj and

Scharfstein (2004) find that agency problems play a large role in predicting pharmaceutical firms' continue-or-terminate decisions.

My analysis suggests avenues for future research. As governments consider investments and policies to spur subsequent innovation, understanding the effects of investments in basic scientific knowledge is essential for structuring policy that encourages the efficient development of effective medical technologies. I study one response to large-scale cancer mapping efforts: firm investments in clinical trials. My findings on cancer mapping and trials testing new uses of approved drugs suggest that cancer mapping may also affect off-label drug use, a widespread practice that is poised to continue to grow in importance over the coming years.[40] Future work should focus on understanding how large-scale cancer mapping initiatives directly shape off-label drug demand among patients and health care providers and how this, in turn, affects firms' investment strategies. Further, the focus in this paper has been on cancer drug development and the paper is motivated by the fact that scientific mapping can improve technological search and innovation. However, the logic of the paper may apply to other diseases, including different types of brain disorders.[41]

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 (Bloom et al., 2018). Future work should examine the extent to which publicly available scientific information can help firms in these industries navigate the research and development process.

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

[41] For example, 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/`)

References

Acemoglu, Daron, and Joshua Linn. 2004. "Market Size in Innovation: Theory and Evidence from the Pharmaceutical Industry." *The Quarterly Journal of Economics*, 119(3): 1049–1090.

Adjei, Alex A., Michaele Christian, and Percy Ivy. 2009. "Novel Designs and End Points for Phase II Clinical Trials." *Clinical Cancer Research*, 15(6): 1866–1872.

American Cancer Society. 2018. "Key Statistics for Ovarian Cancer." `https://www.cancer.org/cancer/ovarian-cancer/about/key-statistics.html`, Last Accessed on 2018-10-01.

Anderson, Monique L. Chiswell, Karen, Eric D. Peterson, Asba Tasneen, James Topping, and Robert M. Califf. 2015. "Compliance with Results Reporting at ClinicalTrials.gov." *The New England Journal of Medicine*, 372(11): 1031–1039.

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

AstraZeneca. 2011. "AstraZeneca Annual Report and Form 20-F Information 2011." AstraZeneca.

AstraZeneca. 2017. "Delivering the Next Wave of Scientific Innovation: 2017 - a Year in Review." AstraZeneca.

Avorn, Jerry. 2015. "The $2.6 Billion Pill—Methodologic and Policy Considerations." *The New England Journal of Medicine*, 372(20): 1877–1879.

Azoulay, Pierre, Joshua S. Graff Zivin, Danielle Li, and Bhaven N. Sampat. 2018. "Public R&D Investments and Private-sector Patenting: Evidence from NIH Funding Rules." *The Review of Economic Studies*, 24(1): 91–114. rdy034.

Bach, Peter. 2015. "Indication-Specific Pricing for Cancer Drugs." *JAMA*, 312(16): 1629–1630.

Barker, Anna D., and Francis S. Collins. 2008. "Mapping the Cancer Genome." *Scientific American.*

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.

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

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

Blumenthal, Gideon. 2017. "Primer on Drug Development." *Presentation at Partners in Progress: Cancer Patient Advocates and FDA Public Workshop.*

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: 913–917.

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.

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: 607–616.

Centers for Disease Control and Prevention. 2018. "Deaths: Leading Causes for 2016." In *National Vita Statistics Reports*. Vol. 67.

Chandra, Amitabh, Craig Garthwaithe, and Ariel Dora Stern. 2018. "Characterizing the Drug Development Pipeline for Precision Medicines." NBER Working Paper No. 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:MIT Press.

Cockburn, Ian, and Rebecca Henderson. 1997. "Public-Private Interaction and the Productivity of Pharmaceutical Research." NBER Working Paper No. 6018.

Cohen, Wesley M., and Daniel A. Levinthal. 1990. "Absorptive Capacity: A New Perspective on Learning and Innovation." *Administration Science Quarterly*, 35(1): 128–152.

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

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): 1134–1139.

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): 419–431.

Danzon, Patricia M., and Eric L. Keuffel. 2014. "Regulation of the Pharmaceutical-Biotechnology Industry." In *Economic Regulation and Its Reform: What Have We Learned?*. , ed. Nancy L. Rose, 407–484. Chicago:University of Chicago Press.

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.

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): 1589–1598.

Dhani, Neesha, Dongsheng Tu, Daniel J. Sargent, Lesley Seymour, and Malcom J. Moore. 2009. "Alternate Endpoints for Screening Phase II Studies." *Clinical Cancer Research*, 15(6): 1873–1882.

DiMasi, Joseph A. 2001. "New Drug Development in the United States from 1963 to 1999." *Clinical Pharmacology and Therapeutics*, 69(5): 286–296.

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): 808–818.

DiMasi, Joseph A., L. Feldman, A. Seckler, and A. Wilson. 2010. "Trends in Risks Associated With New Drug Development: Success Rates for Investigational Drugs." *Clinical Pharmacology and Therapeutics*, 87(3): 272–277.

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): 151–185.

Donelan, Ronan, Stuart Walker, and Sam Salek. 2015. "Factors Influencing Quality Decision-Making: Regulatory and Pharmaceutical Industry Perspectives." *Pharmacoepidemiology and Drug Safety*, 24: 319–328.

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): 1589–1598.

Dubois, Pierre, Olivier de Mouzon, Fiona Scott-Morton, and Paul Seabright. 2015. "Market Size and Pharmaceutical Innovation." *The RAND Journal of Economics*, 46(4): 844–871.

Dulbecco, Renato. 1986. "A Turning Point in Cancer Research: Sequencing the Human Genome." *Science*, 231(4742): 1055–1056.

Eisenberg, Rebecca S. 2005. "The Problem of New Uses." *Yale Journal of Health Policy, Law, and Ethics*, 5(2): 717–740.

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: 917–921.

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: 909–928.

Furman, Jeffrey L., and Scott Stern. 2011. "Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research." *The American Economic Review*, 101(5): 1933–1963.

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): 177–183.

GeneEd: Genetics, Education, Discovery. 2018. "DNA, Genes, Chromosomes." `https://geneed.nlm.nih.gov/topic_subtopic.php?tid=15`, Last Accessed on 2018-09-07.

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): 469–474.

Guedj, Ilan, and David Scharfstein. 2004. "Organizational Scope and Investment: Evidence from Drug Development Strategies and Performance of Biopharmaceutical Firms."

Hall, Bronwyn, and John Van Reenen. 2000. "How Effective are Fiscal Incentives for R&D? A Review of the Evidence." *Research Policy*, 29: 449–469.

Hay, Michael, David W. Thomas, John L. Craighead, Celia Economides, and Jesse Rosenthal. 2014. "Clinical Development Success Rates for Investigational Drugs." *Nature Biotechnology*, 32(1): 40–51.

Henderson, Rebecca, and Iain Cockburn. 1996. "Scale, Scope, and Spillovers: The Determinants of Research Productivity in Drug Discovery." *The RAND Journal of Economics*, 27(1): 32–59.

IQIVIA Institute. 2018. "Global Oncology Trends 2018."

Jayaraj, S. 2018. "Scientific Maps and Innovation: Impact of the Human Genome on Drug Discovery." Doctoral dissertation, Rutgers University.

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).

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.

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): 214–218.

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: 219–221.

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): 1063–1070.

Mardis, Elaine R. 2018. "Insights from Large-Scale Cancer Genome Sequencing." *Annual Reviews of Cancer Biology*, 2: 429–444.

Molitor, David, and Leila Agha. 2012. "Technology Adoption Under Uncertainty: Off-label Prescribing in Cancer Care." Dissertation Chapter.

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.

Murray, Fiona, Philippe Aghion, Mathias Detratripont, Julian Kolev, and Scott Stern. 2016. "Of Mice and Academics: Examining the Effect of Openness on Innovation." *American Economic Journal: Economic Policy*, 8(1): 212–252.

Nagaraj, Abhishek. 2017. "The Private Impact of Public Maps: Landsat Satellite Imagery and Gold Exploration."

NCI Center for Cancer Research. n.d.. "Clinical Trial Design." `http://docplayer.net/15224109-Clinical-trial-design-sponsored-by-center-for-cancer-research-national-cancer-insti.html`, Last Accessed on 2018-10-30.

Nelson, Richard R. 1982. "The Role of Knowledge in R&D Efficiency." *The Quarterly Journal of Economics*, 97(3): 453–470.

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: 663–664.

Pfister, David G. 2012. "Off-Label Use of Oncology Drugs: The Need for More Data and Then Some." *Journal of Clinical Oncology*, 30: 584–586.

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): 1389–1398.

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): 540–556.

Roin, Benjamin. 2014. "Solving the Problem of New Uses." *Michigan State Law Review*.

Rosenberg, Nathan. 1990. "Why Do Firms Do Basic Research (With Their Own Money)?" *Research Policy*, 19(2): 165–174.

