# Data-Driven Search and the Birth of Theory: Evidence from Genome-Wide Association Studies *

Matteo Tranchero
The Wharton School

**Abstract**

How does big data change discovery? Datasets covering broad portions of a scientific landscape enable a data-driven approach to search, uncovering findings whose underlying mechanisms may be unclear. This shift has raised concerns that decoupling discovery from theoretical understanding lowers innovation quality by prioritizing incremental ideas or false positive signals. Even when successful, data-driven search could weaken incentives to develop theory and leave the consequences of new discoveries poorly understood. I examine these issues in human genetics, where genome-wide association studies (GWAS) enable a data-driven search for the genetic roots of disease. Compared with traditional theory-based approaches, GWAS expands the genetic landscape examined, increases outcome variability with a proportionally larger increase in breakthroughs, and stimulates follow-on work aimed at clarifying causal mechanisms. An instrumental variable strategy exploiting a technology-driven decline in the cost of GWAS supports a causal interpretation of these results. Mechanism tests show that these effects arise because data-driven search surfaces more empirical anomalies: valuable discoveries that depart from theoretical expectations and redirect subsequent research toward new theorizing. Together, the results suggest that big data technologies can fuel virtuous cycles of knowledge accumulation by increasing the frequency of findings that challenge existing theories.

*E-mail: mtranc@wharton.upenn.edu. A previous version of this paper was circulated with the title "Data-Driven Search and Innovation". I gratefully acknowledge the financial support of Panmure House's Emergent Thought Award.

# 1 Introduction

Constantin Polychronakos is a leading researcher on the genetic factors behind type 1 diabetes. Pinpointing the genes that carry disease-related mutations is critical because they can reveal targets for therapeutic intervention. Yet, the task is daunting: the human genome is vast, while the time and resources available to investigate it are necessarily limited. Faced with thousands of potential gene-disease combinations, how does Polychronakos decide where to search?

Understanding how Polychronakos searches in the genetic space can offer broader insight into the nature of innovation — not only for individual scientists, but also for organizations and firms. Many scientific and technological problems, from drug discovery to spatial exploration (Kang, 2025; Kao, 2024; Nagaraj, 2022), unfold on large combinatorial landscapes defined by prevailing paradigms (Dosi, 1982; Kuhn, 1962). Prior research suggests that rather than searching at random, actors concentrate on combinations they expect to be most valuable (Felin and Singell, 2025; Kneeland et al., 2020). Actors draw on knowledge embodied in mental models and cause-and-effect beliefs about how components work together to generate salience for a subset of promising directions (Felin et al., 2024; Gavetti and Levinthal, 2000; Sorenson, 2024). This theory-driven process is exactly what Polychronakos followed in a 2007 study. Building on evidence that the IRF5 gene is implicated in lupus, he hypothesized that IRF5 might also play a role in other autoimmune diseases, leading him to investigate its connection with diabetes (Qu et al., 2007).

However, this theory-driven approach appears to be increasingly supplanted by innovation practices powered by machine learning and artificial intelligence (AI) (Allen and McDonald, 2025; Agrawal et al., 2024). In pharmaceutical research, for example, scientists routinely screen data on millions of compounds without theoretical guidance (Jayaraj and Gittelman, 2018). Similarly, researchers use computational models to evaluate large numbers of novel material configurations (Merchant et al., 2023) or to detect patterns in geoscientific data (Bergen et al., 2019). The ability to collect comprehensive datasets that map large portions of a stable landscape makes it feasible to rely on data-driven prediction to prioritize promising leads (Ludwig and Mullainathan, 2024; Kim, 2025). Polychronakos himself adopted this approach in a second 2007 study. Rather than drawing on prior theory, he collected data from thousands of individuals to identify statistical associations between genetic variants and type 1 diabetes. This data-driven strategy pointed to KIAA0350, a gene of unknown function at the time, as a predicted contributor to the disease (Hakonarson et al., 2007).

In settings like genetics, where relatively stable paradigms define the search space (Dosi, 1982; Kuhn, 1962), the core uncertainty is where to look within a vast set of possible combinations (Fleming, 2001). With

large-scale data about that landscape, data signals can help triage which candidates merit follow-on testing, even when causal mechanisms remain unclear (Anderson, 2008; Leonelli, 2014). This creates an epistemic wedge between prediction and understanding: innovations can be identified and advanced despite limited insight into why or how they work (Evans, 2010). Even in drug development, the U.S. Food and Drug Administration requires that treatments be safe and effective, but not that their biological mechanisms be fully understood (Zittrain, 2019). The power of data-driven search lies in its capacity to surface valuable innovations precisely when prior theory is incomplete.

However, the diffusion of data-first approaches in innovation has also raised concerns (Leonelli, 2014; Mazzocchi, 2015). On the one hand, prioritizing candidates based on statistical patterns rather than mechanisms can degrade performance by diverting effort toward false positives (Berman and Van den Bulte, 2022; Calude and Longo, 2017; Tranchero, 2024). Data-driven search is further constrained by what is measurable, which can steer attention toward incremental directions at the expense of more valuable ones (Allen and McDonald, 2025; Hoelzemann et al., 2025). On the other hand, if the yield from data-driven innovation is similar or better than theory-driven approaches, it may ultimately backfire by weakening incentives to invest in explanation (Anderson, 2008; Balasubramanian et al., 2022). The risk is an "intellectual debt" in which deployment runs ahead of understanding, increasing reliance on innovations whose risks and side effects are harder to anticipate (Zittrain, 2019; Heaven, 2023). Whether a more data-driven search process improves innovation outcomes is therefore an open and consequential question.

Empirically examining how data change innovation presents a challenge: observing both theory-driven and data-driven search processes and connecting them to their resulting findings. In most settings, researchers see only the final outcomes without knowing the search process that produced them (Kneeland et al., 2020; Schilling and Green, 2011). I address this challenge by studying how scientists investigate the genetic roots of human diseases. The search for therapeutically useful gene-disease combinations happens with two distinct approaches. In *candidate gene studies*, researchers use existing knowledge to target genes expected to matter for a focal disease. In contrast, *genome-wide association studies* (GWAS) take a data-driven approach, scanning the genome to identify statistical associations between genes and diseases. Crucially, the method reported in the discovery publication reveals the search process that resulted in discovering each new gene-disease pair. By comparing associations introduced by GWAS to those introduced by candidate gene studies, I can estimate the effects of data-driven search in a domain central to drug development.

I assemble a novel dataset capturing the characteristics of gene-disease associations introduced between 2005 and 2016. The raw data come from DisGeNET, a comprehensive aggregator of information on genes

linked to human diseases. For each gene-disease pair, DisGeNET identifies the original publication reporting the association, along with all subsequent articles that further investigate it. I augment these data in two ways. First, I merge the list of articles introducing new associations with the GWAS Catalog, which records studies employing genome-wide methods. This allows me to identify which combinations resulted from a data-driven search process. Second, I merge the follow-on articles studying each association with NIH's PubTator3, which uses advanced AI models to classify the epistemic nature of scientific claims. This enables me to separately count subsequent studies that seek to elucidate how the gene is *causally* involved in the disease from those adding generic evidence about the gene-disease combination. The former serves as my proxy for the development of new theoretical understanding.

To ground the discussion, consider again Polychronakos's two studies on the genetic basis of type 1 diabetes. In one, Qu et al. (2007), he leveraged theoretical knowledge about the IRF5 gene to target it with a candidate gene study. In the other, Hakonarson et al. (2007), he used a genome-wide approach that uncovered an association with KIAA0350, a gene of unknown function at the time and therefore ignored by theory-driven investigations. Unlike the IRF5–diabetes association, the KIAA0350–diabetes link involved an understudied gene that ultimately proved highly valuable for therapeutic development. It also catalyzed substantially more downstream work: 24 follow-on articles, mostly to elucidate how KIAA0350 is causally linked to diabetes, compared to just two for the IRF5-diabetes combination. That these studies were conducted by the same researcher, in the same year, for the same disease underscores the central point: the search process itself shaped the scope and value of the discovery, as well as the extent of subsequent theorizing generated.

My empirical analyses mirror the patterns illustrated in the case study. Compared to candidate gene studies, gene–disease associations introduced by GWAS are twice as likely to involve the least studied human genes. Mechanism tests suggest that genome-wide screening removes discretion over gene choice, loosening the path dependence that concentrates attention on known genes. GWAS findings also skew more heavily toward both extremes of scientific quality, with a notable asymmetry: the increase is proportionally larger for breakthroughs than for false positives. Even among combinations discovered by the same researcher, GWAS associations are 36% more likely to fall in the top tail of scientific importance, translating into a higher average yield. Finally, GWAS discoveries encourage downstream theorizing, rather than crowding it out. They attract substantially more follow-on work, with the increase in attention driven by mechanism-oriented research (+141–221% over the sample mean). These patterns are robust to stringent fixed effects that compare discoveries made by the same laboratory, for the same disease and year, and even for the same gene.

