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

Matteo Tranchero
The Wharton School

May 26, 2026

Abstract

Big data is increasingly used to predict which opportunities are worth pursuing. By making technological search less dependent on prior expertise, data-driven predictions could level the playing field in innovation. Yet predictions based on codified information can be noisy when evaluating opportunities requires tacit and contextual knowledge. In such settings, domain knowledge helps firms screen out false positives, ultimately reinforcing the advantage of firms with deeper expertise. I test this argument in pharmaceutical innovation, where genome-wide association studies (GWAS) identify genes that may serve as drug targets. I find that GWAS stimulate corporate investment, but around one-third of this response targets false positives. Firms lacking gene-specific expertise rely more on GWAS and disproportionately pursue these misleading signals. As a result, the same data-driven opportunities widen performance gaps in drug discovery, with downstream gains concentrated among firms with deeper knowledge bases. Mechanism tests trace this effect to patient-oriented knowledge that helps firms judge whether statistical patterns are biologically meaningful. Together, the results show that domain knowledge remains crucial for competitive advantage, but its role shifts from finding opportunities to assessing those revealed by data.

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

1 Introduction

The data revolution is transforming how companies operate. Firms are increasingly using machine learning and artificial intelligence (AI) to analyze large amounts of data and make predictions about future events or uncertain choices (Agrawal et al., 2018; Brynjolfsson and McElheran, 2016; Brynjolfsson et al., 2021). Beyond their use in informing operational decisions, data-driven approaches are becoming common in innovation (Cockburn et al., 2019). 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, 2025), and pharmaceutical firms try to predict the therapeutic properties of millions of potential drugs (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 enables firms to anticipate the value of alternative choices and recognize opportunities, thereby enhancing performance (Cohen and Levinthal, 1990; Gruber et al., 2008; Shane, 2000). However, data-driven predictions can serve a similar function without first accumulating domain knowledge (Agrawal et al., 2024; Balasubramanian et al., 2022). Since obtaining predictions is now cheaper and faster than developing expertise, these technologies could disproportionately benefit firms lacking domain knowledge, thus narrowing performance gaps across firms. Consistent with this idea, studies of algorithm-augmented work show that predictive tools can compensate for a lack of expertise and generate larger gains for low performers (Brynjolfsson et al., 2024; Doshi and Hauser, 2024; Noy and Zhang, 2023). Extrapolated to firm innovation, this logic would imply that data-driven predictions can substitute for domain knowledge as a source of competitive advantage.

However, this democratization effect is most plausible for well-defined tasks in which problems fall within the technology's capability frontier and predictions can be implemented with limited additional judgment (Dell'Acqua et al., 2026). These conditions are less likely to hold in innovation search, where firms use predictive tools to prioritize unfamiliar technological combinations (Agrawal et al., 2024). In such settings, the challenge is not only one of discovery, but also one of evaluation (Levinthal and Schliesmann, 2025), because the value of a lead may remain uncertain even after it is identified. Especially in more complex domains, predictions can therefore be informative while still highly imprecise. They may generate frequent *false positives*, namely candidates that look statistically compelling but do not correspond to valuable opportunities (Berman and Van den Bulte, 2022). Because technological innovation often requires substantial investment, broader access to predictions shifts the core challenge from finding leads to determining which ones are worth pursuing.

Motivated by this logic, I argue that domain knowledge complements data-driven predictions in innovation search. Predictive technologies infer statistical associations from codified observations, whereas domain knowledge reflects contextual judgment grounded in practice (Anthony, 2021). The two, therefore, rely on epistemically distinct forms of knowledge to assess opportunities (Knorr-Cetina, 1999). This distinction matters because the value of a technological combination depends on mechanisms and boundary conditions that are often revealed only through direct engagement with a domain (Dougherty and Dunne, 2012; Gittelman, 2016). Such contextual understanding is tacit and "sticky" (Polanyi, 1966; Von Hippel, 1994), and remains difficult to encode in predictive models even as they improve. Domain knowledge helps firms interpret statistical signals in light of these omitted features, allowing them to judge whether statistically persuasive associations are also substantively plausible. As a result, when spurious patterns are common, broader access to predictions may ultimately increase the competitive advantage of firms with deeper domain knowledge.

Empirically assessing how data-driven predictions and domain knowledge interact to shape innovation performance poses several challenges. First, it requires observing the risk set of potential opportunities that firms could pursue, which is usually difficult to ascertain. Second, this measurement problem is compounded by the fact that the accuracy of a prediction cannot be learned when no firm chooses to act on it. Finally, the empirical setting must allow identification of which specific predictions were available at the time of decision-making and how they aligned with a given firm's knowledge. This last challenge creates a crucial identification issue: if firms with domain knowledge systematically receive better predictions, it becomes impossible to separate prediction quality from the ability to evaluate them correctly.

I address these challenges by focusing on pharmaceutical innovation, a setting whose empirical features are ideally suited for this paper. The drug development process begins with the challenging task of identifying genes that cause disease, as they can serve as drug targets. The recent emergence of genomics has enabled a new approach to locating 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 associated with the disease. In practice, they yield predictions on which genes could be 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 researchers (Uffelmann et al., 2021). Second, GWAS are primarily conducted by academic researchers. These gene-disease associations are typically unanticipated and become observable to all firms only after publication in scientific journals. Third, GWAS uncover both breakthrough and false positive genetic targets. To separate them, I exploit the fact that diseases are often the subject of multiple GWAS, each with 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, regardless of whether they actually did.

This setting allows me to study how firms leverage data-driven predictions, but it also creates a measurement challenge. The average GWAS in my data introduces 14 new gene-disease associations, 12 of which are false positives. Conventional measures, such as patent-to-paper citations, cannot reveal which specific association within a GWAS publication a firm is pursuing. To address this issue, I use a text-based approach that extracts the gene-disease pairs targeted in each patent and maps them onto an empirical landscape of all possible combinations. This approach tracks firms' investment across the search landscape before and after GWAS findings, without relying on patent citations. I then use firms' publications to identify which genes they had previously researched in relation to other diseases. As confirmed in my interviews, firms with an active research program on a gene can use this knowledge to assess GWAS findings even if that expertise was developed for another disease area. The resulting design estimates the impact of data-driven predictions separately for firms with and without domain knowledge.

I find that patent applications targeting a gene-disease combination more than double after a GWAS reports it. Event study specifications confirm the absence of pre-trends and support the validity of the research design. This first result shows that data-driven predictions alter the direction of firm investments. However, the aggregate response also reveals the evaluative challenge posed by abundant predictions: more than one-third of the increase in patenting is directed toward false positive findings that do not translate into downstream outcomes such as new drug molecules. I then examine how this response varies with firms' domain knowledge. Firms lacking gene-specific expertise react more strongly to GWAS, consistent with the idea that data-driven predictions partly compensate for the absence of prior knowledge in guiding search. Yet these firms are also disproportionately more likely to pursue false positives, unlike firms with stronger genetic expertise.

To unpack the mechanisms, I classify firms' prior research according to the biological system studied, distinguishing *in vitro* and animal model studies from human-based research. Suggestive results show that the ability to avoid false positives is strongest when prior knowledge is more patient-oriented. Additional tests suggest that the results are not driven by domain expert firms being better

at processing statistical evidence or more scientifically capable in general. Rather, such firms possess context-specific knowledge that GWAS do not encode, but that helps them judge whether a statistically significant association is also biologically meaningful. The results suggest that data-driven predictions changed the role, rather than the importance, of domain knowledge: they shifted its value from finding opportunities to evaluating them.

I further investigate the competitive implications of these findings using a firm-level panel design. I construct a measure of GWAS exposure from firms' pre-2005 disease portfolios, capturing the flow of data-driven opportunities relevant to each firm's existing pipeline. Because these prior disease choices are not randomly assigned, I interpret this analysis as descriptive; nevertheless, the panel structure allows me to control for firm and year fixed effects and examine whether the earlier results translate into performance differences. Strikingly, I find that data-driven predictions lead to divergence in innovation outcomes. GWAS exposure increases drug candidates and commercialized therapies among firms with stronger genetic expertise, but it mainly drives additional investment in false positives among firms with limited expertise. Taken together, these findings show that GWAS expanded the supply of plausible targets without equalizing firms' ability to convert them into valuable innovation.

This paper makes three contributions. First, it adds to a growing research agenda on how predictive tools such as AI will shape innovation (Agrawal et al., 2024; Allen and McDonald, 2025; Cockburn et al., 2019; Conti et al., 2024) and their limits (Hoelzemann et al., 2024; Kim, 2025). Rather than necessarily equalizing innovation by making opportunities visible to lower performers (Nagaraj, 2022), data-driven predictions can generate false positives that widen performance gaps. Second, the paper contributes to research on firm capabilities and absorptive capacity (Cohen and Levinthal, 1990; Gambardella, 1992; Nelson and Winter, 1982). The results suggest that domain knowledge is valuable because it provides contextual understanding that predictions may not reflect (Choudhury et al., 2020; Dougherty and Dunne, 2012). This also clarifies the relevance of my findings as predictive tools improve: domain knowledge remains important when evaluating opportunities hinges on sticky or tacit knowledge that cannot be encoded in data. Third, the paper speaks to research on innovation sourcing beyond firm boundaries (Chesbrough, 2003; Laursen and Salter, 2006). While this literature highlights the value of expanding the set of ideas firms consider, it also shows that evaluation challenges shape which external ideas receive attention (Piezunka and Dahlander, 2015). My evidence points to a related risk: data-driven predictions can give such leads a veneer of certainty, encouraging firms to back ideas that ultimately prove unproductive. This insight also relates to emerging work on AI as a source of ideas (Boussioux et al., 2024). The value of these "black-boxed" tools is likely to depend on whether firms possess the judgment needed to separate valuable opportunities from false positives.

2 Theoretical Framework

2.1 Innovation Search as a Prediction Problem

The ability to identify promising opportunities is central to firm performance. This is especially true in innovation, which involves the recombination of technological components into new configurations (Fleming, 2001; Katila and Ahuja, 2002; Nelson and Winter, 1982). A common way to conceptualize this process is as a search over a technological landscape, where each location corresponds to a distinct combination of components and differs in its potential value (Fleming and Sorenson, 2001; Levinthal, 1997). Such landscapes are typically vast, so that much of the opportunity space lies beyond the firm's direct experience. Technological uncertainty arises because firms must evaluate unfamiliar components and novel combinations of unknown value (Fleming, 2001). Since firms cannot experiment with all possible configurations, they must decide how to allocate their limited resources toward the opportunities that appear most promising.

As a result, firms do not search at random but rather in areas where they expect the highest returns (Felin and Singell, 2025; Kneeland et al., 2020). In this sense, innovation in a combinatorial landscape can be understood as involving a *prediction problem*: firms must first predict which technological combinations are most likely to be valuable and use those expectations to focus their search. Put differently, innovation depends on the ability to assess alternatives "offline," distinguishing productive paths from likely dead ends before incurring the cost of testing them (Fleming and Sorenson, 2004; Gavetti and Levinthal, 2000). This echoes a broader idea in strategy, where prior research links performance to superior foresight about the value of alternative courses of action (Csaszar and Laureiro-Martínez, 2018; Kapoor and Wilde, 2025). Applied to innovation, such foresight means predicting where valuable combinations are likely to be found. Firms that make these predictions more accurately can direct investment toward promising regions of the technological landscape and gain a competitive advantage.

Researchers have long examined the sources of predictive capabilities in innovation. Prior work documented how domain knowledge permits decision-makers to understand the mechanisms underlying why some alternatives are more valuable than others (Arora and Gambardella, 1994a; Felin and Zenger, 2017), improving their ability to recognize value ahead of the competition (Camuffo et al., 2024; Gruber et al., 2008; Shane, 2000). At the firm level, the same mechanism underpins absorptive capacity, as deeper domain knowledge permits the identification of opportunities (Cohen and Levinthal, 1990). This logic has long motivated firms to accumulate specialized knowledge bases (Cohen and Levinthal, 1994; Gambardella, 1992). However, a key insight from the open innovation literature is that firms can also search for innovation beyond their knowledge boundaries (Chesbrough,

2003). By sourcing expertise externally, firms can broaden the set of combinations they consider and identify promising ideas (Laursen and Salter, 2006; Leiponen and Helfat, 2010). External search provides an alternative source of predictions about where valuable opportunities lie, thus substituting for the lack of domain knowledge.

2.2 The Impact of Data-Driven Predictions on Innovation

Against this backdrop, the growing availability of big data and predictive technologies[1] offers an alternative approach to predicting the value of untested technological combinations (Agrawal et al., 2024; Cockburn et al., 2019; Lou and Wu, 2021). Instead of relying on human expertise, firms can leverage large datasets to infer which untested combinations are more likely to warrant investment before committing resources (Nagaraj, 2022). Crucially, the usefulness of these predictions does not depend on the focal alternative being already theoretically characterized or of known function. Patterns in the data can be extrapolated to technological combinations whose underlying mechanisms remain poorly understood (Kim, 2025). In this sense, predictive technologies provide a theory-free way of prioritizing search across technological landscapes for which relevant data are available (Tranchero, 2026).

Predictive technologies fundamentally change the epistemic nature of predictions used to guide search. Like externally sourced ideas, they do not require that the firm itself has developed the relevant understanding through prior learning and experimentation. However, they do not arrive as contextualized judgments from outside experts, but as statistical signals inferred from codified observations (Choudhury et al., 2020; Boussioux et al., 2024). This has two implications. First, prediction becomes extremely cheap and scalable, allowing the rapid screening of larger portions of the landscape (Agrawal et al., 2024). Second, the basis for any given prediction becomes harder for the focal actor to interrogate or verify (Anthony, 2021; Lebovitz et al., 2022). Because the prediction is detached from articulated causal reasoning, an opportunity may appear promising without revealing why, rendering it effectively a "black box" (Ke et al., 2024). As a result, users may struggle to assess a prediction's precision, because they cannot easily determine whether it falls within the technology's "jagged frontier" of reliable performance (Dell'Acqua et al., 2026).

These features have opposing implications for innovation. On the one hand, exposure to more leads increases the likelihood that firms will encounter combinations worth investing in (Laursen and Salter, 2006; Leiponen and Helfat, 2010). When the ex ante probability of success is extremely low, as

[1] I define predictive technologies as algorithms designed to use large amounts of data to extrapolate information about options of unknown value (Agrawal et al., 2018; Brynjolfsson et al., 2021). This includes tools ranging from simple correlations to deep learning and AI, as long as they have diagnostic value, that is, they predict the success of an option better than the prior (Arora and Fosfuri, 2005). As such, data-driven predictions are distinct from descriptive uses of data analytics.

in technological innovation, even weak signals can stimulate investment by de-risking choices (Kao, 2025; Nagaraj, 2022). On the other hand, expanding the set of candidates does not necessarily translate into more innovation. While extrapolating leads from past data becomes increasingly cost-effective, turning predictions into outcomes remains an organizational challenge (Kim et al., 2024). Firms are still constrained by the physical (Fontanelli et al., 2026) and cognitive (Piezunka and Dahlander, 2015) resources they can devote to validating and developing ideas. Moreover, predictions may contain false positives (Berman and Van den Bulte, 2022). As a result, the same technologies that broaden search could also divert resources toward combinations that appear deceptively promising in the data. Whether data-driven predictions ultimately lead to more innovation is therefore unclear.

2.3 False Positive Predictions and Competitive Advantage in Innovation

The preceding discussion raises a natural follow-up question: how do data-driven predictions reshape competitive advantage in innovation? To the extent that they can substitute for predictions generated through domain expertise (Yerramilli-Rao et al., 2025), their diffusion should mostly benefit firms with limited knowledge bases. Evidence from algorithm-augmented work, although largely confined to individual performance on specific tasks, is consistent with this possibility. A recurring finding is that algorithmic tools generate larger gains for lower-skilled or less experienced users (Allen and Choudhury, 2022; Brynjolfsson et al., 2024; Noy and Zhang, 2023). In one study examining open-ended creative work, Doshi and Hauser (2024) find that access to generative AI ideas benefits writers who are inherently less creative. If these results extend to firm innovation, data-driven predictions would attenuate performance differences traditionally rooted in domain knowledge (Krakowski et al., 2023).

This logic, however, assumes that data-driven predictions convey the same kind of insight as domain expertise. Research on the epistemology of data-derived knowledge suggests otherwise. Because predictive technologies operate on codified information, they generate a form of knowledge distinct from the understanding accumulated through direct experience in a domain (Knorr-Cetina, 1999; Nelson et al., 2011). Digital sciences produce insight by manipulating data and representations rather than physical materials or organisms (Dougherty and Dunne, 2012). By contrast, hands-on engagement with physical reality can yield valuable knowledge that is difficult to detach from the context in which it was produced (Nightingale, 2004). This knowledge is "sticky" when its interpretation depends on local conditions, making it unlikely to extrapolate cleanly across settings (Von Hippel, 1994). It is also often tacit: the relevant judgments cannot be fully articulated, making them difficult to convert into variables that enter predictive models (Polanyi, 1966; Cowan et al., 2000). Whenever these contextual or uncodified dimensions matter for assessing value, predictions based only on codified data will necessarily remain noisy (Choudhury et al., 2020).

This problem is likely to be especially acute in technological innovation. The value of an opportunity depends on complex interactions among components, which makes most technological landscapes rugged (Fleming and Sorenson, 2001). In such settings, "predictability" emerges only under specific conditions that are often local and difficult to reproduce (Nightingale, 2004). In medicine, for instance, this challenge is salient because predictive search abstracts from the natural complexity and variability of biological systems (Anckaert et al., 2020; Nelson et al., 2011). One chemist in a pharmaceutical company captured the limits of this logic well: "*We thought that by sheer numbers, a gem would pop out. No, that did not happen.*" (Dougherty and Dunne, 2012, p. 1218). As Gittelman (2016) argues, more experiential forms of inquiry preserve the contextual learning needed to judge whether a lead is meaningful in practice. Without incorporating such tacit and contextual understanding, even informative data-driven predictions may remain structurally imprecise, yielding false positives that distort innovation search.