Sampat, Bhaven N. 2012. "Mission-oriented Biomedical Research at the NIH." *Research Policy*, 41(10): 1729–1741.

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

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

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): 1764–1769.

Sharpe, Paul, and John Keelin. 1998. "How Smithkline Beecham Makes Better Resource-Allocation Decisions."

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

Stratton, Michael R., Peter J. Campbell, and P. Andrew Futreal. 2009. "The Cancer Genome." *Nature*, 458: 719–724.

The Cancer Genome Atlas Research Network. 2011*a*. "Integrated Genomic Analyses of Ovarian Carcinoma." *Nature*, 474: 609–615.

The 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`, Last Accessed on 2018-10-01.

The Cancer Genome Atlas Research Network. 2012. "Comprehensive Genomic Characterization of Squamous Cell Lung Cancers." *Nature*, 489: 519–525.

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

Thomas, David W., Justin Burns, John Audette, Adam Carroll, Corey Dow-Hygelund, and Michael Hay. 2016. "Clinical Development Success Rates: 2006-2015." BIO Industry Analysis.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgement under Uncertainty: Heuristics and Biases." *Science*, 185(4157): 1124–1131.

U.S. Food and Drug Administration. 1998*a*. "Guidance for Industry: FDA Approval of New Cancer Treatment Uses for Marketed and Biological Products."

U.S. Food and Drug Administration. 1998*b*. "Guidance for Industry: Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products."

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."

U.S. Food and Drug Administration. 2007. "Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics."

Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz Jr., and Kenneth W. Kinzler. 2013. "Cancer Genome Landscapes." *Science*, 339: 1546–1558.

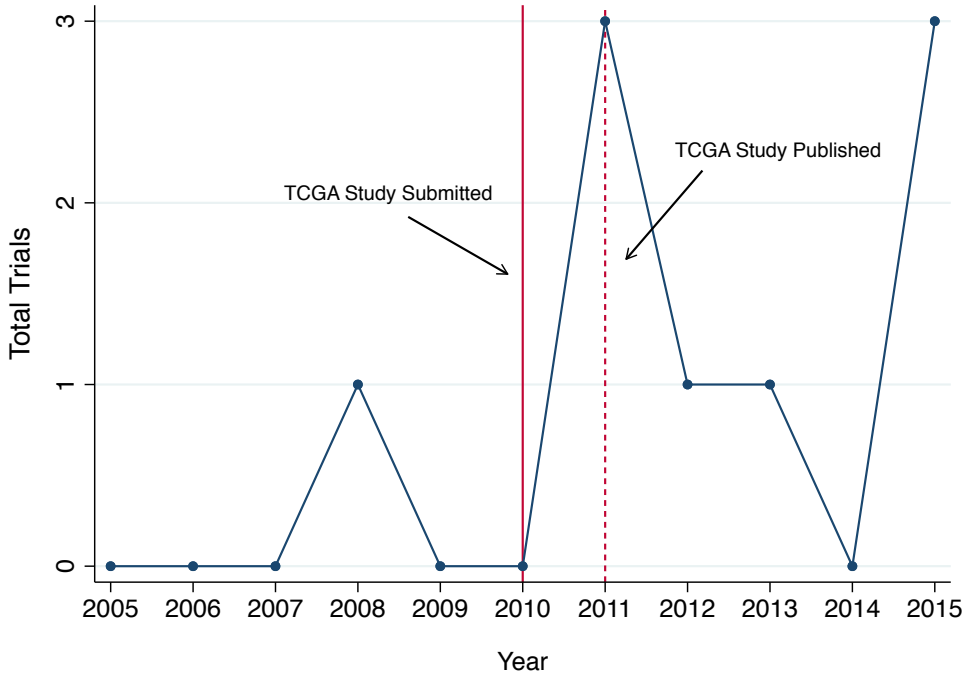
Ward, M.R., and David Dranove. 1995. "The Vertical Chain of Research and Development in the Pharmaceutical Industry." *Economic Inquiry*, 33(1): 70–87.

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: 46–50.

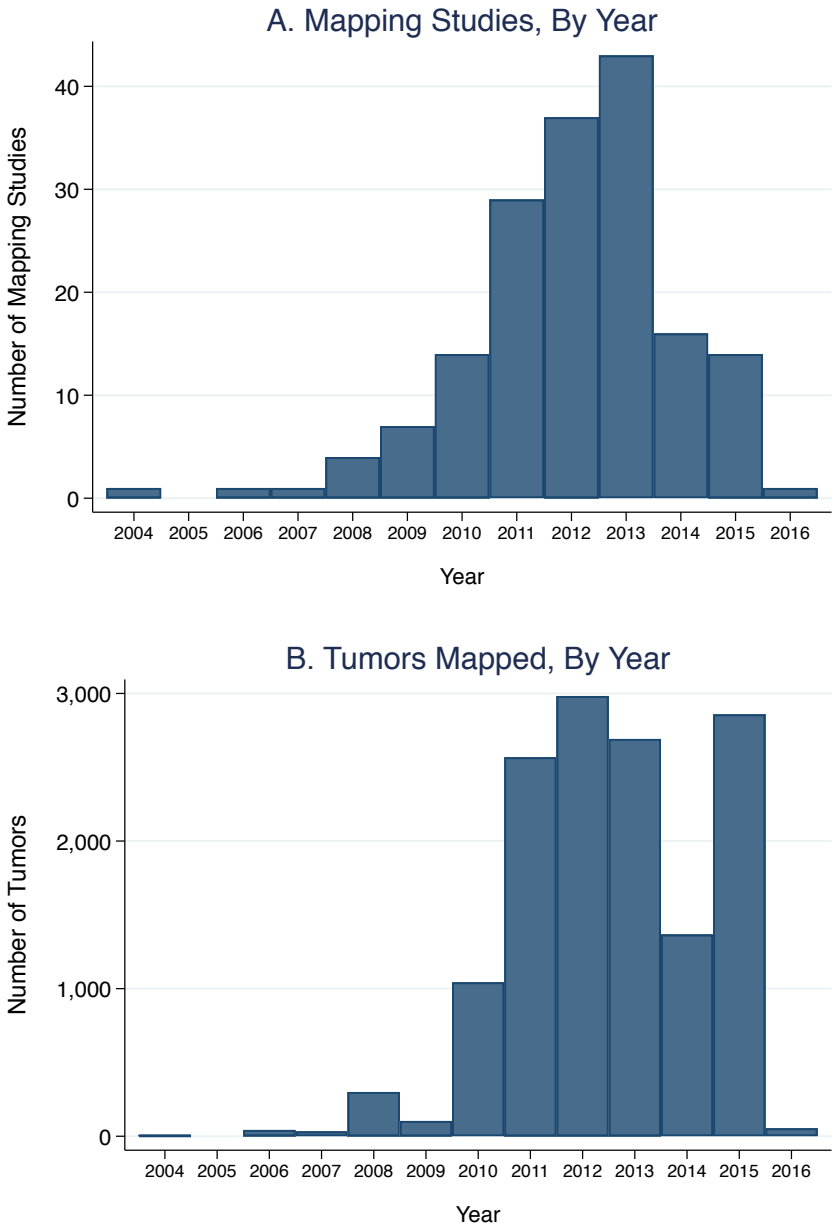
Figures

Figure 1: Trials Enrolling Patients with Ovarian Cancer and BRCA2 Gene Mutations



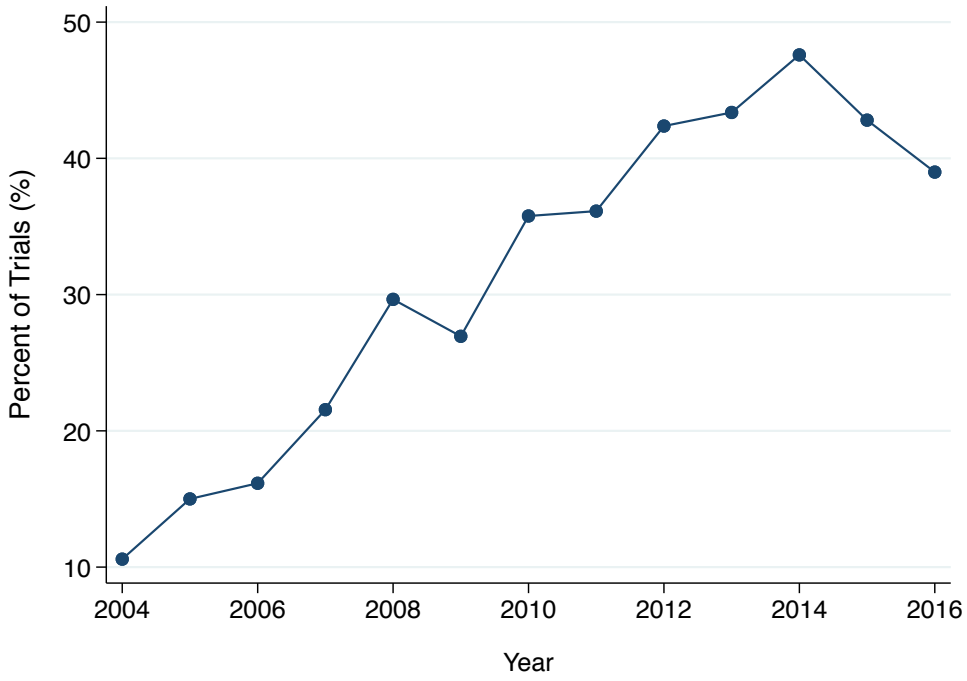
Notes: This figure shows the total number of clinical trials (privately-funded, phase II only) enrolling patients with BRCA2-mutated ovarian cancer in each year from 2005 to 2015. The vertical lines indicated the years in which the TCGA's ovarian cancer study (TCGA, 2011a) was submitted to (solid line) and published in (dashed line) the journal *Nature*. For simplicity, trials testing the drug Olaparib are omitted (see Appendix Figure B.1 for trials testing Olaparib).