A potential concern is that GWAS adoption is a researcher choice and may therefore be endogenous to the outcomes I document. Fixed effects mitigate selection on time-invariant disease and researcher characteristics but cannot rule out time-varying shocks that jointly affect method choice and discovery. I therefore introduce an instrumental variable (IV) strategy that exploits the sharp decline in genotyping costs, that is, the cost of collecting genome-wide genetic data. Because genome-wide studies require large cohorts, this cost decline increased GWAS adoption disproportionately in more prevalent diseases, where large samples are easier to assemble and funding is typically more available (Heeney, 2021; Visscher et al., 2017). I capture this differential exposure with a shift-share instrument that interacts annual genotyping costs with predetermined disease prevalence in 2005. Extensive tests and qualitative evidence indicate that the instrument is both plausible and relevant, with a strong first stage. The resulting 2SLS estimates align with the baseline findings and are, if anything, larger in magnitude.

Finally, I explore why GWAS elicit stronger downstream theorizing than candidate gene studies. Building on the Kuhnian notion of *empirical anomalies* (Kuhn, 1962; Chai, 2017), I propose that genome-wide screening is more likely to surface findings that violate prevailing theoretical expectations in a disease area. I capture this idea with an Anomaly Index that measures how unlikely a new gene–disease association is given what prior knowledge would predict from gene co-occurrence patterns across diseases (Shi and Evans, 2023). The evidence supports this channel. Anomalous discoveries are disproportionately concentrated in both tails of scientific value, attract mostly mechanism-oriented follow-on work, and are more likely to be introduced by GWAS. This mechanism also yields a counterintuitive implication. One might expect data-driven search to matter most where theory is less developed and thus provides little guidance about where to look, yet I find the opposite: GWAS discoveries trigger more theorizing in high-knowledge disease areas because they confront well-developed theory with unexpected empirical patterns that demand new explanation.

This paper makes three contributions. First, it advances our understanding of how data change innovation (Agrawal et al., 2024; Allen and McDonald, 2025; Nagaraj and Tranchero, 2024). When data describe broad portions of a fixed technological landscape (Kang, 2025; Kao, 2024; Nagaraj, 2022), I document that the very nature of the search process can change. Second, my findings contribute to the epistemology of discovery (Chai, 2017; Kuhn, 1962). In combinatorial domains, strong theory can make search efficient yet path dependent, concentrating attention on familiar regions (Fleming and Sorenson, 2004; Knudsen and Levinthal, 2007). I show that new data loosen this constraint by making neglected parts of the landscape visible and testable, clarifying a central function of scientific instruments as data-generating tools (Furman and Teodoridis, 2020; Murray et al., 2016; Price, 1983; Rosenberg, 1982). Third, the paper speaks to

the emerging theory-based view in management (Camuffo et al., 2024; Felin and Zenger, 2017; Felin and Singell, 2025; Ott and Hannah, 2024) by identifying empirical anomalies as a key engine of theorizing (Kuhn, 1962; Mullainathan and Rambachan, 2024). By confronting researchers with patterns that sit uneasily with prevailing explanations, I find that data-driven search supplies puzzles that catalyze theory building.

The paper proceeds as follows. Section 2 outlines the conceptual framework. Section 3 illustrates the empirical setting. Section 4 describes the data and novel measurement approach, while Section 5 and 6 present the empirical results. Section 7 discusses the findings.

# 2 Conceptual Framework

## 2.1 The Growing Role of Data in Research

Historians of science have long emphasized advances in scientific instruments as a driver of knowledge advancement (Mokyr, 2002; Rosenberg, 1982; Price, 1983). A key reason is that such tools expand the capacity to generate *data*: they make previously invisible phenomena observable and measurable, rendering those observations usable for scientific analysis (Barbosu and Teodoridis, 2025). For example, model organisms such as the OncoMouse permit the collection of data on biological processes that cannot be isolated in humans (Murray et al., 2016). Similarly, motion-sensing sensors collect movement data that was previously difficult to obtain, enabling new lines of research (Furman and Teodoridis, 2020). By extending the observable portions of specific scientific domains, data-generating tools allow researchers to "tune in and turn on" to new aspects of the natural world (Price, 1983, p. 246).

In recent decades, advances in data-generating tools have made it feasible to map entire scientific landscapes (Barbosu and Teodoridis, 2025). Unlike traditional data collection, which relies on sampled observations, these mapping efforts aim for systematic coverage of a domain (Kao, 2024), enabling empirical inquiry across the full knowledge space. This shift is especially salient when the landscape is literal. Satellite programs such as Landsat provide continuous measurement of the Earth's surface (Nagaraj, 2022), and the Sloan Digital Sky Survey has produced a similarly comprehensive map of the sky (York et al., 2000). However, the same logic also applies to more conceptual spaces. In medicine, efforts such as the Human Genome Project or The Cancer Genome Atlas systematically measured the genetic makeup of humans, fostering new discoveries and drug development (Kang, 2025; Kao, 2024; Williams, 2013). Similar patterns appear across fields as diverse as biology and economics, where research increasingly relies on this type of comprehensive data (Leonelli, 2014; Nagaraj and Tranchero, 2024).

The ability to map entire landscapes represents more than a quantitative increase in data, because it can

*qualitatively* change how search is organized. When data are costly to produce, innovators face tight limits on experimentation and observation (Adner and Levinthal, 2024; Barbosu and Teodoridis, 2025; Price, 1983). In such settings, success depends on the ability to articulate theories that guide attention toward the most promising regions of the landscape to explore (Camuffo et al., 2024; Felin and Singell, 2025; Sorenson, 2024). Theory generates salience by identifying a limited set of plausible hypotheses and focusing data collection efforts toward ideas believed to offer the highest expected returns (Felin and Singell, 2025). Search is therefore inherently *theory-driven*, because abstract reasoning plays an upstream role in logically narrowing the relevant search space (Arora and Gambardella, 1994; Felin et al., 2024; Sorenson, 2024). Consistent with this view, empirical evidence shows the performance advantage of using theory as a scientific "map" to efficiently navigate the problem landscape (Fleming and Sorenson, 2004).

However, advances in data-generating tools have made it inexpensive to collect observations over large swaths of a landscape, and this shift can alter the organization of search itself (Anderson, 2008; Leonelli, 2014). Rather than beginning with theory and using data primarily to test pre-specified hypotheses, researchers can deploy predictive methods to prioritize where to look (Ludwig and Mullainathan, 2024; Jayaraj and Gittelman, 2018). In well-defined search spaces, large datasets can be mined to surface promising ideas or technological combinations by detecting statistical regularities and new associations (Choudhury et al., 2021; Shrestha et al., 2021). This logic is most applicable in scientific and technological domains governed by a meta-theoretical paradigm that delineates the relevant objects, their feasible recombinations, and the criteria by which progress is evaluated (Dosi, 1982; Kuhn, 1962). In such settings, uncertainty is primarily combinatorial (Fleming, 2001), so prediction can serve as a screening device within an otherwise stable landscape.[1] The growing availability of big data and predictive technologies therefore enables an alternative strategy, which I refer to as *data-driven search*. Critically, this prioritization can proceed without a fully articulated causal account of why a given combination should work, because initial guidance comes from "black-box" prediction rather than mechanism-based understanding (Agrawal et al., 2024; Kim, 2025).

This prediction-first logic changes what it means to explore a landscape (Anderson, 2008; Ludwig and Mullainathan, 2024). Because data-driven search is not tethered to existing understanding, it can push beyond the boundaries of current knowledge and surface promising leads outside established search trajectories (Choudhury et al., 2020; Price, 1983; Shrestha et al., 2021). A widely discussed illustration is the discovery of a new antibiotic, halicin, identified by mining chemical data before researchers could explain its mechanism

[1]This separates the phenomenon studied here from settings of radical uncertainty, such as entrepreneurial and market contexts, where the opportunity set is not given and the payoff-relevant landscape is shaped endogenously by actors' choices and strategic interaction (Felin and Zenger, 2017; Felin et al., 2024). In this sense, the focus of this paper is on alternative ways to generate salience and prioritize effort allocation within a pre-defined (but enormously vast) opportunity set (Felin and Singell, 2025).

of action (Stokes et al., 2020). Cases like halicin foreground two fundamental questions. How do such discoveries compare to theory-guided discoveries in novelty and scientific value? And do they redirect effort away from the mechanism-building work that has historically underpinned technological progress (Arora and Gambardella, 1994; Mokyr, 2002)? The following sections examine whether, and how, a more data-driven search process impacts scientific progress.

