Finding Diamonds in the Rough: Data-Driven Opportunities and Pharmaceutical Innovation *

Matteo Tranchero
The Wharton School

December 23, 2024

Abstract

Big data are increasingly used to make predictions about the value of uncertain investments, thereby helping firms identify innovation opportunities without the need for domain knowledge. This trend has raised questions about which firms will primarily benefit from the availability of these data-driven predictions. Contrary to existing research suggesting that data-driven predictions level the playing field for firms lacking domain knowledge, I argue—using a simple theoretical framework—that these predictions reinforce the competitive advantage of firms with domain knowledge. In high-stakes contexts like innovation, where returns are skewed and only a few leads can be pursued, domain knowledge helps evaluate predictions and avoid false positives. I test this idea using novel data on the pharmaceutical industry, exploiting the features of genome-wide association studies (GWAS) that provide data-driven predictions about new drug targets. The results show that GWAS stimulate corporate investments in innovation, yet around one-third of these efforts are misallocated toward false positive predictions. Companies lacking domain knowledge react more strongly but are disproportionally likely to fall into the trap of false positives. Instead, domain knowledge helps firms pursue fewer alternatives that are more likely to be the best opportunities. Together, the results show that even if data-driven predictions are valuable in innovation, domain knowledge remains a crucial source of competitive advantage in the age of big data technologies.

*E-mail: mtranc@wharton.upenn.edu. Website: https://www.matteotranchero.com/.

1 Introduction

The data revolution is transforming how companies operate, with firms increasingly using machine learning and artificial intelligence (AI) to analyze large amounts of data and make predictions about future events or uncertain choices (Brynjolfsson and McElheran, 2016; Brynjolfsson et al., 2021; Cockburn et al., 2019). Beyond their use in informing operational decisions, data-driven approaches are becoming common when making strategic decisions in high-risk contexts such as innovation or entrepreneurship (Agrawal et al., 2018b). For instance, venture capitalists use predictive algorithms to identify successful start-ups (Bhatia and Dushnitsky, 2023), consumer goods corporations rely on social media data to forecast the revenues of prospective products (Allen and McDonald, 2024), and pharmaceutical firms try to predict the therapeutic properties of millions of potential drugs (Heaven, 2023; Lou and Wu, 2021). In both traditional and high-tech sectors, data-driven predictions are reshaping how firms find innovation opportunities.

Given the prevalence of this phenomenon, a central question is how predictive technologies affect competitive dynamics. Traditionally, scholars have argued that domain knowledge allows firms to foresee the value of alternative choices and better recognize opportunities, thus enhancing performance (Cohen and Levinthal, 1990; Gruber et al., 2008; Shane, 2000). However, data-driven predictions can serve the same role without the need for domain knowledge (Arora and Fosfuri, 2005; Agrawal et al., 2024; Balasubramanian et al., 2022; Kim, 2024b). Since obtaining predictions is now cheaper and faster than accumulating expertise, performance gains should favor those lacking domain knowledge. While we still lack direct evidence at the firm level, recent research at the individual level would support this idea: predictive technologies help mostly novices, hence substituting for their lack of knowledge (Brynjolfsson et al., 2024; Dell'Acqua et al., 2023; Noy and Zhang, 2023). It is thus plausible to expect a similar democratization effect when extrapolating this logic to firm innovation. Data-driven predictions might erode the role of domain knowledge as a source of competitive advantage, leveling the playing field through what some have hailed as the "death of theory" (Anderson, 2008).

Even though these arguments suggest that domain knowledge might no longer be necessary in the age of big data, there are reasons to think otherwise. While the limited existing evidence does show that data-driven predictions can yield significant competitive gains (Conti et al., 2024; Kao, 2023), even data-driven predictions that are valuable on average can mask several false positive findings (Berman and Van den Bulte, 2022). When validating predictions is relatively inexpensive, such as in operational decisions, firms can afford a trial and error approach to experimentally learn which ones unlock performance improvements (Koning et al., 2022). However, this logic might not extend to innovative contexts, where testing a prediction is costly and the distribution of innovation returns is highly skewed. Financial and organizational constraints preclude the firm from implementing

every prediction, and betting on the wrong ones can dramatically hurt performance and dynamically foreclose other courses of action (Eggers, 2012; Gans et al., 2019). These features of the innovation process imply that achieving gains from predictive technologies poses unique challenges.

Motivated by this logic, I argue that domain knowledge complements data-driven predictions in the context of innovation. While it is true that predictions can identify opportunities without requiring comprehension, domain knowledge becomes essential in judging their plausibility to rule out false positives (Agrawal et al., 2018a; Allen and Choudhury, 2022; Arora and Gambardella, 1994b). I suggest that this can be visualized as a two-stage process: predictive algorithms generate a shortlist of potential opportunities, while domain knowledge is used to evaluate them and funnel resources to the most promising ones. This simple framework clarifies how the same data-driven predictions can have heterogeneous performance implications among firms depending on their domain knowledge. Firms without expertise in the domain will rely more on data-driven predictions to guide their investments; however, such firms will be less likely to uncover actual opportunities in the data despite pursuing a larger number of predictions. Acknowledging the consequences of false positives implies the opposite of what existing research would suggest: data-driven predictions in innovation reinforce the advantage of incumbents with the domain knowledge to evaluate them rather than help novices catch up.

The validity of my theoretical framework is ultimately an empirical question. However, assessing how data-driven predictions and domain knowledge interact in shaping innovation performance presents several empirical issues. First, it requires observing the risk set of potential opportunities that firms could pursue, which is usually difficult to ascertain. Furthermore, this measurement challenge is compounded by the inherent impossibility of learning the actual value of any counterfactual investments that were not made. Second, there also needs to be some way of knowing which specific data predictions were available at the time of decision-making and how they aligned with any given firm's knowledge. The last challenge presents a crucial identification issue: if the quality of data-driven predictions correlates with firms' domain knowledge, then it becomes impossible to distinguish the prediction's quality from the ability to evaluate it correctly.

I address these challenges by focusing on the setting of pharmaceutical innovation, a field whose empirical features are ideally suited for this paper. The drug development process is a lengthy and costly endeavor, starting with the challenging task of identifying genes that can act as drug targets. The recent emergence of genomics has enabled a new approach to locate the genetic roots of human diseases, called genome-wide association studies (GWAS). GWAS are studies that compare large genomic databases of individuals with and without a specific disease to identify genetic mutations correlated with the presence of a disease. In practice, they yield predictions on the value of genes as drug targets, bypassing the need to comprehend the gene's role in the disease. This allows me

to explore how pharmaceutical firms react to the arrival of data-driven predictions that inform their selection of genetic targets. I use patent applications as an indicator of early-stage investments and the discovery of new drugs to measure successful downstream outcomes following a GWAS.

To identify the causal role of data-driven predictions, I exploit three unique features of GWAS. First, these studies scan the entire genome and do not target specific genes ex-ante, ensuring that any correlation between genes and the disease of interest is not the result of endogenous selection by the researcher (Uffelmann et al., 2021). Second, GWAS are primarily conducted by academic researchers, who make their findings publicly available to every firm. New gene-disease associations are usually not anticipated by researchers and especially not by firms, who only learn about them once published in scientific journals. Third, GWAS uncover both breakthrough opportunities and false positive genetic targets. To separate them, I exploit the fact that diseases are often subject to multiple GWAS of increasing statistical power over time (Marigorta et al., 2018). Subsequent GWAS provide an intuitive way to identify findings that are not robust to replications, revealing that they were likely false positives that firms should have avoided.

This setting allows me to test how firms leverage data-driven predictions. However, the average GWAS study in my data introduces 14 new gene-disease associations, of which 12 are false positives. This presents a challenge because one cannot link investments to individual predictions within the same publication using conventional measurements, such as patent-to-paper citations. To overcome this challenge, I adopt a novel text-based approach to extract the specific associations targeted in each patent and map them onto an empirical landscape of all possible gene-disease combinations. This approach enables me to track firms' investments over the search landscape before and after the arrival of data-driven GWAS findings while bypassing the use of patent citations. Then, I use firms' publications to measure the genes researched before GWAS' arrival. As confirmed in my interviews, firms with an active research program on any given gene can leverage this knowledge to assess GWAS findings even if the genetic expertise was developed for another disease (Cattani, 2005). This allows me to study the impact of data-driven predictions separately for firms with and without domain knowledge.

I find that the average number of patent applications for innovations targeting a gene-disease combination more than doubles after a GWAS reports it. Event study specifications confirm the absence of pre-trends and the validity of the research design. However, I also find that over a third of patent applications are directed to pursuing false positive findings. Investments based on these false positives fail to yield downstream outcomes such as highly cited patents or new drug molecules, highlighting that pursuing false positive investments is a misallocation of firm resources. Next, I explore the interaction between firms' domain knowledge and the impact of data-driven predictions. The results indicate that firms lacking domain knowledge react more strongly after a GWAS is published. However, they are

disproportionally likely to select false positive findings in their innovation efforts. In contrast, firms with domain knowledge exhibit a more discerning approach, making fewer investments that target the most promising opportunities in the data. While both commission (type I) and omission (type II) errors decrease, firms with domain knowledge seem relatively more effective in ruling out low potential gene-disease associations rather than ruling in the very best ones in the data.

I further investigate these findings with a firm-level research design. In particular, I leverage within-firm heterogeneity in genetic expertise resulting from their past research choices. While these analyses are descriptive in nature, this alternative design allows me to control for tighter firm and firm-by-disease fixed effects. The analysis confirms that genetic knowledge enables firms to become more efficient and targeted in their investments. To explore the mechanisms underlying my results, I categorize the specific nature of firms' domain knowledge. Regression results show that the ability to avoid false positives without committing resources to test them does not automatically stem from experience; instead, it seems to reflect the capacity to translate theoretical scientific principles into the evaluation of GWAS associations. Additional tests rule out social connections with scientists, generic organizational capabilities, or different firm-level strategies as potential alternative mechanisms. Taken together, these findings provide intriguing evidence that the ability of firms to assess data-driven predictions derives from understanding the mechanisms of what makes an opportunity valuable.

This paper makes several contributions. First, it contributes to a developing research agenda on how predictive tools like AI will shape innovation (Agrawal et al., 2024; Allen and McDonald, 2024; Bessen et al., 2022; Cockburn et al., 2019) and their limits (Cao et al., 2024; Hoelzemann et al., 2024; Kim, 2024b). By shedding light on the implications of false positive predictions, my results show the importance of looking beyond the average effects of predictive technologies to understand better what complementary assets will shape firm-level heterogeneity (Conti et al., 2024; Kao, 2023; Krakowski et al., 2023). In particular, my framework highlights how domain knowledge can be a critical determinant of who benefits from data-driven predictions in innovation (Agrawal et al., 2018a; Allen and Choudhury, 2022; Toner-Rodgers, 2024). Relative to research showing a democratizing effect of data availability (Galdon-Sanchez et al., 2024; Jin and McElheran, 2024; Nagaraj, 2022), accounting for false positives implies a complementarity with domain knowledge that reinforces incumbents' advantage. Greater attention to the mechanisms and boundary conditions of data-driven innovation may provide insights into why the promise of big data seems slow to materialize in the context of innovation (Brynjolfsson et al., 2021; Lou and Wu, 2021).

Moreover, this paper contributes to the literature on firm capabilities (Cattani, 2005; Gambardella, 1992; Nelson and Winter, 1982) and their interaction with predictive technologies (Balasubramanian et al., 2022). Using a new empirical approach to map firms' investments onto an empirical search

landscape, I show the continuing importance of absorptive capacity, defined as the use of domain knowledge to both recognize and exploit opportunities (Cohen and Levinthal, 1990, 1994). My results suggest that recognizing opportunities, in particular, becomes more important in a world awash with data-driven leads (Knudsen and Levinthal, 2007), thus changing the relative importance of the two sides of absorptive capacity (Arora and Gambardella, 1994b). The paper also relates to the theory-based view in management (Camuffo et al., 2024; Felin and Zenger, 2017), since separating the wheat from the chaff in the data is the purview of firms with a better grasp of theoretical principles. Lastly, the paper builds on research about using predictive tools in hiring, marketing, and other firm operations (Hoffman et al., 2018; Kim, 2024a). I extend this line of work to technological innovation, where domain knowledge is pivotal due to the many false positives and the higher testing costs.

The paper proceeds as follows. Section 2 introduces the theoretical framework. Section 3 presents the drug discovery setting, the measurement strategy, and the data. Section 4 discusses the research design. Section 5 reports the results, while Section 6 explores mechanisms. Section 7 concludes.

2 Theoretical Framework

Survey evidence shows that most U.S. firms use predictive tools, even in traditional sectors like manufacturing (Brynjolfsson et al., 2021). Yet, little is known about their impact on innovation, which entails searching for opportunities in complex technological spaces. This section addresses this gap by theorizing about the mechanisms through which data-driven predictions change innovation, as well as which firms benefit from them.

2.1 Innovation as a Prediction Problem

A firm's ability to find opportunities is critical to its performance. This is especially true in the innovation process, often depicted as searching for novel combinations of technological components (Fleming, 2001; Kang, 2024; Katila and Ahuja, 2002; Nelson and Winter, 1982). Since technological landscapes are generally too vast to experiment with all the potential combinations, firms have to evaluate potential investments to allocate their limited resources (Arora and Fosfuri, 2005; Krieger et al., 2023). This means that they do not search at random but rather in areas where they expect the highest returns (Arora and Gambardella, 1994a; Fleming and Sorenson, 2004; Kneeland et al., 2020). In practice, the innovation search process can be seen as akin to a *prediction problem*: firms try to predict which technological combinations are most valuable and allocate their resources accordingly. Firms that can better prioritize valuable projects will thus enjoy a competitive advantage.

But what is the source of superior predictive capabilities in innovation? A large body of work has documented how domain knowledge enables scientists and entrepreneurs to foresee opportunities

(Chatterji et al., 2023; Li, 2017; Gruber et al., 2008; Shane, 2000). Knowledge in any given domain permits decision-makers to understand the underlying mechanisms and principles behind what makes an opportunity valuable (Felin and Zenger, 2017), allowing them to recognize value ahead of the competition (Agarwal et al., 2023; Camuffo et al., 2024; Gavetti and Levinthal, 2000). The same logic applies at the firm level, with deeper domain knowledge as a key determinant of absorptive capacity (Arora and Gambardella, 1994b; Cohen and Levinthal, 1990). For example, discovering a new drug involves searching in a chemical space of $\sim 10^{60}$ molecules, far too vast to test all possible options. Therefore, pharmaceutical firms traditionally leveraged chemical and pharmacological knowledge from internal research (Gambardella, 1992) and academic collaborations (Cockburn and Henderson, 1998) to invest in the leads expected to be more promising.

Against this backdrop, the recent availability of big data and predictive technologies[1] offers an alternative way to make predictions about risky choices (Agrawal et al., 2018b; Brynjolfsson and McElheran, 2016; Lou and Wu, 2021). Data on past successful combinations can be analyzed to find patterns that help assess the viability of potential investments (Kim, 2024b), thus enabling firms to perform a triage before committing any actual resource (Kao, 2023; Nagaraj, 2022). The advantage is that data-driven predictions are not confined to technological components that are theoretically characterized or of known function (Cockburn et al., 2019; Toner-Rodgers, 2024; Tranchero, 2024). Instead, algorithms can identify association patterns between variables that are useful in making predictions even if the underlying mechanisms for success are unknown to the innovator. The result is an effectively theory-free approach to prioritizing search in vast technological landscapes for which data are available (Agrawal et al., 2024).

Predictive technologies fundamentally reshape the source of predictions used to identify opportunities, with data-driven predictions forming the basis for innovation decisions instead of domain expertise. This change has potentially important implications for strategy and competitive dynamics. Firms may stop investing in domain knowledge since data-driven predictions could be a cheaper and more fungible substitute (Balasubramanian et al., 2022; Cohen and Levinthal, 1994; Puranam, 2019). This could result in what the popular press has called the death of theory (Anderson, 2008; Schmidt, 2023), potentially eroding the role of domain knowledge in innovation. On the other hand, however, it could result in the democratization of competition: new entrants could use data-driven predictions to disrupt incumbents that have traditionally relied on large knowledge bases to maintain their competitive edge. In summary, the rise of predictive technologies raises the question of how they will shape

[1] I define predictive technologies as algorithms designed to use large amounts of data to make predictive statements about options of unknown value (Brynjolfsson et al., 2021). This includes tools of varying levels of sophistication, from simple correlations to deep learning and AI (Kim, 2024b). However, data-driven predictions are conceptually different from descriptive uses of data analytics (Berman and Israeli, 2022).

innovation and who will benefit most from these trends.

2.2 The Impact of Data-Driven Predictions on Innovation

Conceptualizing innovation as a prediction problem is useful to visualize how data technologies affect the search process: by shaping how firms find potentially more promising technological combinations. However, the value of such data-driven predictions depends on the nature of the innovation problem. In contexts where the investment required to validate a prediction is relatively low, parallel experimentation and trial and error are viable approaches. For instance, A/B testing is effective in guiding the search for product specifications in software companies (Koning et al., 2022). On the other hand, when it comes to research and development (R&D) in traditional sectors, most choices do not share these features (Camuffo et al., 2024; Gans et al., 2019). Discovering the ground truth state of novel technological combinations can be very costly and may dynamically foreclose other paths (Adner and Levinthal, 2024). Therefore, data-driven predictions enable the possibility of large-scale "offline" learning, i.e., before incurring the costs of testing (Gavetti and Levinthal, 2000).

Data-driven predictions are best intended as noisy and imperfect signals even when informative on average (Agrawal et al., 2018b; Arora and Fosfuri, 2005). This is likely magnified in technological search, where the precision of predictive tools is lower due to the complexity of the prediction task and the scarcity of training data (Camuffo et al., 2023; Choudhury et al., 2020; Kim, 2024b). Nonetheless, there are good reasons to expect that even imprecise data-driven predictions can largely affect the willingness to invest in a given project. Firms are likely to hold uninformative priors on large swaths of the search landscape, meaning that predictions can have a significant "surprise effect" that sways their posterior beliefs (Camuffo et al., 2024; Harrison and March, 1984). When the probability of success is extremely low ex-ante, as in technological innovation, such signals help de-risk investment choices (Ewens et al., 2018; Nagaraj, 2022).

But what about the downstream innovation outcomes resulting from those investments? Studies showing the benefits of predictive tools tend to focus on tasks where predictions can be quite accurate, such as in board games or resume screening (Choi et al., 2024; Hoffman et al., 2018). However, when applied to technological innovation, data-driven predictions frequently conceal *false positives* (Berman and Van den Bulte, 2022). Components interact in complex ways, meaning that associations between variables are often spurious (Tranchero, 2024). Furthermore, returns from innovation are extremely skewed (Fleming and Sorenson, 2001). When technologies require sizable investments and take years to mature, such as in deep tech or drug discovery, the price of a mistake far outweighs any positive spillovers from learning (Eggers, 2012). Together, this reasoning leads to the following baseline hypothesis:

Hypothesis 1: *Data-driven predictions stimulate firm investments in innovation, but only true positive predictions lead to better firm innovation outcomes*

The preceding discussion clarifies an important and often overlooked feature of predictive technologies: while extrapolating leads from past data is increasingly cheap, this does not change the other features of the innovation search process. In particular, firms remain limited in the resources they can allocate to validate and develop ideas. Whether a firm can gain a competitive advantage from data-driven predictions will thus crucially hinge on its ability to avoid false positives.

2.3 Data-Driven Predictions and Domain Knowledge: A Simple Conceptual Framework

The earlier discussion focused on the aggregate effects of a data-driven innovation process. Yet, a crucial question for management is which kinds of firms benefit more from the diffusion of big data technologies. The existing results are ambiguous. Some research has found that new entrants and small firms can use data to recognize opportunities ahead of experienced competitors (Conti et al., 2024; Galdon-Sanchez et al., 2024; Nagaraj, 2022), while others have reached the opposite conclusion (Kao, 2023; Otis et al., 2024).[2] These conflicting results are puzzling if one thinks of data-driven predictions as reducing the cost of locating opportunities, which would imply that they substitute domain knowledge and benefit smaller and inexperienced firms. The difficulty in reconciling existing evidence points to the need for a new conceptual framework to understand the competitive effects of predictive technologies.

The presence of false positives among data-driven predictions suggests the need to separate the quantitative increase in investments responding to those leads from the actual innovation outcomes (Conti et al., 2014). Starting with the former, past evidence suggests that inexperienced firms rely comparatively more on data to find opportunities (Conti et al., 2024; Furman et al., 2021). New entrants and smaller firms, which are more likely to lack domain knowledge, tend to face higher cognitive (Galdon-Sanchez et al., 2024) and investment costs (Nagaraj, 2022). Insofar lower domain knowledge translates into weaker priors about the potential of an opportunity, data-driven signals will greatly change investment behaviors (Chavda et al., 2024). Instead, domain experts can more easily form accurate predictions of a project's feasibility and potential (Lane et al., 2022a; Li, 2017). Firms with greater in-house expertise thus need to rely proportionally less on external information to find and recognize promising leads (Arora and Gambardella, 1994b; Cohen et al., 2002). In sum, one would expect the following:

[2]Most research in this area has not investigated the mechanisms behind the effects documented, further increasing the ambiguity of existing results. A recent exception is the work by Conti et al. (2024), who highlighted how small firms seem to leverage big data to develop innovations while big firms use them to achieve cost efficiencies.

Hypothesis 2: *After observing the same valuable but imprecise data-driven predictions, firms without domain knowledge invest proportionally more than firms with domain knowledge*

Nevertheless, the presence of false positives implies that investing based on data-driven predictions is no guarantee of success. The challenge for firms is thus evaluating predictions and deciding how to allocate their scarce resources (Agrawal et al., 2018a). This suggests that rather than being substituted away, domain knowledge could become a complement in choosing between potential opportunities in the data (Allen and Choudhury, 2022). One way to visualize this argument is through a two-stage process. First, data-driven predictions provide a shortlist of technological combinations that are valuable on average but too numerous to test individually. Then, domain knowledge helps firms triage the perceived value of potential opportunities, thus channeling resources to the most promising predictions. For instance, domain knowledge permits to understand the interdependencies between technological components, allowing them to rule out implausible associations (Chatterji et al., 2023; Fleming and Sorenson, 2004) or interpret data signals in light of existing theories (Agarwal et al., 2023; Felin and Zenger, 2017).

This simple framework implies that firms with domain expertise will benefit more from data-driven predictions. On the one hand, data-driven predictions might counteract organizational inertia and uncover breakthroughs outside domains mastered by the firm, offsetting the proclivity of experts to succumb to competency traps (Denrell and March, 2001; Levinthal and March, 1993). On the other hand, domain experts tend to be more discerning and accurate when reacting to novel information (Allen and Choudhury, 2022; Boudreau et al., 2016; Lane et al., 2022a). Instead, firms without the knowledge to judge predictions will make commission (or type I) and omission (or type II) errors simultaneously; said differently, they will mistakenly select low-value combinations at the expense of true opportunities. Compared to firms with domain knowledge, those without it should be less likely to find the best opportunities among data-driven predictions despite investing quantitatively more. This argument can be synthesized as follows:

Hypothesis 3: *After observing the same valuable but imprecise data-driven predictions, firms without domain knowledge make proportionally more type I and II errors than firms with domain knowledge*

Overall, this simple theoretical framework offers insights into how domain knowledge interacts with data-driven predictions. The emergence of predictive technologies changes the role of knowledge in innovation, but not by rendering it obsolete, as some have argued (Anderson, 2008). In a world where obtaining data-driven predictions is cheap, the source of competitive advantage shifts from idea generation to filtering out false positive leads (Knudsen and Levinthal, 2007). Firms with robust knowledge bases may thus continue to outperform in innovation, even if that knowledge was initially

developed for other purposes (Cattani, 2005; Cohen and Levinthal, 1994). The rest of the paper tests whether the empirical evidence supports these implications in the context of pharmaceutical innovation. In particular, I investigate which firms can benefit from data-driven predictions that identify potential genetic targets for drug discovery.

3 Empirical Setting, Measurement Strategy, and Data

3.1 Drug Discovery and GWAS

The first crucial step in drug discovery is choosing the right genetic target: namely, firms must identify genes that can be modulated by a drug to affect the outcomes of specific diseases (Nelson et al., 2015).[3] This task is complex because, in principle, each genetic disease could be caused by mutations in any of the over 19,000 protein-coding human genes. Common diseases are often polygenic—for instance, diabetes has been tied to over 150 DNA mutations. The result is a vast combinatorial space of tens of millions of potential gene-disease combinations, of which only a minor fraction has actual therapeutic value. Simply testing all of them is not feasible: firms can pursue only a limited number of alternatives because developing a new drug takes 10-15 years and costs $2.6 billion on average (Kao, 2023). Therefore, it is unsurprising that pharmaceutical firms often rely on university research to reduce the risks involved in target selection (Arora and Gambardella, 1994a).

Starting from the early 2000s, the completion of the Human Genome Project and the steep decline in the cost of collecting genetic data prompted the emergence of GWAS. GWAS are case-control studies where researchers sequence human genomes to find genetic mutations that are more likely to appear among subjects with a specific condition as compared to healthy subjects (Uffelmann et al., 2021; Visscher et al., 2017). Figure 1 shows the abstract of one early GWAS and a stylized depiction of how a typical study unfolds. Researchers start by collecting DNA samples from several subjects, some affected by the disease of interest and some not. They then use microarrays to sequence DNA locations (called markers) to reconstruct the genetic constitution (or genotype) of the subjects in the sample. Lastly, they test for statistically significant differences in genotypes that correlate with the presence of the disease.[4] Intuitively, by comparing the genetic makeup of people with and without a condition, one can predict which genes might be involved in a given disease and thus serve as drug targets (see Appendix A for more details).

[3] Genes are sequences of DNA bases that encode the "instructions" to synthesize gene products (e.g., proteins) that allow the organism to function. When genes acquire mutations in their sequence, they might alter their behavior with significant consequences for human health. Knowing the genetic roots of diseases is vital in designing pharmaceutical drugs because genes that cause disease can be targeted for therapeutic purposes (see Appendix A for more details).

[4] Association tests are adjusted for multiple hypothesis testing, usually imposing a high threshold for statistical significance. The results are often graphically represented as a "Manhattan plot" and show the p-value of multiple statistical tests between DNAs in the case and control groups. The y-axis usually reports -$\log_{10}$(p-value), and hence higher values correspond to stronger associations. See Figure 1 for an example.