Figure 2: Total Cancer Mapping Studies and Mapped Tumors by Year



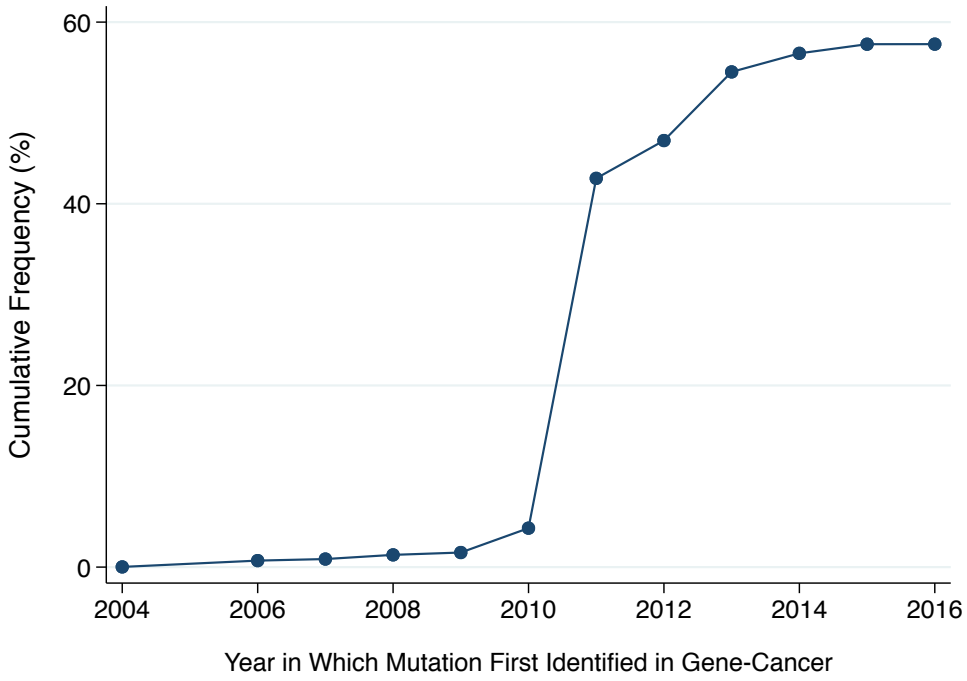
Notes: These figures plot the total number of cancer mapping studies (Panel A) and mapped tumors (Panel B) in each year from 2004 to 2016. The x-axis indicates the year in which the mapping study was submitted to the journal it was ultimately published in. Mapping studies are large and published in a top 25 Genetic journal, based on rankings between 1999-2004. The increase in mapped tumors in 2015 is driven by a single study that sequenced 1144 lung cancer tumors and was submitted to the journal *Nature Genetics* in 2015 (Campbell et al., 2016).

Figure 3: Share of Cancer Trials that Enroll Patients Based on Genes, by Year



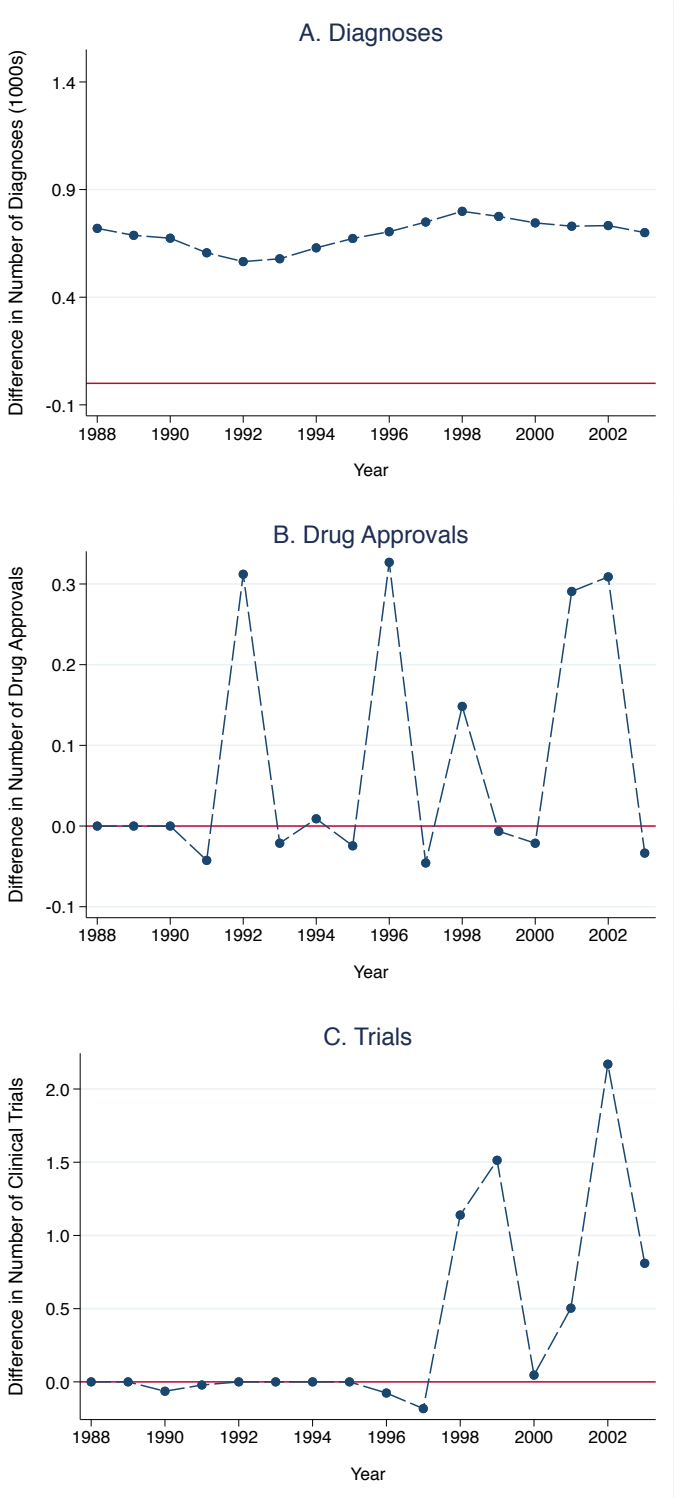
Notes: This figure plots the percent of cancer clinical trials (privately-funded, phase II only) that are gene-related—i.e., genes are used to enroll patients. Observations are at the trial-cancer level.

Figure 4: Cumulative Share of Gene-Cancer Pairs with Mutations Identified by Mapping Studies