## 2.2 Data-Driven Search and Innovation Quality

Innovation is often characterized as search over a vast combinatorial landscape, where most combinations have low value and a few correspond to high peaks (Levinthal, 1997; Schilling and Green, 2011). Borrowing this imagery, theory operates as a filter in this space. Cause-and-effect knowledge enables forward-looking inference, focusing attention on components more likely to solve a focal problem (Ehrig and Zenger, 2024; Felin and Singell, 2025; Sorenson, 2024). Theory also helps triage ideas, either by directing data collection toward the most informative tests (Adner and Levinthal, 2024; Camuffo et al., 2024) or by enabling "offline" assessments without direct trial (Gavetti and Levinthal, 2000). As Popper (1963, p. 36) remarked, *"theory is a prohibition"*: by ruling out broad regions of the space, it defines paradigmatic boundaries around what is seen as plausible and therefore worth investigating (Dosi, 1982; Kuhn, 1962). In turn, narrowing attention toward theoretically grounded candidates increases efficiency and reduces outcome variability (Fleming and Sorenson, 2004), consistent with Popper's idea that *"the more a theory forbids, the better it is"* (Popper 1963, p. 36).

Because theory improves search by screening candidates through mechanism-based reasoning, a shift toward data-driven search raises natural concerns. Ranking candidates based on predictive patterns rather than articulated mechanisms weakens the quality filter that theory provides (Felin et al., 2024; Popper, 1963). A first implication is a higher risk of false positives (Berman and Van den Bulte, 2022). This risk is amplified in large combinatorial landscapes, where the sheer number of potential relationships makes it easy to mistake chance regularities for robust signals (Calude and Longo, 2017). Consider drug discovery, where large-scale screening is routinely used to identify initial "hits." Although high-throughput screening (HTS) accelerates lead identification, it also produces many false positive candidates that appear promising but later prove to be dead ends (Sink et al., 2010). Similar warnings arise in scientific discovery, including genetics. As MacArthur (2012) cautions, *"it has never been easier to generate high-impact false positives than in the genomic era, in which massive, complex biological data sets are cheap and widely available"* (p. 427). Spurious leads can end up diverting substantial resources away from valuable targets (Tranchero, 2024).

A second risk is the "streetlight effect" (Hoelzemann et al., 2025). Because data-driven search is constrained by what is measurable, it can prioritize what is easy to quantify rather than what is scientifically important. As Merton (1948, p. 513) warned long ago, new data can *"divert attention from problems with larger implications onto those for which there is the promise of immediate solutions."* Recent evidence is consistent with this mechanism, showing that expanded data availability can tilt search inwards and toward less novel directions (Kang, 2025; Kao, 2024; Kim, 2025). The same logic also operates endogenously inside organizations: as data analytics infrastructures expand, they channel attention toward ideas that fit existing metrics, which favors incremental innovation (Benner and Tushman, 2002). By contrast, breakthroughs are difficult to assess using existing metrics and historical data, making them less likely to be prioritized through data-driven search alone (Allen and McDonald, 2025; Felin et al., 2024).

These concerns highlight what is lost when theory no longer serves as the primary quality filter. The same filter, however, carries a cost that is often overlooked: by reducing variance in what is pursued, it can limit the exploration needed to escape locally attractive but ultimately suboptimal regions of a landscape (Levinthal and March, 1993; March, 1991; Schliesmann, 2025). This tradeoff becomes more consequential as knowledge accumulates, because search trajectories can become entrenched (Chai, 2017; Kuhn, 1962) and overly effective at screening out alternatives that do not fit prevailing explanations (Knudsen and Levinthal, 2007). The consequence is that theory-driven search can become path dependent, producing competency traps that leave actors stuck on local optima (Levinthal and March, 1993). Instead, data-driven search can relax this constraint by prioritizing candidates that would have been screened out under existing theory (Ludwig and Mullainathan, 2024; Shrestha et al., 2021). Barring systematic bias in the data (Cao et al., 2024), this approach can broaden the scope of recombination beyond familiar domains and increase the likelihood of longer jumps in well-defined landscapes (Levinthal, 1997).

It is worth noting that even if broader search has been empirically linked to a higher chance of breakthroughs, expanding scope does not inherently guarantee better outcomes (Kaplan and Vakili, 2015). Theory-driven search may remain more efficient, in the sense of locating high-value innovations with fewer attempts (Felin and Singell, 2025; Fleming and Sorenson, 2004). By contrast, because it relaxes upstream theoretical filtering, data-driven search may primarily reshape the distribution of outcomes, increasing the frequency of both false positives and exceptional discoveries. Which search strategy has higher yield on average is unclear ex ante and likely contingent on context, including the strength of existing theory (Sorenson, 2024). In domains where theory is more established, one might expect less need for data-mined signals. Yet stronger theory can also intensify path dependence, increasing the likelihood of overlooking high-value

opportunities (Chai, 2017; Knudsen and Levinthal, 2007). This leaves open the rather surprising possibility that data-driven search is most valuable precisely where theory is strongest. Whether, and when, this is true remains an empirical question.

### 2.3 Data-Driven Search and Theory Generation

A discovery can be practically valuable even when researchers lack a clear explanation for why it works (Zittrain, 2019). Accordingly, some have argued that if predictive screening yields innovations that perform as well as, or better than, those generated through theory-driven search, incentives to develop new theory may weaken (Anderson, 2008; Leonelli, 2014; Mazzocchi, 2015). Consider again the case of halicin (Stokes et al., 2020). If the goal is to identify an antibiotic drug, a reliable predictive signal that a molecule works can be sufficient to justify further development even when the mechanism remains unclear. More generally, prediction can be valuable absent explanation when the objective is control over a particular outcome (Shmueli, 2010). In econometric terms, theory-driven search emphasizes learning $\hat{\beta}$ in order to act on outcomes $y$, whereas data-driven search directly forecasts $\hat{y}$ to guide choice (Mullainathan and Spiess, 2017). In the limit, sufficiently accurate prediction can substitute for theory, reallocating effort toward improving data and algorithms rather than understanding (Anderson, 2008).

However, commentators worry that this substitution can generate an "intellectual debt" (Mazzocchi, 2015; Zittrain, 2019): innovations adopted on the strength of predictive performance while postponing the understanding, creating a backlog of unanswered "whys" that must be paid down later through costly validation or failures. This debt is reminiscent of earlier, pre-scientific eras of discovery (Arora and Gambardella, 1994). As Mokyr (2002) notes, when useful knowledge is limited, research often proceeds through trial-and-error, and mechanisms are clarified only after adoption. Data-driven search reproduces this sequence, but at a far greater scale (Anderson, 2008; Kim, 2025). It can uncover large numbers of discoveries faster than mechanisms can be established, expanding the stock of black-boxed innovations that remain poorly understood. The social risk is that, without theory as a basis for extrapolation, boundary conditions and side effects are more likely to be discovered only after deployment, sometimes when irreparable harm has already occurred (Zittrain, 2019).

While much of the debate emphasizes that big data may crowd out theory, a reading of the philosophy of science suggests the opposite possibility. Kuhn (1962) argued that paradigms discipline attention by defining what counts as a relevant fact, which implies that observations inconsistent with prevailing assumptions are often dismissed (Chai, 2017). Data-driven methods applied to large-scale datasets can relax this constraint

by surfacing relationships that would be unlikely to appear in data collected solely to test pre-specified hypotheses (Choudhury et al., 2021). By weakening theory's upstream filter, data-driven search can increase the supply of *empirical anomalies*, that is, findings that sit uneasily with existing explanations (Kuhn, 1962; Mullainathan and Rambachan, 2024). Such anomalies can be theoretically productive precisely because they reveal the boundaries of prevailing accounts and create pressure for conceptual revisions (Ott and Hannah, 2024; Shi and Evans, 2023). In this sense, data-driven search may foster a virtuous cycle in which anomalous results stimulate follow-on theorizing rather than displacing it. This mechanism also yields a counterintuitive implication for where the effect should be strongest. When prior theory is thin, expectations are weak, so unexpected correlations are less likely to register as puzzles that demand explanation. Instead, with stronger priors, departures from expectations are more salient (Camuffo et al., 2024), thereby increasing the impetus to invest in mechanism-oriented explanation.