GWAS provide data-driven leads into gene-disease combinations that can be valuable in R&D, and anecdotal evidence points to several drugs developed thanks to GWAS findings (Visscher et al., 2017). Nevertheless, there is still no consensus on the ability of GWAS to identify targets with high therapeutic potential (Struck et al., 2018; Tam et al., 2019; Uffelmann et al., 2021). One major limitation is that GWAS cannot explain the underlying mechanisms of gene-disease associations, raising the possibility of false positives with no actionable causal pathway (Goldstein, 2009; Hermosilla and Lemus, 2019). The average GWAS can find tens of potentially promising gene-disease associations, many of which could be worthless or spurious. This challenges scientists and firms interested in therapeutic applications due to the high costs of following a data-driven lead.[5] As a result, questions remain regarding how many GWAS associations are valuable and whether they generate useful knowledge for drug discovery at all (Goldstein, 2009).

3.2 Measuring the Impact of GWAS with Text-Based Empirical Landscapes

For my empirical analysis, I need a precise method to measure which firms decide to invest in the gene-disease associations uncovered by GWAS. Unfortunately, information on R&D spending is only available from firms at an aggregate level (Cohen and Levinthal, 1990). To get at the project-specific level of investment, I leverage the richness of patent data (Gambardella, 1992). Following Eggers and Kaplan (2009), I use patent applications as an indicator of early-stage investments in a given domain. This is well-suited to my empirical setting since pharmaceutical firms usually apply for patents at the beginning of the long therapeutic development process, before knowing whether the product patented will be successful in the clinic. I then use the information on molecules entering the clinical trial phase to measure the downstream outcomes of R&D investments.[6]

However, linking firm investments and outcomes to the gene-disease associations uncovered by GWAS poses another measurement challenge. Traditional approaches relying on patent-to-paper citations are not well-suited for this context. On the one hand, a typical GWAS identifies multiple genes linked to a disease, yet it is unclear which are true positives among the false positives. Thus, it is impossible to determine from a patent citation to the GWAS paper exactly which association the patent is leveraging. On the other hand, using direct citations is ineffective for tracing the unpredictable path between foundational basic research and innovation. For instance, patents might cite later research that validates a particular GWAS association but neglect to mention the original GWAS because such

[5]This theme emerged prominently in my interviews with industry professionals: "I don't think anyone who is doing good science would just trust a GWAS, even an accurate GWAS, because R&D is so expensive" (Interview, 17 October 2022).

[6]Since I focus on the pre-clinical stage, patent applications are better thought of as a proxy for firm investments (Eggers and Kaplan, 2009). Insofar as firms pay attention to the findings of GWAS and start investing based on them, we should see more patent applications as a byproduct. Yet, only successful investments will reach the clinic, making the number of drug molecules entering the clinical trial stage a proxy of innovation success conditional on upstream investments.

discovery has become common knowledge in the field.

To address these challenges, I develop a novel measurement strategy (see Appendix B).[7] Taking a recombinant view of innovation (Fleming, 2001), each GWAS association can be considered as a prediction of a valuable gene-disease combination. This suggests that the innovation impact of a GWAS is proxied by the change in firm patenting behavior for gene-disease pairs "treated" compared to otherwise similar combinations that were not reported by the GWAS. Next, named entity recognition algorithms can be used to extract the genes and diseases from the text of each patent application, providing a method of inferring which gene-disease combinations a firm is targeting in its investments. Figure 2 offers a representation of how the increase in patents targeting specific gene-disease pairs captures the heterogeneous impact of individual GWAS associations. Assuming a standard parallel trends assumption, the before-and-after changes in the number of patents mentioning treated and control combinations provide an estimate of the GWAS' effect. Notably, this approach permits me to empirically study firms' exploration decisions in a combinatorial landscape of genes and diseases.

This novel measurement approach offers three significant advantages over traditional citation-based methods. First, employing text-based knowledge entities allows for distinguishing between the impact of different contributions made in the same paper. Traditional citations lack such granularity at the gene-disease level. Second, by tracking mentions of genes and diseases directly in the patent's text description, this approach is well suited to objectively capturing the impact of basic research findings that might not be explicitly acknowledged in patent applications. Third, one can similarly collect data on the gene and the disease targeted by the drugs tested in each clinical trial (Kang, 2024; Kao, 2023). This allows me to similarly map drug discovery in the same gene-disease combination landscape, and trace the innovation process from the initial investments to the outcomes. Appendix B presents conceptual details of how this measurement strategy can be generalized to other applications, bypassing the limitations of patent citations (Nagaraj and Tranchero, 2024).

3.3 Data Sources

To empirically test how firms respond to the arrival of data-driven predictions, three key ingredients are required. First, I need data on new gene-disease associations identified by GWAS. Second, I need information on firms' domain knowledge, early-stage investment decisions, and drug development outcomes. Third, I need to know the specific genes and diseases targeted by each GWAS, patent, and drug in order to link them. Below I summarize the main data and how I collect these key elements

[7] Nagaraj (2022) uses a similar idea to assess the impact of satellite images on gold discovery at the level of individual blocks of the earth. Kao (2023) adopts the same logic to study the effect of large-scale cancer maps on the number of clinical trials on 627 genes. The approach detailed in Appendix B formalizes their intuition and expands it to additional use cases, including quantifying science-to-technology spillovers without the limitations of citation data.

(see Appendix C and D for details):

GWAS Catalog: To identify the data-driven predictions available to firms, I rely on the GWAS Catalog, a manually curated source managed by the European Bioinformatics Institute (EBI) and the National Human Genome Research Institute (NHGRI). The GWAS Catalog is a comprehensive list of GWAS published in peer-reviewed journals, starting from the first one published in 2005. Studies are eligible for inclusion in the GWAS Catalog if they include an array-based genome-wide scan that does not target any specific gene ex-ante. The GWAS Catalog also collects the details of the findings in the original study. In particular, each gene is identified by its NCBI Gene IDs, while diseases are reported according to the Experimental Factor Ontology (EFO). I use the crosswalk available on the EFO website to map diseases into the corresponding MeSH Unique IDs.[8] My sample includes 17,965 gene-disease associations first reported by 1,259 distinct GWAS between 2005 and 2019 (Appendix Figure D.1). These associations span 404 unique diseases and 5,080 protein-coding genes.

SciBite/EBI Patent Data: Through a partnership with EBI, I obtained proprietary data to measure the gene-disease pairs mentioned in the text of each USPTO patent application (2001–2019).[9] The data have been compiled from complete patent texts using TERMite (TERM identification, tagging, and extraction), a named entity recognition software developed by the Elsevier-owned start-up SciBite. TERMite directly maps the entities extracted into NCBI Gene IDs and MeSH Unique IDs. A manual validation of 200 random patents finds that SciBite's entity recognition algorithm has a precision rate of 95%–97% and a recall rate of 91%–92% and is thus highly reliable. I merge these data with information on assignees and patent characteristics taken from PatentsView. Commonly used indicators of patent quality are from the OECD Patent Quality Indicators Database (Squicciarini et al., 2013) and the data of Kogan et al. (2017). My final sample includes all pharmaceutical companies applying for at least one USPTO patent between 2001 and 2019.

Cortellis Drug Data: I supplement my data with proprietary information on drug molecules collected by the Clarivate Analytics Cortellis Competitive Intelligence Database (Krieger, 2021). For this study, I use the drug development records in Cortellis up to July 2020, which contain information for over 70,000 drugs. Cortellis aggregates information from various sources to provide a comprehensive

[8]Since the MeSH taxonomy is a hierarchical tree, in my analysis I include all EFO diseases matched with MeSH IDs at level four of the tree. If a more specific disease was matched (i.e., at level five or above), I assigned it to its parent branches up to level four. Vice versa, if the disease matched was coarser (i.e., at level three or below), I assigned the finding to all its descending level four branches. This procedure permits harmonizing the diseases targeted in GWAS at the same level of specificity. The sample is further restricted to only diseases that receive more than one GWAS because I exploit subsequent GWAS to code which gene-disease associations are not replicated and are thus likely to be false positives (as explained in Section 3.4).

[9]Note that these are *not* gene patents, i.e., the exclusive rights to a specific sequence of DNA. Rather, they are patents for genetic tests, method-of-use of molecules, or new drug molecules that target a specific gene to treat a given disease. See Appendix C for details.

list of historical development milestones for each drug molecule. I use data about new molecules observed entering the earliest phases of drug development (what Cortellis records as the "discovery phase" and the "pre-clinical phase"). For each drug, I match the genetic target to NCBI Gene IDs and the condition addressed to the corresponding MeSH Unique IDs using string matching.

PubTator Central Publication Data: I use publicly available data from PubTator Central, which provides computer-annotated genes and diseases for each paper published in PubMed. The PubTator Central team maps genes and diseases to NCBI Gene IDs and MeSH Unique IDs. Next, I match each patenting firm in my sample with their respective publication records using the information on authors' affiliations. Doing this enables me to code which firms had prior knowledge of the biology of a specific human gene before deciding whether to invest in a relevant GWAS finding. This procedure yields a very granular measure of gene-specific expertise, an improvement over previous studies that used corporate publications as a generic firm-level proxy for absorptive capacity. Additional bibliographic characteristics of firms' publications are from the National Institutes of Health's iCite database.

Open Targets Score: Open Targets is a public-private partnership that collects all the available evidence on the strength of gene-disease associations and summarizes it in a synthetic score. The data can be downloaded from `https://platform.opentargets.org/` (see also Appendix D.2). Experts in the field consider Open Targets to provide the most comprehensive assessment of the genetic roots of human diseases based on the available knowledge. I download the Open Targets scores and merge them with my data. Once more, genes and diseases are already provided with NCBI Gene IDs and MeSH IDs. The score is available for 594,353 gene-disease pairs (8% of my sample), spanning 17,437 genes and 366 diseases.

By putting together these data sources, I can trace an empirical search landscape that captures all potential combinations of genes and diseases that pharmaceutical firms could select. Table 1 reports summary statistics at the gene-disease combination level, constituting the paper's primary unit of analysis.[10] Around 21.8% of the 7,223,924 gene-disease pairs have received investments, as proxied by their appearance in at least one patent application. Yet, only 0.19% of these pairs have advanced to drug development stage, confirming the extremely skewed distribution of innovation returns. GWAS have uncovered 17,965 new associations, constituting 0.25% of this vast combinatorial landscape. The average gene-disease pair receives 0.13 patent applications per year. However, there is a considerable variation, with some genetic targets receiving over 1,000 annual patents. Around three-quarters of patents are filed by firms that have not previously published research on the targeted gene. Descriptive evidence confirms that firms are more likely to target gene-disease combinations

[10]More precisely, each observation unit is a combination of NCBI Gene ID and MeSH Unique ID (at level four of the MeSH tree).

with higher therapeutic potential when previously researching the gene involved (Appendix D.3). This pattern suggests that domain knowledge is related to the ability to focus investments on more promising targets, even in a cross-section of my data (Nelson et al., 2015).

3.4 Identifying False Positive Associations

GWAS conduct an unbiased, atheoretical search for genetic mutations associated with human diseases. This feature is both a strength and a weakness: despite very stringent statistical significance thresholds, it is well-known that many GWAS findings turn out to be statistical noise or spurious correlations (Goldstein, 2009). For this paper, I need to empirically measure which GWAS predictions prove to be false. The main challenge is that ex-post direct assessments of gene-disease associations' quality are only available for findings that received investments, resulting in a classic missing data problem. Specifically, researchers can only observe the ground truth value of the prediction when a decision maker chooses to act on it (Hoffman et al., 2018). In my setting, this issue would make it impossible to understand if GWAS associations that received little attention were correctly avoided due to low potential or were mistakenly overlooked by firms failing to recognize their value.

To distinguish true positives from false positives, I exploit the fact that diseases are often subject to multiple GWAS over time. In particular, I categorize gene-disease associations as true positives only if they are initially reported by a GWAS and later confirmed by at least one subsequent GWAS focusing on the same disease. This provides an intuitive way to identify likely false positives regardless of the level of firm investment they received.[11] Using this approach, I find that 84.3% of the 17,965 gene-disease associations in my sample fail subsequent replications using different samples.[12] This high number aligns with accounts from experts who have long warned about the risks posed by big data in genomics (MacArthur, 2012). Non-replicable findings can lead to inefficient use of R&D resources, misdirecting firms' limited resources to targets without therapeutic potential (Freedman et al., 2015).

I validate this measure of false positives through three methods. First, I compare my coding with

[11] Put otherwise, the genome-wide design of subsequent GWAS replications serves as a retest for all findings, thereby providing assessments even for associations that have not received investments from firms. Note also that this approach provides a conservative metric of which findings are false positive in a therapeutic sense since even replicable associations might not offer avenues for treatment (Goldstein, 2009; Hermosilla and Lemus, 2019).

[12] The low applicability of GWAS findings is well-known in the field (MacArthur, 2012). Marigorta et al. (2018) find that around 40% of associations are replicable when including also findings in intergenic regions (i.e., non-protein coding). The discrepancy with my analysis is mostly due to the fact that I focus on protein-coding regions because of their higher relevance for drug development. One reason behind the lower replicability for findings related to protein-coding genes is the frequent mistakes made by GWAS' authors in mapping mutations into the correct corresponding genes (Visscher et al., 2017). This type of mistake in reporting GWAS results provides an alternative empirical strategy that I leverage in Subsection 5.3 as a robustness test.

the Open Targets score to measure the ground truth value of each gene-disease pair.[13] Appendix Table E.2 shows that replicable gene-disease associations are 114%–191% more likely to rank in the top decile of the Open Targets score relative to non-replicable ones. Second, I investigate the characteristics of the study design that predict which GWAS discoveries will replicate in subsequent studies. The data in Appendix Table E.3 support the intuition that more robustly designed GWAS are less likely to report associations classified as false positives. Finally, I present evidence in Appendix Table E.4 to confirm that true positives are more likely to appear in articles that receive more citations in clinical papers and a lower share of citations with a negative tone (Catalini et al., 2015), as captured by the Scite data (Nicholson et al., 2021). Together, these tests corroborate the approach used to identify false positives in GWAS associations.

4 Research Design

There are two main challenges to assessing the impact of data-driven predictions on innovation: measurement and endogeneity. The first challenge arises from the difficulty of determining which predictions a firm has access to. This becomes problematic when firms with domain knowledge obtain systematically better predictions, thus conflating the prediction's quality with the ability to interpret it. The second challenge relates to identifying the causal effects of data-driven predictions. In practice, firms are likely to mine for predictions where they expect the highest returns (e.g., for genes known to be druggable), potentially leading to upwardly biased estimates. An ideal experiment would bypass these issues by assigning all firms identical information on gene-disease associations. The causal effect of data-driven predictions would then be evident from changes in patenting involving "treated" gene-disease pairs relative to the others. Subsequently, I could then test my theoretical framework with heterogeneity analyses by firms' pre-existing domain knowledge.

I approximate this ideal experiment using the staggered publication of GWAS associations in scientific journals. First, GWAS are mainly conducted by academic research teams,[14] and their findings are publicly accessible upon publication. As such, their findings are available to every pharmaceutical firm simultaneously and are not driven by unobservable proprietary data or expertise that they might have. Second, GWAS scan for genetic variants across the whole genome (Uffelmann et al., 2021). This method, by design, avoids the issue of endogenous sorting since it does not focus on specific genes. GWAS discoveries are, therefore, entirely unforeseen by the research team and even more so by firms that later read about the findings in journals. Importantly, unbeknownst to the scientists conducting the GWAS, only some of their findings will be confirmed in future publications.

[13] Out of the 17,965 gene-disease associations in my data, there are 16,298 pairs with an Open Targets Score.

[14] In my sample, 324 GWAS are co-authored by corporate scientists, and another 117 of them acknowledge funding from pharmaceutical firms. However, all my results are robust to the exclusion of those GWAS (Appendix Table E.9).

This presents a unique opportunity to study how firms react to true and false positive associations and how their responses vary based on domain knowledge.

I use OLS to estimate the following specification at the gene-disease-year level:

$$Y_{i,j,t} = \alpha + \beta Post_t \times GWAS_{i,j} + \gamma GD_{i,j} + \delta_t \times Gene_i + \omega_t \times Disease_j + \epsilon_{i,j,t}, \quad (1)$$

where $Y_{i,j,t}$ is either a measure of investments or innovation outcomes by firms in year t involving gene i and disease j. In alternative specifications, $Y_{i,j,t}$ is measured separately for firms with domain knowledge ($k = 1$) and without it ($k = 0$). $Post_t \times GWAS_{i,j}$ equals one after the first GWAS reports a gene-disease association and equals zero otherwise. $GD_{i,j}$ are fixed effects for the combination of gene i and disease j, which account for pair-specific differentials in research potential. I control for time-varying disease-level differences in market size by including disease-year fixed effects ($\omega_t \times Disease_j$). Similarly, gene-specific time trends ($\delta_t \times Gene_i$) consider the growth of interest in specific genes. Estimates report standard errors clustered two-way by gene and disease.[15] The coefficient of interest β captures the change in $Y_{i,j,t}$ for genes and diseases found correlated in early GWAS relative to those reported later (or never within my sample period).

The main threat to identification is the potential sorting of data-driven predictions on gene-disease pairs that would receive investments even without GWAS findings. However, I provide evidence that this is not a concern in my context. First, GWAS are distinct from targeted scientific studies in that they scan every human gene. The genome-wide nature of GWAS ensures that scientists there is no gene-level selection conditional on the choice of disease to study (Appendix Figure E.1). Second, the timing of the findings is plausibly exogenous to trends in the dependent variable. Indeed, the temporal order in which gene-disease pairs are reported in a GWAS is not influenced by previous inventive activity related to those pairs (Appendix Table E.1 and Figure E.2). In what follows, I exploit this plausibly exogenous timing variation between genes within a disease to assess the causal effect of GWAS on firms' early-stage innovation investments. Directly testing for the absence of pre-trends in the outcome variables will further confirm the validity of my identification strategy.

5 Results

5.1 Do GWAS Lead to More Innovation?

I begin by examining the aggregate impact of GWAS findings on pharmaceutical innovation. Given their uncertain and debated reliability, it is unclear whether the publication of gene-disease associations will increase firms' willingness to invest in a gene-disease combination. Columns 1 and 2 of Table

[15]The precision of the results is robust to alternate methods of clustering standard errors, such as by gene, by disease, or by gene-disease pairs.

2 present the baseline results, using patent applications as an indicator that a firm is investing in a given gene-disease pair. The main finding is a positive impact of GWAS, amounting to an average increase of 125% in patents filed. The design of GWAS should ensure that the effects reported are not due to researchers endogenously targeting the most promising gene-disease pairs. Figure 3 directly checks the validity of this assumption with an event study version of Equation 1. The plot confirms flat pre-trends and a persistent effect after the GWAS is published, with the estimates stabilizing around the value of the primary estimate of Table 2. Additional heterogeneity analyses in Appendix Table E.5 reveal greater increases in patenting for associations that are statistically more robust or larger in magnitude.

GWAS are known to harbor many false positive associations (MacArthur, 2012). In my sample, 84.3% of the associations involving a protein-coding gene fail to replicate in follow-up GWAS targeting the same disease. The high incidence of misleading predictions would not be problematic if firms could recognize and avoid them. Unfortunately, as shown in Columns 3 and 4 of Table 2, this is not the case. The estimates imply that around 35% of the total increase in patent applications is directed at non-replicable gene-disease associations. Nevertheless, GWAS findings that were later confirmed, albeit a minority, mustered up to three times more investments than the average effect. This large increase is likely compounded by firms initially investing in false positives and then pivoting away after learning about their mistake.

A case study exemplifies these dynamics. Panel (a) of Figure 4 shows patenting activity for the PLA2G7-myocardial infarction pair after it was reported in a GWAS by Suchindran et al. (2010). While GlaxoSmithKline was already investigating this specific pair, the GWAS led to a spike in investments by new firms. Unbeknownst to them, the finding would later fail to translate into therapeutic advances.[16] In 2014, GlaxoSmithKline announced the failure of two clinical trials for darapladib, a molecule targeting PLA2G7. This led other firms to redirect their investments toward alternative targets (Krieger, 2021). The same dynamics can be seen in an event study regression where I keep only false positive GWAS associations in the sample (Panel (b) of Figure 4). Patenting initially rises and then declines, most likely when firms discover that the finding is not robust after costly investments.

The allocation of investments toward false positives hurts firm performance because only valid associations yield successful innovation outcomes. Table 3 indicates that highly cited and high-value patents increase for GWAS associations that are later replicated, but not for the others. The estimates relating to drug discovery are also positive but noisier, probably because successful drug development

[16]To date, the precise function of PLA2G7 is still unclear. The consensus is that it may be a non-causal biomarker (i.e., an indicator) for the risk of cardiovascular disease, which could contribute to explaining the low Open Targets score (0.045) and the failure to be confirmed in GWAS replications. By comparison, the IL23R-Crohn's disease pair uncovered by Duerr et al. (2006), and discussed in Appendix A.2, has an Open Targets score of 0.468 and has been robustly replicated.

is an infrequent outcome. In Column 4, I follow the approach of Dranove et al. (2022) and weigh each drug by its scientific novelty, captured by the number of previous drugs that adopted the same molecular-targeting design.[17] The coefficient becomes more precise, suggesting that GWAS leads are especially helpful in designing drugs that leverage new scientific approaches. These results suggest that GWAS can boost pharmaceutical investments, but their impact on innovation is more nuanced because they produce several false positive leads. If these are not avoided, they can waste resources and reduce firms' technological performance. Overall, this evidence is consistent with my baseline hypotheses.

5.2 Data-Driven Predictions, Domain Knowledge, and Firm Investments

This subsection explores how data-driven predictions interact with firms' domain knowledge. Previous studies have used R&D expenditures (Cohen and Levinthal, 1990) or the number of corporate publications (Cockburn and Henderson, 1998; Gambardella, 1992) to identify a firm's knowledge. However, these proxies do not capture the specific domains in which the firm has the knowledge to recognize opportunities. I make progress on this issue by recording the specific genes that each firm has researched in its publications. This empirical approximation is grounded in the specificities of the setting. In my interviews with industry professionals, they confirmed that firms in the bio-pharma sector tend to specialize in genetic space.[18] Some firms go as far as making the selection of their target genes their fundamental competitive hypothesis. For instance, the biotech company Denali Therapeutics even mentioned its strategic focus on what it calls "degenogenes" in its IPO filings (Appendix C.3). My empirical approach captures the fact that when a new gene-disease association involving one of these genes is published, Denali's scientists will have the expertise to assess it.[19]

My research design exploits between-firm variation in genetic knowledge predating each GWAS (Cattani, 2005), as schematically illustrated in Appendix Figure E.3. Columns 1 and 3 of Table 4 present results comparing patent applications filed by firms with and without previous publications on the gene involved in a GWAS finding. While both groups of firms significantly increase their investments, there is a stronger reaction by firms lacking gene-specific knowledge. Compared to the sample mean, the regression coefficients imply that patenting increases by one-third more among firms

[17] Dranove et al. (2022) refer to the term "molecular-targeting design" as the mechanism by which drugs produce a pharmacological effect (e.g., darapladib inhibits Lp-PLA_2, thus being an "Lp-PLA_2 inhibitor"). This measure captures which drugs are more novel from a scientific perspective. Other approaches in the literature include considering the molecule's chemical novelty instead of its biological mechanism (Krieger et al., 2022).

[18] The patent portfolio of the median firm in my sample spans only 19 genes, a number that further decreases to 11 if considering only patents that will eventually be granted (Appendix C).

[19] As one of my interviewees at Denali put it, whether to pursue or not a new genetic target appearing in the literature is an assessment based on their extensive domain expertise: "We are so entrenched in the neuro field that we kind of know off-the-cuff" (Interview, 17 October 2022).

that had not previously researched the gene before the GWAS was published. This empirical evidence confirms the second implication of my theoretical framework: firms without domain knowledge will invest more after observing identical data-driven predictions.[20]

Next, I investigate if firms with relevant domain knowledge can correctly triage data-driven predictions and avoid those likely to be false positives (Arora and Gambardella, 1994b; Cohen and Levinthal, 1994). Appendix Figure E.5 descriptively shows the reaction of firms with and without domain knowledge to newly published gene-disease associations. Strikingly, the increase in patenting based on a false positive is over three times larger for firms without past research on the gene. I validate the statistical significance of this pattern with difference-in-differences regressions (Columns 2 and 4 of Table 4). The results confirm that the increase in patenting on false positive associations is significant only for firms without prior research on the gene, supporting the idea that such firms lack the knowledge to evaluate the findings. Mapping this result back to my theoretical framework, patent applications building on false positive associations constitute type I errors that distort resources away from pursuing valuable leads.

However, the above results leave open the possibility that domain knowledge simply increases skepticism of atheoretical data-driven predictions (Allen and Choudhury, 2022; Boudreau et al., 2016). Under this alternative explanation, avoiding false positives might mechanically result from lower investments made by conservative firms, potentially at the cost of curtailing the exploration of valuable opportunities. I rule out this possibility using the Open Targets data to capture the underlying therapeutic value of gene-disease pairs with a continuous score.[21] Figure 5 shows that firms with domain knowledge are also better at selecting the most valuable opportunities among the data-driven predictions.[22] This result implies that firms without domain expertise are more likely to make omission (or type II) errors despite investing quantitatively more.

Overall, I find that domain knowledge increases the efficiency of investments by focusing them on promising data-driven opportunities, supporting my third hypothesis. Firms with genetic expertise see a reduction of both type I and type II errors, but Figure 5 shows an interesting asymmetry. Domain

[20]Since larger firms also have broader knowledge bases, one might worry that this result reflects organizational inertia. To rule out this concern, I repeat my analysis for small and large firms separately. I find the same result in both sub-samples: GWAS generate stronger reactions from firms without domain knowledge regardless of their size (Appendix Figure E.4). This evidence is consistent with non-expert firms updating their belief more strongly because of weaker ex ante priors rather than their leaner organizational structure (Camuffo et al., 2024; Chavda et al., 2024).

[21]While available only for a subset of gene-disease pairs in my sample, this metric offers a way to compare the "ground truth" value of the combinations highlighted by GWAS associations. See Appendix D.2 for details.

[22]The better allocation of early-stage investments by firms with domain knowledge translates into more drug molecules entering the discovery stage (Appendix Table E.6). However, these results should be interpreted cautiously since I do not have information on licensing deals and firms without domain knowledge may license promising compounds for later stages of drug development.

knowledge seems relatively more effective in ruling out low potential gene-disease associations rather than ruling in the best ones in the data. While this finding was not explicitly hypothesized in advance, it is consistent with my theoretical framework and echoes the findings from past research (Boudreau et al., 2016). In particular, evidence shows that experts tend to be better at distinguishing between high and low-quality projects rather than singling out true breakthroughs (Krieger et al., 2023; Lane et al., 2022b). Understanding under which conditions these results are confirmed with black-boxed predictive technologies like generative AI is an interesting avenue for future research.