Taking the precision of data-driven predictions seriously changes expectations about which firms benefit most from their diffusion. When predictions are highly accurate, they may indeed substitute for domain expertise in opportunity recognition and reduce performance heterogeneity across firms. When predictions are noisy, however, they ease the problem of discovery while complicating the problem of evaluation (Levinthal and Schliesmann, 2025). In such settings, domain knowledge can help firms interpret statistical signals in context and screen out leads that appear promising in the data but are in fact false positives. To the extent that domain knowledge informs both the identification and evaluation of opportunities (Cohen and Levinthal, 1990, 1994), firms with deeper knowledge bases may continue to outperform in innovation, even when that knowledge was accumulated for other purposes. This advantage may grow as predictive signals proliferate, because a larger volume of plausible leads intensifies the evaluation problem (Piezunka and Dahlander, 2015) and increases the returns to firms with stronger evaluative capabilities (Arora and Gambardella, 1994b). The same data-driven predictions that broaden access to opportunities may thus increase performance differences among firms that differ in their ability to evaluate them.

Overall, this discussion highlights that when data-driven predictions become cheap and abundant, competitive advantage in innovation hinges on selecting among them. The rest of the paper tests this argument in the pharmaceutical setting by examining which firms benefit more from data-driven predictions of potential drug targets.

3 Empirical Setting

The first step in drug discovery is choosing the right genetic target: firms must identify genes whose products can be modulated by a drug to treat a specific disease (Nelson et al., 2015).[2] This task is complex because, in principle, diseases could be caused by mutations in any of the over 19,000 protein-coding genes. Common diseases are often polygenic—for instance, diabetes has been linked to over 150 mutations across multiple genes. The result is a combinatorial space of tens of millions of potential gene-disease combinations, of which very few have actual therapeutic value. 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, 2025). Therefore, pharmaceutical firms rely on scientific research to guide target selection (Arora and Gambardella, 1994a; Fleming and Sorenson, 2004).

Starting in the mid-2000s, the completion of the Human Genome Project and the decline in the cost of collecting genetic data prompted the emergence of GWAS (Visscher et al., 2017). GWAS are case-control studies that compare human genomes to identify genetic mutations that occur more frequently in individuals with a given condition than in healthy controls (Uffelmann et al., 2021). Figure 1 shows the abstract of a GWAS and a 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 those subjects. Lastly, they test for statistically significant differences in genotypes that correlate with the disease. Association tests are adjusted for multiple hypothesis testing by imposing a high threshold for statistical significance. Intuitively, by comparing the genetic makeup of people with and without a condition, one can predict which genes are involved in a given disease and thus serve as drug targets (see Appendix A for details).

GWAS generate data-driven signals about gene-disease combinations that may prove valuable for R&D, and anecdotal evidence points to several drugs developed on the basis of such findings (Visscher et al., 2017). Nevertheless, there is ongoing controversy over their ability to identify targets with therapeutic potential (Tam et al., 2019; Uffelmann et al., 2021). One major limitation is that GWAS cannot explain the mechanisms underlying gene-disease associations, raising the possibility of false positive discoveries with no biological meaning (Goldstein, 2009). A single GWAS can identify dozens of associations, many of which are spurious or therapeutically irrelevant (MacArthur, 2012). This creates a practical challenge for scientists and firms pursuing therapeutic applications, because testing any given lead is costly. The same concern surfaced repeatedly in my interviews with industry

[2]Genes are segments of DNA that carry the instructions that encode products, such as proteins, that carry out biological functions. Mutations in these sequences can alter gene behavior with important consequences for human health. Identifying the genetic basis of disease is crucial for drug discovery because genes whose products contribute to disease processes can be targeted by therapeutic interventions.

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). In practice, it remains unclear whether GWAS generate useful knowledge for drug discovery.

4 Data and Measurement Strategy

4.1 Measuring the Impact of GWAS on an Empirical Search Landscape

The empirical analysis requires identifying which firms invest in the associations uncovered by GWAS. However, information on R&D spending is typically available only at the firm level, making it unsuitable for tracing investment in specific gene-disease pairs. Instead, I rely on patent applications to recover project-level information in firms' investment choices (Gambardella, 1992). Following Eggers and Kaplan (2009), I treat patenting as a proxy for early-stage investment in a technological domain. This measure is particularly appropriate for pharmaceuticals, where firms typically apply for patents well before it is known whether a drug candidate will succeed in the clinic. Patent applications thus capture where firms direct investment in response to GWAS associations.[3] I then use data on new drug development to examine which of these upstream investments ultimately translate into innovation.

However, linking firm patents to the specific gene-disease associations uncovered by GWAS poses an additional measurement challenge. Traditional approaches based on patent-to-paper citations are not suited to my setting for two reasons. First, a typical GWAS reports multiple gene-disease associations in the same publication. A citation from a patent to the GWAS paper does not reveal which specific association the invention builds on. Second, citations are incomplete measures of how foundational science shapes innovation. A patent may rely on a GWAS finding indirectly, for example, by citing later validation work rather than the original GWAS publication. Thus, citation-based measures are too coarse to measure which gene-disease prediction a firm is pursuing and to trace the downstream impact of GWAS discoveries.

To address these challenges, I draw on the recombinant view of innovation (Fleming, 2001) and treat each GWAS association as a prediction of a valuable gene-disease combination. This suggests that the impact of a GWAS is proxied by the change in firm patenting for gene-disease pairs "treated" compared to otherwise similar pairs that were not reported. Named-entity recognition algorithms can then be used to extract genes and diseases from each patent application, allowing me to identify which gene-disease combinations a firm is targeting in its investments. Figure 2 offers an intuitive example

[3] A natural concern is that patent applications omit failed projects or abandoned research. For my purposes, this makes the measure conservative: if GWAS findings redirect firms toward particular gene-disease pairs, the resulting change in patenting should be interpreted as a lower bound on the research effort devoted to those opportunities (Eggers and Kaplan, 2009). Because pharmaceutical firms patent before clinical success is known, patent applications are better viewed as indicators of early-stage commitment than as measures of successful innovation. Consistent with this interpretation, only 49.4% of patent applications in my data result in a patent grant.

of how GWAS shape firm investments in a combinatorial landscape of genes and diseases. Under a standard parallel trends assumption, the before-and-after changes in the number of patents mentioning treated and control combinations are an estimate of the causal effect of GWAS.[4]

This measurement approach offers three advantages over traditional citation-based methods. First, using text-based knowledge entities allows me to distinguish the different contributions made in the same paper. This granularity is essential for comparing the differential effect of false positives and therapeutically meaningful associations reported in the same study. Second, by tracking mentions of genes and diseases directly in the patent text, this approach can capture the impact of basic research findings that are not explicitly cited by patent applications. Third, the same logic can be applied to newly discovered drugs by collecting data on the gene and disease targeted by each molecule (Kang, 2024; Kao, 2025). This allows me to map drug discovery within the same combinatorial landscape and trace the innovation process from GWAS predictions to downstream outcomes.

4.2 Data Sources

To study how firms respond to the arrival of data-driven predictions, three ingredients are required. First, I need data on the 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 identify the specific genes and diseases targeted by each patent and drug in order to link them to GWAS findings. I summarize each data source below and provide details in Appendix C and D.

GWAS Catalog: Information on GWAS findings comes from the GWAS Catalog, a source manually curated by the European Bioinformatics Institute (EBI) and the National Human Genome Research Institute (NHGRI). The Catalog is a comprehensive list of GWAS published in peer-reviewed journals, starting with the first in 2005. Studies are eligible for inclusion in the GWAS Catalog only if they include a genome-wide scan that does not target any specific gene ex-ante. The GWAS Catalog collects details of each gene-disease association. Genes are identified by their 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.[5] My sample includes 17,965 gene-disease associations first reported by 1,259 distinct GWAS between 2005 and

[4]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 (2025) adopts the same logic to study the effect of large-scale cancer maps on the number of clinical trials. The approach detailed in Appendix B formalizes their intuition.

[5]Since the MeSH taxonomy is a hierarchical tree, 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), I assigned it to its parent branches up to level four. Vice versa, if the matched disease was coarser (i.e., at level three), I assigned the finding to all its descending level-four branches. This procedure allows harmonizing GWAS diseases to the same level of specificity. The sample is further restricted to diseases that receive more than one GWAS because I exploit subsequent GWAS to code which associations are not replicated and are thus likely to be false positives (as explained in Section 4.3).

2019. These associations span 404 unique diseases and 5,080 protein-coding genes.

Patent Data: Through a partnership with EBI, I obtained proprietary data on the gene-disease pairs mentioned in the text of each USPTO patent application (2001–2019).[6] The data have been compiled from complete patent texts using TERMite (TERM identification, tagging, and extraction), a named entity recognition software developed by SciBite. TERMite directly maps the extracted entities to 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%, indicating that it is highly reliable. I merge these data with information on assignees and patent characteristics from PatentsView. Commonly used indicators of patent quality come from the OECD Patent Quality Indicators Database. My final sample includes 4,027 companies applying for at least one USPTO patent between 2001 and 2019.

Drug Discovery Data: I use proprietary information on drug molecules collected by Clarivate's Cortellis Database (Krieger, 2021). My data include 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 list of development milestones for each drug molecule. I consider new discoveries to be the molecules observed entering the earliest phases of drug development (which 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. Next, I use the FDA's *Orange Book* to identify patents linked to approved small-molecule drugs, thereby revealing which molecules reached commercialization by 2025.

Publication Data: I use data from PubTator Central, which provides computer-annotated gene and disease information for each PubMed paper. PubTator also maps genes and diseases to NCBI Gene IDs and MeSH Unique IDs. Next, I match each patenting firm in my sample to its publication records using information on authors' affiliations. Doing this enables me to code which firms had biological expertise on a specific human gene before deciding whether to invest based on a GWAS finding. This procedure yields a granular measure of gene-specific knowledge, 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 iCite database curated by the National Institutes of Health (NIH).

These data allow me to reconstruct the broader search landscape faced by pharmaceutical firms and locate each firm within it. Table 1 shows substantial heterogeneity. The median firm targets 19

[6]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, methods of use, or new drug molecules that target a specific gene to treat a given disease. See Appendix C for details.

genes and 27 diseases in its patents, while its scientific publications span a broader portfolio of 52 genes and 65 diseases. I then define the relevant search landscape as the Cartesian product of the 404 diseases and 17,881 protein-coding genes in my data, yielding 7,223,924 possible gene-disease pairs. Around 21.8% of these possible pairs receive investment, proxied by their appearance in at least one patent application, while only 0.19% advance to the drug development stage. The average gene-disease pair receives 0.13 patent applications per year. Around three-quarters of patents are filed by firms that had not previously published research on the targeted gene, consistent with prior work showing that public data are disproportionately used by firms entering new areas (Nagaraj, 2022).

4.3 Identifying False Positive Associations

GWAS perform a data-driven search for genetic mutations associated with human diseases (Tranchero, 2026). This feature is both a strength and a weakness: despite stringent significance thresholds, many GWAS findings turn out to be statistical noise or spurious correlations. For this paper, I need to determine empirically which GWAS associations are false positives. The challenge is that ex post assessments of association quality are typically available only for findings that receive investment, creating a missing-data problem. If association quality became observable only when firms invested, I would observe the realized value of a prediction only for the GWAS findings that firms chose to pursue. In my setting, this would make it impossible to know whether GWAS associations that receive no attention were correctly avoided because they had low potential or mistakenly overlooked because firms failed to recognize their value.

To distinguish true positives from false positives, I exploit the fact that many diseases receive multiple GWAS over time. Specifically, I classify a gene-disease association as a likely true positive if it is later confirmed by at least one subsequent GWAS on the same disease. This provides an intuitive way to identify false positives regardless of the level of investment they received. Put differently, each subsequent GWAS serves as a retest of prior findings at the genome-wide level, allowing me to assess even associations that firms never pursued.[7] Using this approach, I find that 84.3% of the 17,965 gene-disease associations in my sample fail to replicate in subsequent studies. This high rate is consistent with concerns among experts about the risks of false positives in genomic big data (MacArthur, 2012).[8] Non-replicable findings can lead firms to misallocate scarce R&D resources

[7] This approach also yields a conservative measure of false positives in a therapeutic sense, because even replicable associations may still fail to offer viable avenues for treatment (Goldstein, 2009).

[8] The limited applicability of GWAS findings is well recognized in the field. For example, Marigorta et al. (2018) estimate that roughly 40% of associations can replicate when also considering findings in intergenic, non-protein-coding regions. The lower rate in my analysis is driven by my focus on protein-coding genes, which are more relevant to drug development. One reason replicability can be lower in this subset is that GWAS authors often face difficulties in mapping mutations to the correct causal genes (Visscher et al., 2017). I exploit this type of reporting error as an alternative empirical strategy in Subsection 5.3.

toward targets with no therapeutic potential.

I validate my new measure of false positives in three ways. First, I examine which study-design characteristics predict whether GWAS discoveries replicate in subsequent studies. Appendix Table E.1 supports the intuition that more robustly designed GWAS are less likely to report associations classified as false positives. Second, I compare my coding with a score summarizing the strength of experimental evidence for a specific gene-disease association as of 2024 (details in Appendix D.5). Appendix Table E.2 shows that replicable gene-disease associations are 114%–191% more likely than non-replicable ones to rank in the top decile of this measure. Finally, I present evidence in Appendix Table E.3 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, as captured by the Scite data (Nicholson et al., 2021). Together, these tests corroborate the approach used to identify false positives in GWAS associations.

5 Empirical Analysis

5.1 Research Design

There are two challenges to assessing the impact of data-driven predictions on innovation. The first challenge arises from the difficulty of determining which predictions a firm has access to. This is problematic when firms with domain knowledge systematically obtain better predictions, thereby conflating the prediction quality with their ability to interpret them. The second challenge relates to identifying the causal effects of data-driven predictions. Firms are likely to apply their predictive tools where they expect the highest returns, potentially leading to upwardly biased estimates. An ideal experiment would bypass these issues by assigning to every firm identical predictions about gene-disease associations. Their causal effect would then be evident from changes in patenting involving "treated" gene-disease pairs relative to the others.

I approximate this ideal experiment using the staggered publication of GWAS in scientific journals. First, GWAS are mainly conducted by academic research teams, and their findings are publicly accessible upon publication. As such, their associations are available to every pharmaceutical firm simultaneously and are not driven by unobservable proprietary data or expertise.[9] Second, GWAS scan for genetic mutations across the whole genome. This method, by design, avoids endogenous sorting because it does not focus on specific genes (Uffelmann et al., 2021). New GWAS discoveries are unforeseen by the research team and even more so by firms that learn about them in journals.

[9] In my sample, 324 GWAS are coauthored by corporate scientists, and another 117 acknowledge funding from companies. All results are robust to excluding these studies, as well as those authored by academics with frequent industry collaborations (Appendix Table E.8).

Importantly, unbeknownst to the scientists conducting the GWAS, only some of their findings will be confirmed in future GWAS. This presents an opportunity to study how firms react to true and false positive associations and how it depends on their 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 $Y_{i,j,t}$ are patent applications by firms in year t involving gene i and disease j. $Post_t \times GWAS_{i,j}$ 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 underlying potential. I control for time-varying disease-level differences 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. Standard errors are clustered two-way by gene and disease.[10] 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 that GWAS may uncover gene-disease pairs that would have attracted investment even in the absence of these findings. However, GWAS differ from targeted scientific studies in that they scan the full genome (Tranchero, 2026). Conditional on the disease under study, their genome-wide design implies, by definition, no gene-level selection in which associations are uncovered. The timing of these findings is thus plausibly exogenous to trends in the dependent variables. Indeed, the temporal order in which gene-disease pairs are reported by GWAS is not shaped by prior inventive activity on those same pairs (Appendix Figure E.1). In what follows, I exploit within-disease variation in the timing of GWAS publications to estimate the causal effect of GWAS on firms' early-stage innovation investments. Tests for pre-trends in the outcome variables further assess the validity of this identification strategy.

5.2 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 publishing gene-disease associations would increase firms' willingness to invest in the corresponding combinations. Column 1 of Table 2 shows the baseline results. The main finding is that GWAS have a large impact, increasing patent applications targeting treated gene-disease pairs by 125% on average. The design of GWAS should ensure that the effects reported are not due to researchers endogenously reporting the most promising gene-disease pairs. Figure 3 directly checks the validity of this assumption with an event study version of Equation

[10] 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.

1. Panel (a) confirms flat pre-trends and a persistent effect after the GWAS is published, with the estimates stabilizing around the coefficient of Table 2. Heterogeneity analyses in Appendix Table E.4 reveal greater increases in patenting for associations that are statistically more robust.

This aggregate effect, however, masks substantial heterogeneity. Because GWAS are known to contain many false positive associations, the key question is whether firms can avoid investing in findings that later prove not robust. Column 2 of Table 2 shows that false positive GWAS findings increase patenting by 0.0676 applications per gene-disease-year, corresponding to a 51% increase relative to the sample mean. Given that 84.3% of the associations in my sample fail to replicate in subsequent GWAS, this estimate implies that roughly 35% of the total increase in patenting is directed toward false positive associations.[11] Panel (b) of Figure 3 helps interpret this pattern. Patenting on false positives rises immediately after publication but then declines, consistent with firms initially treating these findings as promising and later redirecting efforts as their lack of robustness becomes apparent. Results are robust to the use of alternative difference-in-differences estimators (Appendix Figure E.3).

How costly is it for a firm to invest on the basis of a false positive finding? While patent fees and legal costs are limited to a few tens of thousands of dollars (de Rassenfosse and Jaffe, 2025), the R&D spending behind each patent application is likely much larger. Tabulations from the 2022 Business Enterprise Research and Development (BERD) Survey indicate that U.S. pharmaceutical firms spent $92.25 billion on R&D and filed 13,486 patents, implying roughly $6.8 million of R&D per application.[12] Whether these investments translate into innovation, however, depends on the validity of the genetic target pursued. Appendix Table E.6 shows that innovation rises only for GWAS associations that are later replicated. In particular, true positive findings increase new drugs discovered and commercialized therapies, whereas false positive findings do not yield comparable downstream outcomes. Taken together, these results suggest that GWAS stimulate investment, but their contribution to innovation is more nuanced because they also yield many false positives. When such leads are not screened out, they can absorb substantial resources and hurt firms' innovative performance.