This virtuous cycle, however, is not guaranteed. Kuhn's own account implies that anomalies often fail to trigger theorizing because the initial response is to discount them. Consistent with this view, Chai (2017) shows that researchers near a breakthrough sometimes overlooked anomalies because attention was directed toward confirming the established paradigm. Even when anomalies were noticed, researchers often talked themselves out of pursuing them, reframing them as curiosities rather than phenomena requiring explanation (Chai, 2017). Data-driven search can amplify this tendency by increasing the volume of anomalous signals that compete for attention. In an age of data deluge, the bottleneck shifts from generating leads to allocating scarce attention to them (Bikard, 2018; Piezunka and Dahlander, 2015). When options proliferate, anomalies may be ignored or pursued superficially, limiting their contribution to theory building. These dynamics make the relationship between data-driven search and theory generation an open question, which I examine empirically in the remainder of the paper.

# 3 Empirical Setting

## 3.1 Genetic Research Before GWAS

Genes are sequences of DNA bases that encode the "instructions" for synthesizing molecules fundamental for the organism's functioning. Genes often harbor DNA mutations, some of which can drive the emergence of diseases. Most common conditions, such as diabetes or hypertension, are polygenic: they result from mutations across multiple genes and their interaction with environmental factors (Uffelmann et al., 2021). Understanding the genetic roots of such diseases is critical because causative genes can become targets for drug intervention. However, identifying the genes involved in thousands of polygenic diseases requires searching a massive combinatorial space of over $\sim 19,000$ known human genes.

Scientists have traditionally searched for drug targets using a *candidate gene approach*, which involves three main steps. First, they choose a disease to study, often guided by its prevalence in the population. Second, they hypothesize which genes might play a role in the disease's etiology. Third, they test these hypotheses using a case-control design: namely, by examining whether subjects with the disease are more likely to carry mutations in the candidate genes than control subjects. Importantly, the selection of candidate genes reflects scientists' theoretical understanding of which genes are likely to matter for the disease, and why. For example, after the IRF5 gene was associated with lupus, Qu et al. (2007) drew on their knowledge of autoimmune diseases to hypothesize that the same gene could also play a role in type 1 diabetes. This led to a candidate gene study that collected targeted data on the IRF5-type 1 diabetes combination for the first time (Panel (a) of Figure 1).

Candidate gene studies epitomize a theory-driven model of scientific inquiry. Yet despite notable successes, these studies inherently steer scientists toward genes for which functional hypotheses can be formulated based on existing knowledge (Visscher et al., 2017). This has led to an extreme concentration of attention on a small subset of theoretically well-understood genes (Stoeger et al., 2018). The focus on exploiting a handful of "superstar" genes might be justified by the risks involved in exploring the remaining gene pool, which likely contains many dead ends. Still, the lack of broader exploration across genetic space is arguably suboptimal since the chances of finding treatments for polygenic diseases would improve by casting a wider net (Uffelmann et al., 2021).

### 3.2 The Emergence of GWAS

In the early 2000s, two major developments converged to create an alternative to candidate gene studies. First, the completion of reference genomes such as the Human Genome Project and the HapMap Project provided a comprehensive map of human genes. Second, the cost of collecting genetic data started to drop rapidly. Together, these advances paved the way for an alternative approach to search for drug targets: genome-wide association studies (GWAS) (Visscher et al., 2017). As the name suggests, GWAS are case-control studies that use genetic data to identify mutations associated with disease across the entire genome.

Like candidate gene studies, researchers conducting a GWAS begin with the selection of a disease to study and then look for mutations that appear more frequently in individuals with the condition than in those without it. However, the two approaches diverge sharply in how genetic targets are selected (Uffelmann et al., 2021). Rather than focusing on specific gene-disease pairs informed by existing theory or prior evidence, GWAS rely on a fully data-driven approach. Researchers collect DNA samples from large numbers of cases and

controls and systematically test for differences between their genetic make-up in every gene. Genes found to harbor mutations more common in cases than in controls are flagged as potentially implicated in the disease's etiology, regardless of whether their functional role in the disease is understood. As such, they become candidates for further investigation or pharmaceutical intervention, even in the absence of a clear causal explanation (Tranchero, 2024).

For example, consider the GWAS conducted by Hakonarson et al. (2007). The researchers analyzed a study population of 563 patients with type 1 diabetes and 1,146 controls without the condition. Using genotyping microarrays, they collected data about the genetic features of their study subjects. The comparison between cases and controls highlighted several significant differences across the genome: some in genes already known to be linked to diabetes (such as the insulin gene INS), but others in less familiar regions, including the gene KIAA0350 (Panel (b) of Figure 1). At the time, the function of KIAA0350 was unknown, which likely explains why it had been overlooked in candidate gene studies. Since the publication of Hakonarson et al. (2007), the KIAA0350–type 1 diabetes association has proven highly impactful and has been the focus of multiple follow-up studies and clinical trials. Additional details on this case are provided in Appendix B.

GWAS represent a prime example of data-driven search (Visscher et al., 2017). A genome-wide search scans the entire set of human genes for associations with a given disease, highlighting promising gene-disease combinations (Panel (c) of Figure 1). Unlike candidate gene studies, where researchers select targets based on theoretical considerations, GWAS generate discoveries purely from patterns that emerge in the data. Yet, this very feature also makes GWAS a source of ongoing controversy within the scientific community, even years after their diffusion (Callaway, 2017). GWAS produce high rates of false positives (MacArthur, 2012; Tranchero, 2024), and they cannot explain the underlying biological mechanisms of the associations uncovered (Uffelmann et al., 2021). Moreover, achieving sufficient statistical power across thousands of genes requires large sample sizes, making GWAS much more expensive than targeted gene-candidate studies. The large cost of data collection posed a significant barrier to their diffusion,[2] leading some skeptics to dismiss GWAS as a costly detour in the quest for genetic insight (Loos, 2020). This is why candidate gene approaches remain in use and questions persist about the value of data-driven genetics.

[2]Over time, the decrease in sequencing costs led to a greater diffusion of GWAS. In Section 5.3, I exploit the cost reduction as an instrument for adoption of the GWAS approach. For a critical discussion of early GWAS costs, see: *www.blog.goldenhelix.com/have-we-wasted-7-years-and-100-million-dollars-on-gwas*.

# 4 Measurement Strategy and Data

## 4.1 Defining the Search Landscape: Gene-Disease Combinations

While search is often conceptualized as occurring within a landscape (Levinthal, 1997), such landscapes are typically difficult to study empirically (Fleming and Sorenson, 2004; Schilling and Green, 2011). This paper leverages the setting of genetic research to address that challenge. A central goal in genetics is to identify which human genes could serve as drug targets for a given disease. In this context, genes and diseases constitute the relevant components of the search problem. Any gene can, in principle, be tied to any disease, creating a well-defined space of millions of potential gene-disease combinations (see Appendix C.1 for an illustration). Importantly, with the completion of the Human Genome Project in the early 2000s, all protein-coding genes have been mapped. This allows the entire landscape of possible gene-disease combinations to be characterized separately from how search unfolds over it.

I obtain data on links between human diseases and their genetic causes from DisGeNET v7.0 (Piñero et al., 2020), an aggregator of scientific sources considered one of the most comprehensive repositories of genetic knowledge (Hermosilla and Lemus, 2019). The database compiles gene-disease associations (GDAs) from specialized sources, including curated datasets and publications indexed in PubMed. The units of observation are gene-disease pairs first introduced in scientific articles published between 2005 and 2016. For each pair, I assemble both the first publication reporting it and all subsequent articles that investigated it. I focus on associations that link a protein-coding gene to a disease, syndrome, or abnormality with health implications. The final dataset includes 369,302 newly introduced gene-disease combinations, encompassing a total of 14,136 genes and 9,863 disease categories.[3] Bibliometric information of papers introducing novel gene-disease associations is taken from NIH's iCite data. To identify the principal investigator (PI), I use data on the last author of each publication from the Author-ity database (Torvik and Smalheiser, 2021).

## 4.2 Data-Driven and Theory-Driven Search Processes

Empirically investigating how alternative search strategies shape innovation requires linking the search process to its resulting outputs. This is difficult because researchers typically observe outcomes but not the search process that generated them (Kneeland et al., 2020; Maggitti et al., 2013). A common workaround is to infer search strategies from observable features of outputs. For example, the breadth of a paper's references or the diversity of a patent's USPC classes have been used as indicators of how inventors searched (Fleming,

[3] Since the analysis focuses on comparing GWAS and candidate gene studies, the sample is restricted to new gene-disease combinations introduced on or after 2005 (the year of the first GWAS). As a result, the number of genes in the dataset is lower than the total number of human genes, as some were not implicated in any disease during the sample period.

2001; Schilling and Green, 2011). The problem is that these proxies are inherently ex-post and therefore conflate the search strategy with the realized outcome. Therefore, the goal is to find a setting where the search strategy adopted is known without conditioning on observing successful outcomes.