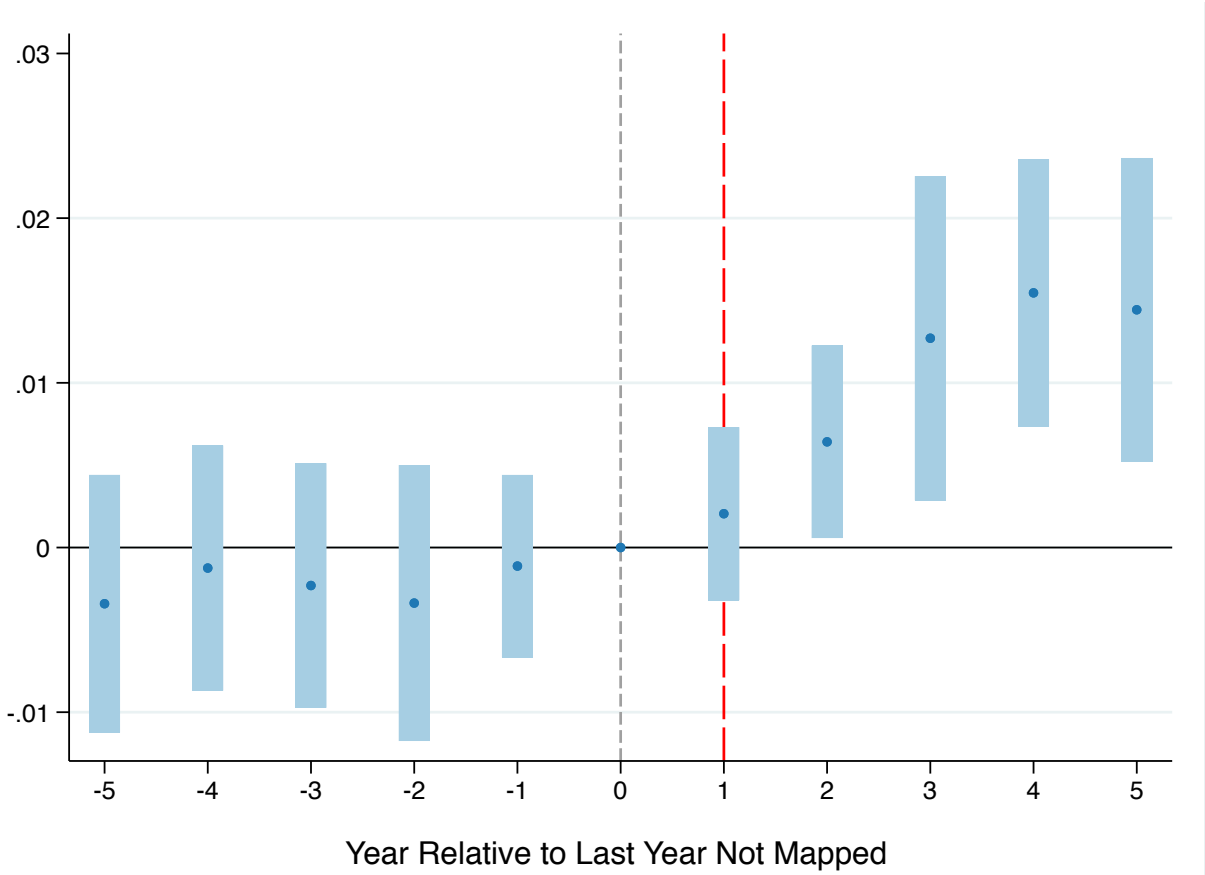
Notes: This figure plots the cumulative share of gene-cancer pairs with mutations identified by cancer mapping studies. As discussed in Section 3, there are 49,542 gene-cancer pairs possible. The period of analysis is 2004-2016. Cancer mapping studies are large and published in a top 25 Genetic journal, based on rankings between 1999-2004.

Figure 5: Examining Cancer-Level Selection



Notes: This figure examines baseline differences between cancers that are first sequenced early (before 2011) and cancers that are first sequenced 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 1998 (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 phase 2, privately-funded clinical trials (Panel C).

Figure 6: Event Study Estimates—Impact of Cancer Mapping Information on Trial Quantity



Notes: This figure plots coefficients (and 95 percent confidence intervals) from the event study specification described in Equation 2 and listed in Table 3, Column 2. On the x-axis are years z relative to a "zero" relative year that marks the last year the gene-cancer was not known to be mutated based on the cancer mapping studies (i.e., year 1 marks the first year a mutation in a gene-cancer was publicly disclosed by a cancer mapping study). As in the specifications in Table 3, this specification is based on gene-cancer-year level observations, the coefficients are estimates from OLS models, the sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004-2016, and the standard errors are robust and clustered at the gene and cancer level. Gene-cancer fixed effects, year fixed effects, and cancer-year linear time trends are included. All trials are privately-funded phase II trials. Cancer mapping studies are large and published in a top 25 Genetic journal, based on rankings between 1999-2004.

Tables

Table 1: Overview of Gene-Cancer-Year Panel Construction

	Count
# of genes (e.g., BRCA1, BRCA2)	627
# of cancer (e.g., ovarian, small intestine)	80
# of cancer groups (e.g., digestive)	19
# of gene - cancer (e.g., BRCA2 - prostate)	50,160
# of gene - cancer, excl. gene-cancer known in 2004	49,542
# of years (2004 to 2016)	13
# of gene - cancer - year (e.g., BRCA2 - prostate - 2004)	652,080
Final Panel: # of gene - cancer - year, excl. gene-cancer known in 2004	**644,046**

Notes: This table provides an overview of how the gene-cancer-year panel was constructed. See Section 4 and the Appendix for more details.

Table 2: Summary Statistics: Gene-Cancer Level Data

	Mean	Standard Deviation	Minimum	Maximum
A. Sequencing				
Share With Mutation (%)	57.58	49.42	0	100
Share With Mutation: Driver Mutation (%)	9.48	29.29	0	100
B. Sequencing Timing				
Year First Mutation	2011.36	1.26	2004	2016
Year First Mutation: Driver Mutation	2012.10	1.23	2008	2016
C. Outcome Variables				
Any Trial (%)	8.99	28.60	0	100
Any Trial With Approved Drug (%)	0.65	8.01	0	100
Any Trial With Pipeline Drug (%)	7.73	26.70	0	100
Any Trial With Novel Drug (%)	5.38	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. The period of analysis is 2004-2016. Share With Mutation: 0/1 = 1 for gene-cancer pairs with mutations identified by cancer mapping studies. Share With Mutation: Driver Mutation: 0/1 = 1 for gene-cancer pairs with driver mutations identified by cancer mapping studies. Cancer mapping studies are large and published in a top 25 Genetic journal, based on rankings between 1999-2004. All trials are privately-funded phase II trials. Any Trial With Approved Drug: 0/1 = 1 for trials testing drugs that have already been approved in the same gene. Any Trial With Pipeline Drug: 0/1 = 1 for trials testing drugs that have not been approved in the same gene, but have been tested previously. Any Trial With Novel Drug: 0/1 = 1 for trials testing drugs that have not been approved in the same gene and have never previously been tested. See text and the appendix for more detailed data and variable descriptions.

Table 3: Does Cancer Mapping Information Influence the Quantity of Trials?

	Dependent Variable: 1(Any Clinical Trials)		
	(1)	(2)	(3)
1(PostDisclGeneCancer)	0.00585**	0.00874***	0.00915**
	(0.00173)	(0.00255)	(0.00302)
Mean of Dep. Var.	0.017	0.017	0.017
Percent Gain	34.41%	51.43%	53.82%
Gene-cancer FEs	yes	yes	yes
Year FEs	yes	yes	yes
Cancer × Linear Year Trend	no	yes	no
Cancer × Year FEs	no	no	yes
Observations	644,046	644,046	644,046

Notes: This table shows the relationship between cancer mapping information and the quantity of subsequent trials. Gene-cancer-year level observations. All estimates are from OLS models. The sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004-2016 (N = 644,046 gene-cancer-year observations). Controls include gene-cancer fixed effects, year fixed effects, and cancer-year linear time trends. Robust standard errors, clustered at the gene and cancer level, are shown in parenthesis. Cancer mapping studies are large and published in a top 25 Genetic journal, based on rankings between 1999-2004. Outcomes: 0/1 = 1 if a privately-funded phase II clinical trial is reported in a gene-cancer-year. PostDisclGeneCancer: 0/1 = 1 for the year after the mutation was disclosed. Mean of Dep. Var. is the mean number of trials in a gene-cancer before the first disclosure of a mutation. Percent Gain is calculated using the pre-mutation information trial mean.
*p<0.10, p**<0.05, ***p<0.01.

Table 4: Impact on Trial Quantity:
Heterogeneity by Clinical Relevance of Cancer Mapping Information

	Dependent Variable: 1(Any Clinical Trials)	
	Driver Mutation (Strong Clinical Relevance)	Passenger Mutation (Weak Clinical Relevance)
	(1)	(2)
1(PostDisclGeneCancer)	0.0392***	0.00511**
	(0.00883)	(0.00233)
Mean of Dep. Var.	0.037	0.017
Percent Gain	106.1%	30.96%
Gene-cancer FEs	yes	yes
Year FEs	yes	yes
Cancer × Linear Year Trend	yes	yes
Observations	644,046	644,046

Notes: This table examines how the relationship between cancer mapping information and quantity of subsequent trials varies across mapping information with differing levels of clinical relevance. Gene-cancer-year level observations. All estimates are from OLS models. The sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004-2016 (N = 644,046 gene-cancer-year observations). Controls include gene-cancer fixed effects, year fixed effects, and cancer-year linear time trends. Robust standard errors, clustered at the gene and cancer level, are shown in parenthesis. Cancer mapping studies are large studies and published in a top 25 Genetic journal, based on rankings between 1999-2004. Outcomes: 0/1 = 1 if a privately-funded phase II clinical trial is reported in a gene-cancer-year. PostDisclGeneCancer: 0/1 = 1 for the year after the mutation was disclosed. Mean of Dep. Var. is the mean number of trials in a gene-cancer before the first disclosure of a mutation. Percent Gain is calculated using the pre-mutation information trial mean. Column 1 shows the effect of the first driver mutation in a gene-cancer, where driver mutations are identified in two ways: 1) the mapping authors list the mutation as a likely driver mutation, or 2) the gene-cancer mutation has occurred in at least ten patients of the same cancer type. All remaining mutations are classified as passenger mutations. Column 2 shows the effect of the first passenger mutation in a gene-cancer.
*p<0.10, p**<0.05, ***p<0.01.