[11] In principle, patenting on false positive leads may be strategically valuable even if it does not result in downstream innovation. Appendix Table E.5 suggests otherwise. Relative to patents based on true positive GWAS findings, patents induced by false positive predictions are less likely to involve strategic continuation behavior, be litigated, or even be renewed. This pattern is more consistent with lower-value innovative effort than with strategic patenting.

[12] These figures are only intended as a rough benchmark to contextualize the cost of false positive GWAS leads. Reassuringly, an industry-sponsored report by NDP Analytics estimates that, during 2012–2022, the pharmaceutical industry invested an average of $12.7 million per patent grant (Pham, 2025). Given that only 49.4% of patent applications in my data are eventually granted, this estimate aligns with my back-of-the-envelope calculation based on BERD.

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

Next, I examine whether firms respond differently to the same data-driven predictions depending on their prior domain knowledge. I measure this knowledge at the gene level using PubTator. Firms differ in the regions of genetic space they study, with the median firm's publication portfolio spanning 52 genes. For each gene-disease association, I code a firm as having domain knowledge if it had published on the focal gene before the GWAS reports it. Since GWAS introduce new associations, this measure captures prior knowledge of the gene, not of the specific gene-disease pair. Such knowledge is developed in other disease contexts but can still help firms assess whether the new association is meaningful. I then compare how firms react to GWAS findings involving genes inside versus outside their prior knowledge base.[13] In practice, I estimate the effect of GWAS on patenting separately for firms with genetic knowledge on the focal gene and firms without it.

In Table 2, Columns 3 and 5 compare the patenting response of firms with and without prior publications involving the focal gene. Both groups increase investment following a GWAS finding, but firms lacking gene-specific knowledge respond more strongly. Relative to the sample mean, the estimates imply that the proportional increase in patenting is about one-third larger for firms that had not previously researched the gene. More importantly, Columns 4 and 6 show that the increase in patenting directed toward false positive associations is significant only for firms without prior publications on the gene. By contrast, domain-expert firms respond primarily to findings that are later replicated. This pattern is consistent with the idea that gene-specific domain knowledge helps firms evaluate the substance of a GWAS rather than merely react to its publication. In the absence of such knowledge, firms are more prone to type I errors, directing inventive effort toward associations that appear promising in the data but ultimately do not represent valuable opportunities.

Several robustness checks reinforce this interpretation. First, the heterogeneity results are robust to using alternative difference-in-difference estimators (Appendix Table E.7). Second, the value of domain knowledge should be greater when GWAS findings are harder to evaluate. Consistent with this, firms without domain knowledge are more likely to invest in false positives when the focal genes are less studied (Appendix Table E.9), but they make fewer mistakes when a GWAS reports fewer associations to assess (Appendix Table E.10). Third, I validate the main findings using an alternative measurement strategy. After identifying mutations associated with a disease, GWAS authors must map those mutations to the correct gene (Visscher et al., 2017). This step was imprecise in earlier studies

[13] For example, Denali Therapeutics focuses on what it calls "degenogenes" (Appendix C.3). When a new gene-disease association involving one of these genes is published, Denali's scientists are well-positioned to assess it, even if the specific association is new. As one interviewee at Denali put it, whether to pursue a new target appearing in the literature is an assessment based on extensive domain expertise: "*We are so entrenched in the neuro field that we kind of know off-the-cuff*" (Interview, 17 October 2022). Appendix Figure E.2 illustrates the corresponding research design.

and associations could be attributed to the wrong gene (usually a neighboring one). In my data, there are 9,273 cases in which the original gene assignment was later corrected by GWAS Catalog curators. Confirming my primary results, Table E.11 shows that only firms without domain knowledge increase investments in response to these wrongly reported findings. This additional test has the advantage of not depending on my method for identifying false positives in the primary analysis.

5.4 Mechanisms

Having shown that gene-specific knowledge improves investment allocation, I next explore the underlying mechanisms. Prior work on drug discovery emphasizes epistemic differences in the form that knowledge takes, distinguishing between a data-based mode of knowing and one rooted in direct engagement with biological phenomena (Dougherty and Dunne, 2012; Knorr-Cetina, 1999). Table 3 provides a first indication that my proxy of domain knowledge captures the latter. Firms without prior gene-specific knowledge are especially likely to respond to false positive GWAS findings when those findings carry markers of statistical credibility, such as larger sample sizes or smaller p-values. Expert firms, by contrast, are less likely to pursue these same false positives even when they look more compelling. This pattern is difficult to reconcile with an interpretation based solely on superior statistical sophistication. Instead, it suggests that these firms evaluate GWAS findings against criteria that are not contained in the signal itself and, on that basis, discount biologically implausible associations.

To characterize this type of knowledge more precisely, I classify firms' publications based on the translational profile of each article, as coded by NIH's iCite (Hutchins et al., 2019). This measure is constructed from the MeSH terms assigned to each publication and records the biological system in which knowledge is produced: molecular and cellular experiments, research using animal models, or human-based work.[14] In my setting, it distinguishes gene-specific knowledge developed through reductionist experimental methods from knowledge produced in a translational and clinical mode (Gittelman, 2016). This distinction aligns with Anckaert et al. (2020)'s argument that biomedical research becomes more valuable when it also considers insights from the bedside. If this interpretation is correct, the advantage of firms with gene-specific knowledge should be strongest when their prior research is rooted in human-based settings, which is better positioned to assess whether a statistically persuasive association is also meaningful for therapeutic purposes.

Figure 4 supports this conjecture. Patenting on false positive GWAS findings declines as firms' genetic expertise becomes more patient-oriented: it is positive for molecular research, attenuates for animal research, and approaches zero for human-based research. This gradient suggests that not all

[14] More specifically, NIH's iCite assigns each publication three scores, Human, Animal, and Molecular/Cellular, based on the share of MeSH terms falling within the corresponding branches of the MeSH tree. These scores place each article along a translational spectrum running from bench research to clinical work. See Appendix D.3 for details.

prior genetic expertise is equally useful. Consistent with this interpretation, publication activity on unrelated genes does not have a comparable effect, helping rule out that my measure proxies for generic scientific capability (Appendix Table E.12). These results suggest that firms avoid false positives by mapping statistical associations onto clinically grounded knowledge of the focal gene. More broadly, they also qualify the scope of the complementarity I document. Even as predictive tools improve, firms must still judge whether data-driven associations are meaningful for the underlying innovation problem. When that judgment depends on contextual knowledge, as in biomedicine, better prediction is unlikely to eliminate the advantage of firms that possess it.

6 Do Data-Driven Predictions Narrow Performance Gaps?

6.1 Translating GWAS to a Firm-Level Analysis

The previous section showed that domain knowledge helps firms avoid false positive predictions. However, evidence at the level of individual predictions does not establish whether this narrows or widens performance gaps across firms. Even as firms without domain knowledge pursue more false positives, GWAS could still improve their relative performance if it helps them uncover more valuable opportunities than they would have found on their own. Firms entered the genomics era from markedly different starting points, consistent with earlier research documenting the value of knowledge capabilities (Gambardella, 1992). As Appendix Figure E.5 shows, firms with stronger genetic expertise were already outperforming others before the emergence of GWAS. In what follows, I explore how the diffusion of data-driven predictions affected these differences, moving to a design that links GWAS opportunities to downstream innovation outcomes at the firm level.

I translate GWAS associations into a firm-level variable by exploiting pre-GWAS differences in firms' disease portfolios. Following the logic of shift-share designs, I combine common GWAS discoveries with predetermined exposure weights: firms are plausibly more "exposed" to GWAS findings in diseases in which they were already active before GWAS diffused. This assumption is consistent with my interviews with industry researchers: "*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). For each firm-year, I count the number of new GWAS associations in the firm's pre-2005 disease areas. This yields a time-varying measure of GWAS exposure determined by firms' ex ante therapeutic positioning. I then distinguish between firms' overall exposure to GWAS and exposure to findings involving genes the firm had previously studied. This allows me to test whether the same flow of data-driven opportunities has different consequences when it falls inside versus outside the firm's knowledge base.

I next separate firms by the depth of their genetic expertise before the emergence of genomics.

Specifically, I classify firms according to whether their pre-2005 stock of genetic publications lies above or below the sample median. This distinction captures whether firms entered the genomics era with stronger or weaker knowledge bases, allowing me to examine whether GWAS exposure has heterogeneous firm-level effects. I focus on three variables that capture downstream outcomes. First, I examine granted patents weighted by their citations, which proxies for the technological value of inventive output. Second, I measure drug discovery by tracking both new drug candidates entering the development pipeline and molecules that ultimately reach commercialization. Together, these outcomes indicate whether data-driven predictions help firms with weaker knowledge bases catch up, substituting for domain knowledge and reducing performance differences.

6.2 Research Design

I use the panel nature of the data to estimate the following specification at the firm-year level:

$$Y_{f,t} = \alpha + \beta_1 \textit{ GWAS Exposure}_{f,t} + \beta_2 \left(\textit{GWAS Exposure}_{f,t} \times \textit{Firm w/ Genetic Expertise}_f\right) + \gamma_t + \mu_f + \delta_{p(f,t)} + \varepsilon_{f,t} \tag{2}$$

where $Y_{f,t}$ is a firm-level innovation outcome in year t and $\textit{GWAS Exposure}_{f,t}$ counts the new GWAS findings in disease areas in which firm f was active before 2005. $\textit{Firm w/ Genetic Expertise}_f$ is an indicator equal to one if the firm's stock of genetic publications before 2005 lies above the sample median. Firm fixed effects (μ_f) absorb time-invariant differences across firms, while year fixed effects (γ_t) control for common shocks in a given year. $\delta_{p(f,t)}$ controls for the firm's overall patenting intensity in year t. The coefficient of interest, β_2, captures whether GWAS exposure has different performance implications for firms that entered the genomics era with deeper scientific expertise. Summary statistics are reported in Appendix Table E.13.

Identification comes from within-firm changes in GWAS exposure, driven by the staggered arrival of new GWAS findings across the diseases in which firms were already active. This variation is plausibly exogenous to the firm, since GWAS findings arise from external scientific advances rather than from firms' own choices. The design combines this shock with predetermined exposure weights from firms' pre-GWAS disease portfolios, while firm fixed effects absorb stable differences across firms. At the same time, pre-GWAS disease choices are not randomly assigned, leaving open the possibility that they are correlated with unobserved factors that also shape subsequent innovation performance. For this reason, I conservatively interpret the firm-level estimates as descriptive rather than causal. Even so, they remain informative about whether data-driven opportunities tend to narrow or widen performance gaps across firms.

6.3 Data-Driven Predictions and Firm Performance

I begin by replicating the findings of Section 5.3 in Appendix Table E.14. The firm-level analog of the earlier results shows that GWAS exposure does not increase patenting on false positives when the implicated genes were already inside the firm's knowledge base. Instead, investments in false positive associations rise only when GWAS findings involve genes the firm had not previously researched. Table 4 shows how this pattern translates into firm-level outcomes. On average, greater exposure to GWAS is associated with a higher share of patent applications directed toward false positive findings. This aggregate pattern, however, is driven entirely by firms with weaker genetic expertise. For these firms, GWAS exposure significantly increases investment in non-replicable associations, while the corresponding coefficients for firms with stronger genetic expertise are small and statistically indistinguishable from zero.

Next, I examine the effect on downstream innovative performance. Figure 5 shows that greater exposure to GWAS increases citation-weighted patents, new drugs entering discovery, and commercialized therapies—but primarily for firms with stronger knowledge bases. In each case, the gains are concentrated among companies that were already outperforming others before GWAS emerged. These estimates suggest that GWAS expanded the set of potential targets without equalizing firms' ability to identify which ones were worth pursuing. Because leading firms were better positioned to recognize false positives, they benefited disproportionately from the same flow of opportunities. Figure 6 offers a bird's-eye view of how drug discovery evolved after the Human Genome Project. Before GWAS diffused, there were hints of convergence across firms. After their arrival, however, the trajectories diverged: drug discovery rose among firms with stronger capabilities while remaining flat among those without them.[15]

Together, these results help explain why predictive technologies do not always yield larger gains among low performers, as documented in other settings. GWAS did not improve execution on a well-defined task, but they expanded the set of plausible targets while leaving firms to decide which ones to pursue. As recent research suggests, performance can diverge when tasks fall only partially within the technology's frontier (Dell'Acqua et al., 2026) and users must choose which suggestions to implement (Kim et al., 2024; Otis et al., 2025). In drug discovery, that evaluative step requires clinically grounded knowledge (Nelson et al., 2011), a form of absorptive capacity that is costly to build. GWAS thus made target discovery less dependent on prior expertise, but also made target evaluation more central to performance. Domain-expert firms benefited more because the source of competitive advantage ended up remaining the same, albeit with a different role.

[15] As an alternative research design, Appendix Table E.15 reports a difference-in-differences specification comparing firms before and after the emergence of GWAS, confirming that the gains are concentrated among firms with stronger expertise.

7 Conclusion

While scholars have extensively studied how firms exploit innovation opportunities, the rapid diffusion of big data and predictive tools is transforming how firms identify them. Data-driven predictions allow firms to estimate the value of potential investments without relying on domain knowledge. This paper provides some of the first empirical evidence on the consequences of this phenomenon. Leveraging the unique features of GWAS, I show that data-driven predictions have heterogeneous effects depending on firms' prior expertise. Firms with relevant genetic knowledge respond more selectively to GWAS findings and are more likely to avoid the false positives they contain. This heterogeneity translates downstream: rather than leveling the competitive field, GWAS widen performance gaps because firms with deeper expertise are better able to convert the same flow of predictions into valuable innovation. The evidence suggests that this advantage is rooted in clinical understanding that helps firms judge whether a gene-disease association is therapeutically meaningful and thus filter out false positives.

My results underscore the continuing importance of domain knowledge in innovation. Because abundant data-driven predictions do not imply a corresponding reduction in the cost of validating them (Fontanelli et al., 2026), expertise remains essential to avoid wasting resources on dead ends. At the same time, the paper clarifies the boundary conditions of this claim. When predictive tools are sufficiently accurate, they may substitute for expertise and disproportionately benefit less experienced actors, as recent studies of AI-augmented work suggest (Brynjolfsson et al., 2024; Dell'Acqua et al., 2026; Noy and Zhang, 2023). The complementarity I document arises when predictions remain noisy because the mechanisms and boundary conditions that determine value are only partially captured in data (Choudhury et al., 2020). This is likely to be common in technological innovation, where the value of novel combinations depends on complex interactions and contextual knowledge that are difficult to encode.

This distinction also helps reconcile conflicting findings in prior research. When firms face primarily a discovery problem, data-driven predictions can reveal promising leads they would otherwise miss, allowing less experienced actors to narrow performance gaps (Nagaraj, 2022). By contrast, when uncertainty persists after a lead has been identified (Otis et al., 2025), predictive tools and domain knowledge become complements rather than substitutes. In such settings, AI and related predictive tools may unbundle discovery from evaluation: they can expand the set of plausible opportunities without resolving the harder task of deciding which ones deserve commitment (Levinthal and Schliesmann, 2025). Because assessing opportunities often requires sticky and tacit expertise, predictive tools may ultimately reinforce the competitive advantage of firms that possess it.

My research also has practical implications. To cut through the hype surrounding big data, it is

essential to understand what predictive technologies can and cannot do (Agrawal et al., 2018). This paper suggests that, while data-driven predictions are valuable for shortlisting potential opportunities, they do not eliminate the need for judgment. This is especially true for what Camuffo et al. (2023) call "low-frequency/high-impact" strategic decisions. In these contexts, 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 greatest in organizations with a strong domain knowledge base. Policymakers, in turn, should consider how predictive tools may reshape competition. Rather than fostering entry and market dynamism, AI and similar technologies may instead stifle competition if they depend on complementary assets concentrated among incumbents.

These findings also raise broader concerns. If firms respond to predictive tools by retreating from investments in domain knowledge, they may erode the very knowledge base that makes those tools useful (Balasubramanian et al., 2022). The risk is that firms rely on data-driven predictions while weakening the contextual and experiential understanding needed to judge whether a statistical pattern is meaningful. This creates the possibility of an organizational form of "cognitive surrender" (Shaw and Nave, 2026): firms may come to treat AI-generated predictions as substitutes for expertise (Yerramilli-Rao et al., 2025), even if their value depends on the evaluative capabilities needed to interrogate them critically (Anthony, 2021; Lebovitz et al., 2022). Future research should examine whether these technologies reshape not only innovation outcomes, but also the accumulation of complementary knowledge capabilities.

Finally, a few limitations should be noted. First, the aggregate effects of GWAS on innovation are beyond the scope of this paper. Data-driven predictions may give rise to a "streetlight effect" (Hoelzemann et al., 2024), diverting investment toward suboptimal targets and potentially reducing welfare. Second, the paper does not examine how firms without domain knowledge might respond strategically to their disadvantage. One possibility is the emergence of markets for expertise, in which they outsource the judgment needed to assess the data-driven opportunities uncovered. Third, my empirical design holds prediction content fixed to isolate differences in firms' ability to evaluate the same signals. This is useful for identification, but limits generalizability. When firms also shape where and how prediction is conducted, domain knowledge may affect not only evaluation, but also the quality of the predictions generated. More broadly, whether superior predictions stem from proprietary data, model-building capabilities, or organizational structures that better connect prediction to expertise remains a first-order question for future research.

References

Agrawal, A., J. Gans, and A. Goldfarb (2018): *Prediction machines: The simple economics of artificial intelligence*, Harvard Business Press.

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 (2025): "Methodological pluralism and innovation in data-driven organizations," *Administrative Science Quarterly*, 70, 403–443.

Anckaert, P.-E., D. Cassiman, and B. Cassiman (2020): "Fostering practice-oriented and use-inspired science in biomedical research," *Research Policy*, 49, 103900.

Anthony, C. (2021): "When knowledge work and analytical technologies collide: The practices and consequences of black boxing algorithmic technologies," *Administrative Science Quarterly*, 66, 1173–1212.

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.

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

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

Boussioux, L., J. N. Lane, M. Zhang, V. Jacimovic, and K. R. Lakhani (2024): "The crowdless future? Generative AI and creative problem-solving," *Organization Science*, 35, 1589–1607.

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*, 140, 889–942.

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*, 9, 382–396.

Chesbrough, H. W. (2003): *Open innovation: The new imperative for creating and profiting from technology*, Harvard Business Press.

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.

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.