In the context of genetics, I can address this issue because the search for new drug targets happens in two distinct ways. Scientists studying a disease can carry out candidate gene studies using their theoretical priors to target specific genes. Alternatively, they can use GWAS, which scan the entire genome to identify genetic mutations correlated with the disease. Candidate gene studies vary in how targeted they are, ranging from testing a single gene-disease pair (as in the paper of Qu et al., 2007) to examining broader regions that may contain several genes. GWAS, by contrast, involve no genetic targeting at all. Crucially, I can infer which search process led to a given gene-disease combination by coding the method used in the paper that first introduced it. This information comes from the GWAS Catalog, a manually curated database of all studies that adopt a genome-wide approach (MacArthur et al., 2017). Studies are eligible for inclusion in the GWAS Catalog only if they use a DNA microarray to scan the entire genome without any genetic targeting. In total, the DisGeNET data include 8,655 gene-disease associations introduced by 1,250 distinct GWAS, while the remainder of my data comes from studies involving genetic targeting.[4]

### 4.3 Characteristics of Gene-Disease Combinations

The empirical section examines the cross-sectional features of discoveries introduced through a data-driven process. Specifically, I ask: how do gene-disease associations established by GWAS differ from those established by candidate gene studies? I address this question by focusing on the following dependent variables.

**A. Novelty of combination:** I use two alternative dependent variables to capture discoveries involving genes that had received scant attention before the emergence of GWAS. The first is a dummy variable that equals one if the gene in the newly introduced gene-disease association had never been linked to any disease in the prior literature. Returning to the example from the introduction, the KIAA0350 gene is coded as underexplored because it had not been associated with any disease prior to the study by Hakonarson et al.

[4]In this paper, "GWAS" refers to studies that genotype the entire genome without ex ante gene selection, consistent with the GWAS Catalog's inclusion criteria. Studies not recorded in the Catalog include a range of targeted and hybrid designs, which I group under the candidate gene category. Any genome-wide study missed by the Catalog would be coded in this comparison group, which would attenuate differences between GWAS and targeted studies and therefore make my estimates conservative. Also, note that the GWAS Catalog records all statistically reported gene-disease associations in each GWAS paper, even when those associations are not novel. By construction, GWAS studies often re-identify previously documented associations, and about 67% of the associations reported in the Catalog are rediscoveries (Appendix Figure A.2). As a result, my data are not biased by strategic reporting of scientists (Appendix A.3). The GWAS Catalog applies a Bonferroni correction and reports as significant only associations that meet a stringent threshold (p-value $< 1.0 \times 10^{-5}$).

(2007). The second proxy is the gene's discovery date, since genes mapped more recently are still relatively less understood (Stoeger et al., 2018). Accordingly, I test whether GWAS are more likely to introduce combinations involving genes discovered after 2000, the year the first draft of the human genome was released. For example, the KIAA0350 gene was discovered in 2002, but remained ignored by candidate gene studies until Hakonarson et al. (2007) linked it to type 1 diabetes in 2007. Both dependent variables are coded as dichotomous to allow for straightforward interpretation of the OLS coefficients as linear probability models. I also replicate all analyses using continuous versions of these variables in the Appendix.

**B. Quality of combination:** Researchers typically rely on paper-to-paper citation counts to proxy for quality, but that approach would be inappropriate in my setting because scientific articles often report multiple discoveries. For example, the GWAS by Hakonarson et al. (2007) reported associations between type 1 diabetes and five genes (KIAA0350, INS, COL1A2, LPHN2, and PTPN22), making it impossible to assign a clear share of the article's citations to each individual combination.[5] To avoid this limitation, I use a unique metric available at the association level: the DisGeNET Score (Piñero et al., 2020). The DisGeNET Score is a composite index that reflects the "ground truth" scientific value of each gene-disease pair, constructed from independent experimental evidence in the genetics literature (see Appendix C.2 for details). Supporting its validity, Appendix Table C.1 shows that articles introducing combinations with a higher Score receive a larger number of citations. Furthermore, gene-disease combinations with higher DisGeNET Scores have substantially greater clinical relevance and downstream impact on drug discovery (Appendix Table C.2). This metric allows me to assess how search methods shape the quality of gene-disease associations discovered, as well as the likelihood of generating discoveries in both the top and bottom tails of scientific value.

**C. Follow-on research on the combination:** To evaluate whether data-driven discoveries crowd out theory development, I need a measure of downstream work that builds directly on a specific finding. I do so by counting the number of subsequent papers that investigate the same gene-disease pair, regardless of whether they cite the original study. For example, Appendix Figure B.2 shows that the KIAA0350–type 1 diabetes association was investigated by 24 articles following the GWAS by Hakonarson et al. (2007). Although many of these studies cite the original GWAS, some do not, and would therefore be missed by citation-based impact measures (Arts et al., 2025). I then characterize whether this follow-on work is aimed at explanation by using NIH's PubTator3 (Wei et al., 2024). PubTator3 applies an AI-based classifier to the full text to identify

[5] A second reason citation counts are misleading in my setting is that citations capture also attention rather than only scientific quality (Bikard, 2018). GWAS papers are, on average, much more cited than candidate gene studies ($Cites_{GWAS}$ = 170 vs. $Cites_{Candidate\ Gene\ Study}$ = 41), but much of that attention comes from factors unrelated to the quality of the findings — such as reviews, critiques, or debate on the merits of the genome-wide approach, all of which inflate citations without necessarily indicating higher-quality discoveries.

the type of gene-disease relationship a paper studies. In particular, it distinguishes studies that investigate causal relationships from those that provide correlational evidence (see Appendix E). I use this information to separately count follow-on articles that probe causal mechanisms, which I interpret as theory development, versus those that primarily add correlational support. Supporting the face validity of the classifier, Appendix Table E.1 shows that clinical trial papers are more likely to be coded as causal than correlational. Returning to the example, Appendix Figure B.2 shows that a large share of the papers building on the KIAA0350–type 1 diabetes association focus on its causal basis. Relative to prior work that infers recombination from co-occurrence patterns in publications or patents, PubTator3 allows me to identify the *type* of relationship under study and, specifically, whether follow-on research is directed toward causal understanding.

### 4.4 Summary Statistics

Table 1 lists the key variables along with summary statistics for the sample used in the analysis. Panel A presents statistics at the publication level, focusing on papers that introduce new gene-disease combinations during the study period. Compared to candidate gene papers, GWAS tend to introduce more associations and cover a larger number of genes on average, consistent with their untargeted nature. Panel B provides summary statistics at the gene-disease combination level. Previewing the following analysis, GWAS and candidate gene studies introduce new discoveries with similar average DisGeNET Scores. However, GWAS combinations are more likely to appear in both the bottom and top deciles of the Score distribution, indicating greater variability in their scientific value. Combinations introduced by GWAS attract more follow-on research, much of which focuses on clarifying the causal relationship between the gene and the disease.

## 5 Results

### 5.1 How Data Change Genetic Search

In this section, I begin by examining how gene-disease associations discovered through a data-driven search process differ from those introduced via a theory-driven one. I use regression analysis to compare the outcomes of the two search strategies, controlling for disease and scientist characteristics. More specifically, I estimate the following cross-sectional OLS model using gene-disease level data:

$$(Features\ of\ gene\text{-}disease\ combination)_{i,j} = \alpha + \beta GWAS_{i,j} + \mathbf{Z}'_{i,j}\gamma + \epsilon_{i,j} \tag{1}$$

where various features of the gene-disease pair are regressed on the indicator variable $GWAS_{i,j}$, which equals one for combinations discovered by a GWAS and zero for those discovered by candidate gene studies. The vector $\mathbf{Z}'_{i,j}$ contains fixed effects for targeted disease, PI, and year of first appearance of the combination

between gene $i$ and disease $j$. With PI fixed effects, identification comes from comparing the same scientist's discoveries when using GWAS versus candidate gene studies, while holding constant all time-invariant individual characteristics. If data-driven search systematically leads to different outcomes, I should observe a statistically significant estimate of $\beta$. All specifications cluster standard errors two ways at the gene and the disease level.[6]

Figure 2 provides an intuitive visualization of the combinatorial space of potential gene–disease pairs. Panel (a) shows that candidate gene studies largely build on existing knowledge, reproducing familiar research patterns. Panel (b) shows a sharp contrast: conditional on the disease being studied, GWAS discoveries effectively span the entire genome. Importantly, GWAS continue to concentrate on historically well-studied diseases, since the choice of disease remains with the PI under both approaches. This pattern indicates that the increased breadth in discovery occurs because the genome-wide method removes discretion over gene targeting, not because researchers shift their disease focus. Table 2 quantifies these differences. The estimate of $\beta$ in Column 1 shows that GWAS increase by 26 percentage points the probability of combining a gene never associated with any disease before 2005, compared to the baseline of 13 percentage points. Column 3 corroborates the result using the gene's discovery date as an alternative proxy. Even with scientist fixed effects, GWAS are 105–147% more likely to introduce associations involving historically understudied genes, relative to the sample means in Table 1.[7]

These findings speak to a long-standing puzzle in genetics: why research remains concentrated on few genes despite potentially large gains from broader exploration (Stoeger et al., 2018). My results point to a mechanism inherent to theory-driven search. When prior theory is required to formulate new functional hypotheses, as in candidate gene studies, attention is channeled toward familiar targets even when returns are lower. I provide suggestive evidence of this in the Appendix using data on *gene families*. Because genes in a family share related functions and are numbered in order of discovery (e.g., BRCA1 and BRCA2, discovered in 1994 and 1995, respectively), differences in attention across members are mostly driven by path dependence in knowledge accumulation rather than scientific promise (Daugherty et al., 2012).[8] Figure G.1 shows that candidate gene studies disproportionately introduce associations involving the first-discovered family

[6]These estimates are descriptive and may still be affected by unobserved, time-varying confounders. Section 5.3 discusses this concern and addresses it with an instrumental variable strategy to mitigate endogeneity in research method choice.