5.3 Robustness Checks

Several robustness checks further validate these findings. First, I replicate the main results by exploiting a unique feature of GWAS. After spotting genetic mutations correlated with the disease, scientists must map the mutation to the correct gene (Visscher et al., 2017).[23] Especially at the beginning of my sample, the gene reported in the original study was sometimes incorrect but later reclassified with more precision by the curators of the GWAS Catalog. In my data, there are 9,273 such cases. Confirming my primary results, Table 5 shows that only firms without domain knowledge increase investments based on these wrongly reported data-driven associations. This additional test has the advantage of not depending on my method for identifying false positives in the primary analysis.

Second, suppose domain knowledge truly helps assess the value of ambiguous data-driven predictions. In that case, its beneficial effects should be proportionally higher when the GWAS findings are harder to evaluate. For instance, evaluating an association should be more difficult when firms cannot leverage publicly available scientific information or when the GWAS reports too many new findings to validate. Indeed, firms without domain knowledge are more likely to fall for false positives when the findings involve less-studied genes (Appendix Table E.7), but they make fewer mistakes when the GWAS reports fewer associations to evaluate (Appendix Table E.8). Results are also robust to the exclusion of GWAS co-authored with industry researchers, suggesting that social connections with scientists are not a mechanism driving my findings (Appendix Table E.9).

Third, I examine whether patenting based on false positive findings is motivated by strategic reasons, implying that doing so is not detrimental to firms' performance, even if it does not result in discovering new drugs. A few pieces of evidence help rule out this possible explanation. First, using two alternative methods to identify patents likely to be motivated by strategic concerns (Righi and Simcoe, 2023), I find the firms are more likely to apply for such patents based on true positive

[23] More specifically, they have to identify the location of molecular markers on the genome relative to the coordinates of known genes (Vaughan and Srinivasasainagendra, 2013). For earlier GWAS, this step was not fully routinized, and it was not uncommon to attribute the mutation to the wrong gene (usually a neighboring one). Wrongly mapped associations are indeed 62.3% less likely to be replicated than correctly reported findings.

associations (Columns 1 and 2 of Appendix Table E.10). Intuitively, this finding is consistent with the idea that there is little strategic value in targeting associations that do not have therapeutic value. Second, patents targeting false positive leads are less likely to end up in litigation or being renewed by the firm (Columns 3 and 4 of Appendix Table E.10). By revealed preferences, this implies that the actual strategic value of such patents is very low.

Finally, it could be that pursuing false positive predictions results in learning spillovers for the firms. This possibility would imply that the cost of selecting the wrong target is much lower for the firm, consistent with recent work in a similar context (Frankel et al., 2024). To test this, I examine innovation outcomes for gene-disease pairs involving the same gene and a similar disease of the GWAS associations. The hierarchical nature of the MeSH disease taxonomy provides an easy way to find diseases that share the same etiology and biological mechanisms. More specifically, I consider a disease (i.e., a four-digit MeSH code) similar to the one targeted by the GWAS if it shares the same "parent" disease (i.e., the same three-digit MeSH code). Appendix Table E.11 shows no evidence of spillover effects on innovation outcomes, suggesting that the beneficial effects of experimentally learning from mistakes are likely low in this context.

6 Firm-Level Mechanisms

The previous section showed that domain knowledge allows firms to recognize false positive predictions and focus investments on valuable opportunities. In this section, I investigate the mechanisms behind these results. First, although my analysis used corporate publications to find the specific genes on which firms had expertise, the results might reflect broader organizational capabilities also applying to genes not in the publications. Second, the ability to recognize opportunities could stem from different organizational learning mechanisms (Di Stefano et al., 2024). Domain knowledge could enable offline learning either because it reflects practical experience in the domain or because it permits the application of theoretical principles to evaluate predictions.

6.1 Empirical Strategy

The findings in the previous section are drawn from split-sample regressions at the gene-disease level that exploit between-firm variation in gene knowledge. In this section, I use an alternative source of variation: the within-firm heterogeneity in the available knowledge about different human genes. Panel (a) of Appendix Figure E.6 provides a graphical representation of this research design. The basic intuition is akin to the preceding section since firms might better understand the biological mechanisms of specific genes than others due to their past research choices. However, this research design benefits from the ability to include firm fixed effects in the regression models. As a result, the

estimates effectively control for unobservable firm characteristics that correlate with their ability to recognize valuable opportunities.

For this analysis, I construct a dataset where each observation represents a potential firm investment in a gene-disease pair supported by GWAS evidence. This data structure allows me to investigate the investment decisions of firms that observe data-driven predictions. I assume that firms consider all GWAS findings about diseases they have previously invested in.[24] Pharmaceutical firms are usually active in specific diseases (that can be thought of as markets), which gives a convenient way to assemble the risk set of GWAS associations that are likely to catch their attention (Bikard, 2018; Krieger, 2021). While the earlier regressions at the gene-disease level offer a more accurate estimate of GWAS' aggregate impact, the within-firm design provides a tighter way to isolate the effect of domain knowledge on firm investments. However, variation in firms' past research portfolios is not random, so the results at the firm level should be interpreted as suggestive correlations.

I use OLS to estimate the following specification at the firm-gene-disease level:

$$Y_{f,i,j} = \alpha + \beta\, Domain\, Knowledge_{f,i,j} + \mu_{f,j} + \theta_{i,j} + \epsilon_{f,i,j}, \tag{2}$$

where $Y_{f,i,j}$ is the number of patent applications by firm f for innovations targeting the gene-disease combination $< i, j >$ appearing in a GWAS. $Domain\, Knowledge_{f,i,j}$ is a dummy that equals one if the firm possesses domain knowledge about the gene involved in the gene-disease association (as proxied by previous publications on that gene). $\mu_{f,j}$ are firm-by-disease dummies, accounting for firm specialization in specific disease areas as well as potential differences in how firms weigh errors of omission against errors of commission.[25] $\theta_{i,j}$ is a time effect that controls for the year when the gene-disease association is published. The coefficient of interest β captures how a firm's investment decisions change when it can leverage genetic knowledge to evaluate the GWAS finding. Estimates report standard errors clustered at the firm level.

[24] This choice is consistent with what emerged from my interviews: "I would say we definitely pay attention to GWAS studies. Any time there is a GWAS study that, you know, that chose an indication that we're interested in, we definitely pay attention" (Interview, 9 December 2022). In the appendix, I experiment with alternative ways to define the appropriate risk set of GWAS evaluated by a firm, finding consistent results (Appendix Tables E.14 and E.15).

[25] For instance, an alternative explanation of my previous findings could be that firms follow different strategies regarding false positives. Firms that prioritize minimizing the risk of missing a valuable target might be willing to incur several commission errors, and vice versa. Including firm fixed effects helps rule out that the results are driven by firms pursuing different strategies regarding the weight given to type I vs. type II errors. This is because if the relative disutility of a false negative relative to a false positive is a fixed firm-specific parameter, the firm-fixed effect will absorb it. The additional inclusion of firm-by-disease fixed effects would also consider the more likely case that this parameter depends on the specific disease (e.g., for Abbvie, missing out on a Crohn's disease target might be worse than doing so for other diseases where it does not specialize).

6.2 How Does Domain Knowledge Help?

I begin by replicating the main results of the paper using within-firm variation. The results are consistent with the gene-disease-level evidence and reported in Appendix Table E.13. Column 1 shows that firms are less likely to invest in a false positive when they can draw upon their knowledge of gene biology. Note that including firm fixed effects changes the interpretation of the estimates. In this case, the coefficient captures the decrease in patenting targeting false positive associations when a firm has domain knowledge, compared to other false positives where the firm lacks such expertise. This result is robust to including stringent firm-by-disease fixed effects that better account for firm specialization in specific disease areas (Column 2). Columns 3 and 4 symmetrically show that domain knowledge increases the likelihood of recognizing valuable data-driven opportunities. In addition to confirming earlier analyses, these findings suggest that the main results are not due to organizational-level capabilities but rather to the specific distribution of knowledge across genes.

Next, I explore why domain knowledge is conducive to my results. A long tradition in the management literature has emphasized how organizations can learn routines to perform even complex operations, such as selecting innovation projects (Nelson and Winter, 1982). This line of work emphasizes these heuristics' tacit and often poorly understood nature. In contrast, a growing strand of research highlights the importance of developing theory-based understandings to guide action (Camuffo et al., 2024; Chavda et al., 2024; Wuebker et al., 2023). Recent evidence shows that experience unlocks higher performance only when accompanied by an understanding of the causal relationships between antecedents and outcomes (Di Stefano et al., 2024). In my setting, this would suggest that the ability to recognize valuable GWAS findings stems from understanding the underlying mechanisms of what makes an opportunity valuable rather than from mere practice or generic organizational routines.

I categorize firms' knowledge types to shed some light on this issue by distinguishing firms' publications that denote testing capabilities from those focused on the genetic mechanisms of diseases. Following the approach of Azoulay et al. (2021), I record which firm publications are translational research that applies theoretical genetic research to therapeutic purposes (Appendix D.3). I also code firms that have conducted clinical or disease-oriented research that was not aimed at elucidating genetic mechanisms. I then run separate regressions using each of these proxies for domain knowledge. Figure 6 shows that only translational knowledge aids in correctly avoiding false positive associations among the GWAS findings. Instead, clinical research or generic experience with the disease do not confer any advantage when assessing GWAS findings. Coupled with my earlier findings, this evidence suggests that offline learning is driven by the ability to apply basic scientific principles to rule out spurious gene-disease associations (Di Stefano et al., 2024; Fleming and Sorenson, 2004).

Finally, one might question why firms do not wait for the uncertainty surrounding a gene-disease

association to dissipate. For instance, a firm could hold back until further research validates the finding or strategically delay investments to observe competitors' actions (Krieger, 2021). Appendix Table E.16 shows descriptive evidence that waiting before investing in a GWAS finding indeed correlates with a lower likelihood of targeting a false positive association. However, it seems to come at the cost of lower-impact patents and lower chances of discovering a drug (albeit the results are noisier in this case). Since patenting results in claiming intellectual property space, being late might result in fewer avenues for future innovation (Eggers, 2012). This result illustrates how data-driven predictions generate a tension between carefully validating them and investing ahead of competitors. Evaluating alternatives offline is not only cheaper but also likely faster, suggesting another advantage of domain knowledge that future research should explore more.

7 Conclusion

While the field of strategic management has extensively researched how firms can exploit innovation opportunities, the rapid diffusion of big data and predictive tools is transforming how firms identify them. In particular, data-driven predictions offer new avenues for assessing the value of potential investments instead of leveraging domain knowledge. This paper provides some of the first empirical evidence of the consequences of this phenomenon. I introduce a new conceptual and empirical framework to analyze how data-driven predictions and firms' domain knowledge interact to shape innovation performance. Leveraging the unique features of GWAS, I show that data-driven predictions can have heterogeneous effects on firms depending on their expertise. Firms with relevant genetic knowledge rely less on GWAS findings but disproportionally avoid investing in the false positives uncovered by GWAS. Together, these patterns are consistent with those firms being better able to evaluate the data-driven findings.

My results underscore the continuing importance of domain knowledge in the innovation process but with a new role: since the abundance of data-driven predictions is not accompanied by a generalized reduction in the cost of validating them, domain knowledge becomes essential to avoid wasting resources on dead ends. My framework also highlights the boundary conditions of this reasoning. If predictive tools were perfectly accurate or validating their predictions was costless, the need to prioritize among data-driven predictions would be lower. However, this is not the case in many innovation and strategy contexts. My theory reconciles conflicting results in previous research by suggesting that when data-driven findings are straightforward to interpret, firms can triage them without relying on domain knowledge, thus benefiting new entrants or smaller firms (Nagaraj, 2022; Galdon-Sanchez et al., 2024). Conversely, one should expect that understanding the domain is a necessary complement to benefit from data predictions in more complex domains where predictions

are likely noisier (Kao, 2023; Otis et al., 2024).[26]

An important contribution of this paper is methodological. The innovation process is often depicted as a search through landscapes of technological combinations (Kneeland et al., 2020; Fleming, 2001). Traditional research in this area frequently uses computer simulations to characterize optimal exploration strategies and their organizational implications (e.g., Knudsen and Levinthal 2007). In contrast, this paper joins a recent effort to try to empirically characterize organizational search (Kang, 2024; Kao, 2023; Nagaraj, 2022) and how different search strategies map into innovation outcomes (Tranchero, 2024). My work conceptualizes and exemplifies a new measurement strategy to depict empirical landscapes using text-based methods, opening up new research avenues for an empirical re-examination of the seminal contributions in the field of organizational search (Denrell and March, 2001; Levinthal and March, 1993; Nelson and Winter, 1982).

My research also has practical implications for managers and policymakers navigating the changing landscape of predictive technologies such as AI. To cut through the hype surrounding big data, it is essential to understand what predictive technologies can and cannot do (Agrawal et al., 2018b). Specifically, this paper proposes that while data-driven predictions are valuable in shortlisting potential opportunities—which is no small feat—they are not without limitations. This is especially true for what Camuffo et al. (2023) call "low-frequency/high-impact" organizational decisions. In these contexts, the source of competitive advantage lies in the ability to interpret and evaluate what the data suggest. Managers should be mindful that the benefits of predictive technologies may be most evident to organizations with a solid domain knowledge base. Similarly, governments funding large-scale data efforts, which are especially common in the bio-pharmaceutical world, should note how they might affect competition dynamics. Instead of being a panacea to foster entry and market dynamism, large-scale mapping efforts might contribute to stifling competition if they require complementary assets that are the purview of incumbents.

More broadly, predictive technologies allow innovators to identify promising areas of the technological landscape based on past observations (Kim, 2024b). As a result, data-based correlations and extrapolations increasingly guide organizational decision-making instead of logical reasoning and domain knowledge. While the popular press often hails the benefits of these developments (Anderson, 2008; Schmidt, 2023), the risk of not understanding why innovations work could be the accumulation of an "intellectual debt" (Zittrain, 2019). This risk is especially evident in fields like drug discovery, where AI is used to find drugs with unknown mechanisms of action, thus preventing the anticipation of

[26]Recent work has started building on the ideas of the present paper and offering additional confirmation at the individual level. Toner-Rodgers (2024) shows that researchers in material sciences have become more innovative thanks to AI-assisted tools. However, the effects are heterogeneous, and the greatest benefit occurs for the most knowledgeable scientists, who can prioritize promising predictions and avoid false positives.

potential side effects (Heaven, 2023). With firms hungry to find innovation opportunities while potentially shying away from investing in domain knowledge (Balasubramanian et al., 2022), more research is needed to understand the nuanced implications of putting data ahead of theoretical understanding. This paper serves as a first step in this direction.

Finally, despite the contributions outlined above, a few limitations of this paper should be acknowledged. First, fully capturing the aggregate effects of GWAS on innovation is beyond its scope. Data-driven predictions in genomics might give rise to a "streetlight effect" (Hoelzemann et al., 2024), potentially diverting investments into suboptimal targets with negative consequences for social welfare. Moreover, since I do not observe costs directly, the corresponding implications for firm profitability are also beyond the scope of this study. Second, this paper does not explore the strategic responses of firms without domain knowledge. It is conceivable that a market for expertise could emerge, where firms outsource not the generation of ideas but expert judgment to assess the potential of data-driven opportunities and avoid false positives (Agrawal et al., 2021; Luo et al., 2021). Finally, while my research design focuses on holding the content of data-driven predictions constant to study firm responses, another critical dimension to explore is heterogeneity in prediction quality. Identifying which firms can obtain better predictions and understanding how organizational factors influence this skill are both first-order questions. Further exploration of these ideas is an exciting avenue for future research.

References

Adner, R. and D. A. Levinthal (2024): "Strategy Experiments in Nonexperimental Settings: Challenges of Theory, Inference, and Persuasion in Business Strategy," *Strategy Science*, Forthcoming.

Agarwal, R., F. Bacco, A. Camuffo, A. Coali, A. Gambardella, et al. (2023): "Does a theory-of-value add value? Evidence from a randomized control trial with Tanzanian entrepreneurs," *Bocconi University Research Paper*.

Agrawal, A., J. Gans, and A. Goldfarb (2018a): "Prediction, judgment, and complexity: a theory of decision-making and artificial intelligence," in *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press, 89–110.

——— (2018b): *Prediction machines: The simple economics of artificial intelligence*, Harvard Business Press.

Agrawal, A., J. S. Gans, and S. Stern (2021): "Enabling entrepreneurial choice," *Management Science*, 67, 5510–5524.

Agrawal, A., J. McHale, and A. Oettl (2024): "Artificial intelligence and scientific discovery: A model of prioritized search," *Research Policy*, 53, 104989.

Allen, R. and P. Choudhury (2022): "Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion," *Organization Science*, 33, 149–169.

Allen, R. and R. McDonald (2024): "Methodological pluralism and innovation in data-driven organizations," *Administrative Science Quarterly*, Forthcoming.

Anderson, C. (2008): "The end of theory: The data deluge makes the scientific method obsolete," *Wired*, 16, 16–07.

Arora, A. and A. Fosfuri (2005): "Pricing diagnostic information," *Management Science*, 51, 1092–1100.

Arora, A. and A. Gambardella (1994a): "The changing technology of technological change: General and abstract knowledge and the division of innovative labour," *Research Policy*, 23, 523–532.

——— (1994b): "Evaluating technological information and utilizing it: Scientific knowledge, technological capability, and external linkages in biotechnology," *Journal of Economic Behavior & Organization*, 24, 91–114.

Azoulay, P., W. H. Greenblatt, and M. L. Heggeness (2021): "Long-term effects from early exposure to research: Evidence from the NIH "Yellow Berets"," *Research Policy*, 50, 104332.

Balasubramanian, N., Y. Ye, and M. Xu (2022): "Substituting human decision-making with machine learning: Implications for organizational learning," *Academy of Management Review*, 47, 448–465.

Berman, R. and A. Israeli (2022): "The value of descriptive analytics: Evidence from online retailers," *Marketing Science*, 41, 1074–1096.

Berman, R. and C. Van den Bulte (2022): "False discovery in A/B testing," *Management Science*, 68, 6762–6782.

Bessen, J., S. M. Impink, L. Reichensperger, and R. Seamans (2022): "The role of data for AI startup growth," *Research Policy*, 51, 104513.

Bhatia, A. and G. Dushnitsky (2023): "The future of venture capital? Insights into data-driven VCs," *California Management Review*.

Bikard, M. (2018): "Made in academia: The effect of institutional origin on inventors' attention to science," *Organization Science*, 29, 818–836.

Boudreau, K. J., E. C. Guinan, K. R. Lakhani, and C. Riedl (2016): "Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science," *Management Science*, 62, 2765–2783.

Brynjolfsson, E., W. Jin, and K. McElheran (2021): "The power of prediction: Predictive analytics, workplace complements, and business performance," *Business Economics*, 56, 217–239.

Brynjolfsson, E., D. Li, and L. R. Raymond (2024): "Generative AI at work," *Quarterly Journal of Economics*.

Brynjolfsson, E. and K. McElheran (2016): "The rapid adoption of data-driven decision-making," *American Economic Review*, 106, 133–39.

Camuffo, A., A. Gambardella, and A. Pignataro (2023): "Framing strategic decisions in the digital world," *Strategic Management Review*, 4, 127–160.

——— (2024): "Theory-driven strategic management decisions," *Strategy Science*, Forthcoming.

Cao, R., R. Koning, and R. Nanda (2024): "Sampling bias in entrepreneurial experiments," *Management Science*, 70, 7283–7307.

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

Cattani, G. (2005): "Preadaptation, firm heterogeneity, and technological performance: A study on the evolution of fiber optics, 1970–1995," *Organization Science*, 16, 563–580.

Chatterji, A., S. Hasan, R. P. Larrick, and R. Masclans (2023): "Taste Before Production: The Role of Judgment in Entrepreneurial Idea Generation," *Available at SSRN 4324964*.

Chavda, A., J. S. Gans, and S. Stern (2024): "Theory-based entrepreneurial search," *Strategy Science*, Forthcoming.

Choi, S., H. Kang, N. Kim, and J. Kim (2024): "How does AI improve human decision-making? Evidence from the AI-powered go program," *Strategic Management Journal*, Forthcoming.

Choudhury, P., E. Starr, and R. Agarwal (2020): "Machine learning and human capital complementarities: Experimental evidence on bias mitigation," *Strategic Management Journal*, 41, 1381–1411.

Cockburn, I. M., R. Henderson, and S. Stern (2019): "The impact of artificial intelligence on innovation," in *The Economics of Artificial Intelligence: An Agenda*, Chicago: Chicago University Press, 115–146.

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

Cohen, W. M. and D. A. Levinthal (1990): "Absorptive capacity: A new perspective on learning and innovation," *Administrative Science Quarterly*, 35, 128–152.

——— (1994): "Fortune favors the prepared firm," *Management Science*, 40, 227–251.

Cohen, W. M., R. R. Nelson, and J. P. Walsh (2002): "Links and impacts: The influence of public research on industrial R&D," *Management Science*, 48, 1–23.

Conti, R., M. G. de Matos, and G. Valentini (2024): "Big data analytics, firm size, and performance," *Strategy Science*, 9, 135–151.

Conti, R., A. Gambardella, and M. Mariani (2014): "Learning to be Edison: Inventors, organizations, and breakthrough inventions," *Organization Science*, 25, 833–849.

Dell'Acqua, F., E. McFowland III, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, et al. (2023): "Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality," *HBS Working Paper 24-013*.

Denrell, J. and J. G. March (2001): "Adaptation as information restriction: The hot stove effect," *Organization Science*, 12, 523–538.

Di Stefano, G., F. Gino, G. P. Pisano, and B. R. Staats (2024): "Learning by thinking: How reflection can spur progress along the learning curve," *Workling paper*.

Dranove, D., C. Garthwaite, and M. Hermosilla (2022): "Does consumer demand pull scientifically novel drug innovation?" *The RAND Journal of Economics*, 53, 590–638.

Duerr, R. H., K. D. Taylor, S. R. Brant, J. D. Rioux, M. S. Silverberg, M. J. Daly, et al. (2006): "A genome-wide association study identifies IL23R as an inflammatory bowel disease gene," *Science*, 314, 1461–1463.

Eggers, J. P. (2012): "Falling flat: Failed technologies and investment under uncertainty," *Administrative Science Quarterly*, 57, 47–80.

Eggers, J. P. and S. Kaplan (2009): "Cognition and renewal: Comparing CEO and organizational effects on incumbent adaptation to technical change," *Organization Science*, 20, 461–477.

Ewens, M., R. Nanda, and M. Rhodes-Kropf (2018): "Cost of experimentation and the evolution of venture capital," *Journal of Financial Economics*, 128, 422–442.

Felin, T. and T. R. Zenger (2017): "The theory-based view: Economic actors as theorists," *Strategy Science*, 2, 258–271.

Fleming, L. (2001): "Recombinant uncertainty in technological search," *Management Science*, 47, 117–132.

Fleming, L. and O. Sorenson (2001): "Technology as a complex adaptive system: evidence from patent data," *Research Policy*, 30, 1019–1039.

——— (2004): "Science as a map in technological search," *Strategic Management Journal*, 25, 909–928.

Frankel, A. P., J. L. Krieger, D. Li, and D. Papanikolaou (2024): "Evaluation and learning in R&D investment," *NBER wp 31290*.

Freedman, L. P., I. M. Cockburn, and T. S. Simcoe (2015): "The economics of reproducibility in preclinical research," *PLoS biology*, 13, e1002165.

Furman, J. L., M. Nagler, and M. Watzinger (2021): "Disclosure and subsequent innovation: Evidence from the patent depository library program," *American Economic Journal: Economic Policy*, 13, 239–70.

Galdon-Sanchez, J. E., R. Gil, and G. Uriz Uharte (2024): "The Value of Information in Competitive Markets: Evidence from Small and Medium Enterprises," *Journal of Political Economy*.

Gambardella, A. (1992): "Competitive advantages from in-house scientific research: The US pharmaceutical industry in the 1980s," *Research Policy*, 21, 391–407.

Gans, J. S., S. Stern, and J. Wu (2019): "Foundations of entrepreneurial strategy," *Strategic Management Journal*, 40, 736–756.

Gavetti, G. and D. Levinthal (2000): "Looking forward and looking backward: Cognitive and experiential search," *Administrative Science Quarterly*, 45, 113–137.

Goldstein, D. B. (2009): "Common genetic variation and human traits," *New England Journal of Medicine*, 360, 1696.

Gruber, M., I. C. MacMillan, and J. D. Thompson (2008): "Look before you leap: Market opportunity identification in emerging technology firms," *Management Science*, 54, 1652–1665.

Harrison, J. R. and J. G. March (1984): "Decision making and postdecision surprises," *Administrative Science Quarterly*, 26–42.

Heaven, W. D. (2023): "AI is dreaming up drugs that no one has ever seen. Now we've got to see if they work," *MIT Technology Review*.

Hermosilla, M. and J. Lemus (2019): "Therapeutic translation of genomic science," in *Economic Dimensions of Personalized and Precision Medicine*, Chicago University Press.

Hoelzemann, J., G. Manso, A. Nagaraj, and M. Tranchero (2024): "The streetlight effect in data-driven exploration," *NBER wp 32401*.

Hoffman, M., L. B. Kahn, and D. Li (2018): "Discretion in hiring," *The Quarterly Journal of Economics*, 133, 765–800.

Jin, W. and K. McElheran (2024): "Economies before scale: IT strategy and performance dynamics of young U.S. businesses," *Management Science*.

Kang, S. (2024): "From outward to inward: Reframing search with new mapping criteria," *UC Santa Barbara*.

Kao, J. (2023): "Charted territory: Mapping the cancer genome and R&D decisions in the pharmaceutical industry," *UCLA Anderson*.

Katila, R. and G. Ahuja (2002): "Something old, something new: A longitudinal study of search behavior and new product introduction," *Academy of Management Journal*, 45, 1183–1194.

Kim, H. (2024a): "The value of competitor information: Evidence from a field experiment," *Management Science*, Forthcoming.

Kim, S. (2024b): "Shortcuts to innovation: The use of analogies in knowledge production," *Columbia Business School*.

Kneeland, M. K., M. A. Schilling, and B. S. Aharonson (2020): "Exploring uncharted territory: Knowledge search processes in the origination of outlier innovation," *Organization Science*, 31, 535–557.

Knudsen, T. and D. A. Levinthal (2007): "Two faces of search: Alternative generation and alternative evaluation," *Organization Science*, 18, 39–54.

Koning, R., S. Hasan, and A. Chatterji (2022): "Experimentation and start-up performance: Evidence from A/B testing," *Management Science*, 68, 6434–6453.

Krakowski, S., J. Luger, and S. Raisch (2023): "Artificial intelligence and the changing sources of competitive advantage," *Strategic Management Journal*, 44, 1425–1452.