Table 5: What Types of Clinical Trials: New Uses or Novel Drugs?

	Dependent Variable: 1(Any Clinical Trials)		
Clinical Trial Testing:	Previously-Tested Drugs		Novel Drugs
	Approved	Non-Approved	
	(1)	(2)	(3)
1(PostDisclGeneCancer)	0.000920*	0.00632**	0.00249
	(0.000467)	(0.00250)	(0.00164)
Mean of Dep. Var.	0.0006	0.012	0.007
Percent gain	141.98%	53.86%	36.41%
Gene-cancer FEs	yes	yes	yes
Year FEs	yes	yes	yes
Cancer × Linear Year Trend	yes	yes	yes
Observations	644,046	644,046	644,046

Notes: This table examines how the relationship between mapping information and quantity of subsequent trials varies across trials testing previously-tested (approved, pipeline) drugs in additional diseases and trials testing novel drugs. Gene-cancer-year level observations. All estimates are from OLS models. The sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004-2016 (N = 644,046 gene-cancer-year observations). Controls include gene-cancer fixed effects, year fixed effects, and cancer-year linear time trends. Robust standard errors, clustered at the gene-cancer level, are shown in parenthesis. Cancer mapping studies are large studies and published in a top 25 Genetic journal, based on rankings between 1999-2004. Mean of Dep. Var. is the mean number of trials in a gene-cancer before the first disclosure of a mutation. Percent Gain is calculated using the pre-mutation information trial mean. Outcomes: 0/1 = 1 if a phase II clinical trial is reported in a gene-cancer-year and tests: an approved drug (column 1); a pipeline drug (column 2); and a novel drug (column 3). PostDisclGeneCancer: 0/1 = 1 for the year after the mutation was disclosed.
*p<0.10, p**<0.05, ***p<0.01.

Table 6: Summary Statistics: Trial-Gene-Cancer Level Data

	Full (1)	Trials With Info (2)	Trials With No Info (3)	Difference (2)-(3) (4)
A. Phase 2 (N = 2,354)				
Trial Outcome: Log(Response Rate)	2.58	2.63	2.58	0.05
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 (Log(# Clinical Trials))	31.13	80.10	27.51	52.59***
B. Phase 3 (N = 422)				
Trial Outcome: 1(Overall Survival)	0.54	0.69	0.52	0.17**
Firm Experience (Log(# Clinical Trials))	21.93	25.74	21.43	4.30

Notes: This table shows summary statistics at the trial-gene-cancer level. The sample includes trial-gene-cancers that are between from 2004-2016 (excluding gene-cancer pairs known in 2004), have available clinical outcomes data, and are 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. Cancer mapping studies are large studies and published in a top 25 Genetic journal, based on rankings between 1999-2004. Response Rate is the trial's objective response rate. Advance to Phase III: 0/1 = 1 for phase II trials that successfully advance to phase III. Advance to Phase III, Within 4 Years: 0/1 = 1 for phase II trials that successfully advance to phase III, within four years of the trial start date. FirmExperience is the log of the total number of clinical trials the trial sponsor has conducted in the focal cancer, one month prior to the trial start date. Overall Survival: 0/1 = 1 if the trial's treatment group demonstrates a statistically significant (p-value < 0.05) improvement in overall survival relative to the trial's control group or a historical control.
*p<0.10, p**<0.05, ***p<0.01.

Table 7: Impact of Cancer Mapping Information on Phase II Clinical Outcomes

	Dependent Variable: Log(Response Rate)	
	(1)	(2)
1(PostDisclGeneCancer)	0.343	0.350
	(0.258)	(0.261)
Firm Experience (Log (# Clinical Trials))		-0.112
		(0.139)
Mean of Dep. Var.	2.583	2.583
Cancer FEs	yes	yes
Gene FEs	yes	yes
Linear Year Trend	yes	yes
Nb. Trial-Gene-Cancers	2,323	2,323
Nb. Trials	159	159
Nb. Genes	61	61
Nb. Cancers	80	80
R^2	0.666	0.672

Notes: This table shows the relationship between mapping information and phase II clinical outcomes, as measured by objective response rate. Trial-gene-cancer level observations. All estimates are from OLS regressions. The sample includes all phase II trial-gene-cancers that are between from 2004-2016 (excluding gene-cancer pairs known in 2004), have available clinical outcomes data, and are completed or terminated as of July 14, 2017. There are fewer than 2,354 observations because the estimation drops observations with a gene or cancer that just shows up once. Controls include cancer fixed effects and gene fixed effects. Robust standard errors, clustered at the gene and cancer level, are shown in parenthesis. Cancer mapping studies are large studies and published in a top 25 Genetic journal, based on rankings between 1999-2004. PostDisclGeneCancer: 0/1 = 1 for whether driver (clinically-relevant) mutation information was disclosed for the gene-cancer by the start of the clinical trial. FirmExperience is the log of the total number of clinical trials the trial sponsor has conducted in the focal cancer, one month prior to the trial start date.
*p<0.10, p**<0.05, ***p<0.01.

Table 8: Impact of Cancer Mapping Information on Phase II to Phase III Transitions

	Dependent Variable: Advancing from Phase II to Phase III				
	Full Sample			Split Sample	
				Response ≤ Median	Response > Median
	(1)	(2)	(3)	(4)	(5)
1(PostDisclGeneCancer)	-0.696	-0.616	-0.672*	-0.913**	-0.0803
	(0.478)	(0.436)	(0.394)	(0.451)	(0.453)
Firm Experience (Log (# Clinical Trials))		-0.257*	-0.187	-0.330	-0.00479
		(0.155)	(0.135)	(0.238)	(0.136)
Phase II Outcome (Log(Response Rate))			0.235***	-0.0365	1.717***
			(0.0832)	(0.155)	(0.387)
Percent Change	-50.15%	-46.00%	-48.93%	-59.86%	-7.72%
Linear Year Trend	yes	yes	yes	yes	yes
Nb. Observations	2,354	2,354	2,354	1,287	1,067
Nb. Trials	164	164	164	78	94
Nb. Genes	92	92	92	61	77
Nb. Cancers	80	80	80	80	74
Log Likelihood	-3520	-3500	-3436	-1559	-1258

Notes: This table shows the relationship between mapping information and phase II transition rates. Trial-gene-cancer level observations. Estimates are from Cox proportional hazard models, stratified by cancer and large firm status. The sample includes all phase II trial-gene-cancers that are between from 2004-2016 (excluding gene-cancer pairs known in 2004), have available clinical outcomes data, and are completed or terminated as of July 14, 2017. Column 4 shows estimates for phase II trials that have a response rate equal to or below the median of the cancer-specific distribution of response rates. Column 5 show estimates for phase II trials that have a response rate above the median of the cancer-specific distribution of response rates. Controls include a linear year time trend. Robust standard errors, clustered at the gene and cancer level, are shown in parenthesis. Cancer mapping studies are large studies and published in a top 25 Genetic journal, based on rankings between 1999-2004. PostDisclGeneCancer: 0/1 = 1 for whether driver (clinically-relevant) mutation information was disclosed for the gene-cancer by the start of the clinical trial. FirmExperience is the log of the total number of clinical trials the trial sponsor has conducted in the focal cancer, one month prior to the trial start date. Response Rate refers to the trial's objective response rate, or the share of patients who respond to treatment.
*p<0.10, p**<0.05, ***p<0.01.

Table 9: Impact of Mapping Information on Phase III Clinical Outcomes

	Dependent Variable: 1(Increase in Overall Survival))	
	(1)	(2)
1(PostDisclGeneCancer)	0.183**	0.217**
	(0.0767)	(0.0936)
Firm Experience (Log (# Clinical Trials))		0.0413
		(0.0572)
Mean of Dep. Var.	0.540	0.540
Percent Gain	33.87%	40.21%
Gene FEs	yes	yes
Cancer FEs	yes	yes
Year Linear Trends	yes	yes
Nb. Trial-Gene-Cancers	394	394
Nb. Trials	71	71
Nb. Genes	31	31
Nb. Cancers	31	31
R^2	0.410	0.417

Notes: This table shows the relationship between mapping information and phase III clinical outcomes, as measured by whether the phase III trial's treatment group demonstrates a statistically significant increase in overall survival. Trial-gene-cancer level observations. All estimates are from OLS regressions. The sample includes all phase II trial-gene-cancers that are between from 2004-2016 (excluding gene-cancer pairs known in 2004), have available clinical outcomes data, and are completed or terminated as of July 14, 2017. There are fewer than 422 observations because the estimation drops observations with a gene or cancer that just shows up once. Controls include gene fixed effects and cancer fixed effects. Robust standard errors, clustered at the gene and cancer level, are shown in parenthesis. Cancer mapping studies are large studies and published in a top 25 Genetic journal, based on rankings between 1999-2004. Outcome: $0/1 = 1$ if the trial's treatment group demonstrates a statistically significant (p-value < 0.05) improvement in overall survival relative to the trial's control group or a historical control. PostDisclGeneCancer: $0/1 = 1$ for whether driver (clinically-relevant) mutation information was disclosed for the gene-cancer by the start of the clinical trial. Firm-Experience is the log of the total number of clinical trials the trial sponsor has conducted in the focal cancer, one month prior to the trial start date.
*p<0.10, p**<0.05, ***p<0.01.