[7]Appendix Table G.1 shows consistent results using continuous versions of the dependent variables. GWAS are more likely to recombine genes that had received 29-51% fewer publications before 2005, and that were discovered 2.1-3 years later.

[8]Supporting this idea, DisGeNET data show that the first gene in a family receives, on average, 49% more publications than the second. Yet, associations involving the second gene tend to have higher average DisGeNET Scores ($DisGeNET\ Score_{First\ Member}$=0.059 vs. $DisGeNET\ Score_{Second\ Member}$=0.063) and are more likely to fall in the top 10% of all associations ($Top\ 10\%\ Score_{First\ Member}$=0.098 vs. $Top\ 10\%\ Score_{Second\ Member}$=0.112).

member, whereas this bias disappears for GWAS. Table G.2 confirms the pattern controlling for disease, year, and PI fixed effects. This suggests that GWAS advances discovery by loosening path-dependent gene targeting that constrains exploration of equally plausible, but less theoretically characterized, genes.

### 5.2 The Impact of GWAS on Scientific Discovery

While data-driven search appears to broaden the scope of search, there is no guarantee that the resulting discoveries are more valuable or generative of new understanding. If anything, one might reasonably expect the opposite. Scientists should already be focusing on what they consider the most promising areas of the scientific landscape, making it less likely that a pure data mining exercise results in more valuable discoveries. I examine this question in Table 3. Panel A shows that gene-disease combinations introduced by GWAS have, on average, a DisGeNET Score that is 19–20% higher. Interestingly, this effect is driven by a thickening of both tails: GWAS combinations are more likely to be either below-median discoveries or breakthroughs. This pattern is consistent with the idea that data-driven search increases variability in outcomes because it lacks the filtering of theory. Notably, the increase in top-tail discoveries is proportionally larger, which explains the overall rise in average DisGeNET Score for GWAS-introduced combinations.[9]

Panel B shifts from the value of discoveries to their downstream scientific consequences. A natural concern is that a prediction-first approach could substitute for explanation, yielding findings that attract less scientific interest or that are applied without prompting work on their mechanisms. Instead, my estimates point to the opposite. Gene–disease associations introduced by GWAS attract substantially more follow-on work. The average association is followed by 0.82 additional papers in specifications without PI fixed effects (+145% over the sample mean) and by 0.54 additional papers when comparing discoveries made by the same investigator (+93% over the sample mean). Importantly, this increase is driven by research aimed at understanding mechanisms. Follow-on articles classified as probing the causal determinants of the gene–disease link rise by 0.29–0.45, a 141–221% increase over the sample mean, whereas the estimates for correlational follow-on work are close to zero and not statistically significant.[10] Taken together, the results suggest that GWAS discoveries crowd in, rather than crowd out, subsequent theorizing.

[9] These results hold under alternative thresholds for defining the tails of the DisGeNET Score distribution (Appendix Tables G.3 and G.4). Further supporting the results, the coefficients grow larger (relative to the sample mean) with more stringent definitions of failure or breakthrough, suggesting that GWAS are especially powerful at uncovering outlier pairs at both quality tails.

[10] One may wonder whether follow-on papers are mostly authored by the same lab (PI) that first reported the gene–disease association. On average, 9.75% of causal follow-on articles are produced by the discovering lab. Importantly, this share is substantially lower when the association was first reported by a GWAS (4.92%), suggesting that the additional theorizing associated with GWAS discoveries is less concentrated within the originating lab and instead reflects broader field-level uptake. Unreported analysis confirm that excluding the follow-on work originating from the same lab that made the original discovery does not alter the estimates.

Two additional analyses help ruling out obvious confounders. First, the results are not driven by lab-specific changes in disease focus. Appendix Table G.5 shows that the results remain similar with tight fixed effects that compare associations introduced for the same disease in the same year, and even within the same laboratory for the same disease and year. Second, the results are not simply a consequence of GWAS drawing from a different pool of genes. When I include gene fixed effects, which compare GWAS and candidate gene discoveries involving the same gene, GWAS associations still have higher DisGeNET scores and elicit more follow-on work (Appendix Table G.6). However, GWAS no longer increases the likelihood of bottom-tail combinations, suggesting that the baseline result partially reflects GWAS associating lower-potential genes that theory-driven search tends to (correctly) screen out. Overall, these checks confirm that the higher scientific value of GWAS findings are not due to compositional differences in diseases or genes studied.

### 5.3 Instrumental Variable (IV) Estimates

The adoption of GWAS is a researcher choice and may therefore be endogenous to the discovery outcomes documented in the previous section. A first concern is selection. Researchers who are systematically better or worse at identifying breakthroughs may differentially adopt GWAS, which could bias the estimates.[11] Similarly, GWAS could be deployed in disease areas with more "low-hanging fruits," which would bias estimates upward, but it could also be adopted precisely where theory-driven approaches are running into diminishing returns, which would bias estimates downward. To mitigate these concerns, each table includes specifications with disease and PI fixed effects, where coefficients are identified from within-disease and within-laboratory comparisons. However, a more serious issue would be time-varying drivers of GWAS adoption. Unobserved shocks, such as disease-specific increases in funding, could both increase the likelihood of switching to GWAS and affect the outcomes of subsequent discoveries. Fixed effect estimation does not eliminate this type of confounding.

To address these concerns, I adopt an instrumental variable (IV) strategy that isolates plausibly exogenous variation in GWAS adoption. The approach exploits a technology-driven shock to the feasibility of genome-wide screening. Panel (a) of Figure 3 shows the sharp decline in the cost of collecting genetic data beginning in the mid-2000s, following the diffusion of lower-cost microarray technologies (Visscher et al., 2017). This drop in genotyping costs acts as an exogenous supply-side shift that increases the attractiveness of a genome-wide approach relative to targeted candidate gene studies.[12] Importantly, the same cost decline did not

[11] The direction of this bias is ambiguous. If scientists adopt GWAS after struggling to generate valuable combinations through theory driven approaches, the bias would be downward. If, instead, the most capable scientists are early adopters, the bias would be upward. Appendix Figure G.2 allays these concerns by showing that, prior to their first GWAS paper, eventual adopters do not introduce gene-disease associations of systematically different scientific quality.

[12] I thank an anonymous referee for this helpful suggestion.

translate into uniform adoption across disease areas. GWAS require large cohorts that are easier to recruit for high-burden diseases, and funding agencies prioritized those same diseases given their public-health salience. As a result, the cost decline generated larger increases in GWAS use precisely in disease areas with higher baseline prevalence in the population (Panel (b) of Figure 3).

To operationalize this intuition, I construct a shift-share instrument that interacts annual genotyping costs with predetermined disease prevalence measured before GWAS emerged. I obtain yearly genotyping cost data from the NIH and measure disease prevalence in 2005 using public data from UMLS and Orphanet (details in Appendix F). Prevalence information is available for 1,724 diseases, which defines my IV sample. The instrument varies at the disease–year level and predicts GWAS adoption through differential exposure to the common decline in genotyping costs:

$$\text{Genotyping cost shift-share}_{j,t} \;=\; \text{Genotyping Cost}_t \times \text{Disease Prevalence}_{j,2005}. \tag{2}$$

I then estimate the first stage as the likelihood that a combination $< i, j >$ is discovered in year $t$ using a GWAS rather than a candidate gene study:

$$\text{GWAS}_{i,j} \;=\; \pi_0 + \pi_1 \,\text{Genotyping cost shift-share}_{j,t} + \mathbf{Z}'_{i,j}\pi_3 + u_{i,j}, \tag{3}$$

where $i$ indexes genes, $j$ indexes diseases, and $\mathbf{Z}_{i,j}$ includes the baseline controls and fixed effects from the main specification. The second stage replaces the endogenous GWAS indicator with its fitted value from the first stage:

$$(Features\ of\ gene\text{-}disease\ combination)_{i,j} \;=\; \alpha + \beta_{2SLS}\,\widehat{\text{GWAS}}_{i,j} + \mathbf{Z}'_{i,j}\gamma + \varepsilon_{i,j}, \tag{4}$$

where $\beta_{2SLS}$ is the IV estimate identified from differential changes in GWAS adoption across diseases as genotyping costs fall. Because the instrument has a shift-share structure, Appendix F.4 reports the diagnostics recommended by Borusyak et al. (2025). In particular, I show that diseases with higher versus lower baseline prevalence exhibit similar pre-GWAS trends in the outcomes and that my estimates are stable in "leave-one-disease-out" specifications (Goldsmith-Pinkham et al., 2020).