Krieger, J., D. Li, and D. Papanikolaou (2022): "Missing novelty in drug development," *The Review of Financial Studies*, 35, 636–679.

Krieger, J., R. Nanda, I. Hunt, A. Reynolds, and P. Tarsa (2023): "Scoring and funding breakthrough ideas: Evidence from a global pharmaceutical company," *HBS Working Paper 23-014*.

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

Lane, J. N., Z. Szajnfarber, J. Crusan, M. Menietti, and K. R. Lakhani (2022a): "Are experts blinded by feasibility? Experimental evidence from a NASA robotics challenge," *HBS Working Paper 22-071*.

Lane, J. N., M. Teplitskiy, G. Gray, H. Ranu, M. Menietti, E. C. Guinan, and K. R. Lakhani (2022b): "Conservatism gets funded? A field experiment on the role of negative information in novel project evaluation," *Management Science*, 68, 4478–4495.

Levinthal, D. A. and J. G. March (1993): "The myopia of learning," *Strategic Management Journal*, 14, 95–112.

Li, D. (2017): "Expertise versus Bias in Evaluation: Evidence from the NIH," *American Economic Journal: Applied Economics*, 9, 60–92.

Lou, B. and L. Wu (2021): "AI on drugs: Can artificial intelligence accelerate drug development? Evidence from a large-scale examination of bio-pharma firms," *MIS Quarterly*, 45.

Luo, H., J. Macher, and M. Wahlen (2021): "Judgment aggregation in creative production: Evidence from the movie industry," *Management Science*, 67, 6358–6377.

MacArthur, D. (2012): "Face up to false positives," *Nature*, 487, 427–428.

Marigorta, U. M., J. A. Rodríguez, G. Gibson, and A. Navarro (2018): "Replicability and prediction: Lessons and challenges from GWAS," *Trends in Genetics*, 34, 504–517.

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

Nagaraj, A. and M. Tranchero (2024): "Empirical search landscapes," *UC Berkeley and University of Pennsylvania*.

Nelson, M. R., H. Tipney, J. L. Painter, J. Shen, P. Nicoletti, et al. (2015): "The support of human genetic evidence for approved drug indications," *Nature Genetics*, 47, 856–860.

Nelson, R. R. and S. Winter (1982): *An evolutionary theory of economic change*, Belknap Press.

Nicholson, J. M., M. Mordaunt, P. Lopez, A. Uppala, et al. (2021): "Scite: A smart citation index that displays the context of citations and classifies their intent using deep learning," *Quantitative Science Studies*, 2, 882–898.

Noy, S. and W. Zhang (2023): "Experimental evidence on the productivity effects of generative artificial intelligence," *Science*, 381, 187–192.

Otis, N., R. P. Clarke, S. Delecourt, D. Holtz, and R. Koning (2024): "The uneven impact of generative AI on entrepreneurial performance," *Available at SSRN 4671369*.

Puranam, P. (2019): "The theorist as an endangered species?" *Journal of Marketing Behavior*, 4, 43–48.

Righi, C. and T. Simcoe (2023): "Patenting inventions or inventing patents? Continuation practice at the USPTO," *The RAND Journal of Economics*, 54, 416–442.

Schmidt, E. (2023): "This is how AI will transform the way science gets done," *MIT Technology Review*.

Shane, S. (2000): "Prior knowledge and the discovery of entrepreneurial opportunities," *Organization Science*, 11, 448–469.

Squicciarini, M., H. Dernis, and C. Criscuolo (2013): "Measuring patent quality: Indicators of technological and economic value," *OECD STI Working Papers, No. 2013/03*.

Struck, T. J., B. K. Mannakee, and R. N. Gutenkunst (2018): "The impact of genome-wide association studies on biomedical research publications," *Human Genomics*, 12, 1–9.

Suchindran, S., D. Rivedal, J. R. Guyton, T. Milledge, X. Gao, et al. (2010): "Genome-wide association study of Lp-PLA2 activity and mass in the Framingham Heart Study," *PLoS genetics*, 6, e1000928.

Tam, V., N. Patel, M. Turcotte, Y. Bossé, G. Paré, and D. Meyre (2019): "Benefits and limitations of genome-wide association studies," *Nature Reviews Genetics*, 20, 467–484.

Toner-Rodgers, A. (2024): "Artificial Intelligence, Scientific Discovery, and Product Innovation," *MIT*.

Tranchero, M. (2024): "Data-driven search and innovation: Evidence from genome-wide association studies," *University of Pennsylvania*.

Uffelmann, E., Q. Q. Huang, N. S. Munung, J. de Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen, and D. Posthuma (2021): "Genome-wide association studies," *Nature Reviews Methods Primers*, 1, 1–21.

Vaughan, L. K. and V. Srinivasasainagendra (2013): "Where in the genome are we? A cautionary tale of database use in genomics research," *Frontiers in Genetics*, 4, 38.

Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang (2017): "10 years of GWAS discovery: Biology, function, and translation," *The American Journal of Human Genetics*, 101, 5–22.

Wuebker, R., T. Zenger, and T. Felin (2023): "The theory-based view: Entrepreneurial microfoundations, resources, and choices," *Strategic Management Journal*, 44, 2922–2949.

Zittrain, J. (2019): "The hidden costs of automated thinking," *The New Yorker*.

8 Figures and Tables

Figure 1: Example and schema of a typical GWAS

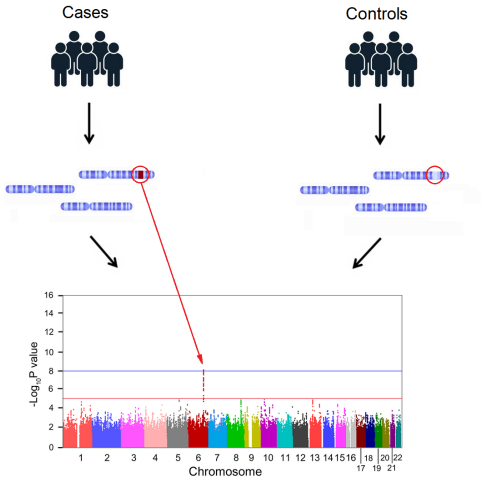
(a) *Abstract of the GWAS by Duerr et al. (2006) that identified IL23R as a therapeutic target for Crohn's disease*

A Genome-Wide Association Study Identifies *IL23R* as an Inflammatory Bowel Disease Gene

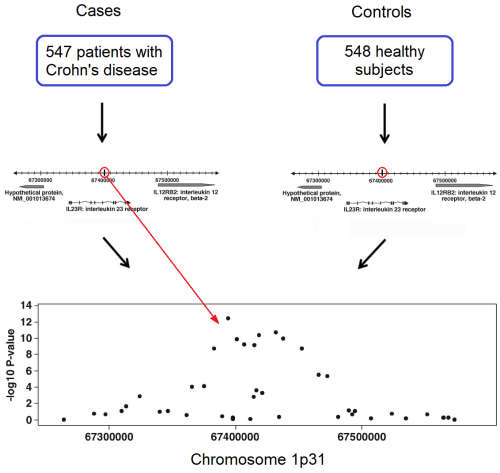
Richard H. Duerr,[1,2] Kent D. Taylor,[3,4] Steven R. Brant,[5,6] John D. Rioux,[7,8] Mark S. Silverberg,[9] Mark J. Daly,[8,10] A. Hillary Steinhart,[9] Clara Abraham,[11] Miguel Regueiro,[1] Anne Griffiths,[12] Themistocles Dassopoulos,[5] Alain Bitton,[13] Huiying Yang,[3,4] Stephan Targan,[4,14] Lisa Wu Datta,[5] Emily O. Kistner,[15] L. Philip Schumm,[15] Annette T. Lee,[16] Peter K. Gregersen,[16] M. Michael Barmada,[2] Jerome I. Rotter,[3,4] Dan L. Nicolae,[11,17] Judy H. Cho[18]*

The inflammatory bowel diseases Crohn's disease and ulcerative colitis are common, chronic disorders that cause abdominal pain, diarrhea, and gastrointestinal bleeding. To identify genetic factors that might contribute to these disorders, we performed a genome-wide association study. We found a highly significant association between Crohn's disease and the *IL23R* gene on chromosome 1p31, which encodes a subunit of the receptor for the proinflammatory cytokine interleukin-23. An uncommon coding variant (rs11209026, c.1142G>A, p.Arg381Gln) confers strong protection against Crohn's disease, and additional noncoding *IL23R* variants are independently associated. Replication studies confirmed *IL23R* associations in independent cohorts of patients with Crohn's disease or ulcerative colitis. These results and previous studies on the proinflammatory role of IL-23 prioritize this signaling pathway as a therapeutic target in inflammatory bowel disease.

(b) *Schema of a typical GWAS*

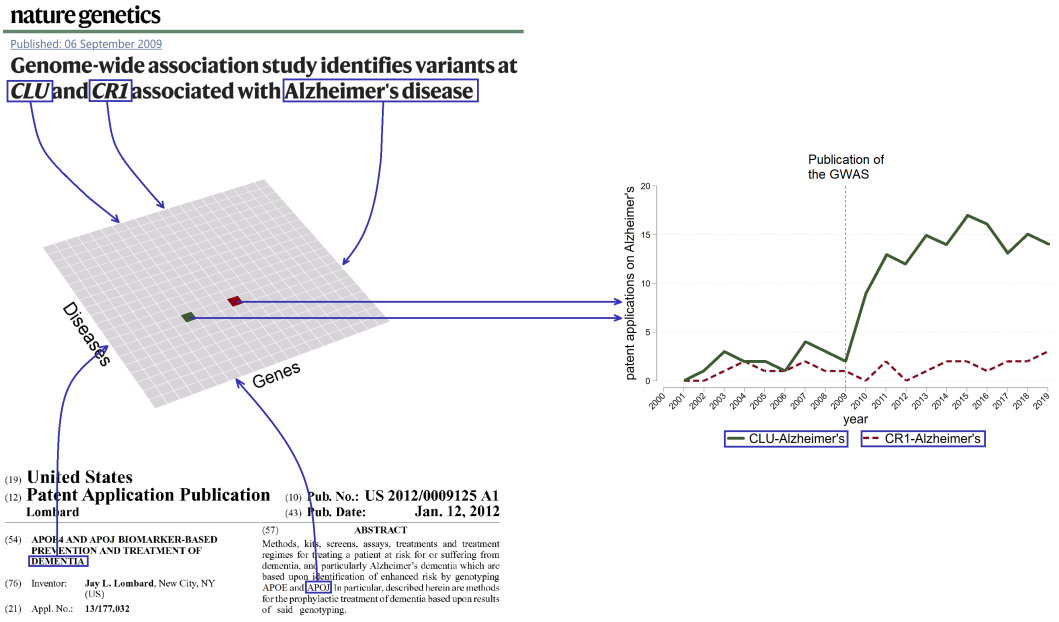


(c) *Schema of Duerr et al. (2006)*



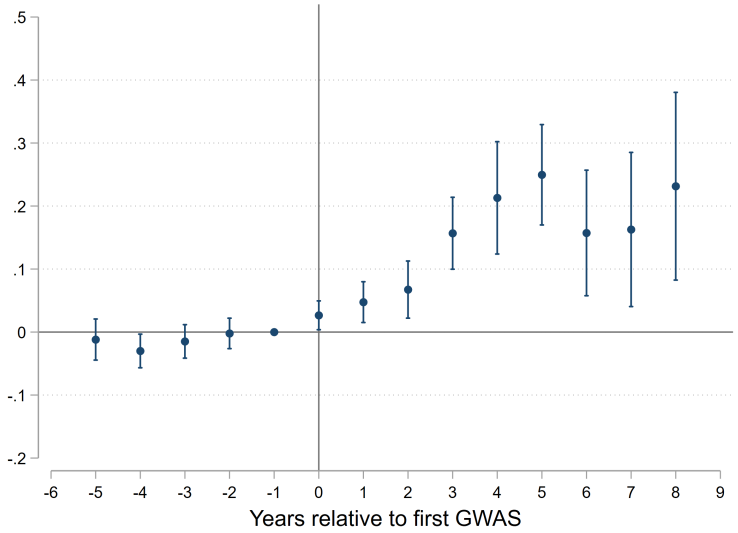
Note: Panel (a) shows the abstract of the Duerr et al. (2006) GWAS published in *Science*. The abstract highlights the potential implications of using the IL23R gene as a target to cure inflammatory bowel diseases. Panel (b) shows a schema of how a GWAS unfolds. First, researchers select the disease of interest and assemble a group of cases (subjects showing the condition) and one of controls (subjects without the condition). Then, the genome of people with and without the condition is genotyped in search of differences. Finally, statistical methods are used to test the association between any genetic mutation and the disease of interest. The panel at the bottom is the characteristic "Manhattan plot," which indicates the location of the statistically significant genetic variants in the chromosome. On the Y axis, there is the strength of the finding expressed as -$\log_{10}$(p-value), hence higher values correspond to stronger associations. Panel (c) shows the same schema for the GWAS of Duerr et al. (2006). See Appendix A.2 for a detailed case study on this GWAS.

Figure 2: Pharmaceutical firms search for drugs in an empirical landscape of gene-disease pairs



Note: The figure exemplifies the novel measurement approach adopted in this paper. Machine learning tools extracting genes and diseases from patent and paper texts allow to map them onto an empirical landscape of gene-disease combinations. The heterogeneous impact of GWAS on firm investments can be measured by tracking patenting at the gene-disease combination level, avoiding the use of patent citations. APOJ is an alternative name for the CLU gene. See the conceptual details in Appendix B.

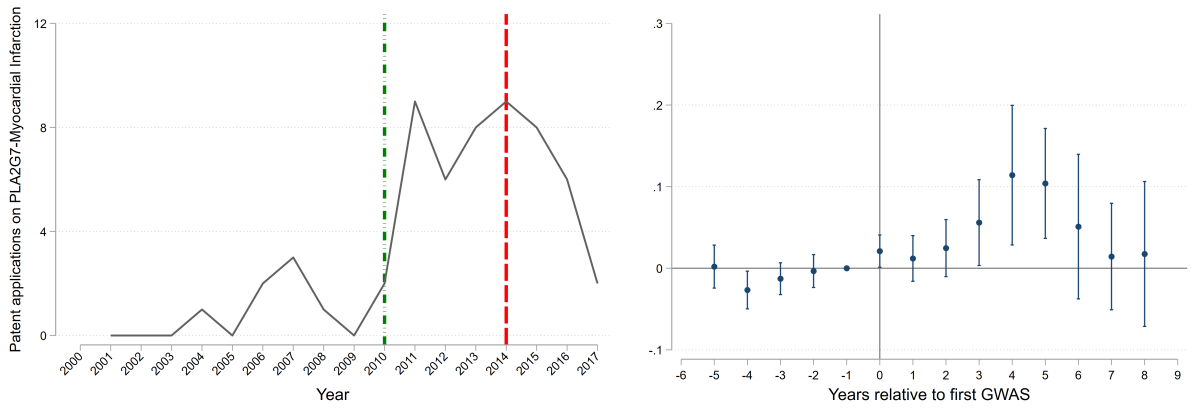
Figure 3: GWAS findings have a large and persistent impact on firm innovation investments



Note: The figure shows the event study coefficients estimated from the following specification: $Patent\ Applications_{i,j,t} = \alpha + \sum_z \beta_z GWAS_{i,j} \times 1(z) + \gamma GD_{i,j} + \delta_t \times Gene_i + \omega_t \times Disease_j + \epsilon_{i,j,t}$. The dependent variable is the number of USPTO patent applications for innovations targeting a specific gene-disease combination. The chart plots values of β_z for different lags z before and after the publication of the first GWAS reporting the gene-disease pair. Regressions include gene $\times$ year and disease $\times$ year fixed effects, as well as gene-disease combination fixed effects. Standard errors are clustered two-way at the gene and disease level.

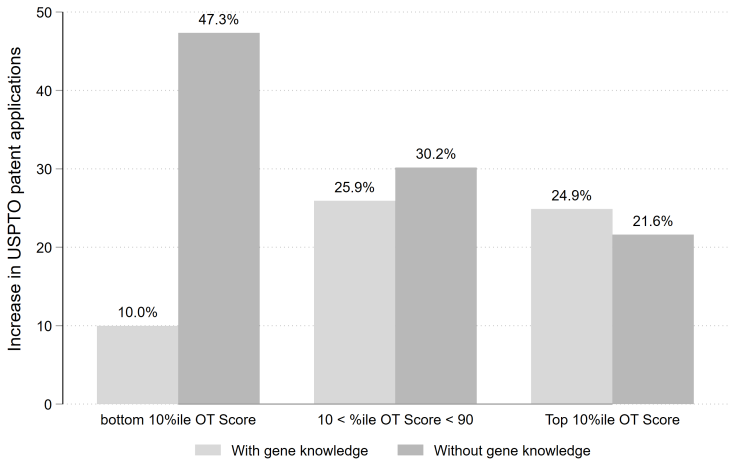
Figure 4: Firms investing in innovations based on false positive GWAS findings pivot away over time

(a) *PLA2G7 - myocardial infarction* (b) *Time-varying impact of false positive GWAS*



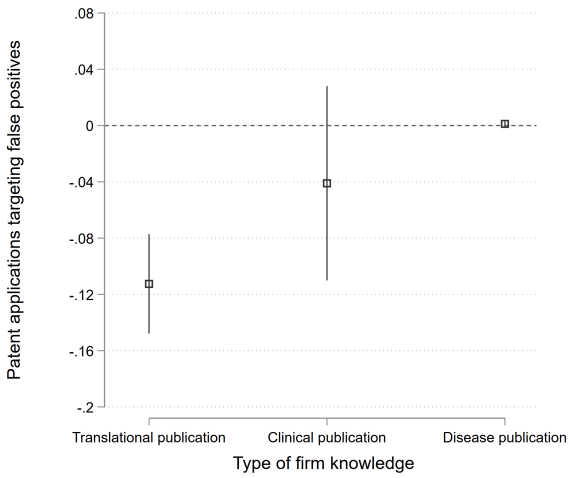
Note: Panel (a) shows the count of yearly USPTO patent applications for innovations targeting the PLA2G7 - myocardial infarction gene-disease pair. The green line in 2010 marks the publication of Suchindran et al. (2010), while the red line in 2014 marks the failure of clinical trials for Darapladib, a GlaxoSmithKline drug targeting this combination. Panel (b) shows the event study coefficients from the same specification of Figure 3. The dependent variable is the number of USPTO patent applications for innovations targeting a specific gene-disease combination. The chart plots values of β_z for different lags z before and after the publication of a GWAS finding about the gene-disease pair that is not subsequently replicated. Regressions include gene $\times$ year and disease $\times$ year fixed effects, as well as gene-disease combination fixed effects. Standard errors are clustered two-way at the gene and disease level.

Figure 5: Firms with domain knowledge invest proportionally more in the top tail of data-driven opportunities, while those lacking it invest mostly in gene-disease associations of low-value



Note: This figure shows the increase in USPTO patent applications for innovations on a gene-disease pair after the publication of a GWAS, as a share of the sample average ($\hat{\beta_{OLS}}/\mu$). The increase is estimated with split-sample regressions that consider gene-disease pairs of different therapeutic potential among those uncovered by GWAS, as proxied by the Open Targets score. Using the Open Targets score to proxy for prediction accuracy permits investigating firms' ability to recognize the most valuable gene-disease combinations, even among replicable associations. Details on the regressions are reported in Appendix Table E.12.

Figure 6: The ability to avoid false positives is driven by translational knowledge that permits a scientific evaluation of GWAS findings, but not by clinical knowledge or generic disease experience.



Note: This figure shows results separately for firms with different types of domain knowledge. Each coefficient is estimated from a separate regression with the same dependent variable. *Patent applications targeting false positives*: count of USPTO patent applications for innovations targeting a specific gene-disease combination that is not replicated by subsequent GWAS. *Translational publication*: 0/1 = 1 if the firm has published at least one previous translational paper involving the gene, according to the definition by Azoulay et al. (2021) (Appendix D.3). *Clinical publication*: 0/1 = 1 if the firm has published at least one previous clinical paper involving the gene, using the classification of NIH iCite. *Disease publication*: 0/1 = 1 if the firm has published at least one previous paper involving the disease but not the specific gene.

Table 1: Summary statistics at the gene-disease combination level

	Panel A: cross-sectional descriptives					
	mean	median	st d	min	max	N
Ever a patent application (0/1)	0.2182	0	0.4131	0	1	7,223,924
...by firms with gene knowledge (0/1)	0.0478	0	0.2134	0	1	7,223,924
...by firms without gene knowledge (0/1)	0.2102	0	0.4074	0	1	7,223,924
Ever a drug (0/1)	0.0019	0	0.0433	0	1	7,223,924
Ever treated by GWAS (0/1)	0.0025	0	0.0498	0	1	7,223,924
Has Open Targets score (0/1)	0.0823	0	0.2748	0	1	7,223,924
Open Targets score	0.0401	0.0074	0.0872	0.00004	0.8975	594,353
	Panel B: panel-level descriptives					
	mean	median	st d	min	max	N
Patent applications	0.1314	0	1.6831	0	1346	137,254,556
...by firms with gene knowledge	0.0376	0	1.0875	0	1346	137,254,556
...by firms without gene knowledge	0.0938	0	0.9243	0	1257	137,254,556
Cit-weighted patents	7.1358	0	85.5509	0	39,899	137,254,556
New drugs (total)	0.0002	0	0.0203	0	22	137,254,556
New drugs (weighted)	0.0001	0	0.0064	0	3.25	137,254,556
Treated by GWAS (0/1)	0.0005	0	0.0234	0	1	137,254,556
True positives (0/1)	0.0001	0	0.0013	0	1	137,254,556
Year	2010	2010	5.4772	2001	2019	137,254,556

Note: This table lists summary statistics at the gene-disease combination level for 7,223,924 combinations (Panel A) and at the gene-disease-year level for a balanced panel of 137,254,556 observations (Panel B). *Ever a patent application*: 0/1 =1 if the gene-disease combination appeared in at least one patent application. *Ever a patent application by firms with gene knowledge*: 0/1 =1 if the gene-disease combination appeared in at least one patent application by firms with previous publications on the gene involved. *Ever a patent application by firms without gene knowledge*: 0/1 = 1 if the gene-disease combination appeared in at least one patent application by firms without previous publications on the gene involved. *Ever a drug*: 0/1 = 1 if the gene-disease combination is targeted by at least one drug in the discovery stage. *Ever treated by GWAS*: 0/1 = 1 if the gene-disease combination is ever reported by a GWAS. *Has Open Targets score*: 0/1 = 1 if the gene-disease combination has an Open Targets score. *Open Targets score*: average value of the Open Target score (for gene-disease pairs that have it). *Patent applications*: count of USPTO patent applications for innovations that target a specific gene-disease combination. *Patent applications by firms with gene knowledge*: count of USPTO patent applications for innovations that target a specific gene-disease combination filed by firms with previous publications on the gene involved. *Patent applications by firms without gene knowledge*: count of USPTO patent applications for innovations that target a specific gene-disease combination filed by firms without previous publications on the gene involved. *Cit-weighted patents*: count of USPTO patent applications for innovations that target a specific gene-disease combination weighted by citations received up to 7 years after their publication. *New drugs (total)*: number of molecules on a gene-disease combination entering the discovery stage. *New drugs (weighted)*: number of molecules on a gene-disease combination entering the discovery stage weighted by their scientific novelty (i.e., by the number of times that the same mechanism of action has been used before, following Dranove et al. 2022). *Treated by GWAS*: 0/1 = 1 in all years after a gene-disease pair is reported by its first GWAS. *True positives*: 0/1 = 1 for GWAS findings that are later replicated by another GWAS about the same disease. *Year*: average year of observations in the panel.

Table 2: Pharmaceutical firms heavily increase their investments in gene-disease combinations that receive a GWAS, even if the association is a false positive finding

Dependent Variable:	USPTO patent applications			
Post × GWAS	0.2681***	0.1637***	0.1532***	0.0676***
	(0.02206)	(0.01985)	(0.01779)	(0.01605)
...× True Positive			0.5739***	0.4802***
			(0.08545)	(0.07693)
Gene-Disease FE	YES	YES	YES	YES
Disease FE	YES	NO	YES	NO
Gene FE	YES	NO	YES	NO
Year FE	YES	NO	YES	NO
Disease-Year FE	NO	YES	NO	YES
Gene-Year FE	NO	YES	NO	YES
N of Observations	137,254,556	137,254,556	137,254,556	137,254,556
N of Gene-Diseases	7,223,924	7,223,924	7,223,924	7,223,924
Mean of Dep Var:	0.1314	0.1314	0.1314	0.1314

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the gene-disease-year level. Std. err. clustered two-ways at the disease and gene level. *USPTO patent applications*: count of USPTO patent applications (by all firms) filed in a given year that target a specific gene-disease combination. *Post × GWAS*: 0/1 = 1 in all years after a gene-disease pair is reported by its first GWAS. *True Positive*: 0/1 = 1 for GWAS findings that are later replicated by another GWAS about the same disease. Given that 84.3% of GWAS findings are non-replicable, the estimate in Column 4 implies that around 34.8% of the total increase in patent applications is directed at non-replicable gene-disease associations.

Table 3: Only investments based on true positive GWAS associations yield downstream outcomes like valuable patents or new drugs entering the discovery stage

Dependent Variable:	Cit-weighted patents	Patent market value	Drugs (total)	Drugs (weighted)
Post × GWAS	-0.2738	0.2338	-0.0004	-0.0000
	(0.51521)	(0.14649)	(0.00026)	(0.00008)
...× True Positive	5.2464**	2.6963***	0.0011	0.0009*
	(1.83822)	(0.74601)	(0.00393)	(0.00043)
Gene-Disease FE	YES	YES	YES	YES
Disease-Year FE	YES	YES	YES	YES
Gene-Year FE	YES	YES	YES	YES
N of Observations	137,254,556	137,254,556	137,254,556	137,254,556 7
N of Gene-Diseases	7,223,924	7,223,924	7,223,924	7,223,924
Mean of Dep Var:	7.1358	0.6317	0.0002	0.0001

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the gene-disease-year level. Std. err. clustered two-ways at the disease and gene level. *Post × GWAS*: 0/1 = 1 in all years after a gene-disease pair is reported by its first GWAS. *True Positive*: 0/1 = 1 for GWAS findings that are later replicated by another GWAS about the same disease. *Cit-weighted patents*: count of USPTO patent applications filed in a given year that target a specific gene-disease combination, weighted by the number of patent citations received up to seven years after patent publication. *Patent value*: estimated stock market value (in constant USD) of patents granted to public firms using data from Kogan et al. (2017). *Drugs (total)*: number of molecules on a gene-disease combination entering the discovery stage. *Drugs (weighted)*: number of molecules on a gene-disease combination entering the discovery stage weighted by their scientific novelty (i.e., by the number of times that the same mechanism of action has been used before, following Dranove et al. 2022).