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

Cowan, R., P. A. David, and D. Foray (2000): "The explicit economics of knowledge codification and tacitness," *Industrial and Corporate Change*, 9, 211–253.

Csaszar, F. A. and D. Laureiro-Martínez (2018): "Individual and organizational antecedents of strategic foresight: A representational approach," *Strategy Science*, 3, 513–532.

de Rassenfosse, G. and A. B. Jaffe (2025): "The Effect of Application Fees on Entry into Patenting," *NBER wp 33492*.

Dell'Acqua, F., E. McFowland III, E. Mollick, H. Lifshitz, K. C. Kellogg, et al. (2026): "Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality," *Organization Science*, 37, 403–423.

Doshi, A. R. and O. P. Hauser (2024): "Generative AI enhances individual creativity but reduces the collective diversity of novel content," *Science Advances*, 10, eadn5290.

Dougherty, D. and D. D. Dunne (2012): "Digital science and knowledge boundaries in complex innovation," *Organization Science*, 23, 1467–1484.

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

Felin, T. and M. Singell (2025): "Technology: Theory-driven experimentation and combinatorial salience," *European Economic Review*, 105186.

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.

Fontanelli, L., K. McElheran, B. Caldarola, and E. Verdolini (2026): "From Bits to Atoms: 3D Printing, Physical Validation, and Firm Growth," *Università di Brescia, University of Toronto, European Commission and Maastricht University*.

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

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

Gittelman, M. (2016): "The revolution re-visited: Clinical and genetics research paradigms and the productivity paradox in drug discovery," *Research Policy*, 45, 1570–1585.

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.

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

Hutchins, B. I., M. T. Davis, R. A. Meseroll, and G. M. Santangelo (2019): "Predicting translational progress in biomedical research," *PLoS biology*, 17, e3000416.

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

Kao, J. (2025): "Charted territory: Mapping the cancer genome and R&D decisions in the pharmaceutical industry," *UCLA Anderson*.

Kapoor, R. and D. Wilde (2025): "Forecasting as a Problem of Cognitive Search: Experimental Evidence from Forecasting Tournaments in the Context of the Auto Industry," *The Wharton School and Indiana University*.

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.

Ke, S., B. Wu, and C. Zhao (2024): "Learning from a black box," *Journal of Economic Theory*, 221, 105886.

Kim, H., E. L. Glaeser, A. Hillis, S. D. Kominers, and M. Luca (2024): "Decision authority and the returns to algorithms," *Strategic Management Journal*, 45, 619–648.

Kim, S. (2025): "Navigating the rugged data landscape: The impact of data-extrapolation technologies on 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.

Knorr-Cetina, K. K. (1999): *Epistemic cultures: How the sciences make knowledge*, Harvard University Press.

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

Laursen, K. and A. Salter (2006): "Open for innovation: The role of openness in explaining innovation performance among UK manufacturing firms," *Strategic Management Journal*, 27, 131–150.

Lebovitz, S., H. Lifshitz-Assaf, and N. Levina (2022): "To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis," *Organization Science*, 33, 126–148.

Leiponen, A. and C. E. Helfat (2010): "Innovation objectives, knowledge sources, and the benefits of breadth," *Strategic Management Journal*, 31, 224–236.

Levinthal, D. A. (1997): "Adaptation on rugged landscapes," *Management Science*, 43, 934–950.

Levinthal, D. A. and D. Schliesmann (2025): "Cautious exploitation: Learning and search in problems of evaluation and discovery," *Organization Science*, 36, 903–917.

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.

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.

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., K. Buterbaugh, M. Perl, and A. Gelijns (2011): "How medical know-how progresses," *Research Policy*, 40, 1339–1344.

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.

Nightingale, P. (2004): "Technological capabilities, invisible infrastructure and the un-social construction of predictability: The overlooked fixed costs of useful research," *Research Policy*, 33, 1259–1284.

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 (2025): "The uneven impact of generative AI on entrepreneurial performance," *Management Science*.

Pham, N. D. (2025): "The Economic Performance of IP-Intensive Manufacturing and Service Industries in the United States, 2012–22," *ndp analytics report*.

Piezunka, H. and L. Dahlander (2015): "Distant search, narrow attention: How crowding alters organizations' filtering of suggestions in crowdsourcing," *Academy of Management Journal*, 58, 856–880.

Polanyi, M. (1966): *The tacit dimension*, Chicago and London: The University of Chicago Press.

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

Shaw, S. D. and G. Nave (2026): "Thinking-Fast, Slow, and Artificial: How AI is Reshaping Human Reasoning and the Rise of Cognitive Surrender," *Available at SSRN 6097646*.

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.

Tranchero, M. (2026): "Data-Driven Search and the Birth of Theory: 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.

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.

Von Hippel, E. (1994): ""Sticky information" and the locus of problem solving: implications for innovation," *Management Science*, 40, 429–439.

Yerramilli-Rao, B., J. Corwin, Y. Li, and K. R. Lakhani (2025): "Strategy in an era of abundant expertise," *Harvard Business Review*, 104, 72–81.

8 Figures and Tables

Figure 1: Example and structure of a typical GWAS

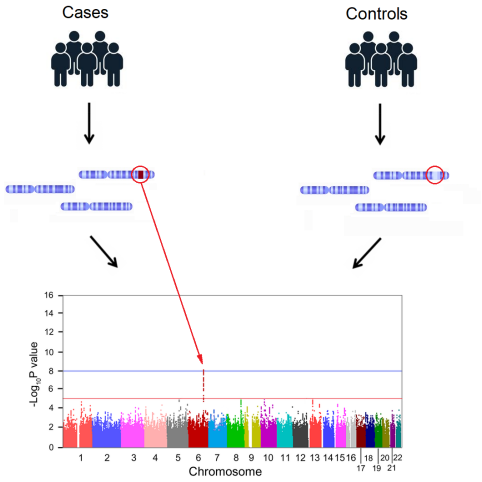
(a) *Abstract of the GWAS by Duerr et al. (2006)*

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

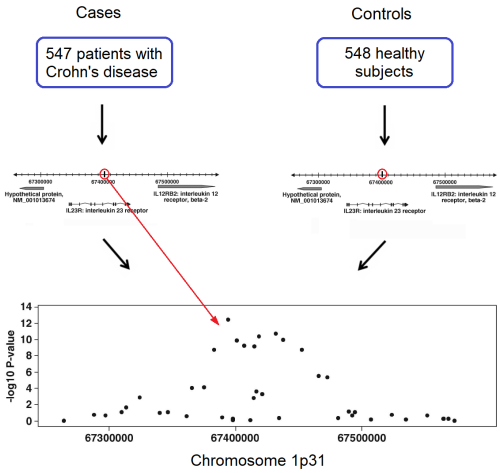
Richard H. Duerr,[1,2] Kent D. Taylor,[3,4] Steven R. Brant,[5,6] John D. Rioux,[7,8] Mark S. Silverberg,[9] Mark J. Daly,[8,10] A. Hillary Steinhart,[9] Clara Abraham,[11] Miguel Regueiro,[1] Anne Griffiths,[12] Themistocles Dassopoulos,[5] Alain Bitton,[13] Huiying Yang,[3,4] Stephan Targan,[4,14] 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*

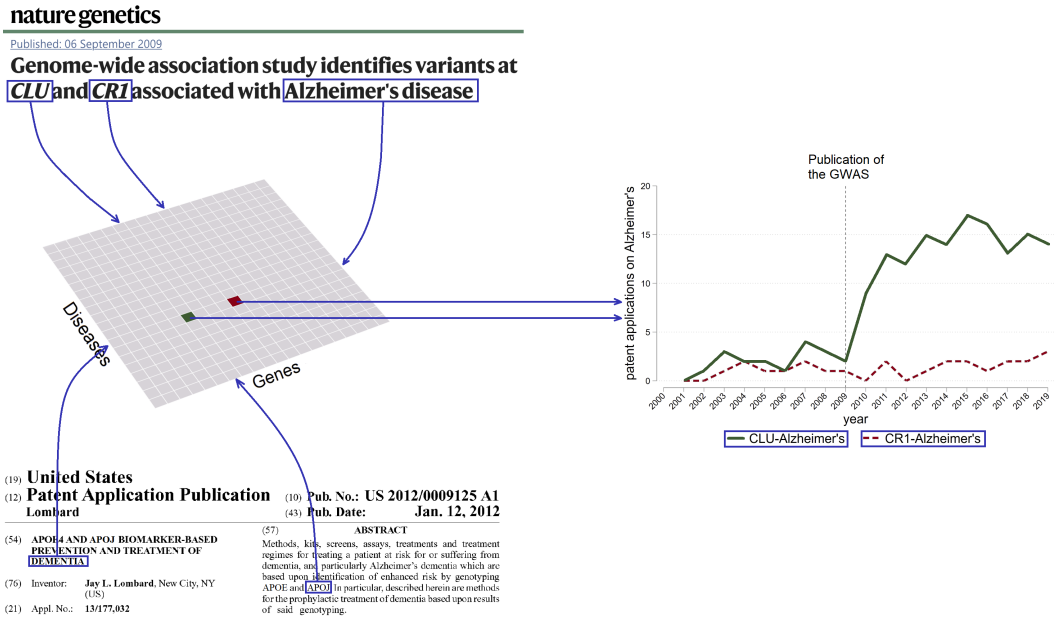


(c) *Schema of Duerr et al. (2006)*



Note: Panel (a) shows the abstract of the GWAS by Duerr et al. (2006). The abstract highlights the implications of using IL23R as a drug target for inflammatory bowel diseases. Panel (b) shows a schema of how a GWAS unfolds. Researchers select the disease of interest and assemble a group of cases (subjects with the condition) and a group of controls (subjects without the condition). Then, the genomes of people with and without the condition are genotyped to identify differences. Finally, statistical methods are used to test the association between any genetic mutation and the disease of interest. The bottom panel shows the characteristic "Manhattan plot," which displays the locations of statistically significant mutations along the chromosome. Panel (c) shows the same schema for the GWAS of Duerr et al. (2006). See Appendix A.2 for a case study on this GWAS.

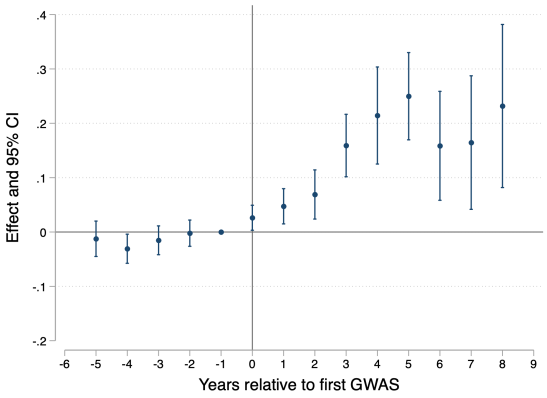
Figure 2: Pharmaceutical firms search for drug targets in an empirical landscape of gene-disease pairs



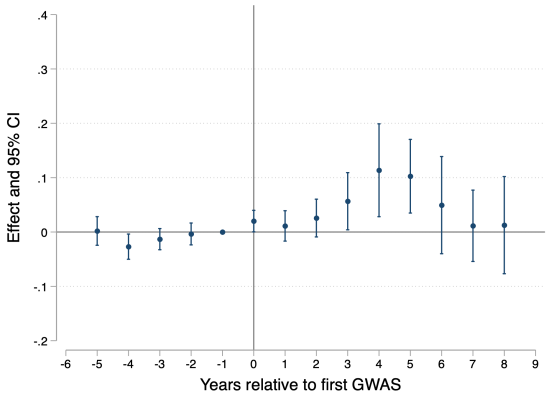
Note: The figure exemplifies the measurement approach adopted in this paper. Extracting genes and diseases from patents and papers enables mapping them onto a landscape of all the potential gene-disease combinations. The heterogeneous impact of GWAS on firm investments can be measured by tracking patenting at the gene-disease combination level, without using patent citations. APOJ is an alternative name for the CLU gene. See Appendix B for details.

Figure 3: GWAS findings stimulate innovation investment, even when the underlying associations are false positives that firms later abandon

(a) *Time-varying impact of GWAS*

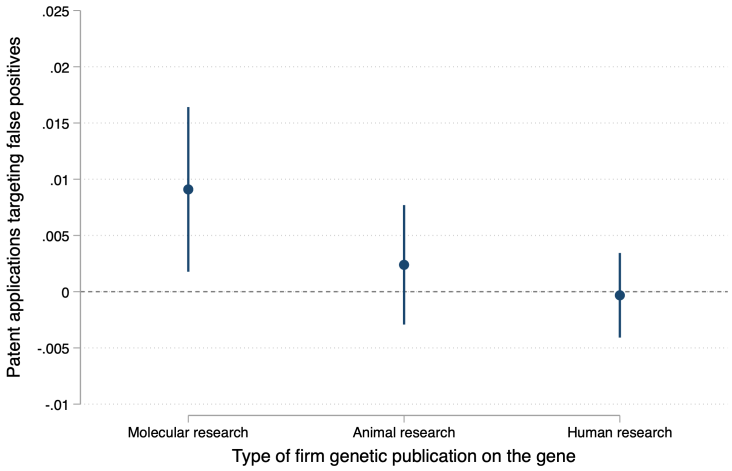


(b) *Time-varying impact of false positive GWAS*



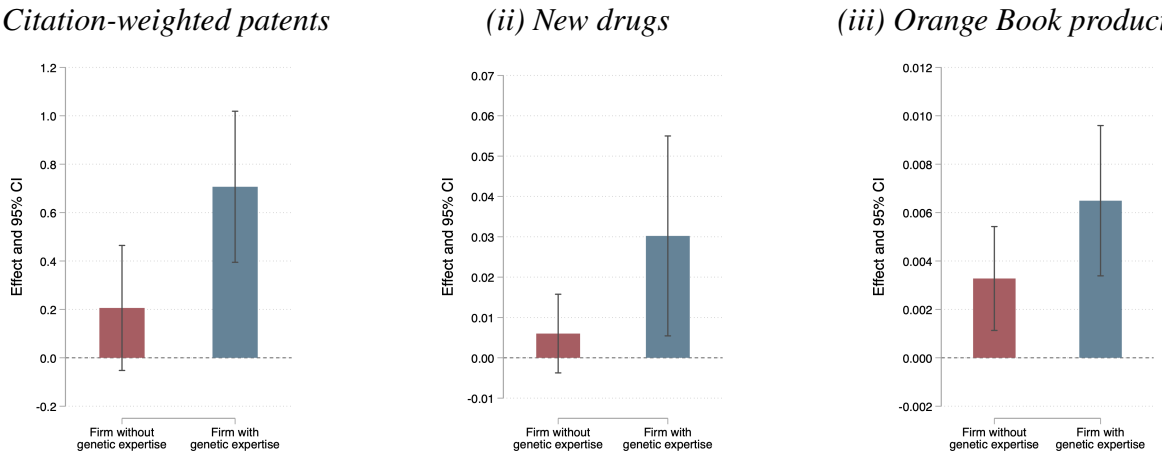
Note: The figure shows the event study coefficients estimated from the following specification: $Patent\ Applications_{i,j,t} = \alpha + \sum_z \beta_z GWAS_{i,j} \times 1(z) + \gamma GD_{i,j} + \delta_t \times Gene_i + \omega_t \times Disease_j + \epsilon_{i,j,t}$. The dependent variable is the number 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. Panel (a) shows the results for all GWAS findings. Panel (b) shows the results for false positive GWAS findings.

Figure 4: Firms with more patient-oriented genetic expertise are less responsive to false positive GWAS discoveries



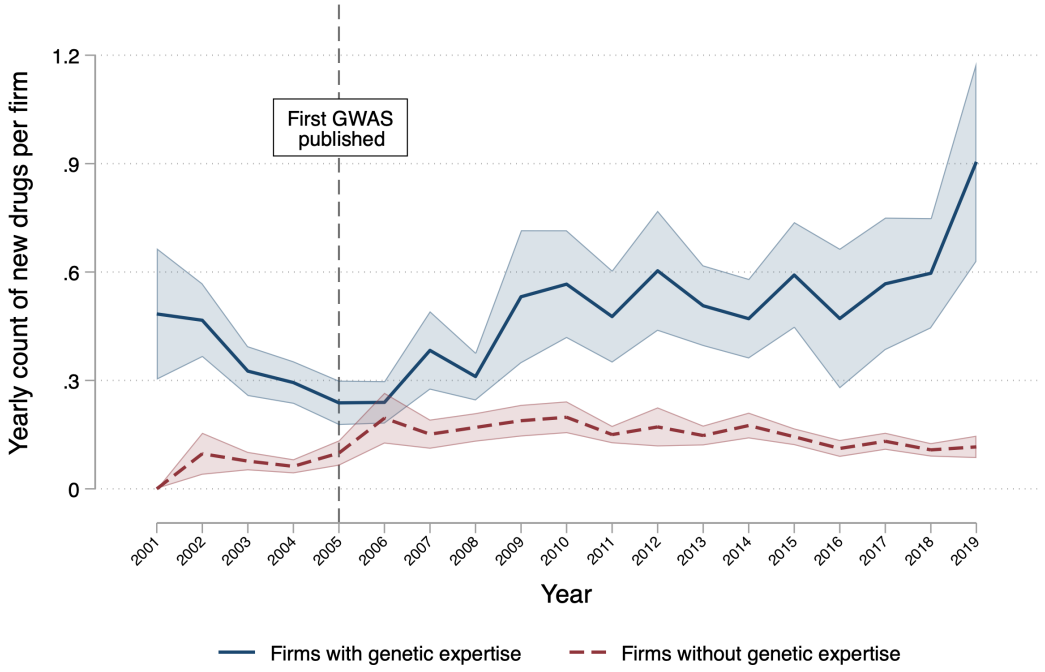
Note: This figure reports coefficient estimates and 95% confidence intervals. Each coefficient is estimated from a separate regression with the same dependent variable, separately by the type of firm genetic expertise that predates the GWAS finding. *Patent applications targeting false positives*: count of USPTO patent applications for innovations targeting a specific gene-disease combination that is not replicated by subsequent GWAS. *Molecular research*: 0/1 = 1 if the firm has at least one pre-GWAS publication on the focal gene classified as molecular research. *Animal research*: 0/1 = 1 if the firm has at least one pre-GWAS publication on the focal gene using animal models. *Human research*: 0/1 = 1 if the firm has at least one pre-GWAS publication on the focal gene based on human subjects. Publication-method classifications are obtained from NIH's iCite data (Appendix D.3).