Appendix

Table of Contents

Appendix A: Data Description

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

A1. Cancer Mapping Data

Mapping Studies

Cancer mapping data comes from cBioPortal for Cancer Genomes[42] and the Catalogue of Somatic Mutations in Cancer.[43] These two publicly available databases contain datasets from many published cancer mapping studies. I focus on the set of cancer mapping studies that are high impact and large (in terms of the number of tumors mapped).

To identify high impact cancer mapping studies, I isolate the list of cancer mapping studies published top genetics journals. A top genetics journal is one that is ranked highly under the Scimago Journal & Country Rank (SJR) system, a yearly ranking scheme that ranks journals using a citation-based algorithm.[44] The SJR measures a journal's influence by looking at the number of citations received by a journal during the past three years (Gonzalez-Pereira et al., 2009). I define a genetics journal to be highly ranked if it is ranked in the top 25 of 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.[45]

A mapping study is defined as "large" if is published in cBioPortal, which focuses on 'large-scale cancer genomics projects" (Cerami et al., 2012) or in COSMIC's "Whole Genome & Large-scale Systematic Screens" sequencing study database.[46] This results in gene-cancer level mutation information from 168 high quality and large cancer mapping studies.

Mutation Data

There are two key facts to note about the mutations analyzed in this paper: first, I focus on mutations that occur in the protein-coding region of the DNA and are non-inherited: non-silent somatic mutations. Somatic mutations are DNA aberrations that occur after conception. According to Stratton, Campbell and Futreal (2009), "all cancers arise as a result of somatically acquired changes in the DNA of cancer cells." Somatic mutations differ from germline mutations, which are inherited.[47] I exclude silent mutations which are mutations that occur in the non-protein coding

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

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

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

[45] 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.

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

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

region of the DNA. The final list of included mutations include: missense, nonsense, insertions, deletions, frameshift, nonstop extension.[48]

Second, this paper focuses primarily on mutations, but other types of genetic alterations may contribute to the progression and growth of cancer. These genetic alterations include: DNA rearrangements, where DNA is broken and then added to a DNA segment from another part of the genome; deletions of small or large parts of the DNA; amplifications or excess copies of a gene. For more details, see Stratton, Campbell and Futreal (2009) and Vogelstein et al. (2013).

A2. Identifying Well-Designed and Non-Well-designed Trials

This section describes how clinical trials were classified into well-designed and non-well-designed trials. Based on recommendations in the scientific literature (Adjei, Christian and Ivy, 2009; Berger and Alperson, 2009; Blumenthal, 2017; Dhani et al., 2009; U.S. Food and Drug Administration, 2007; Grossman et al., 2017; Kemp and Prasad, 2017; NCI Center for Cancer Research, n.d.; Prasad et al., 2015; Seymour et al., 2010), I classify phase II trials as well-designed if they satisfied one of the following three criteria:

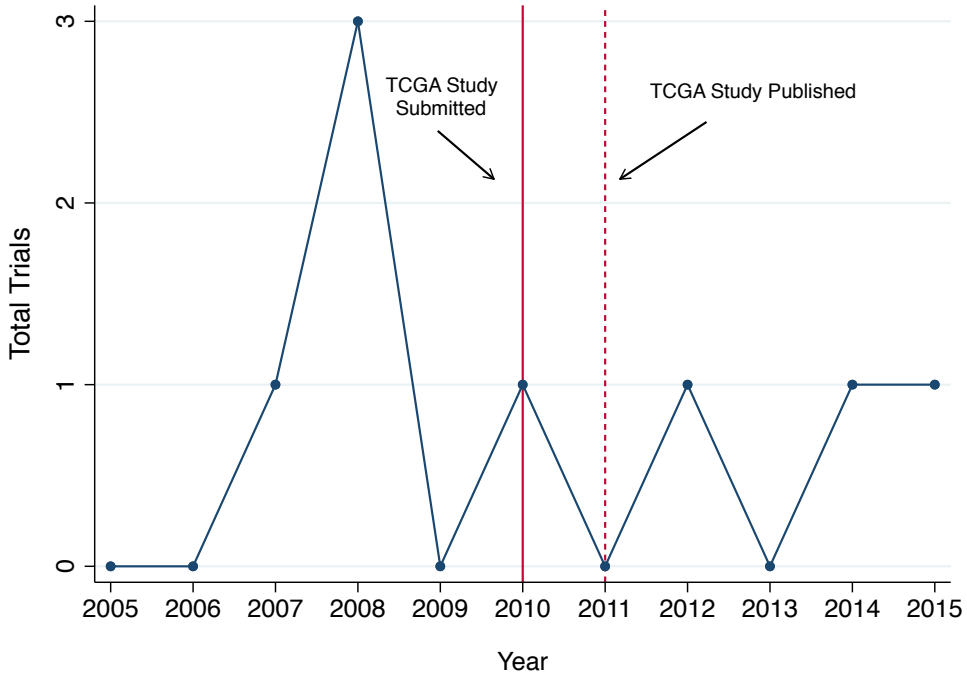
1. Randomized, controlled, overall survival endpoint
2. Randomized, controlled, validated surrogated endpoint
3. Non-randomized, controlled, validated surrogate endpoint

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

[48] For more details, see `https://ghr.nlm.nih.gov/primer/mutationsanddisorders/possiblemutations`

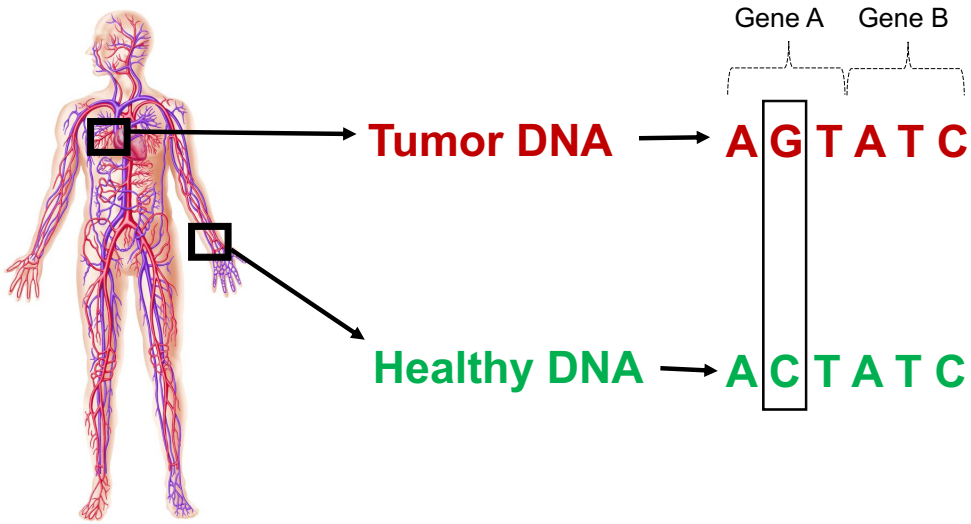
Appendix B: Additional Figures and Tables

Figure B.1: Trials Enrolling Patients with Ovarian Cancer and BRCA2 Gene Mutations, Olaparib Only