Table 4: Firms with genetic domain knowledge invest proportionally less in gene-disease combinations that receive a GWAS but are able to avoid false positive associations

Dependent Variable:	USPTO patent applications by...			
	...firms with gene knowledge		...firms w/out gene knowledge	
Post × GWAS	0.0381***	0.0138	0.1256***	0.0539***
	(0.01009)	(0.00821)	(0.01429)	(0.01052)
...× True Positive		0.1216**		0.3586***
		(0.03917)		(0.05837)
Gene-Disease FE	YES	YES	YES	YES
Disease-Year FE	YES	YES	YES	YES
Gene-Year FE	YES	YES	YES	YES
N	137,254,556	137,254,556	137,254,556	137,254,556
N of Gene-Diseases	7,223,924	7,223,924	7,223,924	7,223,924
Mean of Dep Var:	0.0376	0.0376	0.0938	0.0938

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the gene-disease-year level. Std. err. clustered two-ways at the disease and gene level. *USPTO patent applications*: count of USPTO patent applications filed in a given year for innovations that target a specific gene-disease combination; the count is then divided between firms with and without previous publications on the gene. *Post × GWAS*: 0/1 = 1 in all years after a gene-disease pair is reported by its first GWAS. *True Positive*: 0/1 = 1 for GWAS findings that are later replicated by another GWAS about the same disease.

Table 5: Firms with genetic domain knowledge avoid investing in GWAS associations that mistakenly report the wrong gene in the original publication

Dependent Variable:	USPTO patent applications by...		
	...all firms	...firms with gene knowledge	...firms w/out gene knowledge
Post × Wrong GWAS Gene	0.0986*	-0.0467	0.1453***
	(0.04945)	(0.03994)	(0.02609)
Gene-Disease FE	YES	YES	YES
Disease-Year FE	YES	YES	YES
Gene-Year FE	YES	YES	YES
N	136,935,850	136,935,850	136,935,850
N of Gene-Diseases	7,207,150	7,207,150	7,207,150
Mean of Dep Var:	0.1304	0.0372	0.0931

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the gene-disease-year level. The sample is slightly smaller because it excludes gene-disease pairs that receive a correctly reported GWAS association. Std. err. clustered two-ways at the disease and gene level. *USPTO patent applications*: count of USPTO patent applications filed in a given year that target a specific gene-disease combination; the count is then divided between firms with and without previous publications on the gene. *Post × Wrong GWAS Gene*: 0/1 = 1 in all years after a gene-disease pair is reported by its first GWAS, but only including *wrongly* reported protein-coding genes (i.e., that are later reclassified by the curators of the GWAS Catalog as being mutations in another gene).

Finding Diamonds in the Rough:
Data-Driven Opportunities and Pharmaceutical Innovation

Appendix

Matteo Tranchero
The Wharton School

A Additional Details on GWAS

Genomics is the branch of biological sciences concerned with the study of genomes, i.e., the entire collection of an organism's genes. Genes are sequences of DNA bases that encode the instructions to synthesize gene products, most notably proteins. The normal functioning of a gene can be altered by a mutation, potentially giving rise to severe health conditions. Knowing the genetic roots of diseases has practical consequences for the design of pharmaceutical drugs (Nelson et al., 2015): leveraging the understanding of gene-disease relationships permits the identification of targets that can be inhibited or activated to produce a desired therapeutic effect. Once the target has been found, scientists can design drugs that bind to the malfunctioning gene (or its products).

A.1 A Scientific Primer

Diseases caused by individual gene mutations are called Mendelian disorders. However, Mendelian diseases are typically severe and, hence, rare because they tend to be eliminated by evolutionary pressures. More common are polygenic diseases (also called complex diseases) that are not due to a single genetic factor but rather by many mutations. For polygenic diseases, any genetic mutation can increase the risk of presenting the condition even without being necessary or sufficient for manifesting the disease. Individual mutations are usually responsible for only a tiny proportion of the heritability of complex diseases. Although complex disorders often cluster in families, they do not have a predictable inheritance pattern. Convoluted interactions between genetic predisposition and environmental factors concur in the etiology of such diseases. Therefore, scientists need to search through all of the $\sim 19,000$ protein-coding genes to find the mutations involved in each of the thousands of polygenic diseases (Tranchero, 2024).

Over the years, researchers have concluded that common disorders are influenced by genetic mutations also common in the population (Reich and Lander, 2001). Instead of looking for individual genes with strong effects on phenotypes, the field has moved toward studying common, generic variants that have a negligible impact on the likelihood of having a disease when taken in isolation (Bush and Moore, 2012). But what precisely is a variant? At the most fundamental level, two genomes differ in a specific genetic locus if they present an alternative single nucleotide (adenine, thymine, cytosine, or guanine) in that location. Such mutation in one DNA basis is called single-nucleotide polymorphism (SNP) when it appears in at least 1% of the population. One approach for associating SNPs with a disease relies on the fact that a causative variant should be found more frequently in cases than in control subjects. In practice, this means looking for statistical correlations between specific genetic variants and diseases in large samples of unrelated people.

Building over this logic, genome-wide association studies (GWAS) are hypothesis-free methods for identifying associations between genetic regions and diseases (Visscher et al., 2017). Using genotyping technologies on large samples, GWAS compare genetic differences between affected and unaffected individuals. In a typical GWAS project, researchers obtain DNA from two groups of participants: patients with the disease studied and healthy individuals with comparable demographics. Then,

selected SNPs on the chromosome are scanned using microarrays that can genotype up to millions of SNPs for each individual. Variants significantly more likely to appear in the affected patients could be biologically relevant to the disease and thus potentially be involved with its etiology. Such SNPs might affect gene expression and function, mainly when located within a protein-coding gene. It is essential to underscore that array-based genome-wide studies do not sequence the DNA base by base since they only determine the presence or absence of a relatively small number of SNPs (usually $< 0.1\%$ of the genome). Nevertheless, exploiting the fact that the co-occurrence of variants in proximal genetic loci is not random (a phenomenon called linkage), researchers can use reference genomes (such as the HapMap) to parsimoniously infer the characteristics of the whole genome from the much smaller number of SNPs genotyped (Bush and Moore, 2012).

GWAS findings entail a substantial risk of false positives (Boyle et al., 2017; MacArthur, 2012). Findings often fail to replicate because of demographic biases in the convenience sample used by the original studies (e.g., white men from North America), or due to mistakes in mapping the variant to the right gene when publishing the result (Vaughan and Srinivasasainagendra, 2013). Even if the association is robust to replications, understanding the biological mechanisms through which it affects human health requires additional study. Moreover, most associations explain a small fraction of the variation in disease susceptibility, which means that the therapeutic benefit from intervening in them could be quite limited (Goldstein, 2009). These limitations notwithstanding, GWAS have proven extremely useful in uncovering drug targets that can assist in identifying compounds suitable for drug repurposing (Reay and Cairns, 2021; Visscher et al., 2017). GWAS also permits the identification of new uses for existing drugs by pointing out new conditions that might be addressed acting on a given target (Andriani and Cattani, 2022; Pushpakom et al., 2019).

A.2 Case Study: The GWAS of Duerr et al. (2006)

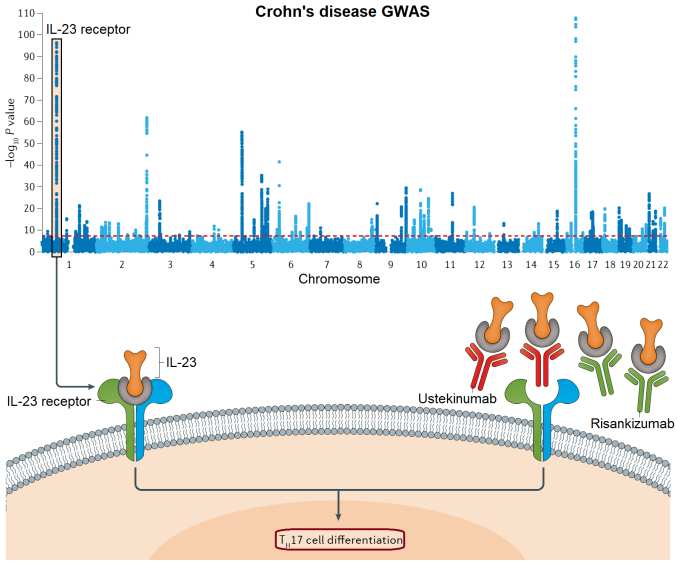
In 2015, an estimated 3 million U.S. adults (around 1.3% of the population) reported being diagnosed with chronic inflammation of the gastrointestinal tract, known as inflammatory bowel disease (IBD). The two most frequent IBD conditions are Crohn's disease and ulcerative colitis (Dahlhamer et al., 2016). Given the prevalence and severity of these diseases, researchers have been intensively studying Crohn's and related diseases. According to the DisGeNET data, IBD was in the top 1% of the diseases with the highest genetic research intensity in the pre-GWAS era (Tranchero, 2024). A few genes, such as NOD2, had been identified as harboring causal mutations by 2005, but without fully explaining the genetic risk for the disease.

In December 2006, Duerr et al. (2006) published one of the very first GWAS. The study involved 567 patients of European ancestry affected by IBD and 571 healthy controls. The GWAS identified genetic mutations in the interleukin 23 receptor gene (IL23R) as significantly associated with Crohn's disease and ulcerative colitis. Before this finding, IL23R was among the least studied human genes. However, many scientists followed up on the promising lead uncovered by the Duerr et al. paper, often to elucidate the causal mechanisms (called "pathways") through which IL23R exerts its effects. It is

now understood that the IL23R gene provides instructions for making a protein called the interleukin 23 receptor, an essential constituent of antibodies that identify foreign substances and defend the body against infections by promoting local inflammations. The IL-23 receptor interacts with a protein called IL-23, binding together like a lock and key (Bianchi and Rogge, 2019). When IL-23 binds to its receptor, it triggers chemical signals to develop and activate Th17 cells, a specific type of lymphocytes that promote inflammation to fight foreign invaders such as viruses. Malfunctioning IL23R genes might misdirect such inflammatory reactions toward human tissues, giving rise to autoinflammatory diseases like Crohn's.

The work of Duerr et al. (2006) uncovered the crucial role that malfunctioning of IL23R has in inducing inflammations in the intestines. Furthermore, the discovery of IL23R's role in IBD suggested that drugs could be engineered to interfere with malfunctioning IL23R genes. This is indeed what ustekinumab, a fully human monoclonal antibody, does by blocking the p40 subunit of IL23 and preventing its binding with the IL23 receptor (Figure A.1). As of 2024, ustekinumab (Stelara, by Janssen Pharmaceuticals) is a drug approved for treating chronic inflammatory diseases in several jurisdictions, including the United States, Europe, and Australia. Several other drugs are being designed to target the IL23 receptor complex, including risankizumab (Skyrizi, by AbbVie), tildrakizumab, and guselkumab. Interestingly, all these molecules have been repositioned as Crohn's disease intervention from their initial indication for psoriasis (Reay and Cairns, 2021). This highlights how uncovering new drug targets through GWAS can enable the rapid repurposing of gene-specific molecules developed for other diseases (Andriani and Cattani, 2022; Kang, 2024), as well as developing new ones.

Figure A.1: GWAS findings on Crohn's disease provided new opportunities for drug development



Note: The figure exemplifies how GWAS can guide drug development by finding new diseases that can be addressed with existing molecules. Once Duerr et al. (2006) located a new variant in IL23R that disrupts its function, monoclonal antibodies that target the IL23 receptor pathway were repositioned as Crohn's disease interventions from their original indication for psoriasis. The image is an edited version of the original figure from Reay and Cairns (2021).

B Measurement Approach

This appendix describes and formalizes the landscape-based measurement approach adopted in the paper to bypass patent-to-paper citations. Two case studies are presented to showcase its advantages relative to bibliometrics.

B.1 Shortcomings of Patent-to-Paper Citations

Traditionally, measuring the technological impact of a scientific project involves counting the patent citations received by the publications where its output is codified (Arora et al., 2021; Azoulay et al., 2019; Fleming et al., 2019). This approach has been validated by methodological studies showing that patent-to-paper citations are a good way to proxy knowledge flows from science to technology (Duguet and MacGarvie, 2005; Narin et al., 1997; Roach and Cohen, 2013). Therefore, it is not surprising that most existing empirical studies use non-patent literature citation counts to measure the applied impact of science. The recent diffusion of open-access databases of patent-to-paper citations has further increased the appeal of this approach (Bryan et al., 2020; Marx and Fuegi, 2020). Nonetheless, this measurement approach suffers from two shortcomings.

First, relying on explicit references to science usually provides a downward biased measure of impact (Myers and Lanahan, 2022). Direct citations, for instance, would not capture foundational intellectual influences that become common knowledge in a field, nor knowledge flowing through more complicated citation patterns (e.g., a patent citing a publication that cites the focal paper of interest). These examples fall under what Roach and Cohen (2013) define as "errors of omission." Such underestimate is possibly magnified for basic research with spillovers in fields very different from where the idea originated and through unpredictable channels (Azoulay et al., 2019; Cohen et al., 2002).

Second, and similarly to academic papers, non-patent literature citations can be made for various reasons (Teplitskiy et al., 2022). Up to half of the citations are possibly devoted to irrelevant prior art (Jaffe et al., 2000), likely for strategic reasons (Kuhn et al., 2020; Lampe, 2012). This practice is what Roach and Cohen (2013) define as "errors of commission," namely citations not corresponding to knowledge relevant to the invention. But even if the reference captures an actual knowledge flow, it is hard to know what part of the study the patent builds on. Academic papers often make multiple contributions; the citation might refer to any of those. Assessing the value of individual contributions made in a given paper is impossible from sheer citation counts.

B.2 Using Entitymetrics to Measure Scientific and Technological Impact

In this paper, I formalize a new measurement approach based on the notion of knowledge entity (Ding et al., 2013). I consider any individual carrier of knowledge as an entity, be it embodied into an artifact (e.g., a piece of computer hardware) or a more abstract unit defined by domain-relevant taxonomies (e.g., biological entities, such as genes and diseases). Each publication or patent can be

characterized by the knowledge entities it studies and recombines (Fleming, 2001). By taking this perspective, one can summarize written documents by compiling a list of their knowledge entities. I propose to assess the technological impact of a paper by quantifying the change in activity experienced by the knowledge entities "treated" by it. The basic idea is that impactful projects generate interest and innovation opportunities involving the entities recombined. The increase in patents including such entity combinations relative to similar control combinations is a measure of impact that has the advantage of also capturing patents that do not directly cite the focal paper (Nagaraj and Tranchero, 2024).[1]

More in detail, my approach combines machine learning with causal inference in three steps. First, one needs to extract knowledge entities from the relevant documents – in the case of science-to-technology linkages, papers and patents. This can be done using automated machine learning procedures, such as state-of-the-art Bidirectional Encoder Representations from Transformers (BERT) or Large Language Models (LLMs).[2] BERT can be tailored to specific domains, such as the life sciences, to increase accuracy and contextual stability (Xu et al., 2020). After extracting accurate knowledge entities using ML algorithms, one can normalize and assign a unique ID to each. The result is that text documents will be characterized by a vector of knowledge entities and their combinations.

Second, one can use the union of all knowledge entities extracted from the relevant document corpus to trace an empirical knowledge landscape (Nagaraj and Tranchero, 2024). The combinatorial landscape constitutes the "ground truth," that is, the space of entities over which researchers and firms carry out their search activities. In some instances, this landscape is known ex ante (e.g., the $\sim 19,000$ genes constituting the human genome), while in other cases, it can be backed out from the entity extraction task (e.g., the landscape of research topics described by Boyack et al. 2020). The advantage of studying research as happening on a landscape is that citations are not needed to track knowledge evolution (Ding et al., 2013). Instead, one can measure the change in follow-on work relating to the entities themselves before and after the project of interest is completed.

Finally, one can quantify the impact of a research project using a difference-in-differences framework with staggered adoption at the level of each entity combination. Said otherwise, combination-level regressions allow assessing the attention change the entity combination $< i, j >$ received after being treated by the article of interest. The basic specification is at the level of individual combination $< i, j >$ and time t and takes the form of the following equation:

$$Y_{i,j,t} = \alpha + \beta Post_t \times Article_{i,j} + \gamma_{i,j} + \delta_{i,t} + \omega_{j,t} + \epsilon_{i,j,t} \tag{B.1}$$

[1]Similarly to my approach, Iaria et al. (2018) trace the diffusion of frontier knowledge measuring the occurrence of new scientific concepts in patent texts. Suh (2024) extracts the chemical compounds mentioned in the body of patent texts to recognize innovations that rely more on technologies where the Soviet Union had a scientific lead. More broadly, Kang (2024), Kao (2023) and Nagaraj (2022) apply the notion of a search landscape when studying pharmaceutical innovation and gold discoveries, respectively.

[2]Advances in natural language processing have been used to capture text similarity between documents (Arts et al., 2018; Kaplan and Vakili, 2015; Younge and Kuhn, 2019). However, these methods have been used to measure the similarity between patent specifications and not to extract knowledge entities.

Where $Y_{i,j,t}$ is any technological impact metric of interest about the treated entities (e.g., patents, clinical trials, or other follow-on outcomes), $Article_{i,j}$ equals one for the combinations in the article and zero for the control ones, $Post_t$ is a binary variable that takes value one only after the project of interest is completed, $\gamma_{i,j}$ are combination fixed effects, while $\delta_{i,t}$ and $\omega_{j,t}$ are entity-specific time trends. Combination fixed effects consider the potential of different combinations, while the entity time trends avoid confounding from heterogeneous changes in attention to individual entities.

The coefficient of interest, β, captures the marginal impact of the project completion on the treated entities. Under the assumption of parallel trends in inventive activities between treated and non-treated entities, this procedure estimates project impact over time that does not depend on patent-to-paper citations. In particular, it solves the measurement issues highlighted in the previous section. First, by capturing the mention of entities directly in the text description of the technological application, this approach is well suited to objectively measure the impact of basic research findings that might not be cited in downstream applications. Second, tracking the reuse of entity combinations permits distinguishing different contributions made in the same article, offering a granularity that citations do not have.

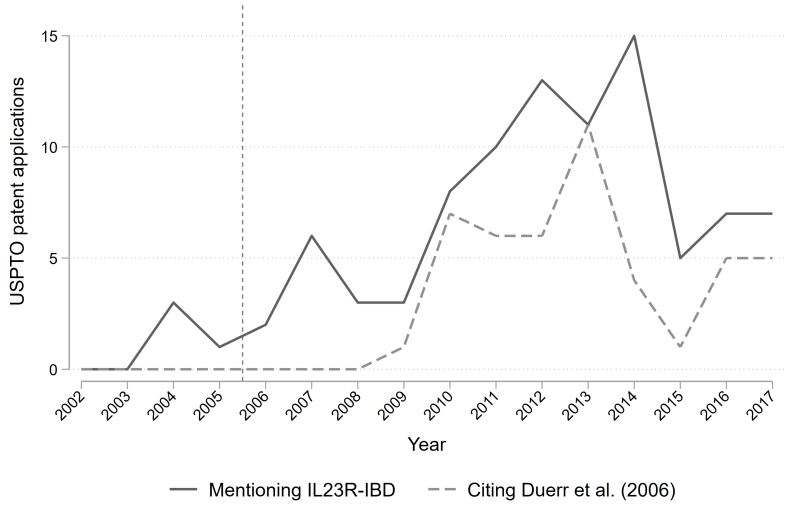
B.3 An Application of Entitymetrics in Pharmaceutical Innovation

The bio-medical sector offers an ideal setting to demonstrate the advantages of entitymetrics. Pharmaceutical innovation heavily depends on science (Cohen et al., 2002), and knowledge entities are well-defined by taxonomies and have clear meanings (e.g., genes and diseases). I leverage European Bioinformatics Institute (EBI) data to measure the gene-disease pairs targeted by each USPTO patent application (2001-2019). The entities from patent texts are extracted using TERMite, the proprietary named entity recognition software developed by SciBite.

Example 1: Undermeasurement of Basic Research Influences: The GWAS of Duerr et al. (2006) was the first to implicate the IL23R gene in the etiology of inflammatory bowel disease (IBD). This finding has proved incredibly impactful both on scientific research and pharmaceutical innovation. Duerr et al. (2006) received over 3,800 citations on Google Scholar (as of 2024) and led to an improved theoretical understanding of how IL23R is involved in IBD (Bianchi and Rogge, 2019). On the therapeutic front, several drug molecules are now available to treat IBD by modulating the IL-23 signaling pathways, including ustekinumab and risankizumab (Reay and Cairns, 2021). Therefore, this offers a perfect case study to compare the ability of patent citations and entitymetrics to capture the considerable technological impact of this paper.

Figure B.1 shows the time series of USPTO patent applications targeting the IL23R-IBD pathway or citing the GWAS by Duerr et al. (2006), respectively. A few things are interesting to note. First, some firms were already exploring the IL23R-IBD nexus before the finding became known in the broader scientific community (Brusoni et al., 2001). Those firms are then the quickest to react, applying for patents that were probably already in the making and thus not citing Duerr et al. (2006), but that found a crucial validation in it (Kao, 2023). Second, it appears that just looking at patent-to-paper citations

Figure B.1: Count of USPTO patent applications that directly cite Duerr et al. (2006) or that target the IL23R-IBD combinations, regardless of whether they cite Duerr et al. (2006)



Note: Data on patent citations for USPTO patent applications are from Google Patents. The vertical line marks the publication of Duerr et al. (2006).

would lead to a substantial underestimate of impact. Only 46 patent applications cite Duerr et al. (2006), compared with the 88 patent applications that exploit its finding. The gap between the two measurements reduces to half among patents filed by universities and research institutions, suggesting that most undercount stems from firms' citation practices.

Direct patent-to-paper citations cannot account for citations to downstream science enabled by Duerr et al. (2006). For instance, patents might cite subsequent studies that experimentally validated the IL23R-IBD association (Ahmadpoor and Jones, 2017). Figure B.2 shows a few claims of the US 2016/0333091 A1 patent application by Boehringer Ingelheim. The patent builds on knowledge of IL23R's role in IBD without acknowledging the GWAS by Duerr and colleagues, instead citing

Figure B.2: Claims 28, 29, and 30 of patent application US 2016/0333091 A1

28) A method for treating an inflammatory disease, an autoimmune disease, a respiratory disease, a metabolic disorder or cancer comprising administering to a subject in need thereof an effective amount of an anti-IL-23p19 antibody or antigen-binding fragment or a pharmaceutical composition comprising an anti-IL-23p19 antibody or antigen-binding fragment and a pharmaceutically acceptable carrier,

wherein the antibody or antigen-binding fragment thereof comprises:

a) a light chain variable region comprising the amino acid sequence of SEQ ID NO:19 (CDR1-L); the amino acid sequence of SEQ ID NO:20 (CDR2-L); and the amino acid sequence of SEQ ID NO:21 (CDR3-L); and

b) a heavy chain variable region comprising the amino acid sequence of SEQ ID NO: 63, 66, 67 or 68 (CDR1-H); the amino acid sequence of SEQ ID NO:64 (CDR2-H); and the amino acid sequence of SEQ ID NO:65 (CDR3-H).

29) The method according to claim **28**, wherein the disease is psoriasis, inflammatory bowel disease, psoriatic arthritis, multiple sclerosis, rheumatoid arthritis, or ankylosing spondylitis.

30) A method for inhibiting the binding of IL-23 to the IL-23 receptor on a mammalian cell, comprising administering to the cell an anti-IL-23p19 antibody or antigen-binding fragment, whereby signaling mediated by the IL-23 receptor is inhibited,

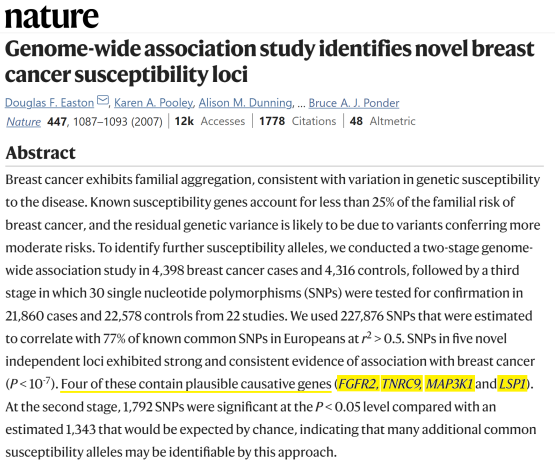
Note: This figure shows selected claims of the patent application titled "Anti-IL-23 Antibodies" filed by Boehringer Ingelheim. The patent does not cite the study of Duerr et al. (2006) even if it builds on its finding; however, it does cite papers that cite Duerr et al. (2006) in turn.

subsequent papers that explain the mechanisms behind the IL23R-IBD correlation (Beyer et al., 2008). In sum, the entity-based approach seems better equipped to capture the impact of fundamental advances triggering extensive follow-on research.

Example 2: Multiple Findings in the Same Paper: Scientific articles often make more than one contribution. For instance, the GWAS of Easton et al. (2007) reported four new genes as correlated with breast cancer (Panel A of Figure B.3). What was the individual impact of these four gene-disease combinations on pharmaceutical innovation? Simply counting how many patents cite this paper would not answer the question. Aveo Pharmaceuticals' patent US 2011/0305687 A1, titled "Anti-FGFR2 antibodies," is a good case in point. This patent cites Easton et al. (2007), and from the title alone, it is clear that it builds on only one of its four findings. Unfortunately, this information is impossible to gather from citation patterns alone.