Figure 5: GWAS increased downstream innovation for firms with deeper genetic expertise, thus widening firm-level performance gaps



Note: Panels (i), (ii), and (iii) report coefficient estimates and 95% confidence intervals from firm-year regressions. Each bar is based on the regression: $Y_{i,t} = \alpha + \beta_1$ *GWAS Exposure*$_{i,t} + \beta_2 \left($*GWAS Exposure*$_{i,t} \times$ *Firm w/ Genetic Expertise*$_i\right) + \gamma_t + \mu_i + \delta_{p(i,t)} + \varepsilon_{i,t}$ where *GWAS exposure* measures the extent to which a firm is exposed to GWAS findings in disease areas in which it was active before 2005. The first bar reports $\hat{\beta}_1$, while the second bar reports $\hat{\beta}_1 + \hat{\beta}_2$. *Firm-level GWAS exposure*: measure of the extent to which GWAS findings emerge in disease areas in which the firm was active before 2005 (scaled by 100 for interpretability). *Citation-weighted patents*: count of firm patents in a given year, weighted by forward citations in the subsequent 5 years. *New drugs*: count of new drugs discovered by the firm in a given year. *Orange Book products*: count of firm patents in a given year that are listed in the FDA Orange Book for approved and commercialized drugs. *Firm with genetic expertise*: 0/1 = 1 if the firm has an above-median number of genetic publications before 2005.

Figure 6: Descriptive evidence on drug discovery patterns following the emergence of GWAS



Note: This figure reports yearly means and standard deviation of new drugs per firm, separately for firms with above and below-median levels of genetic expertise. *New drugs*: count of new drugs discovered by the firm in a given year. *Firm with genetic expertise*: 0/1 = 1 if the firm has an above-median number of genetic publications before 2005.

Table 1: Summary statistics

	Panel A: firm-level descriptives					
	mean	median	std. dev.	min	max	N
Patent applications per firm	36.7845	11	156.2978	2	3,978	4,027
Genes in patent portfolio	64.285	19	195.922	1	5,467	4,027
Diseases in patent portfolio	41.362	27	46.173	1	338	4,027
Publications per firm	59.170	9	390.944	1	10,656	3,346
Genes in publication portfolio	176.563	52	568.922	1	12,267	3,346
Diseases in publication portfolio	109.239	65	156.608	1	1,876	3,346
	Panel B: gene-disease cross-sectional descriptives					
	mean	median	std. dev.	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 treated by GWAS (0/1)	0.0025	0	0.0498	0	1	7,223,924
	Panel C: gene-disease-year panel descriptives					
	mean	median	std. dev.	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
Treated by GWAS (0/1)	0.0005	0	0.0234	0	1	137,254,556
False positive (0/1)	0.0004	0	0.0204	0	1	137,254,556
Year	2010	2010	5.4772	2001	2019	137,254,556

Note: Panel A reports firm-level statistics. Publication-based statistics are computed for firms with at least one matched publication. Panel B reports statistics at the gene-disease combination level for the 7,223,924 possible combinations obtained from the Cartesian product of 404 diseases and 17,881 protein-coding genes. Panel C reports statistics after expanding this cross-section into a balanced 2001–2019 panel of 137,254,556 gene-disease-year observations.

Table 2: 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...					
	...all firms		...firms with gene knowledge		...firms without gene knowledge	
	(1)	(2)	(3)	(4)	(5)	(6)
Post × GWAS	0.1637***		0.0381***		0.1256***	
	(0.01985)		(0.01009)		(0.01429)	
Post × False Positive GWAS		0.0676***		0.0138		0.0539***
		(0.01605)		(0.00821)		(0.01052)
Post × True Positive GWAS		0.5479***		0.1354***		0.4125***
		(0.07524)		(0.03830)		(0.05741)
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	137,254,556	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	7,223,924
Mean of Dep Var:	0.1314	0.1314	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. *False Positive*: 0/1 = 1 for GWAS findings that are never replicated by another GWAS about the same disease. *True Positive*: 0/1 = 1 for GWAS findings that are later replicated by another GWAS about the same disease.

Table 3: Firms with genetic domain knowledge avoid false positive GWAS findings, even when they appear more statistically solid

Dependent Variable:	USPTO patent applications by...					
	...firms with gene knowledge			...firms without gene knowledge		
	(1)	(2)	(3)	(4)	(5)	(6)
Post × False Positive GWAS	0.0169	0.0121	0.0159	0.0272*	0.0322**	0.0343**
	(0.0110)	(0.0071)	(0.0081)	(0.0122)	(0.0115)	(0.0121)
... × Small p-value	-0.0091			0.0673**		
	(0.0163)			(0.0227)		
... × Large sample size		0.0039			0.0604*	
		(0.0209)			(0.0243)	
... × Powerful array			-0.0071			0.0536*
			(0.0198)			(0.0234)
Gene-Disease FE	YES	YES	YES	YES	YES	YES
Gene-Year FE	YES	YES	YES	YES	YES	YES
Disease-Year FE	YES	YES	YES	YES	YES	YES
N	137,254,556	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	7,223,924
Mean of Dep. Var.	0.0376	0.0376	0.0376	0.0938	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. *False Positive*: 0/1 = 1 for GWAS findings that are never replicated by another GWAS about the same disease. *Small P-value*: 0/1 = 1 for associations with a p-value smaller than the median GWAS in my sample. *Large sample*: 0/1 = 1 for associations published in studies with a sample size larger than the median GWAS in my sample. *Powerful array*: 0/1 = 1 if the gene-disease association is obtained using a genotyping microarray with an above-median number of SNPs.

Table 4: Greater exposure to GWAS findings increases false positive patenting for firms without genetic expertise

Dependent Variable:	USPTO patent applications on false positives by...			Share of false positive patents by...		
	...all firms	...firms with genetic expertise	...firms without genetic expertise	...all firms	...firms with genetic expertise	...firms without genetic expertise
	(1)	(2)	(3)	(4)	(5)	(6)
Firm-level GWAS exposure	0.0118	0.0007	0.0192***	0.0030***	0.0017	0.0054***
	(0.0070)	(0.0132)	(0.0057)	(0.0009)	(0.0015)	(0.0012)
Year FE	YES	YES	YES	YES	YES	YES
Assignee FE	YES	YES	YES	YES	YES	YES
Total Patents FE	YES	YES	YES	YES	YES	YES
N	40,924	14,351	26,560	40,924	14,351	26,560
N of Firms	4,027	1,103	2,924	4,027	1,103	2,924

Note: *, **, *** denote significance at the 5%, 1%, and 0.1% levels, respectively. Observations are at the firm-year level. Std. err. clustered at the firm level are reported in parentheses. *USPTO patent applications on false positives*: count of USPTO patent applications filed in a given year for innovations that target a specific gene-disease combination following a false positive GWAS; the count is then divided between firms with and without previous publications on the gene. *Share of false positive patents*: share of USPTO patent applications filed by the firm in a given year that target gene-disease combinations associated with false positive GWAS findings. *Firm-level GWAS exposure*: measure of the extent to which GWAS findings emerge in disease areas in which the firm was active before 2005 (scaled by 100 for interpretability). *Firm with genetic expertise*: 0/1 = 1 if the firm has an above-median number of genetic publications before 2005.

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

Appendix

A Additional Details on GWAS **2**

A.1 A Scientific Primer 2

A.2 Case Study: The GWAS of Duerr et al. (2006) 4

A.3 Case Study: The GWAS of Suchindran et al. (2010) 5

B Measurement Approach **7**

B.1 Shortcomings of Patent-to-Paper Citations 7

B.2 Using a Landscape Approach to Measure Technological Impact 7

B.3 An Application to Pharmaceutical Innovation 9

C SciBite/EBI Patent Data **12**

C.1 Sample and Validation 12

C.2 Descriptive Statistics 13

C.3 Example: Denali Therapeutics 15

D Other Data Sources **17**

D.1 The NHGRI-EBI GWAS Catalog 17

D.2 PubTator Central Publication Data 18

D.3 NIH's iCite Data 19

D.4 Drug Data 19

D.5 Open Targets Score 20

E Additional Figures and Tables **26**

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 instructions for synthesizing products, such as proteins. The normal function of a gene can be altered by a mutation, potentially leading to health conditions. Knowing the genetic roots of diseases has practical consequences for the design of pharmaceutical drugs: leveraging the understanding of gene-disease relationships permits the identification of targets that can be inhibited or activated by a drug to produce a desired therapeutic effect (Nelson et al., 2015).

A.1 A Scientific Primer

Diseases caused by a single genetic mutation are called Mendelian disorders. These diseases are typically severe and, as a consequence, rare because they tend to be eliminated by evolutionary pressures. More common are polygenic diseases (also called complex diseases) that are caused by many mutations. For polygenic diseases, any genetic mutation can increase the risk of presenting the condition without being necessary or sufficient for manifesting the disease. Individual mutations are usually responsible for a small proportion of the heritability of complex diseases. Although such diseases often cluster in families, they do not have a predictable inheritance pattern because complex interactions between genetic predisposition and environmental factors contribute to their etiology. Therefore, scientists need to search through all of the $\sim 19,000$ protein-coding genes to identify the mutations underlying each polygenic disease (Tranchero, 2026).

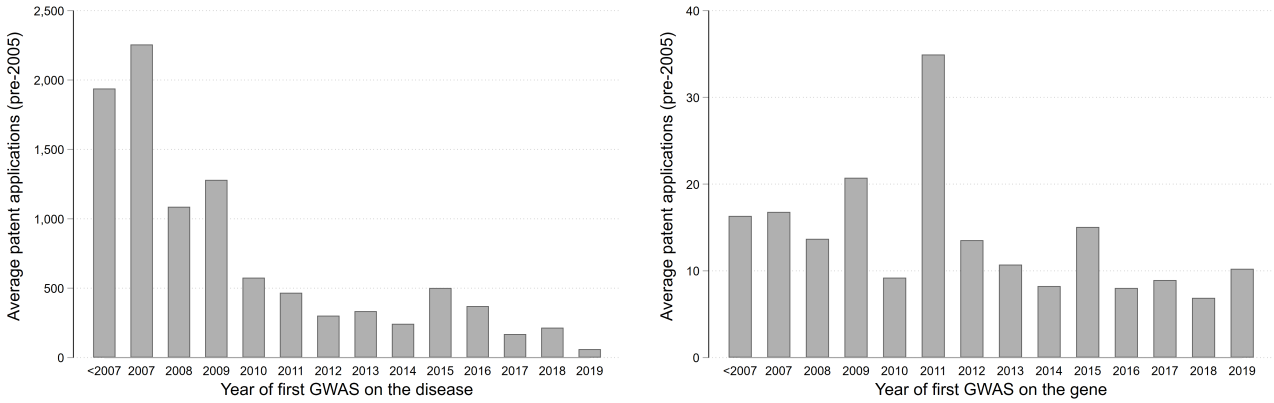
Over the years, researchers have concluded that common diseases are influenced by genetic mutations also common in the population (Reich and Lander, 2001). Instead of looking for rare individual genes with strong effects on phenotypes, the field has moved toward studying common genetic mutations that have a small impact on the likelihood of presenting a disease when taken in isolation (Bush and Moore, 2012). But what precisely is a mutation? 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. A mutation (or variant) in a single DNA base is called a single-nucleotide polymorphism (SNP) when it occurs in at least 1% of the population. One approach to associating SNPs with a disease is to assume that a causative variant will 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 on this logic, genome-wide association studies (GWAS) are a data-driven method for identifying associations between genetic regions and diseases (Visscher et al., 2017). In a typical GWAS project, researchers obtain DNA samples from two groups: patients with the disease under study and healthy individuals with comparable demographics. Then, data on specific SNPs across the entire chromosome are collected using microarrays that can genotype millions of SNPs per individual. Variants that are significantly more likely to appear in affected people may be biologically relevant to the disease and thus potentially involved in its origin. Such SNPs might affect gene expression and func-

tion, 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).

Figure A.1: GWAS timing reflects selection at the disease level, but not at the gene level

(a) *Average past patenting on diseases receiving GWAS* (b) *Average past patenting on genes receiving GWAS*



Note: Panel (a) shows the average number of patent applications before 2005 mentioning each disease, grouped by the year in which that disease first appeared in a GWAS. Panel (b) shows the average number of patent applications before 2005 mentioning each gene, grouped by the year in which that gene was first reported by a GWAS. The distinction is important because GWAS target specific diseases, not specific genes. Conditional on the disease being studied, the genes associated with that disease are revealed by the genome-wide scan rather than selected ex ante by researchers. Consistent with this, the figure shows clear temporal sorting at the disease level, but no comparable pattern appears at the gene level.

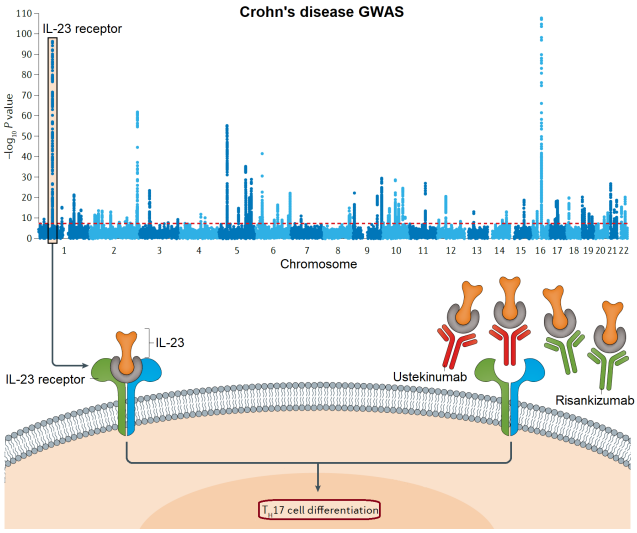
GWAS findings entail a substantial risk of false positives (Boyle et al., 2017; MacArthur, 2012). Findings often fail to replicate because of demographic biases in the convenience samples used by the original studies (e.g., white men from North America), or because of errors in mapping the variant to the correct gene when publishing the results (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 only a small fraction of the variation in disease susceptibility, suggesting that the therapeutic benefit of using them as drug targets could be quite limited (Goldstein, 2009). These limitations notwithstanding, GWAS have proven useful for uncovering new drug targets (Reay and Cairns, 2021; Visscher et al., 2017). GWAS also enable the identification of new uses for existing drugs by highlighting new conditions that might be addressed by acting on a given target (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. Given the prevalence and severity of these diseases, researchers have been intensively studying Crohn's and related diseases. According to the DisGeNET data, IBD ranked among the top 1% of diseases by research intensity in the pre-GWAS era. A few genes, such as NOD2, had been identified as harboring causal mutations by 2005, but they did not fully explain the genetic risk for the disease.

In December 2006, Duerr et al. (2006) published one of the first GWAS. The study included 567 patients of European ancestry with IBD and 571 healthy controls. The GWAS identified genetic mutations in the interleukin-23 receptor (IL23R) gene 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 lead uncovered by Duerr et al., often to elucidate the causal mechanisms (called "pathways") through which IL23R operates in human biology. It is now understood that the IL23R gene provides instructions for making the interleukin-23 receptor, a protein involved in immune signaling and inflammatory responses to defend the body against infections by promoting local inflammation. 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 signaling pathways that develop and activate Th17 cells, a specific type of lymphocyte that promotes inflammation to fight foreign invaders such as viruses. Malfunctioning IL23R genes might misdirect these inflammatory responses toward human tissues, leading to autoinflammatory diseases such as IBD.

Figure A.2: 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 drug targets. Once Duerr et al. (2006) identified a new variant in IL23R that disrupts its function, monoclonal antibodies targeting 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).

The work of Duerr et al. (2006) uncovered the role that IL23R has in inducing inflammation in the intestines. Furthermore, the discovery of IL23R's role in IBD suggested that drugs could be developed to modulate the IL-23/IL23R pathway. 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.2). As of 2024, ustekinumab (marketed as 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 (marketed as Skyrizi by AbbVie), tildrakizumab, and guselkumab. Interestingly, all these molecules have been repositioned as Crohn's disease interventions from their initial indication for psoriasis (Reay and Cairns, 2021). This highlights how uncovering new drug targets through GWAS can enable repurposing of gene-specific molecules developed for other diseases, as well as the development of new ones.

A.3 Case Study: The GWAS of Suchindran et al. (2010)

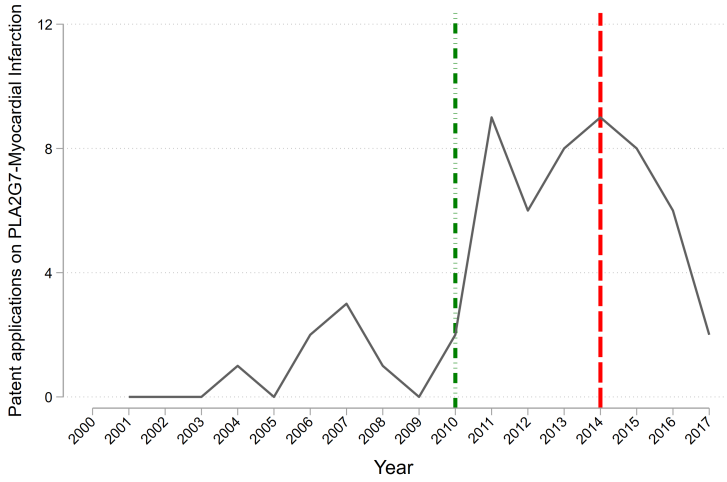
Myocardial infarction is one of the most severe manifestations of coronary heart disease, a condition that affects approximately 5% of U.S. adults. Coronary heart disease typically results from atherosclerosis, a process in which lipid accumulation and inflammation contribute to the formation of plaques in coronary arteries. Because plaque rupture can obstruct blood flow to the heart, researchers have long sought biomarkers that predict cardiovascular risk and molecular targets that could reduce future events. One such candidate was lipoprotein-associated phospholipase A2 (Lp-PLA2), an inflammatory enzyme encoded by the PLA2G7 gene. Before the GWAS era, epidemiological and experimental evidence had linked circulating Lp-PLA2 levels to cardiovascular disease, making this pathway an attractive but uncertain target for therapeutic intervention.