Table 4 reports 2SLS estimates using the genotyping-costs shift-share instrument on the diseases for which disease prevalence data are available. The first column of each panel presents the first stage. As expected, higher instrument values, which correspond to years with higher genotyping costs, predict a lower probability that a newly reported association is discovered via GWAS ($\hat{\pi}_1 = -0.0104$, *p-value* $< .01$). The first stage is strong, with Cragg Donald F statistics $= 53.6$ and Kleibergen Paap F statistics $= 19.5$. The remaining

columns replicate the outcome specifications from Table 3. Across panels, the 2SLS estimates match the OLS results in sign and are, if anything, larger in magnitude.[13] The larger magnitudes are consistent with the instrument identifying the effect of GWAS off the diseases in which declining genotyping costs change research design, that is, a local average treatment effect (LATE). Overall, the IV results strengthen a causal interpretation of the baseline findings: GWAS adoption raises the average scientific value of newly reported gene-disease associations and increases the follow-on research they generate.

### 5.4 Mechanisms

The findings from the previous sections show that *how* scientists search shapes the features of their discoveries. What remains to be explained is why a data-first approach like GWAS appears particularly generative of follow-on theorizing. A potential mechanism, suggested by the philosophy of science, is the uncovering of empirical anomalies (Chai, 2017; Kuhn, 1962). Relative to candidate gene studies, GWAS perform a systematic genome-wide screening, increasing the likelihood of uncovering relationships outside the boundaries of genetic theory. This mechanism can in principle account for my earlier findings. First, anomalies should include a mix of valuable associations and spurious ones, which is consistent with GWAS thickening both tails of the value distribution. Second, anomalies should be theoretically productive because they expose gaps in current theories and heighten the demand for new explanations. This section formalizes and tests this anomalies-based channel.

I define empirical anomalies as new gene–disease associations that are surprising given the state of existing knowledge (Shi and Evans, 2023). Measuring anomalousness is feasible because genes do not operate in isolation, but rather through chains of activation and inhibition. As a result, functionally related genes tend to be implicated in the same diseases, since a mutation in any one gene of a disease-relevant process can trigger the condition (Visscher et al., 2017). For a given disease, the set of associations observed up to a given date therefore provides information about which additional genes are likely to be implicated next, based on co-occurrence patterns in other diseases. I use the inverse of this predicted likelihood to construct

[13] Appendix Table G.7 similarly shows that the IV estimates corroborate the baseline breadth result: GWAS are more likely to introduce associations involving less explored genes. Across outcomes, the only pattern that does not survive the IV specifications is the increase in low-quality combinations. One plausible, though necessarily speculative, explanation is differential selective reporting. Fanelli (2010) shows that studies testing multiple hypotheses are more likely to report at least one null result, consistent with unfavorable findings being easier to publish when they can be bundled alongside stronger results. This logic is particularly relevant here because candidate gene studies typically test few hypotheses. Indeed, bottom tail associations are uncommon in candidate gene papers that report a single association (9.6%), but become much more prevalent when candidate gene papers report multiple associations (25.2%). In contrast, GWAS designs are less conducive to suppressing low-value associations because researchers typically report the full set of genome-wide hits, leading to a higher rate of low quality associations (35.1%). If low-value single-hypothesis results are disproportionately missing from the candidate gene literature, OLS will mechanically exaggerate the bottom-tail gap between GWAS and candidate gene studies, which could possibly explain why this effect attenuates and is not robust in the IV specifications.

an Anomaly Index (details in Appendix D). Two validation exercises support its face validity. Using lexical cues from paper abstracts (Mishra et al., 2023), Appendix Table D.1 shows that discoveries with a higher Anomaly Index are more likely to be described as "novel," but are no more likely to be framed using generic hype terms. Appendix Table D.2 further shows that these same papers are more likely to be tagged as "new findings" by expert reviewers on the Faculty Opinions platform.

Appendix Table G.8 links the anomalousness of a discovery to both its scientific value and the downstream work it attracts. Panel A shows that anomalous gene–disease pairs are concentrated in both the bottom and top tails of the DisGeNET Score distribution, implying that anomalies are more likely to be either false positives or genuine breakthroughs. Panel B shows that anomalous discoveries are not necessarily related to the overall volume of follow-on papers, consistent with Kuhn's view that scientists' first instinct is to ignore anomalies rather than treat them as puzzles demanding attention (Chai, 2017; Kuhn, 1962). However, this result conceals meaningful heterogeneity in the composition of follow-on work. As discoveries become more anomalous, follow-on research shifts toward studies that probe causal explanations, while other types of follow-on research remain flat (Figure G.4). Finally, Appendix Table G.9 shows that GWAS-introduced associations have a substantially higher Anomaly Index. Taken together, these results suggest that data-driven search amplifies discovery and theorizing by expanding the supply of empirical anomalies.

This mechanism also yields implications about when data-driven search should deliver the largest gains. In disease areas with deeper pre-existing genetic knowledge, theory offers better guidance, but it also tightens selection over what seems worth testing, increasing the risk of overly local search (Knudsen and Levinthal, 2007). Appendix Figure G.3 is consistent with this tradeoff, showing that GWAS produces proportionally more high-quality discoveries precisely where prior genetic knowledge is greatest. The same logic applies to subsequent theory generation. Where knowledge is deeper, anomalous associations are at odds with what researchers believe they understand, so they strongly signal a need to revise priors (Camuffo et al., 2024). Figure 4 shows this directly. The follow-on advantage of GWAS is concentrated in disease areas with more known genetic linkages, and is strongest when GWAS discoveries are more anomalous relative to such knowledge base. When GWAS discoveries are not anomalous, or when there is little prior theory to contradict, the follow-on response is muted. Taken together, the results suggest that data-driven search is most generative when it confronts well-developed theory with unexpected empirical patterns – implying that it is more valuable precisely in domains where existing theories appear most satisfactory.

# 6 The Dynamic Effects of Data-Driven Discoveries

The results so far show that GWAS discoveries attract more follow on research, especially work aimed at clarifying causal mechanisms. This section examines whether that response reflects durable theory building or a short lived burst of attention, and whether the additional follow on work is scientifically valuable. To study these dynamics, I move from cross sectional comparisons to a panel that tracks each gene-disease association over time. Using DisGeNET, I construct a gene disease year dataset in which each observation records the number of subsequent articles on a given association in year $t$, excluding the original paper. For example, Appendix Figure B.2 shows that the KIAA0350–type 1 diabetes association was investigated by 24 articles following the GWAS by Hakonarson et al. (2007). By contrast, the same figure shows that the candidate gene study by Qu et al. (2007) generated little follow-up work on the IRF5–type 1 diabetes pair. Summary statistics for this panel are reported in Appendix Table G.10.

Figure 5 shows descriptive evidence on the average number of publications following the introduction of a new gene-disease combination. GWAS-introduced associations draw substantially more follow-on work than candidate gene discoveries, and the increase in scientific attention is large and persistent over time. I quantify these dynamics by estimating the following panel difference-in-differences specification at the gene-disease-year level:

$$(Papers\ on\ gene\text{-}disease\ combination)_{i,j,t} = \alpha + \beta Post_t \times Publication_{i,j} \times GWAS_{i,j} + \lambda GD_{i,j} + \delta_t \times Disease_i + \omega_t \times Gene_j + \epsilon_{i,j,t} \quad (2)$$

where the count of articles exploring a given gene-disease pair $< i, j >$ is regressed on the interaction of $Post_t \times Publication_{i,j}$, which equals one in the years following the initial publication of the gene-disease combination, and $GWAS_{i,j}$, which equals one if the association was first reported in a GWAS. $GD_{i,j}$ are fixed effects for the combination of gene $i$ and disease $j$, which account for pair-specific differentials in research potential. I include fixed effects for each gene-disease pair ($GD_{i,j}$) to account for inherent differences in the research potential of specific combinations. To control for time-varying disease-level shifts in market size, I include disease-year fixed effects ($\delta_t \times Disease_j$). I also account for changes in gene-specific interest over time by including gene-year fixed effects ($\omega_t \times Gene_i$). Standard errors are clustered at the gene and disease level.