Figure B.3: Knowledge entities permit to measure the impact of multiple discoveries in the same paper

(a) *Abstract of the GWAS by Easton et al. (2007)*



nature

Genome-wide association study identifies novel breast cancer susceptibility loci

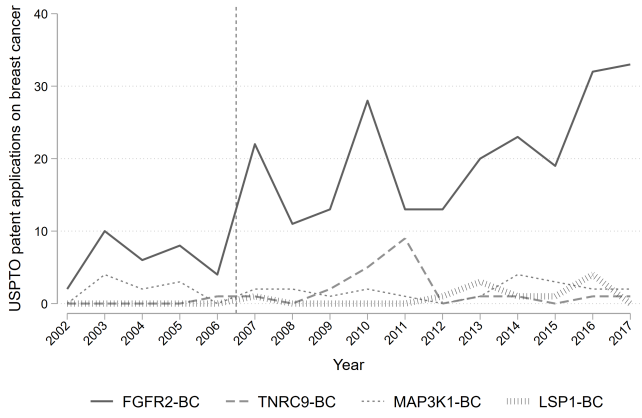
Douglas F. Easton, Karen A. Pooley, Alison M. Dunning, ... Bruce A. J. Ponder

Nature **447**, 1087–1093 (2007) | **12k** Accesses | **1778** Citations | **48** Altmetric

Abstract

Breast cancer exhibits familial aggregation, consistent with variation in genetic susceptibility to the disease. Known susceptibility genes account for less than 25% of the familial risk of breast cancer, and the residual genetic variance is likely to be due to variants conferring more moderate risks. To identify further susceptibility alleles, we conducted a two-stage genome-wide association study in 4,398 breast cancer cases and 4,316 controls, followed by a third stage in which 30 single nucleotide polymorphisms (SNPs) were tested for confirmation in 21,860 cases and 22,578 controls from 22 studies. We used 227,876 SNPs that were estimated to correlate with 77% of known common SNPs in Europeans at $r^2 > 0.5$. SNPs in five novel independent loci exhibited strong and consistent evidence of association with breast cancer ($P < 10^{-7}$). Four of these contain plausible causative genes (*FGFR2*, *TNRC9*, *MAP3K1* and *LSP1*). At the second stage, 1,792 SNPs were significant at the $P < 0.05$ level compared with an estimated 1,343 that would be expected by chance, indicating that many additional common susceptibility alleles may be identifiable by this approach.

(b) *USPTO patent applications after Easton et al. (2007)*



Note: The top panel shows the abstract of the GWAS by Easton et al. (2007). The bottom panel shows USPTO patent applications that target the four findings of the GWAS by Easton et al. (2007). The vertical line marks the publication of Easton et al. (2007), showing substantial heterogeneity in the impact of individual findings.

My measurement strategy can bypass this limitation by extracting each recombination introduced by a given article and then relating it to patenting dynamics on the recombination itself. Figure B.3 shows that the GWAS by Easton et al. (2007) triggered substantial patent applications only on the FGFR2-breast cancer combination. Patenting on this gene-disease combination tripled, with almost no impact on the other combinations. Relying on patent-to-paper citations would completely obscure the heterogeneous effect of the GWAS by Easton et al. (2007). This aspect is all the more important since some of these associations might be false positives. Text-extracted real-world entities are especially suited to measure the impact of basic research with cumulative impact across disparate domains.

C SciBite/EBI Patent Data

This appendix describes the data on the genes and diseases listed in USPTO patent applications (2001-2019).

C.1 Sample and Validation

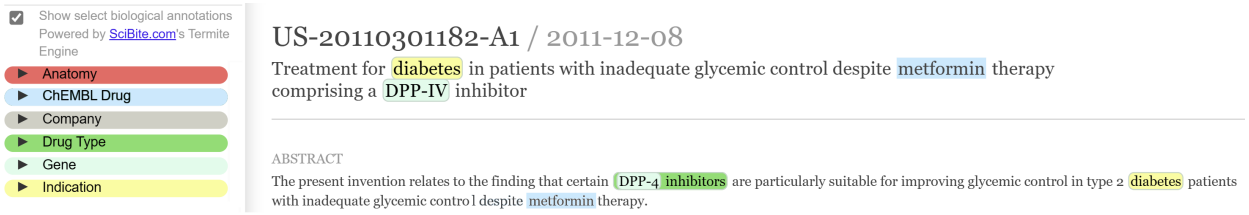
Information on R&D expenditure is usually only available at the organizational level and not for specific projects. In this paper, I follow the approach of Eggers and Kaplan (2009) and use patent records to infer where firms are directing their innovation investments. I leverage information about the genes and diseases mentioned in firms' patents to learn the projects to which firms devote resources. Given the length and uncertainty of the pharmaceutical innovation process, patent applications are better thought of as a byproduct of early-stage investments rather than successful innovations. When a firm starts patenting in a given domain, it is a good indicator that it is investing in that area. However, it is important to note that these are not patents on the gene sequence itself (Williams, 2013), which have been ruled inadmissible by the U.S. Supreme Court with the Myriad ruling. Instead, my sample considers patents for innovations like genetic tests, new drug molecules, or method-of-use of molecules that *target* a specific gene to *treat* a specific disease.

The primary source for the data used in the paper is a proprietary database from the European Bioinformatics Institute (EBI). The data have been compiled using TERMite (TERM identification, tagging & extraction), a named entity recognition software developed by the Elsevier-owned startup SciBite. TERMite scans and semantically annotates raw text with entities from over 50 biopharma and biomedical topics. The entities are drawn from VOCabs, a manually curated vocabulary with over 20 million synonyms specifically tuned for named entity recognition text analytics. For instance, this permits recognizing that SEPT1 is the symbol for the SEPTIN1 gene, not a date. Importantly, TERMite has built-in relevance detection, distinguishing between terms that are casual mentions and those that constitute the critical bio-entities of a document.

The data include all the protein-coding genes and diseases extracted from complete patent texts. Genes are matched to their unique NCBI IDs, while diseases are mapped into MeSH Unique IDs. Figure C.1 shows how the TERMite software works. The figure shows the USPTO patent application US 2011/0301182 A1, a patent granted on September 30, 2014, to Boehringer Ingelheim. This patent is listed in the FDA's *Approved Drug Products with Therapeutic Equivalence Evaluations* (also known as Orange Book) as the intellectual property behind Tradjenta, the brand name for Linagliptin. Tradjenta is a medication used to treat type 2 diabetes by acting as a DPP-4 inhibitor, i.e., by altering the function of the gene DPP4, which plays a significant role in glucose metabolism. Figure C.1 shows how the entity-based approach correctly captures that this patent describes a drug targeting the DPP4-diabetes combination.

I carry out a manual validation to assess the performance of TERMite in terms of recall and precision, following the procedure of Marx and Fuegi (2020). Recall measures how many actual bio-entities of

Figure C.1: Example of SciBite's entity recognition algorithm used to extract genes and diseases targeted by each USPTO patent application



Note: This figure shows an example of how SciBite's entity recognition algorithm (called TERMite) extracts genes and diseases targeted by the drug disclosed in USPTO patent application US 2011/0301182 A1.

a patent the algorithm found. This is equivalent to one minus the percentage of false negatives, i.e., entities mentioned by a patent that the software failed to find. To assess this metric, I sampled 100 patent applications, obtained their full specification from Google Patents, and randomly extracted one gene and one disease they mentioned. Then, I assessed whether the same entities were listed in my data for that specific patent, finding that 91% of genes and 92% of diseases were correctly captured. Symmetrically, precision is given by the share of reported bio-entities that are correct. This metric is computed as one minus the percentage of false positives, i.e., entities mistakenly extracted from the patent. I evaluated precision by extracting one gene and one disease for 100 patents in my data and then manually checking whether the entity in question was present in the patent specification. Overall, 95% of genes and 97% of diseases were true positives. Taken together, the F_1 score is equal to 92.96 in the case of genes and 94.43 in the case of diseases, proving the high reliability of my data.

C.2 Descriptive Statistics

My sample includes 148,232 USPTO patent applications published from 2001 to 2019 inclusive. Of these, 73,255 are eventually granted as of summer 2021. All my primary analyses rely on the entire sample of applications, but I also present descriptive statistics for the subset of granted patents for comparison. Table C.1 presents the main descriptive statistics. Each patent application mentions, on average, 6.3 genes and 12.3 diseases; this number is only slightly smaller for granted patents.[3] However, the sample shows a large variance, with a few patents listing hundreds of genes and diseases as targets. This probably reflects strategic disclosure behaviors and offers an exciting avenue for future research on the conditions under which patent text can or cannot be relied upon to gauge a patent's technological content. Finally, the average patent covers 188 gene-disease pairs, primarily due to a few outliers since the median patent focuses on a much smaller set of 13 gene-disease combinations.

The average number of diseases each patent mentions is roughly constant over my sample period (Figure C.2), while there seems to be a slightly upward trend in the number of genes appearing in the patent text. Patents published in the year 2001 seem to reference an abnormally low number of

[3]This is consistent with evidence showing that patent examiners tend to restrict the scope of patent applications during the granting process (Marco et al., 2019).

Table C.1: Descriptive statistics at the patent level.

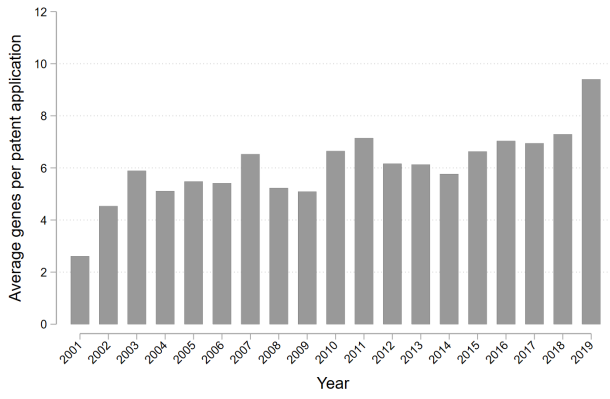
	USPTO patent applications						USPTO granted patents					
	mean	median	st d	min	max	N	mean	median	st d	min	max	N
Genes per patent	6.285	2	31.1786	1	4060	148,232	5.625	2	31.3932	1	4,060	73,255
Diseases per patent	12.257	5	22.5936	1	911	148,232	12.095	5	20.4855	1	853	73,255
Gene-disease pairs per patent	188.416	13	3,671.136	1	318,560	148,232	165.865	12	3645.6	1	289,212	73,255
Year of patent publication	2010.704	2011	5.0519	2001	2019	148,232	2011.062	2011	5.0451	2001	2019	73,255

Note: This table presents descriptive statistics at the level of individual patents in my data. The left part of the table presents descriptives considering all patent applications, while the right panel presents descriptives considering only applications that eventually result in a granted patent as of 2021. *Genes per patent*: count of genes mentioned by the patent; *Diseases per patent*: count of diseases mentioned by the patent; *Gene-disease pairs per patent*: count of gene-disease pairs mentioned by the patent; *Year of patent publication*: year when the patent specification was published.

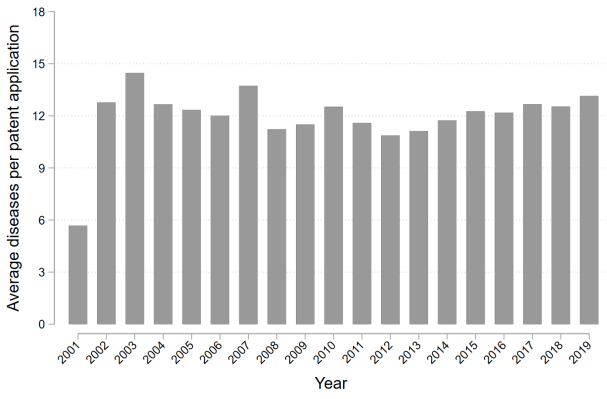
bio-entities, possibly because of idiosyncrasies of that specific year (which is the first when patent applications started to be published and it corresponded to the release of the first draft of the human genome). Together, the two graphs help rule out structural breaks in disclosure practices that could potentially bias my measurement approach.

Figure C.2: The average number of genes and diseases mentioned by each USPTO patent application is relatively constant over time

(a) *Average number of genes referenced per patent*



(b) *Average number of diseases referenced per patent*



Note: Panel (a) shows the average number of genes mentioned by each USPTO patent application per year. Panel (b) shows the average number of diseases (MeSH Unique IDs at level 4) mentioned by each USPTO patent application per year.

Table C.2 shows the descriptive statistics at the firm level. On average, firms' patent portfolio encompass around 64 genes and 41 diseases, but there is significant variation. Some firms span up to thousands of genes in their R&D efforts, while the median firm explored only 19 genes. The dispersion in the number of diseases is generally smaller. Interestingly, the table shows that focusing only on granted patents would lead to missing much of the firms' exploration in genetic space. This validates the choice of looking at patent applications to capture the earliest stages of pharmaceutical innovation and suggests that firms' competencies are much more comprehensive than what traditional research can capture (Brusoni et al., 2001). On average, firms are active on 1,437 gene-disease pairs, but the median firm explored only 162 pairs.

Table C.2: Descriptive statistics at the firm level.

	USPTO patent applications						USPTO granted patents					
	mean	median	st d	min	max	N	mean	median	st d	min	max	N
Genes per firm	64.285	19	195.922	1	5,467	4,117	37.637	11	130.255	1	4,579	3,866
Diseases per firm	41.362	27	46.173	1	338	4,117	31.530	17	40.546	1	305	3,866
Gene-disease pairs per firm	1,436.791	162	8,680.241	1	239,160	4,117	842.248	76	5,821.625	1	166,437	3,866

Note: This table presents descriptive statistics at the level of individual firms. The left part of the table presents descriptives considering all patent applications, while the right panel presents descriptives considering only applications that eventually result in a granted patent as of 2021. *Genes per firm*: count of genes mentioned by a firm in its patents; *Diseases per firm*: count of diseases mentioned by a firm in its patents; *Gene-disease pairs per firm*: count of gene-disease pairs mentioned by a firm in its patents.

Finally, I provide some descriptive evidence of how the bio-entities recombined by a patent relate to common empirical proxies of its value. Simple OLS regressions reveal that the number of diseases is associated with patents of higher economic value (Table C.3). Innovations targeting more diseases have higher market value and larger patent family sizes, indicating that the number of potential applications for a drug is a predictor of its economic value. Instead, the number of genetic targets of a patent seems to be associated with higher technological impact, as proxied by the number of forward patent citations received. These patents are also more likely to end up in litigation, confirming the intuition that they might cover a larger swath of the technological space and block other applications.

Table C.3: USPTO patent applications for innovations targeting multiple diseases have a higher market value, while USPTO patent applications for innovations targeting multiple genes have a higher technological impact.

Dependent Variable:	Patent family size		Market value		Patent citations		Litigated patent (0/1)	
Genes per patent	0.0004		-0.0274		0.04659***		0.0004*	
	(0.00222)		(0.03072)		(0.01253)		(0.00017)	
Diseases per patent		0.0329**		0.1312*		0.0076		0.0002
		(0.01043)		(0.06300)		(0.02196)		(0.00026)
Year of application FE	YES	YES	YES	YES	YES	YES	YES	YES
N	148,226	148,226	30,019	30,019	148,226	148,226	148,226	148,226
Mean of Dep Var:	10.7969	10.7969	28.8594	28.8594	21.0813	21.0813	0.4366	0.4366

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the patent application level. Std. err. clustered at the assignee level. *Genes per patent*: count of genes mentioned by a patent application; *Diseases per patent*: count of diseases mentioned by a patent application; *Patent family size*: number of patent applications in the same patent family; *Market value*: estimate of the market value (in constant USD) of the patent using data from Kogan et al. (2017). Note that this measure is available only for applications that are eventually granted, hence the smaller sample size; *Patent citations*: number of forward patent citations received by the USPTO patent application up to seven years after its publication; *Litigated patent (0/1)*: 0/1 = 1 if the USPTO patent application is involved in litigation.

C.3 Example: Denali Therapeutics

In 2015, three top researchers left Genentech to start a new company: Denali Therapeutics. Aptly named after the tallest mountain in North America, Denali focuses on treating and curing neurodegenerative diseases like Alzheimer's, amyotrophic lateral sclerosis, and Parkinson's. Based in San Francisco, the company has raised over $350 million in venture capital and $250 million from its IPO in 2017. On its website, Denali lists ten compounds at different stages of clinical development as

of summer 2022. Some of them are being developed together with large pharmaceutical companies, including Biogen, Sanofi, and Takeda Pharmaceuticals.

Advances in the genetics, pathology, and cell biology underlying chronic neurodegenerative disease have identified new pathways that trigger neurodegeneration. In particular, researchers have dubbed "degenogenes" a set of genes that, when mutated, have a likely causative role in neurodegenerative disease. Denali Therapeutics was founded based on the idea that such degenogenes could constitute a viable therapeutic avenue to tackle the most common neurodegenerative disorders. The focus of drug discovery activities on a handful of genetic targets constitutes not only the scientific foundation of this company but also its key competitive hypothesis for drug development.

In its IPO filings, Denali explicitly listed the key genetic targets that the company decided to focus on.[4] This offers an opportunity to test the reliability of the SciBite data, that include ten patent applications by Denali Therapeutics as of 2019. Two patterns stand out. First, nine out of ten patents are indeed tagged with Alzheimer's and Parkinson's diseases, showing that the data accurately capture the markets targeted by Denali.[5] Second, 86.3% of the gene-disease combinations mentioned in its patents include the two key genes listed in the SEC files: RIPK1 and LRRK2. The first is a gene with an essential role in driving cell death and inflammation, and Denali was the first company to establish the safety of inhibiting RIPK1 kinase activity in humans with a Phase 1 clinical trial (Mifflin et al., 2020); the latter is a gene whose mutations increase the risk of developing Parkinson's disease, and Denali is leading the way in showing how it can be used for drug targeting (Kingwell, 2022). This example shows the potential of using bio-entities to capture a company's technological portfolio.

[4]The list of genetic pathways and genes targets is at page 3 of the following link: https://www.sec.gov/Archives/Denali

[5]The tenth patent generically addresses lysosomal storage disorders. However, this is also consistent with the IPO filings of Denali: the lysosomal system is associated with several neurodegenerative diseases, including Parkinson's.

D Other Data Sources

The present appendix provides additional details and descriptive statistics about the other data used in this study.

D.1 The NHGRI-EBI GWAS Catalog

Information about genome-wide association studies is from the GWAS Catalog, a publicly available list of all published GWAS and association results (MacArthur et al., 2017). The Catalog includes all eligible GWAS studies since the first published in 2005, with details about associations with a high statistical significance (p-value $< 1.0 \times 10^{-5}$) (Marigorta et al., 2018). Compiling this resource requires manual curation of a large body of diverse and unstructured data from the literature, a task carried out by scientists at the European Bioinformatics Institute (EBI) with the support of the National Human Genome Research Institute Home (NHGRI). Catalog data are routinely used by biologists, bioinformaticians, and researchers aiming to translate scientific findings to medical applications and establish targets for novel therapies.

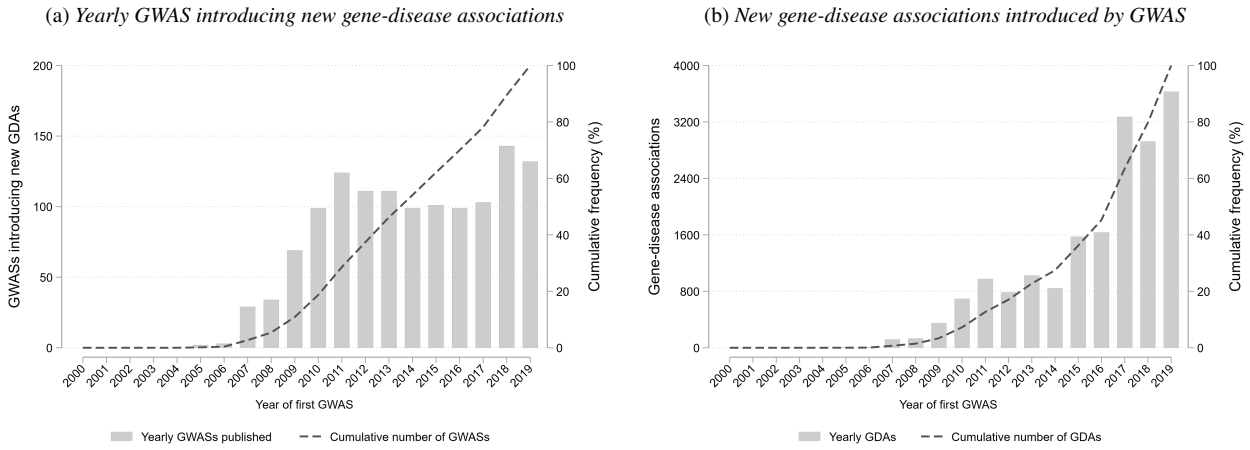
Table D.1: Descriptive statistics of GWAS papers that introduce a gene-disease association involving protein-coding genes

	mean	median	st d	min	max	N
Genes	7.848	3	25.849	1	522	1,259
Diseases	2.053	2	1.335	1	12	1,259
New gene-disease associations	14.269	4	36.276	1	522	1,259
Sample size	104,582.80	11,348	224,424.40	41	1,474,097	1,259
Replication sample (0/1)	0.639	1	0.48	0	1	1,259
Top journal (0/1)	0.314	0	0.464	0	1	1,259
High status PI (0/1)	0.261	0	0.439	0	1	1,259
Co-author industry (0/1)	0.218	0	0.413	0	1	1,259
Scientific citations	138.258	51	311.392	0	6,244	1,259
Cited by clinical trials (0/1)	0.57	1	0.495	0	1	1,259
Patent citations	1.221	0	7.459	0	226	1,259
Year	2015.823	2017	2.979	2005	2019	1,259

Note: This table presents descriptive statistics of GWAS that introduce new gene-disease associations involving protein-coding genes. *Genes*: number of protein-coding genes associated with a disease in the GWAS; *Diseases*: number of diseases studied in the GWAS; *New gene-disease associations*: number of new gene-disease associations introduced by the GWAS; *Sample size*: total number of subjects involved in the GWAS; *Replication sample (0/1)*: 0/1 = 1 for associations reported in GWAS that include also a replication analysis of their result; *Top journal (0/1)*: 0/1 = 1 for associations published in the 15 most prestigious genetics journals or the top 3 generalits scientific journals (*Science*, *Nature*, *PNAS*). *High status PI (0/1)*: 0/1 = 1 for GWAS whose last author is affiliated with one of the 20 most prestigious universities according to the QS World University Rankings for the biological sciences; *Co-author industry (0/1)*: 0/1 = 1 for GWAS with at least one industry co-author; *Scientific citations*: count of scientific citations received by the GWAS (data from NIH ICite); *Cited by clinical trials (0/1)*: 0/1 = 1 if the GWAS has received at least one citation from a clinical trial (data from NIH ICite); *Patent citations*: count of USPTO patent citations received by the GWAS (data for granted patents from Marx and Fuegi 2020); *Year*: year of publication of the GWAS.

Each entry of the GWAS Catalog includes details about the PubMed ID of the paper and the list of associated genes and diseases. Genes are identified by their NCBI IDs, while diseases are mapped into the Experimental Factor Ontology (EFO). I use the crosswalk available on the EFO website to map each disease into the corresponding MeSH Unique IDs. Note that the National Library of Medicine's MeSH semantic keyword tree is a hierarchical tree with 13 levels of increasing specificity. For this study, I map each GWAS to the fourth level of the MeSH tree. If a more specific disease was matched

Figure D.1: Publication of GWAS and introduction of new gene-disease associations by year, 2000-2019



Note: Panel (a) shows the number of GWAS introducing at least one novel gene-disease association that involves a protein-coding gene, both by year and cumulatively over the sample period (2000-2019). Panel (b) shows the number of new gene-disease associations involving a protein-coding gene, both by year and cumulatively over the sample period (2000-2019).

(i.e., at level 5 or above), I assigned it to its parent branches up to level 4. Vice versa, if the disease matched was coarser (i.e., at level 3 or below), I assigned the finding as about all its descending level 4 branches. This procedure permits harmonizing GWAS as uncovering gene-disease associations in a landscape of NCBI IDs (unique genes) and level-4 MeSH IDs (unique diseases).

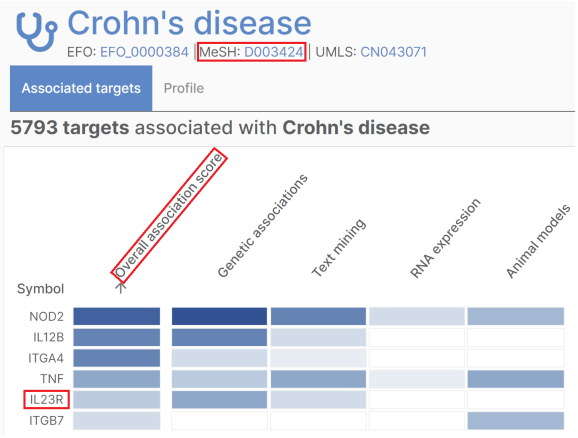
Table D.1 presents the descriptive statistics of the 1,259 GWAS papers that introduce new gene-disease associations (i.e., what constitutes the "treatment" in the primary analysis of the paper).[6] The average GWAS targets two diseases and uncovers 14 associations. However, considerable variability exists, with a few GWAS finding associations with up to 522 genes. Around 64% of the GWAS include a replication sample within the same paper, which is considered a best practice to reduce the risk of spurious results. Around a quarter of studies are carried out by principal investigators affiliated with high-status institutions, and little over a fifth includes an industry co-author. The average GWAS receives 138.3 scientific citations, 1.2 patent citations and has a 57% chance of being cited as the scientific background for a clinical study. Figure D.1 shows the time series of GWAS publications and the arrival of new gene-disease associations. While the number of studies stabilized around 2010, the number of associations reported keeps growing, possibly due to increased sample sizes that allow for more statistically powered analyses (Goldstein, 2009).

[6]The GWAS Catalog contains information on many more GWAS. This paper focuses on GWAS that introduce associations new to the world, but I use information from subsequent GWAS to code the replicability of these findings.

D.2 Open Targets Score

The Open Targets Platform is a public-private partnership that aims to provide evidence for identifying and prioritizing drug targets, one of the fundamental challenges in developing new medicines (Ochoa et al., 2023). Open Targets collects all the available evidence on the strength of gene-disease associations and summarizes it in an open-source synthetic score (Figure D.2 shows an example). To contextualize the relative importance of different pieces of evidence, Open Targets weights documentation on a gene-disease pair according to a scoring framework for each data source. All evidence is then mapped to the genetic target (NCBI Gene identifiers, which I merge to NCBI IDs) and disease (Unique MeSH IDs, which I harmonize to level-4 MeSH IDs). The Open Targets team also makes sure to minimize the presence of duplicates within the same data source. For this paper, I focus only on the Open Targets score aggregating direct evidence on the gene-disease relationship.[7]

Figure D.2: Example of Open Targets Platform's synthetic score for the genes associated with Crohn's disease



Note: This figure shows an example of how the Open Targets Platform the available evidence on the strength of genetic associations with Crohn's disease and summarizes them in an open-source synthetic score.

The Open Targets score is available for 594,353 gene-disease pairs (8% of my sample), spanning 17,437 genes and 366 diseases. However, one must take into consideration some potential limitations. Open Targets scores partially reflect data availability about a given gene-disease pair. This means that under-studied genes or diseases are unlikely to produce high-scoring associations simply due to the lack of available evidence. Vice versa, not all pairs with available evidence can be considered legitimate genetic targets. These limitations explain why I use the replicability of a GWAS finding to find false positives and not the value of the Open Target Score, which could confound the interest in the association with its actual underlying therapeutic value. These limitations notwithstanding, the Open Targets score provides a valuable measure to rank the relative strength of genetic targets within a given disease.