In April 2010, Suchindran et al. (2010) published a GWAS of Lp-PLA2 activity and mass using 6,668 Caucasian participants from the Framingham Heart Study. The study tested whether common genetic variants across the genome were associated with Lp-PLA2 activity or mass. Four loci were significantly associated with Lp-PLA2 activity, while variants clustered around PLA2G7 were significantly associated with Lp-PLA2 mass. The finding was plausible because PLA2G7 encodes Lp-PLA2 itself, and Lp-PLA2 had already been viewed as a biomarker and possible therapeutic target for cardiovascular disease. At the same time, the biological interpretation remained uncertain. The GWAS linked genetic variation in PLA2G7 to variation in an intermediate biomarker, but it did not establish that manipulating PLA2G7 would reduce myocardial infarction risk.

Figure A.3 shows how this signal shaped firm investment. GlaxoSmithKline was already investigating the PLA2G7 pathway through darapladib, an Lp-PLA2 inhibitor, but the publication of Suchindran et al. (2010) was followed by a spike in patenting by additional firms targeting the PLA2G7–myocardial infarction pair. In retrospect, this investment response appears to have followed a misleading lead. In my data, the PLA2G7–myocardial infarction pair failed to replicate in later GWAS, suggesting limited genetic support. Consistent with this interpretation, late-stage clinical trials of darapladib

failed to show a significant reduction in major coronary events, and GlaxoSmithKline announced in 2014 that darapladib did not meet its primary endpoint (Krieger, 2021). This case provides a useful contrast with the IL23R–Crohn's disease example in Appendix A.2: both GWAS findings spurred innovation, but only the latter identified a robust therapeutic opportunity. The PLA2G7 case illustrates how a statistically salient GWAS association can draw firms toward a target that proves therapeutically unproductive.

Figure A.3: A false positive GWAS signal stimulated patenting on PLA2G7–myocardial infarction before firms retreated following a clinical failure



Note: The figure plots the yearly number of USPTO patent applications mapped to the PLA2G7–myocardial infarction pair. The green line marks the publication of Suchindran et al. (2010), which reported genome-wide associations involving Lp-PLA2 traits, including variants near PLA2G7. This association later failed to replicate in subsequent GWAS on myocardial infarction. The red line marks the 2014 failure of late-stage clinical trials for darapladib, a GlaxoSmithKline drug targeting the Lp-PLA2 pathway. The pattern illustrates how a biologically plausible GWAS signal can stimulate patenting around a target that later fails to translate into downstream innovation.

B Measurement Approach

This appendix describes and formalizes the landscape-based measurement approach adopted in the paper to measure the impact of GWAS. Two case studies are presented to showcase the advantages relative to patent-to-paper citations.

B.1 Shortcomings of Patent-to-Paper Citations

Measuring the technological impact of a scientific project traditionally involves counting the patent citations received by publications in which its output is codified (Arora et al., 2021; Fleming et al., 2019). This approach has been validated showing that patent-to-paper citations are a good way to proxy knowledge flows from science to technology (Narin et al., 1997; Roach and Cohen, 2013). Therefore, it is not surprising that most existing studies use patent-to-paper 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 these metrics (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 that flows through more complex 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 an underestimate may be magnified in basic research, where spillovers occur through unpredictable channels that might not be acknowledged in citations (Cohen et al., 2002).

Second, as in academic papers, patent references to science 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 contain multiple contributions, and a citation to the paper does not clarify which one is being used. Citation counts cannot reveal the value of any specific contribution within a paper.

B.2 Using a Landscape Approach to Measure Technological Impact

In this paper, I formalize my measurement approach using the notion of *knowledge entities* (Ding et al., 2013). I define a knowledge entity as any discrete carrier of knowledge, whether a physical artifact, such as a piece of computer hardware, or an abstract unit defined by domain-specific taxonomies, such as a gene or a disease. Publications and patents can then be represented as combinations of the entities they study or recombine (Fleming, 2001). This perspective makes it possible to measure the technological

impact of a paper by tracing whether inventive activity increases around the entity combinations it introduces. The intuition is that when a paper creates valuable innovation opportunities, subsequent patents should mention the same entity combinations more often than comparable combinations not introduced by the paper. This measure captures downstream inventions that build on the focal contribution, including those that do not directly cite the original paper.[1]

More specifically, 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, publications and patents. This can be done using automated machine learning procedures, such as Bidirectional Encoder Representations from Transformers (BERT) or Large Language Models (LLMs). These tools can be tailored to specific domains, such as the life sciences, to increase accuracy and contextual stability. After extracting accurate knowledge entities using natural language processing 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. 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 inferred from the entity extraction task (e.g., the landscape of research topics identified with an LLM). The advantage of studying research as happening in a landscape is that citations are not needed to track knowledge evolution. 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 change in inventive output involving 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:

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

Where $Y_{i,j,t}$ is any 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 account for the stable differences across combinations, while entity time trends avoid confounding from heterogeneous changes across individual entities.

The coefficient of interest, β, captures the change in inventive activity around the treated entities after the project is completed. Under a parallel trends assumption, it estimates the project's technological

[1] Similarly to my approach, Suh (2024) extracts the chemical compounds mentioned in the body of patent texts to identify innovations that rely more on technologies where the Soviet Union had a scientific lead. Kao (2025) and Nagaraj (2022) apply this idea when studying pharmaceutical innovation and gold discoveries, respectively.

impact by comparing treated entity combinations with similar combinations that were not introduced by the project. Importantly, this estimate does not rely on patent-to-paper citations, addressing both measurement issues discussed above. First, because it identifies entities directly in the text of technological applications, it can capture downstream inventions that build on basic research even when they do not cite the original paper. Second, because it tracks specific entity combinations, it can distinguish among different contributions in the same article, providing a level of granularity that citations cannot.

B.3 An Application to Pharmaceutical Innovation

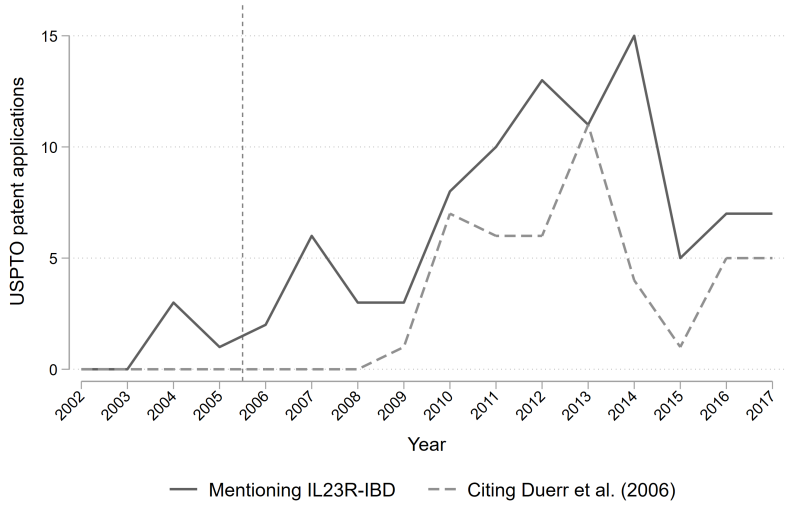
The biomedical sector offers an ideal setting to demonstrate the advantages of a landscape approach. Pharmaceutical innovation heavily depends on science, and knowledge entities are well-defined by taxonomies and have clear meanings (e.g., genes and diseases). I leverage data from the European Bioinformatics Institute (EBI) to measure the gene-disease pairs targeted by each USPTO patent application (2001-2019). Entities from patent texts are extracted using TERMite, SciBite's proprietary named entity recognition software (Appendix C).

Example 1: Undermeasurement of Basic Research Impact: 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 impactful both on scientific research and pharmaceutical innovation. Duerr et al. (2006) received over 4,000 citations on Google Scholar (as of 2026) 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 a landscape-based approach 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 worth noting. First, some firms had already been exploring the IL23R-IBD nexus before the finding became known to the broader scientific community. 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. Second, it appears that just looking at patent-to-paper citations would lead to a substantial underestimate of impact. Only 46 patent applications cite Duerr et al. (2006), compared with 88 that exploit its finding. The gap between the two measurements halves 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. 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

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).

acknowledging the GWAS by Duerr and colleagues, instead citing 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? Counting the number of patents that cite this paper would not answer the question. Aveo Pharmaceuticals' patent US 2011/0305687 A1, titled "Anti-FGFR2

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.

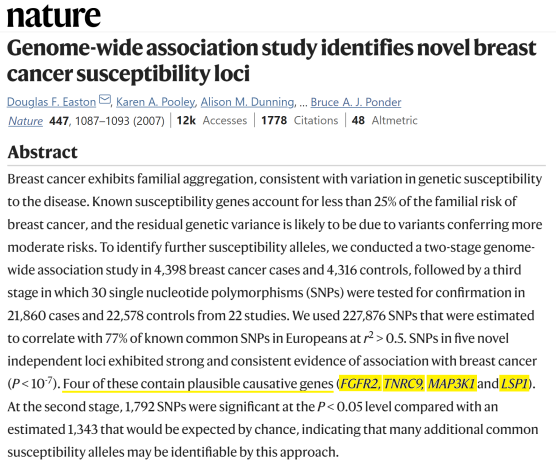
30) A method for inhibiting the binding of IL-23 to the IL-23 receptor on a mammalian cell, comprising administering to the cell an anti-IL-23p19 antibody or antigen-binding fragment, whereby signaling mediated by the IL-23 receptor is inhibited,

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.

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 cannot be inferred from citation patterns alone.

Figure B.3: Knowledge entities permit the measurement of 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

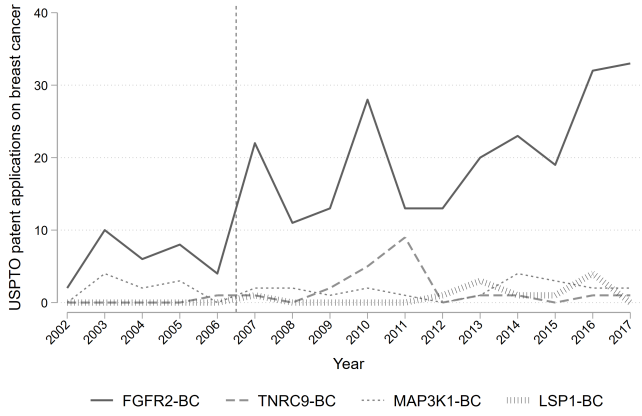
Douglas F. Easton ✉, Karen A. Pooley, Alison M. Dunning, ... Bruce A. J. Ponder

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.

A landscape-based approach bypasses this limitation by extracting each recombination introduced by an article and then tracing patenting dynamics around that specific recombination. Figure B.3 illustrates this logic using the GWAS by Easton et al. (2007). The study reported multiple gene-disease associations, but patenting increased substantially only for the FGFR2–breast cancer combination. Patenting on this pair nearly tripled, while the other combinations experienced little change. A citation-based approach would miss this heterogeneity, treating all patents citing Easton et al. (2007) as the same. This distinction is especially important when some associations may be false positives. By tracking text-extracted entities directly, the landscape-based approach provides a more granular way to measure the technological impact of basic research, especially when a single article contributes to multiple downstream domains.

C SciBite/EBI Patent Data

This appendix describes the data on the biological entities listed in USPTO patent applications (2001-2019).

C.1 Sample and Validation

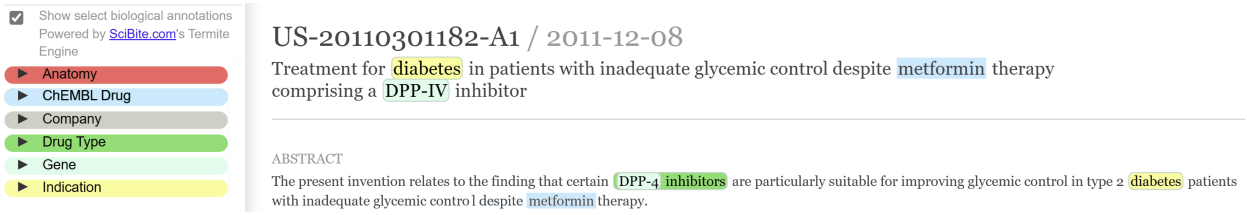
Information on R&D expenditure is usually available only at the organizational level, not for specific projects. In this paper, I follow the approach of Eggers and Kaplan (2009) and use patent records to infer where firms are directing their innovation investments. I infer firms' project choices from the genes and diseases mentioned in their patent applications. Because pharmaceutical innovation is long and uncertain, these applications capture early resource commitments rather than completed innovation outcomes. 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, which were ruled inadmissible by the U.S. Supreme Court in the Myriad ruling. Instead, my sample considers patents for innovations such as genetic tests, new drug molecules, or method-of-use patents for molecules that *target* a specific gene to *treat* a specific disease.

The primary source 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 Gene 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 by 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 manually validate TERMite's performance on recall and precision, following the procedure of Marx and Fuegi (2020). Recall measures the number of actual bio-entities in a patent that 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,

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.

retrieved their full specifications from Google Patents, and randomly selected one gene and one disease mentioned in each. 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 defined as the proportion 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 scores are 92.96 for genes and 94.43 for diseases, demonstrating the high reliability of the data.

C.2 Descriptive Statistics

My sample includes 148,232 USPTO patent applications published from 2001 to 2019. 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. 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). However, the sample shows a large variance, with a few patents listing hundreds of genes and diseases as targets. This extreme disclosure behavior probably reflects strategic considerations and suggests the need for more 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 the number of genes appearing in the patent text shows a slight upward trend. Patents published in 2001 seem to reference an abnormally low number of bio-entities, possibly because of idiosyncrasies of that year (the first year in which patent applications began to be published, and the year the first draft of the human genome was released). Together, the two graphs help rule out

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

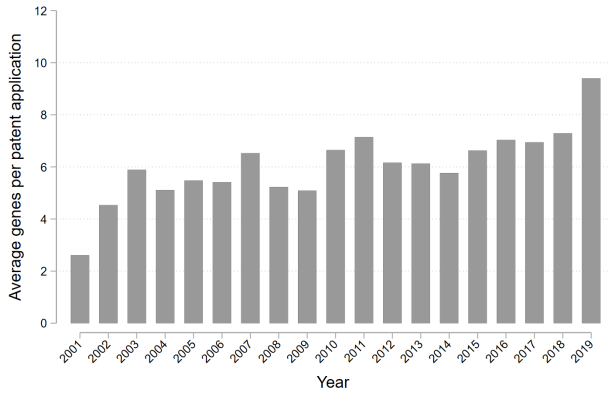
	USPTO patent applications						USPTO granted patents					
	mean	median	st d	min	max	N	mean	median	st d	min	max	N
Genes per patent	6.285	2	31.1786	1	4060	148,232	5.625	2	31.3932	1	4,060	73,255
Diseases per patent	12.257	5	22.5936	1	911	148,232	12.095	5	20.4855	1	853	73,255
Gene-disease pairs per patent	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. The left panel of the table presents descriptives for all patent applications, while the right panel presents descriptives for only those 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.

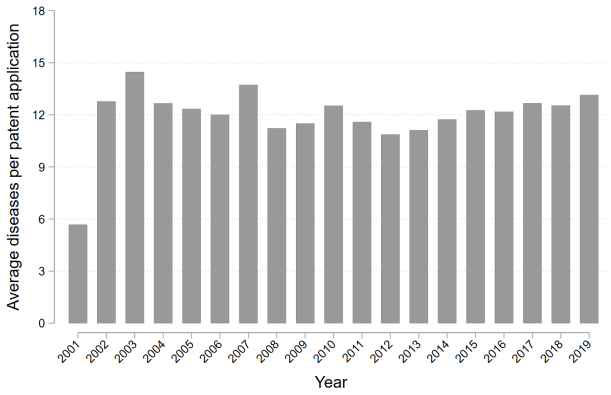
structural breaks in disclosure practices that could potentially bias my measurement approach.

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 portfolios 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 miss much of the firms' exploration in genetic space. This validates the decision to examine patent applications to capture the earliest stages of pharmaceutical innovation. On average, firms are active on 1,437 gene-disease pairs, but the median firm explored only 162 pairs.

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

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	4,027	37.637	11	130.255	1	4,579	3,866
Diseases per firm	41.362	27	46.173	1	338	4,027	31.530	17	40.546	1	305	3,866
Gene-disease pairs per firm	1,436.791	162	8,680.241	1	239,160	4,027	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.

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 to publicly listed firms, 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 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 an additional $250 million from its 2017 IPO. 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 genetics, pathology, and cell biology underlying chronic neurodegenerative disease have identified new pathways that trigger neurodegeneration. In particular, researchers have dubbed

a set of genes "degenogenes" because, when mutated, they are likely to have a causative role in neurodegenerative disease. Denali Therapeutics was founded on the idea that such degenogenes could constitute a viable therapeutic avenue for tackling 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.[2] This offers an opportunity to check the reliability of the SciBite data, which 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.[3] 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 pioneering its use for drug targeting (Kingwell, 2022). This example shows the potential of using bio-entities to capture a company's technological portfolio.

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

[3]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

D.1 The NHGRI-EBI GWAS Catalog

Information about genome-wide association studies is from the GWAS Catalog, a publicly available list of 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 $< 1.0 \times 10^{-5}$). Compiling this resource requires manual curation of a large body of diverse and unstructured data, a task carried out by scientists at the European Bioinformatics Institute (EBI) with the support of the National Human Genome Research Institute (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.