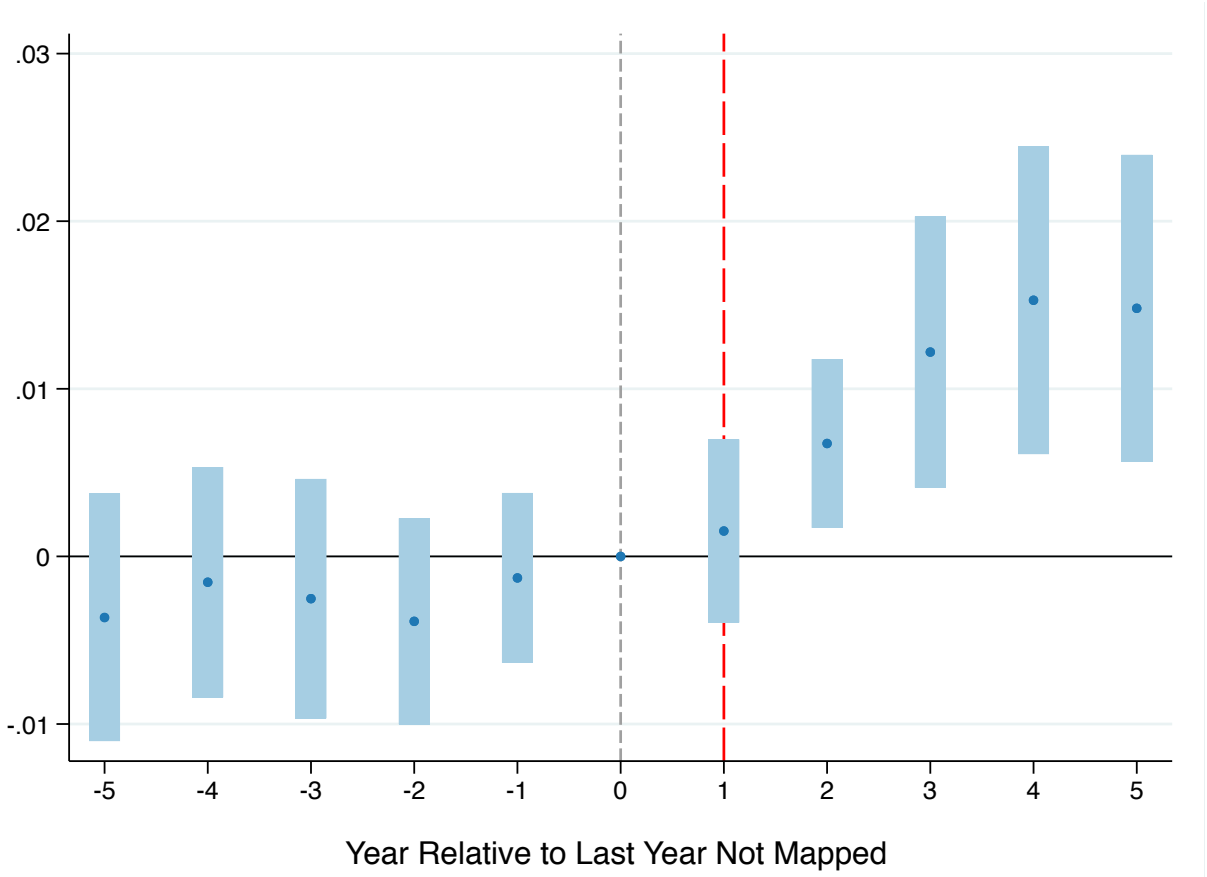
Notes: This figure shows the total number of clinical trials (privately-funded, phase II only) enrolling patients with BRCA2-mutated ovarian cancer and testing Olaparib in each year from 2005 to 2015. The vertical lines indicated the years in which the TCGA's ovarian cancer study (TCGA, 2011a) was submitted to (solid line) and published in (dashed line) the journal *Nature*.

Figure B.2: Overview of Scientific Background on Cancer Genome Sequencing



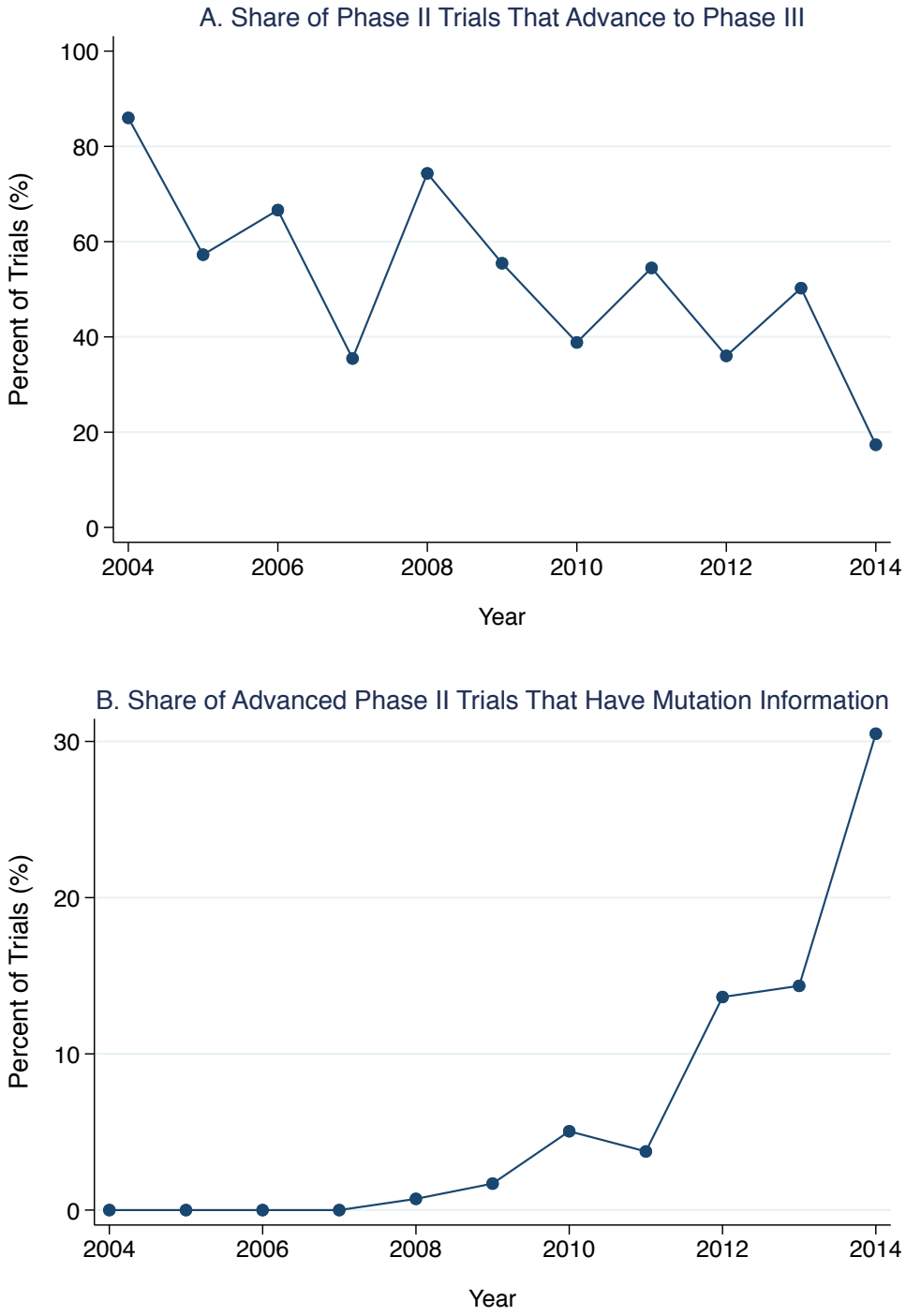
Notes: This figure graphically summarizes the scientific background described in Section 3. A individual's genome is all the DNA contained in a particular a cell. DNA is comprised of four bases : adenine (A), cytosine (C), guanine (G), and thymine (T). The unique sequence of these four DNA bases—A, C, 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 B.3: Event Study Estimates—Impact of Mapping on Trial Quantity, Trials with Non-Missing Intervention



Notes: This figure shows the relationship between cancer mapping and the subsequent quantity of clinical trials using the subset of trials with non-missing intervention data. The figure plots coefficients (and 95 percent confidence intervals) from the event study specification described in Equation 2. On the x-axis are years z relative to a "zero" relative year that marks the last year the gene-cancer was not known to be mutated based on the cancer mapping studies (i.e., year 1 marks the first year a mutation in a gene-cancer was publicly disclosed by a cancer mapping study). As in the specifications in Table 3, this specification is based on gene-cancer-year level observations, the coefficients are estimates from OLS models, the sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004-2016, and the standard errors are robust and clustered at the gene and cancer level. Gene-cancer fixed effects, year fixed effects, and cancer-year linear time trends are included. All trials are privately-funded phase II trials. Cancer mapping studies are large and published in a top 25 Genetic journal, based on rankings between 1999-2004.