[7]The Open Targets Platform also provides scores that consider indirect evidence using the properties of the EFO disease ontology. However, this evidence is considered less robust (Ochoa et al., 2023).

Table D.2: Patent applications for innovations targeting gene-disease pairs with higher Open Target scores are of higher technological and economic value

Dependent Variable:	Patent citations		Patent family size		Market value	
Max(OT score)	4.6822	7.7646*	3.8489***	1.6012***	14.7768**	4.0955*
	(3.4223)	(3.10934)	(0.36592)	(0.31473)	(4.7530)	(2.0507)
Patent year FE	YES	YES	YES	YES	YES	YES
Firm FE	NO	YES	NO	YES	NO	YES
N	137,614	137,614	137,614	137,614	27,213	27,213
Mean of Dep Var:	26.063	26.063	11.081	11.081	30.365	30.365

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the patent application level. Std. err. clustered at the patent assignee level. *Max(OT score)*: maximum value of the Open Target score reached by any gene-disease pair mentioned in a patent application; *Patent citations*: number of forward patent citations received by the USPTO patent application up to seven years after its publication; *Patent family size*: number of patent applications in the same patent family; *Market value*: estimate of the market value (in constant USD) of the patent using data from Kogan et al. (2017); note that this measure is available only for applications that are eventually granted, hence the smaller sample size.

Table D.2 shows that targeting gene-disease pairs with a higher Open Target score has a technologically and economically significant effect on outcomes. Across the board, there is a strong correlation between patent applications for inventions targeting gene-disease combinations with higher Open Target scores and traditional metrics of patent value. In particular, the highest value reached by the Open Target score of the combinations in a patent application correlates with the number of forward patent citations when including assignee fixed effects. Open Target scores also predict the higher economic value of the patents, as captured by the dimension of the patent family and the monetary value of granted patents (using data from Kogan et al. 2017).

D.3 PubTator Central Publication Data

Data on the genes and diseases studied in each scientific publication are taken from PubTator Central, a web-based tool that automatically annotates biomedical concepts in PubMed abstracts and text (Wei et al., 2019). Articles are processed through concept taggers and disambiguation dictionaries to resolve annotation conflicts. The results of this process are publicly available online and include over 29 million abstracts and 3 million full-text documents. The entities extracted are matched to unique identifiers from NCBI and MeSH. The F1 score for the entity extraction pipeline is 86.70%

The coverage of authors' affiliations in PubMed is low and includes only the first author for papers published before 2013. To address this shortcoming, I obtain proprietary information on the authors' affiliations from Dimensions, a data product by Digital Science (Herzog et al., 2020). Compared to PubMed, whose recording of authors' affiliations is often limited to the corresponding author, Dimensions has much broader coverage. I use these data to match each patenting firm in my sample to their publication portfolios using fuzzy string matching on the firm names. Then, thanks to PubTator Central, I record all the genes and diseases studied by the firms. Finally, I use NIH's iCite data to collect additional bibliometric information, including the translational focus of each article and its

impact on clinical development.

Table D.3: Descriptive statistics of firms' publication portfolios

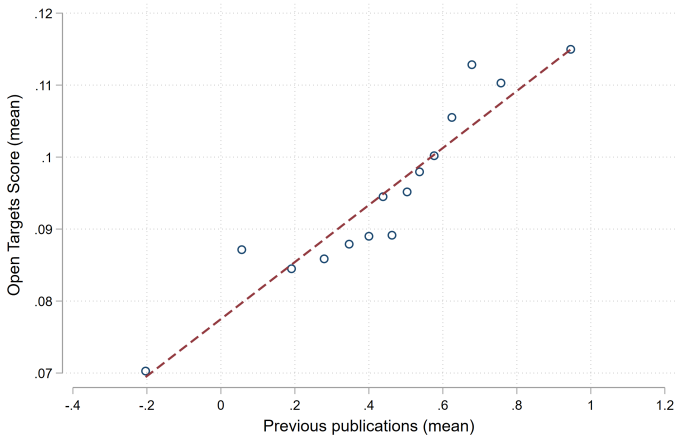
	mean	median	st d	min	max	N
Publications	59.170	9	390.944	1	10,656	3,346
Genes	176.563	52	568.922	1	12,267	3,346
Diseases	109.239	65	156.608	1	1,876	3,346
Gene-disease pairs	5,507.258	923	27,322.490	1	545,193	3,346
Year	2012.97	2014	4.672	1985	2019	3,346

This table presents descriptive statistics of firms' publication portfolios (conditional on having at least one publication). *Publications*: average number of firm publications; *Genes*: average number of genes on which the firm has published; *Diseases*: average number of diseases on which the firm has published; *Gene-disease pairs*: average number of gene-disease pairs on which the firm has published; *Year*: average year of the publications.

Table D.3 shows the descriptive statistics of the resulting dataset. 3,346 of the 4,117 firms in my sample (81.3%) have at least one publication. Each publishing firm has, on average, 59 papers about 177 genes and 109 diseases. The median firm has performed basic research on 923 gene-disease pairs. I then follow the approach of Azoulay et al. (2021) and classify as "translational research" disease-oriented studies that try to apply bench science findings to practical therapies but without being clinical trials. More specifically, I code articles that both mention a disease and are tagged with MeSH terms for molecular biology techniques and model organisms. I do so using data from NIH's iCite database. Intuitively, firms with translational experience can evaluate findings based on their theoretical understanding of basic genetics principles.

Table D.4 reports suggestive cross-sectional evidence that having domain knowledge on a specific gene (as proxied by past publications) correlates to firm patents of higher technological and economic

Figure D.3: The average therapeutic value of gene-disease combinations mentioned by a patent is higher when involving genes on which the firm has published before



Note: This figure shows a binscatter plotting the association between the average Open Target score of gene-disease pairs mentioned in a USPTO patent application and the share of those pairs that include a gene the firm has previously studied. Both publication counts and Open Targets scores are residualized by firm-disease fixed effects.

Table D.4: Patent applications are of higher value when they target gene-disease pairs involving genes where the firm has published before

Dependent Variable:	Patent citations		Patent family size		Market value	
Has gene knowledge (share)	-0.1496	8.0016*	4.0945***	0.9079***	24.7607**	0.488
	(3.1426)	(3.2385)	(0.6110)	(0.2152)	(7.8265)	(1.0171)
Patent year FE	YES	YES	YES	YES	YES	YES
Firm FE	NO	YES	NO	YES	NO	YES
N	134,459	134,330	134,459	134,330	26,498	26,329
Mean of Dep Var:	26.166	26.166	11.107	11.107	30.527	30.527

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the patent application level. Std. err. clustered at the patent assignee level. *Has gene knowledge (share)*: share gene-disease pairs mentioned in a patent that involve genes where the firm has published before; *Patent citations*: number of forward patent citations received by the USPTO patent application up to seven years after its publication; *Patent family size*: number of patent applications in the same patent family; *Market value*: estimate of the market value (in constant USD) of the patent using data from Kogan et al. (2017); note that this measure is available only for applications that are eventually granted, hence the smaller sample size.

value. The only partial exception is market value, where the effect vanishes after including firm fixed effects. However, this is likely due to how the measure of value is constructed from firms' stock market fluctuations, which does not allow for the detection of within-firm differences in patent value (Kogan et al., 2017). Finally, Figure D.3 shows a binscatter plotting the relationship between the average Open Target score of gene-disease pairs referenced in a USPTO patent application and the share of those pairs that include a gene that the firm has previously studied. The figure shows that the average therapeutic value of gene-disease combinations targeted by a patented innovation is higher when involving genes on which the firm has published before.

D.4 Cortellis Drug Data

I obtained drug development records up to July 2020 from Cortellis, which contains development information for 42,896 drugs targeting at least one of the gene-disease pairs in my sample. Cortellis aggregates information from various sources to assemble a list of historical development milestones for each drug molecule. This paper's analyses use those milestones to construct complete drug development histories for each drug whose genetic target and disease indication are recorded (Krieger, 2021). In particular, I focus on new molecules observed entering the earliest phases of drug development (what Cortellis records as the "discovery phase" and the "pre-clinical phase"). This choice is motivated by the short time lag from the first GWAS and the length of the drug development process, which means that it is still too early to observe drug approvals. In additional analyses reported in the main text, I follow the approach of Dranove et al. (2022) and weigh each drug by its relative novelty. The basic intuition of this weighting is that the higher the number of previous drugs that adopted the same molecular-targeting design, the less scientifically novel the drug is.

Table D.5 confirms the intuition that gene-disease combinations with at least one drug molecule have received more investments, as proxied by patent applications. Pairs with successful drug discovery activities appear in 112 patent applications and 2,176 publications, while the others only receive 2

patents and 47 publications on average. Interestingly, the difference is much lower when measured using the Open Target Score. This suggests that research and development inputs might disproportionally target gene-disease pairs known to be druggable but potentially miss out on other therapeutic opportunities (Oprea et al., 2018; Stoeger et al., 2018).

Table D.5: Descriptive statistics for gene-disease pairs with clinical activity.

	Gene-disease pairs with drugs						Gene-disease pairs without drugs					
	mean	median	st d	min	max	N	mean	median	st d	min	max	N
Patents on GDA	112.463	25	319.223	0	13,900	13,582	2.289	0	19.631	0	8,133	7,210,342
Publications on GDA	2,175.96	425	6,192.69	0	173,306	13,582	47.250	2	424.712	0	98,941	7,210,342
Open Targets score	0.099	0.025	0.158	0.00005	0.897	9,782	0.039	0.007	0.085	0.00004	0.874	584,571

Note: This table presents descriptive statistics at the level of gene-disease pairs. The left part of the table presents descriptives considering only pairs with at least one drug molecule listed in the Cortellis data, while the right panel presents descriptives considering only pairs that do not have drug molecules listed in the Cortellis data as of July 2020. *Patents on GDA*: average count of patents for inventions targeting the gene-disease pair; *Publications on GDA*: average count of publications mentioning the gene-disease pair; *Open Targets score*: average Open Targets score of the gene-disease pair.

Appendix References

Abrams, D. S., U. Akcigit, and J. Grennan (2018): "Patent value and citations: Creative destruction or strategic disruption?" *NBER wp 19647*.

Ahmadpoor, M. and B. F. Jones (2017): "The dual frontier: Patented inventions and prior scientific advance," *Science*, 357, 583–587.

Andriani, P. and G. Cattani (2022): "Functional diversification and exaptation: The emergence of new drug uses in the pharma industry," *Industrial and Corporate Change*, 31, 1177–1201.

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

Arts, S., B. Cassiman, and J. C. Gomez (2018): "Text matching to measure patent similarity," *Strategic Management Journal*, 39, 62–84.

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

Azoulay, P., W. H. Greenblatt, and M. L. Heggeness (2021): "Long-term effects from early exposure to research: Evidence from the NIH "Yellow Berets"," *Research Policy*, 50, 104332.

Beyer, B. M., R. Ingram, L. Ramanathan, P. Reichert, H. V. Le, V. Madison, and P. Orth (2008): "Crystal structures of the pro-inflammatory cytokine interleukin-23 and its complex with a high-affinity neutralizing antibody," *Journal of Molecular Biology*, 382, 942–955.

Bianchi, E. and L. Rogge (2019): "The IL-23/IL-17 pathway in human chronic inflammatory diseases—new insight from genetics and targeted therapies," *Genes & Immunity*, 20, 415–425.

Boyack, K. W., C. Smith, and R. Klavans (2020): "A detailed open access model of the PubMed literature," *Scientific Data*, 7, 408.

Boyle, E. A., Y. I. Li, and J. K. Pritchard (2017): "An expanded view of complex traits: From polygenic to omnigenic," *Cell*, 169, 1177–1186.

Brusoni, S., A. Prencipe, and K. Pavitt (2001): "Knowledge specialization, organizational coupling, and the boundaries of the firm: Why do firms know more than they make?" *Administrative Science Quarterly*, 46, 597–621.

Bryan, K. A., Y. Ozcan, and B. Sampat (2020): "In-text patent citations: A user's guide," *Research Policy*, 49, 103946.

Bush, W. S. and J. H. Moore (2012): "Genome-wide association studies," *PLoS Computational Biology*, 8, e1002822.

Cohen, W. M., R. R. Nelson, and J. P. Walsh (2002): "Links and impacts: The influence of public research on industrial R&D," *Management Science*, 48, 1–23.

Dahlhamer, J. M., E. P. Zammitti, B. W. Ward, A. G. Wheaton, and J. B. Croft (2016): "Prevalence of inflammatory bowel disease among adults aged $\leq$ 18 years—United States, 2015," *Morbidity and Mortality Weekly Report*, 65, 1166–1169.

Ding, Y., M. Song, J. Han, Q. Yu, E. Yan, L. Lin, and T. Chambers (2013): "Entitymetrics: Measuring the impact of entities," *PloS one*, 8, e71416.

Dranove, D., C. Garthwaite, and M. Hermosilla (2022): "Does consumer demand pull scientifically novel drug innovation?" *The RAND Journal of Economics*, 53, 590–638.

Duerr, R. H., K. D. Taylor, S. R. Brant, J. D. Rioux, M. S. Silverberg, M. J. Daly, et al. (2006): "A genome-wide association study identifies IL23R as an inflammatory bowel disease gene," *Science*, 314, 1461–1463.

Duguet, E. and M. MacGarvie (2005): "How well do patent citations measure flows of technology? Evidence from French innovation surveys," *Economics of Innovation and New Technology*, 14, 375–393.

Easton, D. F., K. A. Pooley, A. M. Dunning, P. D. Pharoah, et al. (2007): "Genome-wide association study identifies novel breast cancer susceptibility loci," *Nature*, 447, 1087–1093.

Eggers, J. P. and S. Kaplan (2009): "Cognition and renewal: Comparing CEO and organizational effects on incumbent adaptation to technical change," *Organization Science*, 20, 461–477.

Fleming, L. (2001): "Recombinant uncertainty in technological search," *Management Science*, 47, 117–132.

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

Goldstein, D. B. (2009): "Common genetic variation and human traits," *New England Journal of Medicine*, 360, 1696.

Herzog, C., D. Hook, and S. Konkiel (2020): "Dimensions: Bringing down barriers between scientometricians and data," *Quantitative Science Studies*, 1, 387–395.

Iaria, A., C. Schwarz, and F. Waldinger (2018): "Frontier knowledge and scientific production: Evidence from the collapse of international science," *The Quarterly Journal of Economics*, 133, 927–991.

Jaffe, A. B., M. Trajtenberg, and M. S. Fogarty (2000): "Knowledge spillovers and patent citations: Evidence from a survey of inventors," *American Economic Review*, 90, 215–218.

Kang, S. (2024): "From outward to inward: Reframing search with new mapping criteria," *UC Santa Barbara*.

Kao, J. (2023): "Charted territory: Mapping the cancer genome and R&D decisions in the pharmaceutical industry," *UCLA Anderson*.

Kaplan, S. and K. Vakili (2015): "The double-edged sword of recombination in breakthrough innovation," *Strategic Management Journal*, 36, 1435–1457.

Kingwell, K. (2022): "LRRK2-targeted Parkinson disease drug advances into phase III." *Nature Reviews Drug Discovery*.

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

Kuhn, J., K. Younge, and A. Marco (2020): "Patent citations reexamined," *The RAND Journal of Economics*, 51, 109–132.

Lampe, R. (2012): "Strategic citation," *Review of Economics and Statistics*, 94, 320–333.

MacArthur, D. (2012): "Face up to false positives," *Nature*, 487, 427–428.

MacArthur, J., E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, et al. (2017): "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)," *Nucleic Acids Research*, 45, D896–D901.

Marco, A. C., J. D. Sarnoff, and A. Charles (2019): "Patent claims and patent scope," *Research Policy*, 48, 103790.

Marigorta, U. M., J. A. Rodríguez, G. Gibson, and A. Navarro (2018): "Replicability and prediction: Lessons and challenges from GWAS," *Trends in Genetics*, 34, 504–517.

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

Mifflin, L., D. Ofengeim, and J. Yuan (2020): "Receptor-interacting protein kinase 1 (RIPK1) as a therapeutic target," *Nature Reviews Drug Discovery*, 19, 553–571.

Myers, K. R. and L. Lanahan (2022): "Estimating spillovers from publicly funded R&D: Evidence from the US Department of Energy," *American Economic Review*, 112, 2393–2423.

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

Nagaraj, A. and M. Tranchero (2024): "Empirical search landscapes," *UC Berkeley and University of Pennsylvania*.

Narin, F., K. S. Hamilton, and D. Olivastro (1997): "The increasing linkage between US technology and public science," *Research Policy*, 26, 317–330.

Nelson, M. R., H. Tipney, J. L. Painter, J. Shen, P. Nicoletti, Y. Shen, et al. (2015): "The support of human genetic evidence for approved drug indications," *Nature Genetics*, 47, 856–860.

Ochoa, D., A. Hercules, M. Carmona, D. Suveges, J. Baker, et al. (2023): "The next-generation Open Targets Platform: Reimagined, redesigned, rebuilt," *Nucleic Acids Research*, 51, D1353–D1359.

Oprea, T. I., C. G. Bologa, S. Brunak, A. Campbell, G. N. Gan, A. Gaulton, S. M. Gomez, R. Guha, A. Hersey, J. Holmes, et al. (2018): "Unexplored therapeutic opportunities in the human genome," *Nature Reviews Drug Discovery*, 17, 317–332.

Pushpakom, S., F. Iorio, P. A. Eyers, et al. (2019): "Drug repurposing: Progress, challenges and recommendations," *Nature Reviews Drug Discovery*, 18, 41–58.

Reay, W. R. and M. J. Cairns (2021): "Advancing the use of genome-wide association studies for drug repurposing," *Nature Reviews Genetics*, 22, 658–671.

Reich, D. E. and E. S. Lander (2001): "On the allelic spectrum of human disease," *TRENDS in Genetics*, 17, 502–510.

Righi, C. and T. Simcoe (2023): "Patenting inventions or inventing patents? Continuation practice at the USPTO," *The RAND Journal of Economics*, 54, 416–442.

Roach, M. and W. M. Cohen (2013): "Lens or prism? Patent citations as a measure of knowledge flows from public research," *Management Science*, 59, 504–525.

Stoeger, T., M. Gerlach, R. I. Morimoto, and L. A. Nunes Amaral (2018): "Large-scale investigation of the reasons why potentially important genes are ignored," *PLoS Biology*, 16, e2006643.

Suh, J. (2024): "Swinging for the fences: Startup novelty as a response to entry costs," *NYU Stern*.

Teplitskiy, M., E. Duede, M. Menietti, and K. R. Lakhani (2022): "How status of research papers affects the way they are read and cited," *Research Policy*, 51, 104484.

Tranchero, M. (2024): "Data-driven search and innovation: Evidence from genome-wide association studies," *University of Pennsylvania*.

Vaughan, L. K. and V. Srinivasasainagendra (2013): "Where in the genome are we? A cautionary tale of database use in genomics research," *Frontiers in Genetics*, 4, 38.

Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang (2017): "10 years of GWAS discovery: biology, function, and translation," *The American Journal of Human Genetics*, 101, 5–22.

Wei, C.-H., A. Allot, R. Leaman, and Z. Lu (2019): "PubTator central: Automated concept annotation for biomedical full text articles," *Nucleic Acids Research*, 47, W587–W593.

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

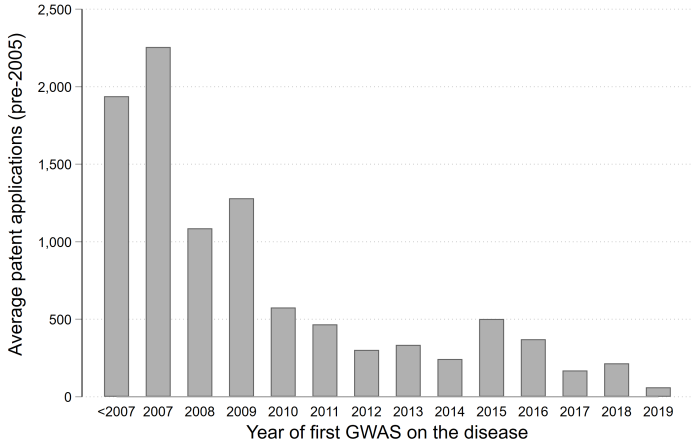
Xu, J., S. Kim, M. Song, M. Jeong, D. Kim, J. Kang, J. F. Rousseau, X. Li, W. Xu, V. I. Torvik, et al. (2020): "Building a PubMed knowledge graph," *Scientific Data*, 7, 1–15.

Younge, K. A. and J. M. Kuhn (2019): "First movers and follow-on invention: Evidence from a vector space model of invention," *Available at SSRN 3354530*.

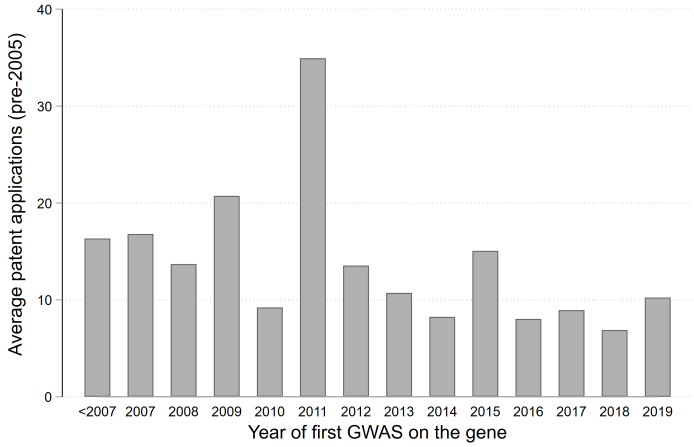
E Additional Figures and Tables

Figure E.1: While there is clear evidence of temporal sorting on the diseases that receive a GWAS, the same is not true for genes

(a) *Average past patents on the diseases by year of their first GWAS*



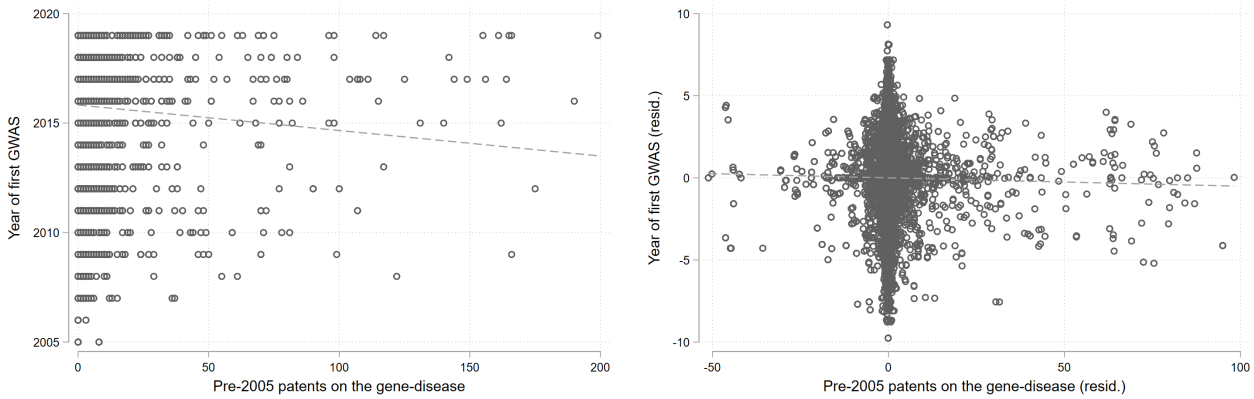
(b) *Average past patents on the genes by year of their first GWAS*



Note: Panel (a) shows the average number of patent applications mentioning each disease before 2005, separately by the year when the disease first appeared in a GWAS. Panel (b) shows the average number of patent applications mentioning each gene before 2005, separately by the year when the gene first appeared in a GWAS. The figure shows that diseases that received a GWAS earlier in time tend to be those with higher pre-GWAS patenting activity, confirming the prioritization of large and important diseases. Reassuringly, the same type of selection is not apparent for genes due to the research design of GWAS.

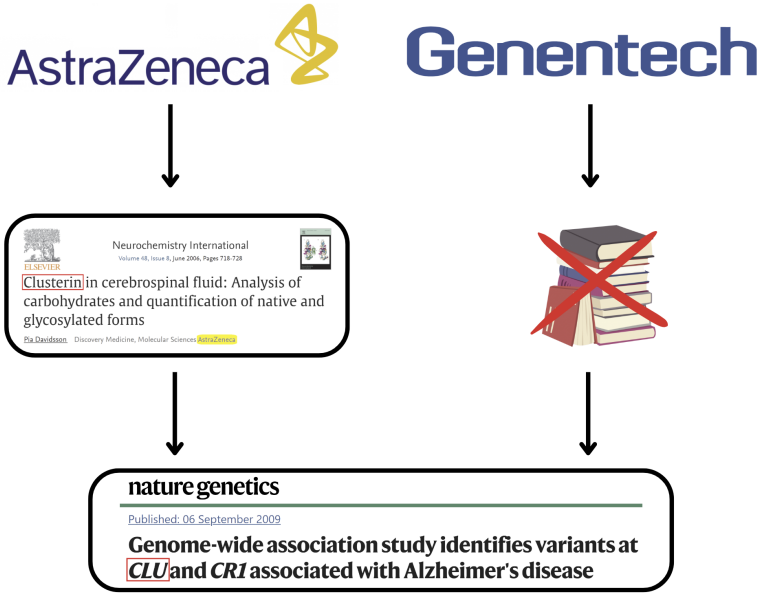
Figure E.2: The timing of GWAS associations is not related to past patenting on gene-diseases pairs, once controlling for disease selection

(a) *Raw association* (b) *Association conditional on disease choice*



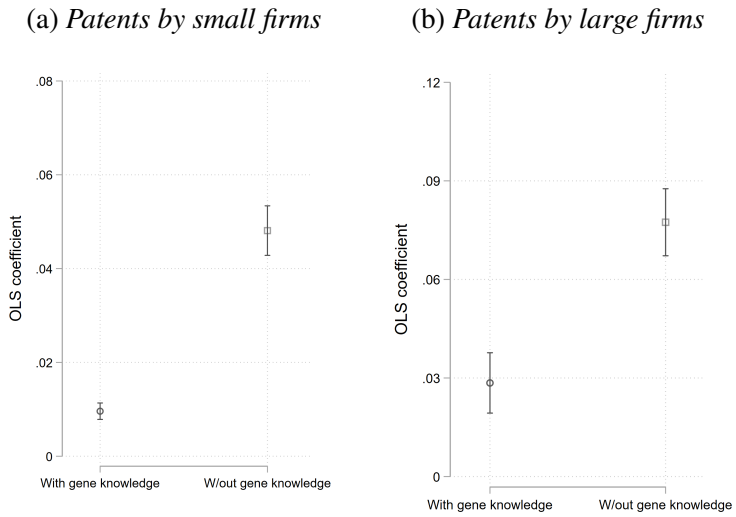
Note: These figures show the correlation between patent applications mentioning one of the 17,965 treated gene-disease pairs before 2005 and the year of the first GWAS association reporting them. Panel (a) presents the raw scatterplot. Panel (b) shows the same scatterplot after residualizing for disease. This figure confirms that gene-disease associations reported by GWAS are a plausibly exogenous shock.