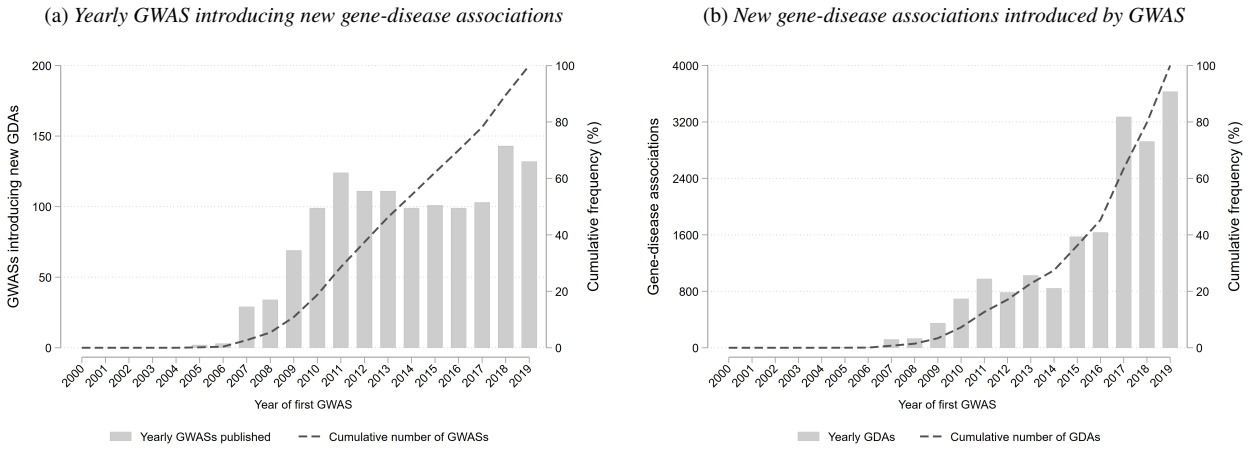
Table D.1: Descriptive statistics of GWAS papers that introduce a gene-disease association involving protein-coding genes

	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
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 generalit 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; *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 in the GWAS Catalog includes details on the paper's PubMed ID 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 to its corresponding MeSH Unique ID. Note that the MeSH taxonomy 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 (i.e., at level 5 or above), I assigned it to its parent branches up to level 4. Vice versa, if the matched disease was coarser (i.e., at level 3 or below), I assigned the finding to all its descending level 4 branches. This procedure permits harmonizing GWAS results by mapping gene-disease associations to a landscape of NCBI Gene IDs (unique genes) and level-4 MeSH IDs (unique diseases).

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).

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). The GWAS Catalog contains information on more GWAS. This paper focuses on GWAS that introduce new-to-the-world associations, but I use information from subsequent GWAS to code the replicability of these findings. 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 universities. 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 continues to grow, due to increased sample sizes that enable more statistically powerful analyses (Goldstein, 2009).

D.2 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% for genes and 83.70% for diseases. 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 information on the authors' affiliations from Dimensions, a data product by Digital Science (Herzog et al., 2020). I use these data to match each patenting firm in my sample to its publication portfolio using fuzzy string matching on firm names. Then, thanks to PubTator Central, I record all the genes and diseases studied by the firms. Overall, 3,346 of the 4,027 firms in my sample 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.

D.3 NIH's iCite Data

I use NIH's iCite data to collect additional bibliometric information for the publications in my sample, including each article's translational profile. iCite's Translation module builds on the Triangle of Biomedicine (Weber, 2013) and assigns every PubMed article three scores (Human, Animal, and Molecular/Cellular) based on the share of Medical Subject Headings (MeSH) attached by the National Library of Medicine that fall in the corresponding branches of the MeSH tree. NIH's Office of Portfolio Analysis implements this framework using fractional counts of MeSH terms, which places articles along a continuum from molecular and cellular bench research to human-focused work (Hutchins et al., 2019).

In this paper, I merge these article-level scores to firms' publication records and use them to classify the biological system in which gene-specific knowledge was produced. Specifically, I assign an article to the human, animal, or molecular/cellular category when its corresponding iCite score exceeds 0.5. This rule distinguishes knowledge rooted primarily in molecular and cellular experiments from knowledge generated in animal models or in more human-oriented research. I then use this variation in the mechanism analyses to test whether the value of prior gene-specific expertise depends on how translational that knowledge is, motivated by research arguing that more patient-oriented knowledge is better suited to evaluate whether a statistically significant GWAS association is also meaningful for therapeutic discovery.

D.4 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, to identify drug discovery, I focus on new molecules observed entering the earliest phases of drug development (recorded by Cortellis as the "discovery phase" and the "pre-clinical phase").

Table D.2 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 Targets Score. This suggests that research and development inputs might disproportionately 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.2: 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.

The final drug data source is the FDA's Approved Drug Products with Therapeutic Equivalence Evaluations, or *Orange Book*, which provides a linkage between approved small-molecule drugs and the patents that protect them (Durvasula et al., 2023). Unlike most settings, where product-patent links are difficult to observe, the Orange Book records this information because firms marketing brand-name drugs must report the relevant patents to the FDA. In this paper, I use these records to identify which patents in my sample are associated with approved products and, on that basis, to code which drug molecules ultimately reached commercialization.

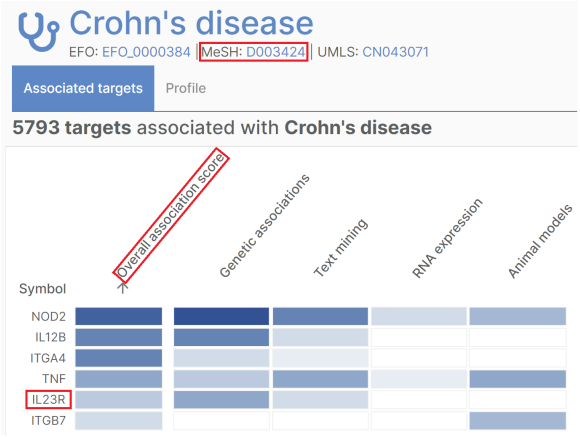
D.5 Open Targets Score

The Open Targets Platform is a public-private partnership that aggregates evidence on gene-disease associations to support drug target identification and prioritization (Ochoa et al., 2023). For each gene-disease pair, it reports a synthetic score summarizing the strength of the available evidence. Figure D.2 reports an example. I merge these data to my sample using NCBI Gene IDs and MeSH disease IDs, harmonized to level-4 MeSH categories. The score is available for 594,353 gene-disease pairs in my data, spanning 17,437 genes and 366 diseases. Throughout this appendix, I focus on the Open Targets measure based on direct evidence for the focal gene-disease relationship.[4]

This score provides a useful way to rank the relative therapeutic promise of gene-disease pairs, but it is not appropriate for identifying false positives. Open Targets partly reflects the amount of evidence accumulated around a pair, so understudied combinations are less likely to receive high scores even when they may be biologically meaningful. Conversely, the presence of evidence does not guarantee

[4]The Open Targets Platform also reports scores that incorporate indirect evidence using the disease ontology. I do not use those measures because they are conceptually farther from the focal gene-disease pair and are generally viewed as less stringent evidence (Ochoa et al., 2023).

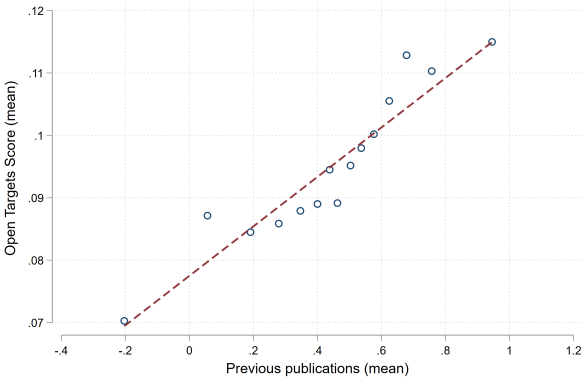
Figure D.2: Example of Open Targets Platform's synthetic score for the genes associated with Crohn's disease



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

that a pair is a valid therapeutic target. For this reason, the paper uses replication across GWAS to classify false positives and treats the Open Targets score only as a complementary, continuous proxy for the underlying value of a target within a disease.

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.

The first use of this measure is descriptive. Figure D.3 shows that the average Open Targets score of the gene-disease pairs mentioned in a patent application rises with the share of those pairs involving genes the firm had previously studied. Consistent with this pattern, Table D.3 shows that patents targeting higher-scoring pairs receive more forward citations and belong to larger patent families. Taken together, these results suggest that firms with prior gene-specific knowledge direct inventive

effort toward targets with higher underlying therapeutic potential.

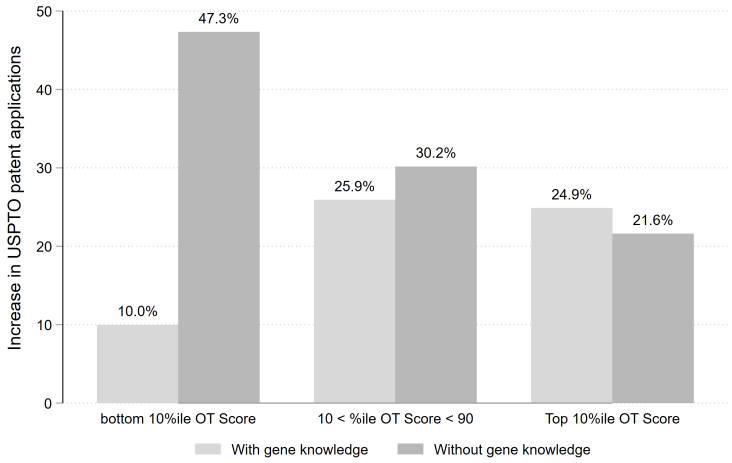
I then use Open Targets to address an alternative interpretation of the main results. One possibility is that domain knowledge simply makes firms more skeptical of atheoretical data-driven predictions (Allen and Choudhury, 2022). Under this view, firms with prior knowledge would avoid false positives only because they respond less to GWAS, potentially at the cost of forgoing valuable opportunities. Figure D.4 does not support this interpretation. Among the same set of GWAS-predicted opportunities, firms with domain knowledge concentrate more strongly on gene-disease pairs with higher Open Targets scores, whereas firms without domain knowledge invest disproportionately in lower-value targets. This pattern suggests that domain knowledge improves selection rather than merely reducing responsiveness. The evidence is more consistent with firms without domain knowledge making omission errors despite investing more broadly than with expert firms being uniformly conservative.

Table D.3: Patent applications targeting gene-disease pairs with higher Open Targets scores are of higher technological value

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

Note: *, **, *** denote significance at 5%, 1%, and 0.1% level respectively. Observations at the patent application level. Standard errors clustered at the patent assignee level are reported in parentheses. *Max(OT score)*: maximum value of the Open Targets 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 publication. *Patent family size*: number of patent applications in the same patent family.

Figure D.4: Firms without genetic domain knowledge invest proportionally more in lower-value GWAS opportunities



Note: This figure shows the increase in USPTO patent applications targeting a gene-disease pair after the publication of a GWAS, expressed as a share of the sample mean ($\hat{\beta}_{OLS}/\mu$). Estimates are from split-sample regressions that compare gene-disease pairs with different levels of therapeutic potential, as proxied by the Open Targets score, among the GWAS-identified associations. This exercise tests whether firms differ not only in how strongly they respond to data-driven predictions, but also in the value of the opportunities they select.

Appendix References

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

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

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.

Borusyak, K., X. Jaravel, and J. Spiess (2024): "Revisiting event-study designs: robust and efficient estimation," *Review of Economic Studies*, 91, 3253–3285.

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.

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.

Callaway, B. and P. H. Sant'Anna (2021): "Difference-in-differences with multiple time periods," *Journal of Econometrics*, 225, 200–230.

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.

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.

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.

Durvasula, M., C. S. Hemphill, L. L. Ouellette, B. Sampat, and H. L. Williams (2023): "The NBER Orange Book dataset: A user's guide," *Research Policy*, 52, 104791.

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.

Hutchins, B. I., M. T. Davis, R. A. Meseroll, and G. M. Santangelo (2019): "Predicting translational progress in biomedical research," *PLoS biology*, 17, e3000416.

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.

Kao, J. (2025): "Charted territory: Mapping the cancer genome and R&D decisions in the pharmaceutical industry," *UCLA Anderson*.

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.

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.

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.

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.

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. (2026): "Data-Driven Search and the Birth of Theory: 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.

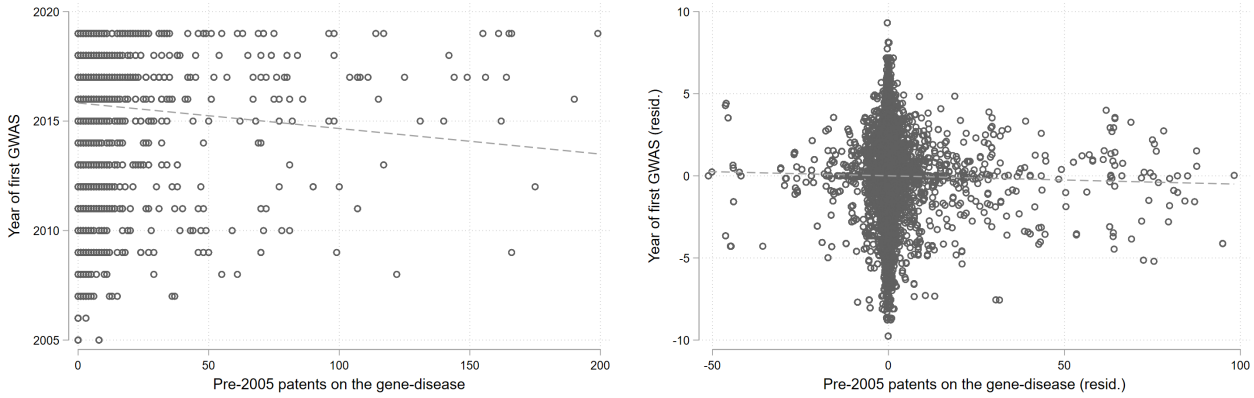
Weber, G. M. (2013): "Identifying translational science within the triangle of biomedicine," *Journal of Translational Medicine*, 11, 126.

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.

E Additional Figures and Tables

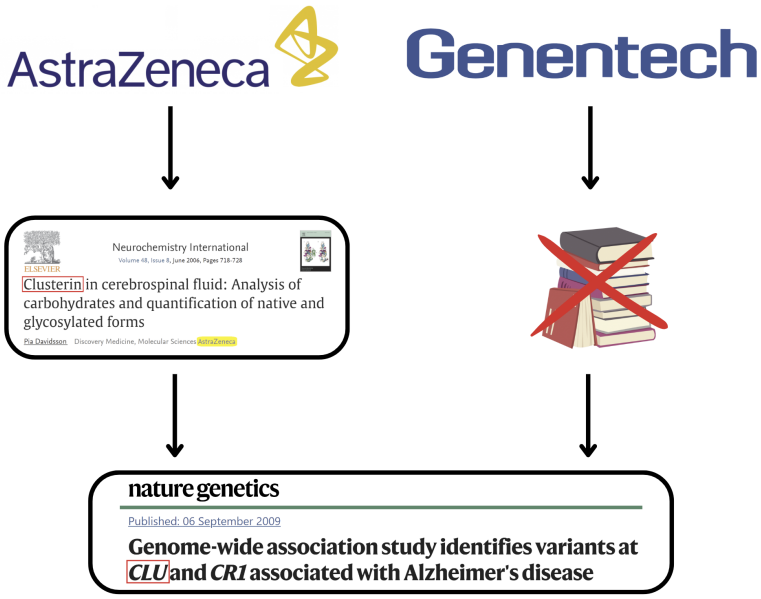
Figure E.1: The timing of GWAS associations is not related to past patenting on gene-disease pairs, once controlling for disease selection

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



Note: These panels 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.

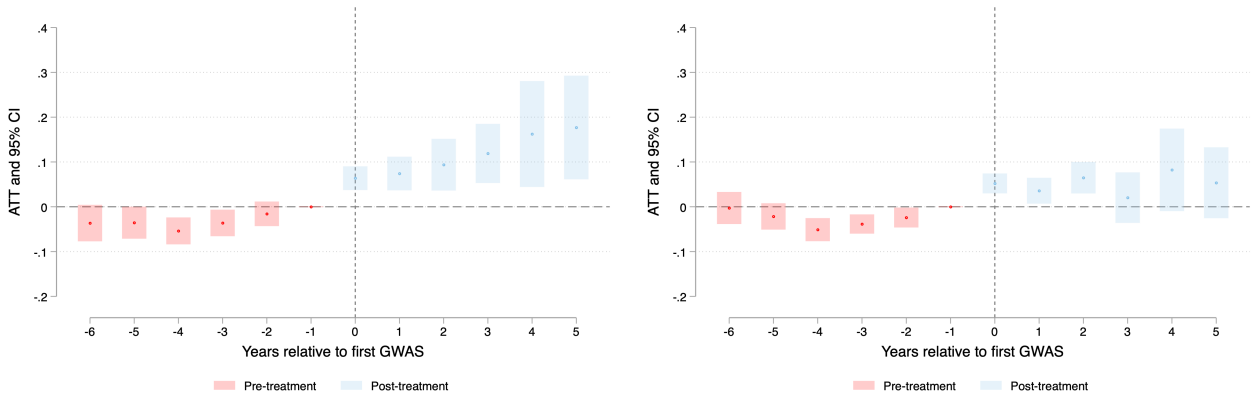
Figure E.2: Schema of the between-firms research design used for gene-disease level analyses



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 prior research on the gene involved, thereby improving their ability to recognize false positive associations.