Figure B.4: Trial Advancement Rates, by Year



Notes: Panel A plots the percent of privately-funded phase II clinical trials that successfully advance to phase III. Panel B plots the percent of privately-funded phase II clinical trials that are initiated in gene-cancer pairs with mutation information, as a share of the total number of privately-funded phase II trials that successfully advance to phase III. In this figure, trials are classified as having successfully advanced to phase III if they transition to phase III within 4 years of the phase II trial start date. 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 B.1: Impact of Mapping Information in Same Gene, Related Cancer

	Dependent Variable: 1(Any Clinical Trials)	
	(1)	(2)
1(PostDisclGeneCancer)		0.00817***
		(0.00230)
1(PostDisclGeneCancerGroup)	0.00502**	0.00129
	(0.00233)	(0.00197)
Mean of Dep. Var.		
Same Gene-Same Cancer	–	0.017
Same Gene-Related Cancer	0.014	0.014
Percent Gain		
PostDisclGeneCancer	–	57.16%
PostDisclGeneCancerGroup	35.10%	9.032%
Gene-cancer FEs	yes	yes
Year FEs	yes	yes
Cancer × Linear Year Trend	yes	yes
Observations	644,046	644,046

Notes: This table examines how the relationship between cancer mapping information and quantity of subsequent trials varies across mapping information with differing levels of clinical relevance. Gene-cancer-year level observations. All estimates are from OLS models. The sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004-2016 (N = 644,046 gene-cancer-year observations). Controls include gene-cancer fixed effects, year fixed effects, and cancer-year linear time trends. Robust standard errors, clustered at the gene and cancer level, are shown in parenthesis. Cancer mapping studies are large studies and published in a top 25 Genetic journal, based on rankings between 1999-2004. Outcomes: 0/1 = 1 if a privately-funded phase II clinical trial is reported in a gene-cancer-year. PostDisclGeneCancer: 0/1 = 1 for the year after the mutation in a same gene-same cancer was disclosed. PostDisclGeneCancerGroup: 0/1 = 1 for the year after the mutation in a same gene-related cancer was disclosed. Cancers are classified as related if they are in the same cancer site group, based on the Surveillance, Epidemiology, and End Results (SEER) classification. Mean of Dep. Var. is the mean number of trials in a gene-cancer before the first disclosure of a mutation in the same-gene, same-cancer or the same-gene, related-cancer. The same-gene, same-cancer trial mean is used to calculate percentage gain for PostDisclGeneCancer. The same-gene, related-cancer trial mean is used to calculate the percent gain for PostDisclGeneCancerGroup.
*p<0.10, p**<0.05, ***p<0.01.

Table B.2: Impact of Mapping Information for Different Types of Firms

	Dependent Variable: 1(Any Clinical Trials)	
	Large Firms	Small Firms
	(1)	(2)
1(PostDisclGeneCancer)	0.00483***	0.00590***
	(0.000764)	(0.000615)
Mean of Dep. Var.	0.013	0.007
Percent Gain	37.46%	83.44%
Gene-cancer FEs	yes	yes
Year FEs	yes	yes
Cancer × Linear Year Trend	yes	yes
Observations	644,046	644,046
Test for Diff. in Percent Gain		
P[$(\beta_1 = \beta_2)$] =	0.00	

Notes: This table examines how the relationship between mapping information and quantity of subsequent trials for different types of firms. Gene-cancer-year level observations. All estimates are from seemingly unrelated models. The sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004-2016 (N = 644,046 gene-cancer-year observations). Controls include gene-cancer fixed effects, year fixed effects, and cancer-year linear time trends. Robust standard errors, clustered at the gene-cancer level, are shown in parenthesis. Cancer mapping studies are large studies and published in a top 25 Genetic journal, based on rankings between 1999-2004. Mean of Dep. Var. is the mean number of trials in a gene-cancer before the first disclosure of a mutation. Percent Gain is calculated using the pre-mutation information trial mean. Outcomes: 0/1 = 1 if a phase II clinical trial conducted by a large firm (column 1) or small firm (column 2) is reported in a gene-cancer-year. Large firms are those with more than 100 patents prior to 2004. All remaining firms are classified as small firms. PostDisclGeneCancer: 0/1 = 1 for the year after the mutation was disclosed. p<0.10, p**<0.05, ***p<0.01.

Table B.3: Impact of Mapping Information for Different Types of Diseases

	Dependent Variable: 1(Any Clinical Trials)			
	Pre-2004 Clinical Trials		Pre-2004 Market Size	
	≤ Median	> Median	≤ Median	> Median
	(1)	(2)	(3)	(4)
PostDisclGeneCancer	0.00675***	0.0873***	0.00828***	0.00885***
	(0.000786)	(0.0165)	(0.00123)	(0.00115)
Mean of Dep. Var.	0.012	0.276	0.015	0.018
Percent Gain	56.27%	31.63%	55.20%	49.16%
Gene-cancer FEs	yes	yes	yes	yes
Year FEs	yes	yes	yes	yes
Cancer × Linear Year Trend	yes	yes	yes	yes
Observations	644046	644046	644046	644046
Tests for Percent Gain				
P[$(\beta_{below} = \beta_{above})$]	0.01		0.56	

Notes: This table examines how the relationship between mapping information and quantity of subsequent trials varies across types of disease. Gene-cancer-year level observations. All estimates are from seemingly unrelated models. The sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004-2016 (N = 644,046 gene-cancer-year observations). Controls include gene-cancer fixed effects, year fixed effects, and cancer-year linear time trends. Robust standard errors, clustered at the gene-cancer level, are shown in parenthesis. Cancer mapping studies are large studies and published in a top 25 Genetic journal, based on rankings between 1999-2004. Mean of Dep. Var. is the mean number of trials in a gene-cancer before the first disclosure of a mutation. Percent Gain is calculated using the pre-mutation information trial mean. Outcomes: 0/1 = 1 if a privately-funded phase II clinical trial is reported in a gene-cancer-year. PostDisclGeneCancer: 0/1 = 1 for the year after the mutation was disclosed. Columns 1 and 2 shows how the effect of mapping information varies across gene-cancers with low and high levels of clinical trial investment (calculated based on pre-2004 clinical trial levels). Columns 3 and 4 shows how the effect of mapping information varies across gene-cancers with low and high levels of market size (measured at the cancer level, calculated based on pre-2004 diagnoses levels). P-values are from 2-sided t-tests.
*p<0.10, p**<0.05, ***p<0.01.

Table B.4: Impact of Mapping Information for Different Trial Design Types

	Dependent Variable: 1(Any Clinical Trials)	
	Well-Designed	Non-Well-Designed
	(1)	(2)
1(PostDisclGeneCancer)	0.00226***	0.00692***
	(0.000410)	(0.000835)
Mean of Dep. Var.	0.004	0.015
Percent Gain	63.63%	46.04%
Gene-cancer FEs	yes	yes
Year FEs	yes	yes
Cancer × Linear Year Trend	yes	yes
Observations	644,046	644,046
Test for Diff. in Percent Gain		
P[($\beta_1 = \beta_2$)] =	0.14	

Notes: This table examines how the relationship between mapping information and quantity of subsequent trials varies across well-designed and non-well-designed trials. Gene-cancer-year level observations. All estimates are from seemingly unrelated models. The sample includes all gene-cancer-years (excluding gene-cancer pairs known in 2004) from 2004-2016 (N = 644,046 gene-cancer-year observations). Controls include gene-cancer fixed effects, year fixed effects, and cancer-year linear time trends. Robust standard errors, clustered at the gene-cancer level, are shown in parenthesis. Cancer mapping studies are large studies and published in a top 25 Genetic journal, based on rankings between 1999-2004. Mean of Dep. Var. is the mean number of trials in a gene-cancer before the first disclosure of a mutation. Percent Gain is calculated using the pre-mutation information trial mean. Outcomes: 0/1 = 1 if a phase II clinical trial is reported in a gene-cancer-year and is a well-designed trial (column 1) or non-well-designed trial (column 2). See Appendix A2 for a description of how well-designed and non-well-designed trials are identified. PostDisclGeneCancer: 0/1 = 1 for the year after the mutation was disclosed.
p<0.10, p**<0.05, ***p<0.01.

Table B.5: Phase II Outcomes and Phase II to Phase III Transitions

	Dependent Variable: Time to Phase III	
	(1)	(2)
Phase II Outcome (Log(Response Rate))	0.245***	0.232***
	(0.0813)	(0.0831)
Firm Experience (Log (# Clinical Trials))		-0.198
		(0.137)
Percent Change	27.784	26.169
Linear Year Trend	yes	yes
Nb. Observations	2,354	2,354
Nb. Trials	164	164
Nb. Genes	92	92
Nb. Cancers	80	80
Log Likelihood	-3455	-3443

Notes: This table shows the relationship between phase II outcomes and phase II transition rates. Trial-gene-cancer level observations. Estimates are from Cox proportional hazard models, stratified by cancer and large firm status. The sample includes all phase II trial-gene-cancers that are between from 2004-2016 (excluding gene-cancer pairs known in 2004), have available clinical outcomes data, and are completed or terminated as of July 14, 2017. Controls include a linear year time trend. Robust standard errors, clustered at the gene and cancer level, are shown in parenthesis. Cancer mapping studies are large studies and published in a top 25 Genetic journal, based on rankings between 1999-2004. DisclGeneCancer: 0/1 = 1 for whether a driver (clinically-relevant) mutation information was disclosed for the gene-cancer by the start of the clinical trial. FirmExperience is the log of the total number of clinical trials the trial sponsor has conducted in the focal cancer, one month prior to the trial start date. Response Rate refers to the trial's objective response rate, or the share of patients who respond to treatment.
*p<0.10, p**<0.05, ***p<0.01.