Figure E.3: Schema of the between-firms research design used for gene-disease level analyses



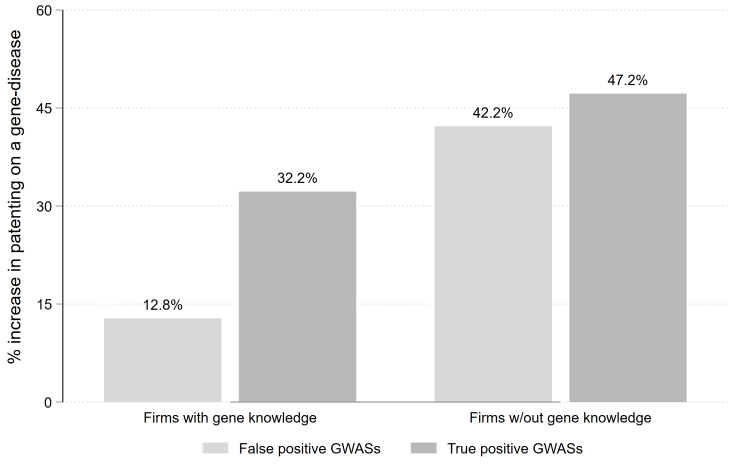
Note: This figure exemplifies the between-firm design used to assess the role of domain knowledge in gene-disease level regressions. Some firms evaluating the plausibility of a GWAS association can leverage their previous research on the gene involved, thus improving their assessment thanks to a deeper understanding of the gene's biology.

Figure E.4: Firms react less to the arrival of GWAS associations when they possess domain knowledge, regardless of their size



Note: The figure reports OLS coefficients capturing the main effect of GWAS on USPTO patent applications, separately by firms with and without previous publications on the gene involved. Each coefficient is estimated with the same specifications reported in Table 4. Panel (a) reports the two estimates for small firms, defined as the ones with a below-median number of patents in my samples. Panel (b) reports the two estimates for large firms, defined as the ones with an above-median number of patents in my samples.

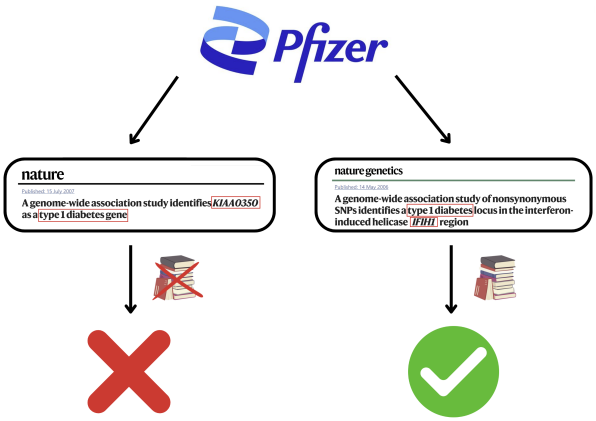
Figure E.5: Average increase in patent applications on treated gene-disease pairs, by type of firm



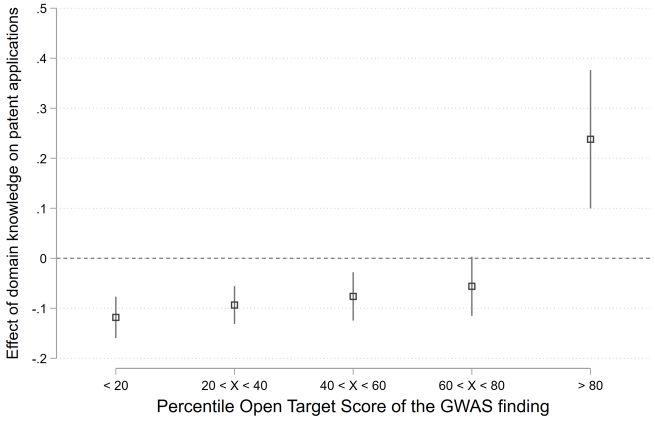
Note: This figure shows the average increase in USPTO patent applications on a gene-disease pair after the publication of a GWAS (considering only treated pairs). The increase is presented separately for GWAS findings that are not replicated in subsequent studies and those that are confirmed by later GWAS. The first two columns report the patenting increase by firms with previous publications on the gene involved, while the latter two columns report the patenting increase by firms without them.

Figure E.6: A within-firm research design confirms that domain knowledge improves the assessment of GWAS predictions and helps to select only the opportunities with higher Open Target scores

(a) *Schema of the within-firm research design*



(b) *Impact of domain knowledge on investments' quality*



Note: Panel (a) exemplifies the within-firms design to isolate the role of domain knowledge in interpreting GWAS and determining performance. The research design exploits the fact that a firm might be evaluating multiple findings for the same disease, but has done genetic research only on a subset of the genes involved. Panel (b) shows coefficients from split-sample regressions evaluating the within-firm impact of genetic knowledge on firms' ability to recognize the best GWAS findings. The ground truth quality of the gene-disease association uncovered by GWAS is proxied by the percentile of the Open Targets score. The regressions include firm × disease and year of discovery dummies.

Table E.1: No evidence of association between past patenting and the timing of the first GWAS for treated gene-disease pairs

Dependent Variable:	Year of first GWAS			
Patents about gene	-0.007	-0.014		
	[-0.65]	[-1.87]		
Patents about gene-disease			-0.040*	-0.020
			[-2.31]	[-1.40]
Disease FE	NO	YES	NO	YES
Gene FE	NO	NO	YES	YES
N of obs	17,925	17,923	16,338	16,298
Mean of Dep Var:	2015.8	2015.8	2015.8	2015.8

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the gene-disease association level. Std. err. clustered two-ways at the disease and gene level. This table shows standardized beta coefficients to ease comparison and reports t-values inside square brackets. *Year of first GWAS*: the year when the gene-disease pair was first reported as significant in one GWAS paper. *Patents about gene*: sum of USPTO patent applications filed before 2005 that mention a specific gene. *Patents about gene-disease*: sum of USPTO patent applications filed before 2005 that mention a specific gene-disease pair.

Table E.2: GWAS findings that are subsequently replicated involve gene-disease pairs with a higher Open Targets score

Dependent Variable:	Open Targets score		Top 90%ile OT score (0/1)	
Replicable association (0/1)	0.0937***	0.0457***	0.1909***	0.1139***
	(0.00900)	(0.00645)	(0.01895)	(0.01442)
Gene FE	YES	YES	YES	YES
Disease FE	YES	YES	YES	YES
Year of GWAS FE	YES	YES	YES	YES
Sources Count FE	NO	YES	NO	YES
N of GDAs	16,298	8,921	16,298	8,921
Mean of Dep Var:	0.0894	0.0894	0.0998	0.0998

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the gene-disease association level. Std. err. clustered at the gene level. *Open Targets score*: value of the Open Targets score of the gene-disease pair. *Top 90%ile OT score*: 0/1 = 1 if the gene-disease pair is in the top decile of Open Targets score in my sample of GWAS associations. *Replicable association*: 0/1 = 1 if the gene-disease association is replicated by subsequent GWAS about the same disease. *Sources Count FE*: number of sources aggregated by Open Targets to compute the score of a given gene-disease pair.

Table E.3: Correlates of replicable GWAS findings: study design quality.

Dependent Variable:	Replicable association (0/1)		
Large sample (0/1)	0.3197***		
	(0.02069)		
Replication sample (0/1)		0.2192***	
		(0.01811)	
Powerful genotyping array (0/1)			0.0993***
			(0.02237)
Disease FE	YES	YES	YES
Year of GWAS FE	YES	YES	YES
N of GDAs	17,923	17,923	16,161
Mean of Dep Var:	0.1574	0.1574	0.1574

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the gene-disease association level. Std. err. clustered at the gene level. *Replicable association*: 0/1 = 1 if the gene-disease association is replicated by subsequent GWAS about the same disease. *Large sample*: 0/1 = 1 for associations published in studies with a sample size larger than the median GWAS in my sample. *Replication sample*: 0/1 = 1 for associations reported in papers that also include a replication analysis of their result. *Powerful genotyping array*: 0/1 = 1 if the gene-disease association is obtained using a microarray with an above-median number of SNPs.

Table E.4: Correlates of replicable GWAS findings: citations to the article introducing them.

Dependent Variable:	Replicable association (0/1)		
Scientific citations to GWAS	0.0002***		
	(0.00003)		
Clinical citations to GWAS		0.0077**	
		(0.00259)	
Share of negative citations			-0.0488**
			(0.01525)
Disease FE	YES	YES	YES
Year of GWAS FE	YES	YES	YES
N of GDAs	17,923	17,923	13,732
Mean of Dep Var:	0.1574	0.1574	0.1574

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the gene-disease association level. Std. err. clustered at the gene level. *Replicable association*: 0/1 = 1 if the gene-disease association is replicated by subsequent GWAS about the same disease. *Scientific citations to GWAS*: number of citations from scientific papers received by the GWAS introducing the gene-disease association. *Clinical citations to GWAS*: number of citations from clinical trials received by the GWAS introducing the gene-disease association. *Share of negative citations*: share of citations with a negative tone from scientific papers received by the GWAS introducing the gene-disease association (data from Scite).

Table E.5: Firms increase their investments more when the associations are statistically stronger or the GWAS have better research designs

Dependent Variable:		USPTO patent applications				
Post × GWAS		0.0432*	0.0779***	0.0581***	0.0828***	0.0432*
		(0.02019)	(0.02019)	(0.01815)	(0.01809)	(0.01871)
	...× High P-value (0/1)	0.2455***				
		(0.03989)				
	...× Large effect (0/1)		0.1970***			
			(0.04246)			
	...× Top journal (0/1)			0.292***		
				(0.04810)		
	...× Large sample (0/1)				0.1830***	
					(0.04271)	
	...× Replication sample (0/1)					0.2449***
						(0.04007)
Gene-Disease FE		YES	YES	YES	YES	YES
Disease-Year FE		YES	YES	YES	YES	YES
Gene-Year FE		YES	YES	YES	YES	YES
N		137,254,556	137,254,556	137,254,556	137,254,556	137,254,556
N of Gene-Diseases		7,223,924	7,223,924	7,223,924	7,223,924	7,223,924
Mean of Dep Var:		0.1314	0.1314	0.1314	0.1314	0.1314

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the gene-disease-year level. Std. err. clustered two-ways at the disease and gene level. *USPTO patent applications*: count of USPTO patent applications filed in a given year for innovations that target a specific gene-disease combination. *Post × GWAS*: 0/1 = 1 in all years after a gene-disease pair is treated by its first GWAS. *High P-value*: 0/1 = 1 for associations with a p-value larger than the median GWAS in my sample. *Large effect*: 0/1 = 1 for associations with an effect size larger than the median GWAS in my sample. *Top journal*: 0/1 = 1 for associations published in the 15 most prestigious genetics journals or the top 3 generalist scientific journals (*Science*, *Nature*, *PNAS*). *Large sample*: 0/1 = 1 for associations published in studies with a sample size larger than the median GWAS in my sample. *Replication sample*: 0/1 = 1 for associations reported in papers that include also a replication analysis of their result.

Table E.6: Firms with genetic domain knowledge introduce new drugs on true positive GWAS findings, while firms lacking domain knowledge miss these opportunities

Dependent Variable:	New drugs by...			
	...firms with gene knowledge		...firms w/out gene knowledge	
Post × GWAS	0.000113	-0.0000653	0.000221	-0.0000121
	(0.0000691)	(0.0000384)	(0.0001569)	(0.0001093)
...× True Positive		0.000892**		0.0011658
		(0.0003058)		(0.0006582)
Gene-Disease FE	YES	YES	YES	YES
Disease-Year FE	YES	YES	YES	YES
Gene-Year FE	YES	YES	YES	YES
N	137,254,556	137,254,556	137,254,556	137,254,556
N of Gene-Diseases	7,223,924	7,223,924	7,223,924	7,223,924
Mean of Dep Var:	0.0000138	0.0000138	0.0000827	0.0000827

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the gene-disease-year level. Std. err. clustered two-ways at the disease and gene level. *New drugs*: number of molecules on a gene-disease combination entering the discovery stage (data from Cortellis); the count is then divided between firms with and without previous publications on the gene. *Post × GWAS*: 0/1 = 1 in all years after a gene-disease pair is treated by its first GWAS. *True Positive*: 0/1 = 1 for GWAS findings that are later replicated by another GWAS about the same disease.

Table E.7: Firms without domain knowledge struggle more to recognize opportunities when the GWAS association involves less-studied genes

Dependent Variable:	USPTO patent applications by...		
	... all firms	...firms with gene knowledge	...firms w/out gene knowledge
Post × GWAS	0.0123**	0.0024**	0.0099*
	(0.00462)	(0.00092)	(0.00431)
...× True Positive	0.0265	0.0202*	0.0063
	(0.01458)	(0.00876)	(0.00891)
Gene-Disease FE	YES	YES	YES
Disease-Year FE	YES	YES	YES
Gene-Year FE	YES	YES	YES
N	66,980,776	66,980,776	66,980,776
N of Gene-Diseases	3,525,304	3,525,304	3,525,304
Mean of Dep Var:	0.02097	0.00072	0.02026

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the gene-disease-year level. Std. err. clustered two-ways at the disease and gene level. *USPTO patent applications*: count of USPTO patent applications filed in a given year that target a specific gene-disease combination; the count is then divided between firms with and without previous publications on the gene. *Post × GWAS*: 0/1 = 1 in all years after a gene-disease pair is treated by its first GWAS. The sample is restricted to gene-disease pairs involving genes that received a below-median number of scientific studies before 2005 (the year of the first GWAS).

Table E.8: Firms without domain knowledge invest proportionally less in false positives when the GWAS report a smaller number of associations

Dependent Variable:	USPTO patent applications by...		
	... all firms	...firms with gene knowledge	...firms w/out gene knowledge
Post × GWAS	0.0670**	0.01935	0.04768**
	(0.02019)	(0.00982)	(0.01385)
...× True Positive	0.6382***	0.1535**	0.4848***
	(0.10028)	(0.05037)	(0.07703)
Gene-Disease FE	YES	YES	YES
Disease-Year FE	YES	YES	YES
Gene-Year FE	YES	YES	YES
N	137,084,183	137,084,183	137,084,183
N of Gene-Diseases	7,214,957	7,214,957	7,214,957
Mean of Dep Var:	0.13088	0.03741	0.09346

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the gene-disease-year level. Std. err. clustered two-ways at the disease and gene level. *USPTO patent applications*: count of USPTO patent applications filed in a given year that target a specific gene-disease combination; the count is then divided between firms with and without previous publications on the gene. *Post × GWAS*: 0/1 = 1 in all years after a gene-disease pair is treated by its first GWAS, excluding studies that report an above-median number of associations (i.e., larger than 55).

Table E.9: Results are robust to the exclusion of GWAS with an industry co-author or that acknowledge funding from a pharmaceutical firm

Dependent Variable:	USPTO patent applications by...					
	...all firms		...firms with gene knowledge		...firms w/out gene knowledge	
Post × GWAS	0.1263***	0.0669***	0.0319**	0.0148	0.0944***	0.0522***
	(0.02182)	(0.01856)	(0.01146)	(0.00979)	(0.01328)	(0.01165)
...× True Positive		0.4433***		0.1278*		0.3155***
		(0.11117)		(0.05833)		(0.06512)
Gene-Disease FE	YES	YES	YES	YES	YES	YES
Disease-Year FE	YES	YES	YES	YES	YES	YES
Gene-Year FE	YES	YES	YES	YES	YES	YES
N	137,167,422	137,167,422	137,167,422	137,167,422	137,167,422	137,167,422
N of Gene-Diseases	7,219,338	7,219,338	7,219,338	7,219,338	7,219,338	7,219,338
Mean of Dep Var:	0.1308	0.1308	0.0373	0.0373	0.0934	0.0934

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the gene-disease-year level. Std. err. clustered two-ways at the disease and gene level. *USPTO patent applications*: count of USPTO patent applications filed in a given year that target a specific gene-disease combination; the count is then divided between firms with and without previous publications on the gene. *Post × GWAS*: 0/1 = 1 in all years after a gene-disease pair is treated by its first GWAS, excluding all GWAS with either a co-author working in a pharmaceutical company or acknowledging financial support from private firms. *True Positive*: 0/1 = 1 for GWAS findings that are later replicated by another GWAS about the same disease.

Table E.10: Investments in false positive GWAS findings seem not driven by strategic considerations, but follow patterns consistent with firms making mistakes

Dependent Variable		Strategic patent (0/1)	Continuation patent (0/1)	Litigated patent (0/1)	Expired patent (0/1)
Post × GWAS		0.0006	0.0094***	0.0022	-0.0039***
		(0.00046)	(0.00139)	(0.00119)	(0.00072)
	...× True Positive	0.0089***	0.0306***	0.0186***	-0.0095***
		(0.00172)	(0.00381)	(0.00340)	(0.00236)
Gene-Disease FE		YES	YES	YES	YES
Disease-Year FE		YES	YES	YES	YES
Gene-Year FE		YES	YES	YES	YES
N		137,254,556	137,254,556	137,254,556	137,254,556
N of Gene-Diseases		7,223,924	7,223,924	7,223,924	7,223,924
Mean of Dep Var:		0.0023	0.0339	0.0282	0.0100

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the gene-disease-year level. Std. err. clustered two-ways at the disease and gene level. *Strategic patent*: 0/1=1 if the gene gene-disease combination received at least one *strategic patent* according to the definition by Abrams et al. (2018). *Continuation patent*: 0/1=1 if the gene gene-disease combination received at least one *continuation patent* according to the definition by Righi and Simcoe (2023). *Litigated patent*: 0/1=1 if the gene gene-disease combination received at least one patent that was subsequently involved in litigation. *Expired patent*: 0/1=1 if the gene gene-disease combination received at least one patent that was not subsequently renewed. *Post × GWAS*: 0/1 = 1 in all years after a gene-disease pair is treated by its first GWAS. *True Positive*: 0/1 = 1 for GWAS findings that are later replicated by another GWAS about the same disease.

Table E.11: No evidence of innovation spillovers on gene-disease pairs proximate to those treated by false positive GWAS associations

Dependent Variable:	Cit-weighted patents	Patent market value	Drugs (total)	Drugs (weighted)
Post × GWAS Spillover	0.2206	-0.0361	-0.000049	-0.000005
	(0.21480)	(0.05840)	(0.00006)	(0.00002)
Gene-Disease FE	YES	YES	YES	YES
Disease-Year FE	YES	YES	YES	YES
Gene-Year FE	YES	YES	YES	YES
N	136,913,221	136,913,221	136,913,221	136,913,221
N of Gene-Diseases	7,205,959	7,205,959	7,205,959	7,205,959
Mean of Dep Var:	7.0851	0.6229	0.0002	0.0001

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the gene-disease-year level. Std. err. clustered two-ways at the disease and gene level. *Post × GWAS Spillover*: 0/1 = 1 in all years after a neighboring gene-disease pair is treated by its first GWAS; neighbor pairs are defined as involving the same gene and a disease sharing the same 3-digit patent MeSH code than the ones in the GWAS. *Cit-weighted patents*: count of USPTO patent applications filed in a given year that target a specific gene-disease combination, weighted by the number of patent citations received up to seven years after patent publication. *Patent value*: estimated stock market value (in constant USD) of patents granted to public firms using data from Kogan et al. (2017). *Drugs (total)*: number of molecules on a gene-disease combination entering the discovery stage. *Drugs (weighted)*: number of molecules on a gene-disease combination entering the discovery stage weighted by their scientific novelty (i.e., by the number of times that the same mechanism of action has been used before, following Dranove et al. 2022).

Table E.12: Firms with domain knowledge invest proportionally more on higher-quality gene-disease combinations

Dependent Variable:	USPTO patent applications by...					
	...firms with gene knowledge			...firms w/out gene knowledge		
Post × GWAS	0.0061	0.0871***	0.1334**	0.0919*	0.1793***	0.1833**
	(0.01962)	(0.02619)	(0.04119)	(0.04219)	(0.03463)	(0.05579)
Gene-Disease FE	YES	YES	YES	YES	YES	YES
Disease-Year FE	YES	YES	YES	YES	YES	YES
Gene-Year FE	YES	YES	YES	YES	YES	YES
N	588,069	9,550,920	1,091,265	588,069	9,550,920	1,091,265
N of Gene-Diseases	30,951	502,680	57,435	30,951	502,680	57,435
Mean of Dep Var:	0.0607	0.3359	0.5363	0.194	0.5944	0.8484
Sample:	<10 OT Score	10 <OT Score <90	>90 OT Score	<10 OT Score	10 <OT Score <90	>90 OT Score

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the gene-disease-year level. Std. err. clustered two-ways at the disease and gene level. The sample is limited to gene-disease pairs for which an Open Targets Score could be matched. Columns 1 and 4 include only gene-disease pairs with an Open Targets Score below the tenth percentile of the sample. Columns 2 and 5 include only gene-disease pairs with an Open Targets Score between the tenth and the ninetieth percentile of the sample. Columns 3 and 6 include only gene-disease pairs with an Open Targets Score above the ninetieth percentile of the sample. *USPTO patent applications*: count of USPTO patent applications filed in a given year that target a specific gene-disease combination; the count is then divided between firms with and without previous publications on the gene. *Post × GWAS*: 0/1 = 1 in all years after a gene-disease pair is treated by its first true positive GWAS.

Table E.13: Genetic domain knowledge allows recognizing both false and true positive GWAS associations in a within firm-design, ruling out that the effects are due to generic firm-level capabilities

Dependent Variable:	Sum of USPTO patent applications			
Sample:	False positive associations		True positive associations	
Has Gene Knowledge (0/1)	-0.1297***	-0.1330***	0.1944***	0.1996***
	(0.01525)	(0.01441)	(0.05453)	(0.05633)
Firm FE	YES	NO	YES	NO
Disease FE	YES	NO	YES	NO
Firm-Disease FE	NO	YES	NO	YES
N of Firms	3,787	3,768	3,651	3,632
N of Observations	4,493,871	4,483,884	1,432,671	1,432,671

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the firm-gene-disease level represent all the potential GWAS opportunities considered by a firm. Std. err. clustered two-ways at the firm level. *Sum of USPTO patent applications*: sum of USPTO patent applications filed in a given year that target a specific gene-disease combination. *Has Gene Knowledge (0/1)*: 0/1 = 1 if the firm has at least one publication about the specific gene before the GWAS associated it with a disease. Columns 1 and 2 include only GWAS associations that fail subsequent replications, while columns 3 and 4 include only GWAS associations that are later replicated.

Table E.14: The results for false positives of Appendix Table E.13 are robust to alternative definitions of the risk set of GWAS evaluated by a given firm

Dependent Variable:	USPTO patent applications				
Share of patent portfolio:	>0.005	>0.01	>0.02	>0.05	>0.1
Has Gene Knowledge (0/1)	-0.1155***	-0.1082***	-0.0895**	-0.1080***	-0.1712***
	(0.01642)	(0.01722)	(0.02690)	(0.02089)	(0.01818)
Firm-Disease FE	YES	YES	YES	YES	YES
N of firms	3765	3752	3666	2784	1562
N of obs	3,546,300	2,618,859	1,515,477	534,200	202,376

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the firm-gene-disease level. Std. err. clustered two-ways at the disease and gene level. Each column is estimated from a separate regression using a progressively more stringent risk set of GWAS evaluated, defined as those involving diseases constituting a certain share of a firm's patent portfolio. The sample is restricted to false-positive GWAS. *USPTO patent applications*: count of USPTO patent applications filed in a given year that target a specific gene-disease combination. *Has Gene Knowledge (0/1)*: 0/1 = 1 if the firm has at least one publication about the specific gene before the GWAS associated it with a disease.

Table E.15: The results for true positives of Appendix Table E.13 are robust to alternative definitions of the risk set of GWAS evaluated by a given firm

Dependent Variable:	USPTO patent applications				
Share of patent portfolio:	>0.005	>0.01	>0.02	>0.05	>0.1
Has Gene Knowledge (0/1)	0.2708***	0.3336***	0.4992***	0.7730**	0.7808
	(0.06344)	(0.06788)	(0.11716)	(0.28434)	(0.44003)
Firm-Disease FE	YES	YES	YES	YES	YES
N of firms	3,629	3,601	3,487	2,485	1,291
N of obs	1,159,770	867,620	514,864	192,353	75,732

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the firm-gene-disease level. Std. err. clustered two-ways at the disease and gene level. Each column is estimated from a separate regression using a progressively more stringent risk set of GWAS evaluated, defined as those involving diseases constituting a certain share of a firm's patent portfolio. The sample is restricted to true positive GWAS. *USPTO patent applications*: count of USPTO patent applications filed in a given year that target a specific gene-disease combination. *Has Gene Knowledge (0/1)*: 0/1 = 1 if the firm has at least one publication about the specific gene before the GWAS associated it with a disease.

Table E.16: Descriptive evidence that firms delaying investments in a GWAS finding are less likely to choose false positives, but also less likely to obtain highly cited patents or drugs

Dependent Variable	False positive	Cit-weighted patents	Drugs (total)	Drugs (weighted)
Years Waited to First Patent	-0.0248***	-1.7812***	-0.0001	-0.0005*
	(0.0028)	(0.1510)	(0.0003)	(0.0002)
Firm FE	YES	YES	YES	YES
Disease FE	YES	YES	YES	YES
N of firms	225	225	225	225
N of obs	30615	30615	30615	30615
Mean of Dep Var:	0.5310	34.861	0.0081	0.0047

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations at the gene-disease association level. Std. err. clustered two-ways at the disease and gene level. *Years Waited to First Patent*: count of years between the GWAS publication and the firm first patent application targeting the gene-disease pair. *False positive*: 0/1=1 if the gene-disease association is not subsequently replicated. *Cit-weighted patent*: count of USPTO patent applications filed in a given year that target a specific gene-disease combination, weighted by the number of patent citations received up to seven years after patent publication. *Drugs (total)*: number of molecules on a gene-disease combination entering the discovery stage. *Drugs (weighted)*: number of molecules on a gene-disease combination entering the discovery stage weighted by their scientific novelty (i.e., by the number of times that the same mechanism of action has been used before, following Dranove et al. 2022).