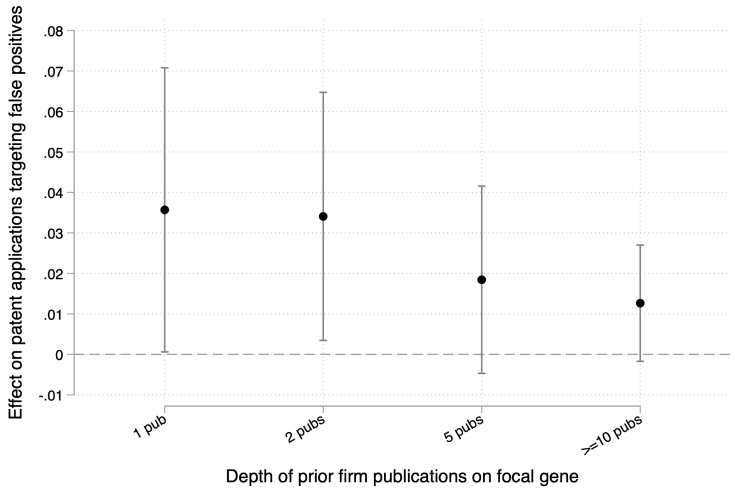
Figure E.3: Alternative event study estimators confirm that GWAS findings stimulate innovation investment, including for false positive associations that firms later abandon

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



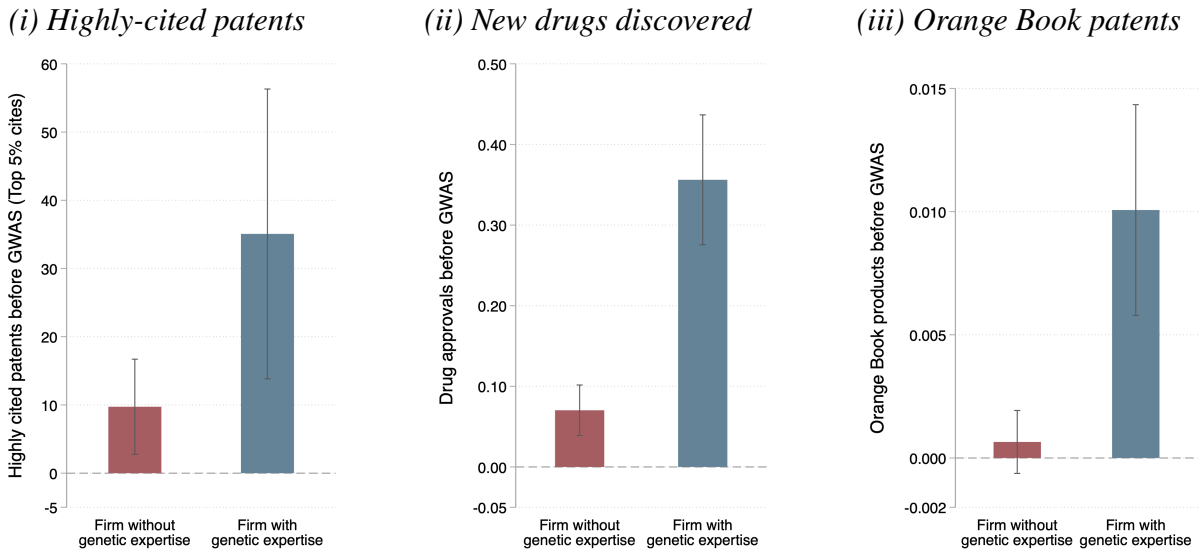
Note: The figure reports ATT estimates and 95% confidence intervals using the doubly robust difference-in-differences estimator of Callaway and Sant'Anna (2021). The dependent variable is the number of USPTO patent applications for innovations targeting a specific gene-disease combination. Treatment timing is defined by the first year in which a gene-disease pair is reported by a GWAS. Panel (a) reports estimates for all GWAS findings. Panel (b) reports estimates for false positive GWAS findings, defined as associations that are not replicated by subsequent GWAS on the same disease.

Figure E.4: The response to false positive GWAS findings declines with the depth of gene-specific knowledge



Note: This figure reports coefficient estimates and confidence intervals from gene-disease-year regressions. The dependent variable is the number of USPTO patent applications targeting false positive GWAS findings, separately measured for firms with different depths of prior publication activity on the focal gene, using increasingly stringent thresholds for the number of pre-GWAS publications on the focal gene. All regressions include gene-disease, gene-year, and disease-year fixed effects. Standard errors are clustered at the gene-disease level.

Figure E.5: Firms with deeper genetic expertise already had stronger innovation outcomes before GWAS



Note: This figure reports pre-2005 firm-year means and 95% confidence intervals, separately for firms with above- and below-median levels of genetic expertise. The figure is descriptive and is not based on the firm-level GWAS exposure regressions. *Firm with genetic expertise*: 0/1 = 1 if the firm has an above-median number of genetic publications before 2005. *Highly-cited patents*: firm-year count of highly cited patents (top 5% in the year). *New drugs discovered*: count of new drugs by the firm in a given year entering the clinical pipeline. *Orange Book patents*: count of firm patents in a given year that are listed in the FDA Orange Book for approved drugs.

Table E.1: Correlates of replicable GWAS findings: study design quality.

Dependent Variable:	Replicable association (0/1)		
	(1)	(2)	(3)
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.2: GWAS findings that are subsequently replicated involve gene-disease pairs with a higher score of genetic potential (Open Targets score)

Dependent Variable:	Open Targets score		Top 90%ile OT score (0/1)	
	(1)	(2)	(3)	(4)
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 (details in Appendix D.5). *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: citations to the article introducing them.

Dependent Variable:	Replicable association (0/1)		
	(1)	(2)	(3)
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.4: Firms increase their investments more when the associations are statistically stronger or the GWAS have better research designs

Dependent Variable:	USPTO patent applications			
	(1)	(2)	(3)	(4)
Post × GWAS	0.0432*	0.0581***	0.0828***	0.0432*
	(0.02019)	(0.01815)	(0.01809)	(0.01871)
...× Small P-value (0/1)	0.2455***			
	(0.03989)			
...× 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
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:	0.1314	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. *Small P-value*: 0/1 = 1 for associations with a p-value smaller 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 a replication analysis of their result.

Table E.5: Investments in false positive GWAS findings do not seem driven by strategic considerations, since those patents are less useful for litigation or less likely to be renewed than patents on true positive findings

Dependent Variable:	Strategic patent (0/1)		Litigated patent (0/1)		Renewed patent (0/1)	
	(1)	(2)	(3)	(4)	(5)	(6)
Post × GWAS	0.0155***		0.0059***		0.0058***	
	(0.00132)		(0.00115)		(0.00073)	
Post × False Positive GWAS		0.0094***		0.0022		0.0039***
		(0.00139)		(0.00119)		(0.00072)
Post × True Positive GWAS		0.0400***		0.0208***		0.0134***
		(0.00355)		(0.00319)		(0.00224)
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	137,254,556	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	7,223,924
Mean of Dep Var:	0.0339	0.0339	0.0282	0.0282	0.9900	0.9900

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-disease combination received at least one continuation patent; continuation patents claim priority to an earlier parent application but introduce new claims, allowing applicants to adjust claim scope after subsequent technological or commercial developments and thus serving as a proxy for strategic patenting (Righi and Simcoe, 2023). *Litigated patent*: 0/1=1 if the gene-disease combination received at least one patent that was subsequently involved in court litigation. *Renewed patent*: 0/1=1 if the gene-disease combination received at least one patent that was subsequently renewed. *Post × GWAS*: 0/1 = 1 in all years after a gene-disease pair is treated by its first GWAS. *False Positive*: 0/1 = 1 for GWAS findings that are never replicated by another GWAS about the same disease. *True Positive*: 0/1 = 1 for GWAS findings that are later replicated by another GWAS about the same disease.

Table E.6: Investments based on GWAS findings lead to successful drug development outcomes, but only for true positive findings

Dependent Variable:	New drugs		Orange Book patents	
	(1)	(2)	(3)	(4)
Post × GWAS	0.000282		0.00230***	
	(0.000165)		(0.000627)	
Post × False Positive GWAS		-0.000083		0.000745
		(0.000113)		(0.000417)
Post × True Positive GWAS		0.00174*		0.00854**
		(0.000689)		(0.00264)
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:	0.0000893	0.0000893	0.000516	0.000516

Note: *, **, *** denote significance at 5%, 1%, and 0.1% level respectively. Observations are at the gene-disease-year level. Std. err. clustered at the gene-disease level. *New drugs*: count of new drugs entering the clinical pipeline in a given year and that target a specific gene-disease combination. *Orange Book drug count*: count of USPTO patents granted in a given year that are listed in the FDA Orange Book for approved drug products and 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. *False Positive*: 0/1 = 1 for GWAS findings that are never replicated by another GWAS about the same disease. *True Positive*: 0/1 = 1 for GWAS findings that are later replicated by another GWAS about the same disease.

Table E.7: Alternative difference-in-differences estimators confirm that false-positive GWAS findings increase patenting mainly among firms without gene knowledge

Dependent Variable:	USPTO patent applications by...		
	... all firms	... firms with gene knowledge	... firms without gene knowledge
	(1)	(2)	(3)
Panel A: Callaway-Sant'Anna doubly-robust DID estimator			
ATT for False Positive GWAS	0.0532***	0.0112	0.0421***
	(0.0162)	(0.0077)	(0.0126)
Gene-Disease FE	YES	YES	YES
Disease-Year FE	YES	YES	YES
Gene-Year FE	YES	YES	YES
N	137,200,824	137,200,824	137,200,824
Panel B: Stacked-regression DID estimator			
Post × False Positive GWAS	0.0196**	-0.0022	0.0218**
	(0.0092)	(0.0044)	(0.0074)
Gene-Disease FE	YES	YES	YES
Disease-Year FE	YES	YES	YES
Gene-Year FE	YES	YES	YES
Event window	$[-3, 3]$	$[-3, 3]$	$[-3, 3]$
N	628,149,848	628,149,848	628,149,848
N of Gene-Diseases	7,223,924	7,223,924	7,223,924
Panel C: Borusyak-Jaravel-Spiess imputation DID estimator			
ATE for False Positive GWAS	0.0633***	0.0114	0.0520***
	(0.0165)	(0.0087)	(0.0109)
Gene-Disease FE	YES	YES	YES
Disease-Year FE	YES	YES	YES
Gene-Year FE	YES	YES	YES
Event window	$[-3, 3]$	$[-3, 3]$	$[-3, 3]$
Post-treatment horizons averaged	0 to 3	0 to 3	0 to 3
N	137,221,775	137,221,775	137,221,775

Note: *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Observations are at the gene-disease-year level. Standard errors are reported in parentheses. Panel A reports ATT estimates using the doubly robust difference-in-differences estimator of Callaway and Sant'Anna (2021). Panel B reports estimates from a stacked-regression difference-in-differences design, using cohort-specific event windows around the first GWAS year. Panel C reports estimates from the imputation-based difference-in-differences estimator of Borusyak et al. (2024); the reported coefficient averages post-treatment horizons 0 through 3. All specifications absorb gene-disease, gene-year, and disease-year fixed effects. *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. *False Positive*: 0/1 = 1 for GWAS findings that are never replicated by another GWAS about the same disease.

Table E.8: Firms without genetic domain knowledge continue to invest more in false positive GWAS findings after excluding industry-related confounds

Sample:	excluding industry-linked PIs (5-year window)		excluding industry-linked PIs (10-year window)		excluding industry-funded GWAS		excluding industry-coauthored GWAS	
Dependent Variable:	USPTO patent applications by...							
	firms with gene knowledge (1)	firms without gene knowledge (2)	firms with gene knowledge (3)	firms without gene knowledge (4)	firms with gene knowledge (5)	firms without gene knowledge (6)	firms with gene knowledge (7)	firms without gene knowledge (8)
Post × False Positive GWAS	0.0114	0.0419***	0.0157	0.0524***	0.0171	0.0589***	0.0120	0.0484***
	(0.00913)	(0.01052)	(0.00955)	(0.01129)	(0.00899)	(0.01146)	(0.00934)	(0.01104)
Post × True Positive GWAS	0.2159***	0.3114***	0.1437**	0.1815***	0.1309**	0.4174***	0.1604**	0.3606***
	(0.05471)	(0.06261)	(0.04577)	(0.04271)	(0.04435)	(0.07015)	(0.05447)	(0.06030)
Gene-Disease FE	YES	YES	YES	YES	YES	YES	YES	YES
Disease-Year FE	YES	YES	YES	YES	YES	YES	YES	YES
Gene-Year FE	YES	YES	YES	YES	YES	YES	YES	YES
N	137,185,852	137,185,852	137,174,433	137,174,433	137,213,592	137,213,592	137,186,688	137,186,688
N of Gene-Diseases	7,220,308	7,220,308	7,219,707	7,219,707	7,221,768	7,221,768	7,220,352	7,220,352

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. Columns 1-4 exclude GWAS led by academic PIs with extensive prior coauthorship ties to industry. Columns 5 and 6 exclude GWAS funded by industry. Columns 7 and 8 exclude GWAS coauthored by industry scientists. *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. *False Positive*: 0/1 = 1 for GWAS findings that are never replicated by another GWAS about the same disease. *True Positive*: 0/1 = 1 for GWAS findings that are later replicated by another GWAS about the same disease.

Table E.9: 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 (1)	...firms with gene knowledge (2)	...firms w/out gene knowledge (3)
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.10: Firms without domain knowledge invest proportionally less in false positives when the GWAS reports a smaller number of associations

Dependent Variable:		USPTO patent applications by...		
		... all firms	...firms with gene knowledge	...firms w/out gene knowledge
		(1)	(2)	(3)
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.11: 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
	(1)	(2)	(3)
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).

Table E.12: Placebo tests comparing GWAS responses by firms with focal gene knowledge versus related publication experience

Dependent Variable:	USPTO patent applications by...					
	... firms with gene knowledge		... firms with publications in other gene		... firms with publications in focal disease	
	(1)	(2)	(3)	(4)	(5)	(6)
Post × GWAS	0.0381***		0.0576***		0.0181**	
	(0.01009)		(0.0082)		(0.0085)	
Post × False Positive GWAS		0.0138		0.0160***		0.0116**
		(0.00821)		(0.0055)		(0.0056)
Post × True Positive GWAS		0.1354***		0.2240***		0.0442
		(0.03830)		(0.0344)		(0.0360)
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	137,254,556	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	7,223,924
Mean of Dep Var:	0.0376	0.0376	0.0456	0.0456	0.0541	0.0541

Note: *, **, and *** denote significance at the 5%, 1%, and 0.1% levels, respectively. Observations are at the gene-disease-year level. Standard errors are clustered two-way at the disease and gene levels. Columns (1)-(2) count applications filed by firms with prior publications on the focal gene (i.e., the main result of Table 2). Columns (3)-(4) count applications filed by firms with prior publications on genes other than the focal gene. Columns (5)-(6) count applications filed by firms with prior publications on the focal disease but not on the focal gene. *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 reported by its first GWAS. *Post* × *False Positive GWAS*: 0/1 = 1 in all post-publication years for GWAS findings that are never replicated by another GWAS about the same disease. *Post* × *True Positive GWAS*: 0/1 = 1 in all post-publication years for GWAS findings that are later replicated by another GWAS about the same disease.

Table E.13: Summary statistics for the firm-level analysis

	Firm-year descriptives					
	mean	median	st d	min	max	N
Total patent applications	3.3851	1	10.6831	0	214	40,924
Patent applications on false positives	0.7141	0	2.9157	0	112	40,924
Share of false positive patents	0.2216	0	0.3965	0	1	40,924
Average 5-year forward citations	10.7874	1	39.4166	0	1835	40,924
New drugs discovered	0.2440	0	2.3440	0	157	40,924
Orange Book patents	0.0344	0	0.2832	0	10	40,924
Firm-level GWAS exposure	277.8794	105	431.4171	0	3283	40,924

Note: This table reports summary statistics for the firm-year panel used to examine whether exposure to GWAS findings translates into differences in downstream innovation outcomes. *Total patent applications*: count of USPTO patent applications filed by the firm in a given year. *Patent applications on false positives*: count of USPTO patent applications filed by the firm in a given year for innovations that target gene-disease combinations associated with false positive GWAS findings. *Share of false positive patents*: share of USPTO patent applications filed by the firm in a given year that target gene-disease combinations associated with false positive GWAS findings. *Average 5-year forward citations*: average forward citations received by the firm's granted patents within five years. *New drugs discovered*: count of new drugs discovered by the firm in a given year. *Orange Book patents*: count of firm patents in a given year that are listed in the FDA Orange Book for approved and commercialized drugs. *Firm-level GWAS exposure*: count of GWAS findings that emerge in disease areas in which the firm was active before 2005.

Table E.14: Firm-level exposure to GWAS findings increases false positive patenting when findings fall outside firms' genetic expertise

Dependent Variable:	Patent applications on false positives			Share of false positive patents		
GWAS exposure:	All	Known genes	Unknown genes	All	Known genes	Unknown genes
	(1)	(2)	(3)	(4)	(5)	(6)
Firm-level GWAS exposure	0.0118	-0.0143	0.0136*	0.0030***	0.0015	0.0033***
	(0.0070)	(0.0786)	(0.0065)	(0.0009)	(0.0036)	(0.0009)
Year FE	YES	YES	YES	YES	YES	YES
Assignee FE	YES	YES	YES	YES	YES	YES
Total Patents FE	YES	YES	YES	YES	YES	YES
N	40,924	40,924	40,924	40,924	40,924	40,924
N of Firms	4,027	4,027	4,027	4,027	4,027	4,027

Note: *, **, *** denote significance at the 5%, 1%, and 0.1% levels, respectively. Observations are at the firm-year level. Std. err. clustered at the firm level are reported in parentheses. *USPTO patent applications on false positives*: count of USPTO patent applications filed in a given year for innovations that target a specific gene-disease combination following a false positive GWAS; the count is then divided between firms with and without previous publications on the gene. *Share of false positive patents*: share of USPTO patent applications filed by the firm in a given year that target gene-disease combinations associated with false positive GWAS findings. *Firm-level GWAS exposure*: measure of the extent to which GWAS findings emerge in disease areas in which the firm was active before 2005 (scaled by 100 for interpretability). *Known genes*: exposure to GWAS findings involving genes on which the firm had prior publications. *Unknown genes*: exposure to GWAS findings involving genes on which the firm had no prior publications.

Table E.15: After the emergence of GWAS, firms with genetic expertise experience larger gains in downstream innovation outcomes

Dependent Variable:	Average 5-year forward citations (1)	New drugs discovered (2)	Orange Book patents (3)
Post × Firm with genetic expertise	16.7737***	0.1195†	0.0307***
	(3.4472)	(0.0683)	(0.0078)
Year FE	YES	YES	YES
Assignee FE	YES	YES	YES
Total Patents FE	YES	YES	YES
N	40,924	40,924	40,924
N of Firms	4,027	4,027	4,027

Note: †, *, **,*** denote significance at 10%, 5%, 1% and 0.1% level respectively. Observations are at the firm-year level. Standard errors clustered at the firm level are reported in parentheses. *Post × Firm with genetic expertise*: 0/1 = 1 for firm-year observations after the emergence of GWAS among firms with above-median levels of genetic expertise before 2005. *Average 5-year forward citations*: average forward citations received by the firm's patents within five years. *New drugs discovered*: count of new drugs discovered by the firm in a given year. *Orange Book patents*: count of firm patents in a given year that are listed in the FDA Orange Book for approved and commercialized drugs. All regressions include year fixed effects, assignee fixed effects, and fixed effects for the total number of patents filed by the firm in that year.