The main results are reported in Appendix Table G.11. Consistent with the cross-sectional evidence, GWAS-introduced associations generate more follow-on work per year than candidate gene discoveries, and the increase is concentrated in studies in their causal mechanisms. Figure 6 reports the corresponding event

study estimates. Panel (a) shows that causal follow-on work rises after the initial GWAS discovery, whereas Panel (b) shows no comparable increase in correlational analyses. The absence of pre-trends strengthens identification, indicating that GWAS triggers a persistent increase in new theorizing that remains sizable even eight years after discovery. Appendix Table G.12 further shows that this additional follow-on work is scientifically consequential: it receives more citations, is more likely to be cited by clinical trials, and appears in higher-prestige journals than follow-on research triggered by candidate gene discoveries. Taken together, this additional evidence indicates that data-driven search through GWAS sustains a durable increase in high-quality theorizing.

# 7 Discussion

This paper studies how data changes scientific search by comparing genome-wide association studies with candidate gene studies in human genetics. GWAS broaden the genetic landscape that researchers empirically examine, uncovering new gene–disease links that are more likely to implicate previously underexplored genes. This wider search increases dispersion in discovery quality, with a pronounced right-tail shift: the growth in breakthroughs outpaces the growth in low-value findings, increasing the average scientific value of new discoveries. GWAS associations also generate stronger downstream interest, with subsequent research mostly focused on clarifying causal mechanisms. An instrumental variable strategy based on the decline in genotyping costs supports a causal interpretation, and mechanism evidence suggests that GWAS stimulate theorizing by uncovering patterns that violate theoretical expectations. Consistent with this mechanism, GWAS has its largest effects in domains with deeper prior theory, where anomalies loosen established research trajectories, and create salient puzzles that prompt theory revision.

My results underscore the nuanced implications of adopting alternative innovation search strategies. On the one hand, data-driven search can generate false positives and incremental leads (Allen and McDonald, 2025; Benner and Tushman, 2002). In strategic decision-making, where choices are often irreversible, pursuing the wrong lead can seriously undermine performance (Tranchero, 2024). Yet on the other hand, those same leads may serve a different function: redirecting exploration toward more open-ended and potentially fruitful directions. While a high rate of false positives poses clear risks in strategy, it may be tolerable or even desirable in innovation. To paraphrase Louis Pasteur, "chance favors the curious mind," and data-driven leads can be exactly what sparks that curiosity. In contrast, false negatives are particularly harmful in innovation contexts, where prematurely dismissing a promising path can halt exploration altogether (Chai, 2017; Fleming and Sorenson, 2004; Hoelzemann et al., 2025). This underscores the asymmetric consequences of false positives and false negatives when using data to explore versus using data to make strategic decisions.

An important contribution of this paper is methodological. While the modeling of innovation search has a storied tradition (Levinthal, 1997; March, 1991; Nelson and Winter, 1982), empirical progress has lagged somewhat behind. Part of the challenge lies in the difficulty of studying a largely unobservable process through the features of its realized outcomes (Fleming, 2001; Jayaraj and Gittelman, 2018; Kneeland et al., 2020; Maggitti et al., 2013; Schilling and Green, 2011). This paper addresses that challenge by showing how deep institutional knowledge enables the identification of the search processes agents actually follow. In turn, this ex-ante characterization can be used to explain how different search strategies shape outcomes on the empirical landscape — directly linking the search process to the innovations it produces. Broader adoption of this approach could meaningfully advance the empirical study of search.

While human genetics provides an unusually clean setting for this study, my findings have boundary conditions. The results speak most directly to innovation search in combinatorial problem spaces, where a stable paradigm defines the objects of search and their feasible recombinations (Dosi, 1982; Kuhn, 1962). In practice, such paradigms operate as meta-theoretical frameworks that make systematic measurement possible by specifying what is worth observing, recording, and comparing (Felin and Singell, 2025). When these conditions hold, data-driven screening can help actors address combinatorial uncertainty (Fleming, 2001) by expanding "offline" investigation of alternatives (Gavetti and Levinthal, 2000) and uncovering patterns that actors would not have prioritized (Knudsen and Levinthal, 2007; Price, 1983). By contrast, in many entrepreneurial and market settings, the opportunity set is not pre-specified and instead must be constructed by the actor (Camuffo et al., 2024; Ehrig and Zenger, 2024; Felin and Zenger, 2017). In such cases, the relevant landscape is endogenous to firms' choices and strategic interaction, which limits any notion of a "ground truth" that data can simply reveal (Felin et al., 2024). Future research should examine whether the gains from data-driven search identified here extend beyond scientific and technological discovery.

In many ways, the history of science and technology has been shaped by episodes similar to those documented in this paper. For instance, Galileo's telescope enabled the observations that fueled the Copernican Revolution, and Rosalind Franklin's X-ray diffraction images prompted the discovery of the structure of DNA. In both cases, data-generating tools made visible anomalies that strained prevailing theories and enabled conceptual breakthroughs. What makes the findings of the present paper especially timely is that AI and related data technologies are rapidly expanding the scale and speed at which such anomalies can be surfaced (Conti and Messinese, 2024; Mullainathan and Rambachan, 2024). By lowering the cost of broad screening, these tools can loosen the path dependence that inevitably accompanies cumulative learning: "standing on the shoulders of giants" is productive but it concentrates attention along narrow trajectories, making unexpected

results easier to overlook or dismiss (Chai, 2017; Levinthal and March, 1993; Knudsen and Levinthal, 2007; Kuhn, 1962). The true promise of AI, however, lies not only in its ability to exhaustively search existing landscapes, but also in reshaping the landscapes themselves by expanding how problems are represented. In this light, big data may not mark the "end of theory," as Anderson (2008) once proclaimed, but rather the beginning of an era in which big data generates big theory too.

# References


Adner, R. and D. A. Levinthal (2024): "Strategy Experiments in Nonexperimental Settings: Challenges of Theory, Inference, and Persuasion in Business Strategy," *Strategy Science*, 9, 311–321.

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

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

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

Arts, S., N. Melluso, and R. Veugelers (2025): "Beyond citations: Measuring novel scientific ideas and their impact in publication text," *Review of Economics and Statistics*, 1–33.

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.

Barbosu, S. and F. Teodoridis (2025): "Catalysts of Discovery. How Technologies are Shaping the Future of Science and Innovation," *Springer Books*.

Benner, M. J. and M. Tushman (2002): "Process management and technological innovation: A longitudinal study of the photography and paint industries," *Administrative Science Quarterly*, 47, 676–707.

Bergen, K. J., P. A. Johnson, M. V. de Hoop, and G. C. Beroza (2019): "Machine learning for data-driven discovery in solid Earth geoscience," *Science*, 363, eaau0323.

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

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

Borusyak, K., P. Hull, and X. Jaravel (2025): "A practical guide to shift-share instruments," *Journal of Economic Perspectives*, 39, 181–204.

Callaway, E. (2017): "New concerns raised over value of genome-wide disease studies." *Nature*, 546, 463–464.

Calude, C. S. and G. Longo (2017): "The deluge of spurious correlations in big data," *Foundations of Science*, 22, 595–612.

Camuffo, A., A. Gambardella, and A. Pignataro (2024): "Theory-driven strategic management decisions," *Strategy Science*, 9, 382–396.

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

Chai, S. (2017): "Near misses in the breakthrough discovery process," *Organization Science*, 28, 411–428.

Choudhury, P., R. T. Allen, and M. G. Endres (2021): "Machine learning for pattern discovery in management research," *Strategic Management Journal*, 42, 30–57.

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.

Conti, A. and D. Messinese (2024): "The Selective Tailwind Effect of Artificial Intelligence," *Available at SSRN*.

Daugherty, L. C., R. L. Seal, M. W. Wright, and E. A. Bruford (2012): "Gene family matters: Expanding the HGNC resource," *Human Genomics*, 6, 1–6.

Dosi, G. (1982): "Technological paradigms and technological trajectories: A suggested interpretation of the determinants and directions of technical change," *Research Policy*, 11, 147–162.

Ehrig, T. and T. Zenger (2024): "Competing with theories: Using awareness and confidence to secure resources and rents," *Strategy Science*, 9, 416–432.

Evans, J. A. (2010): "Industry induces academic science to know less about more," *American Journal of Sociology*, 116, 389–452.

Fanelli, D. (2010): ""Positive" results increase down the hierarchy of the sciences," *PloS One*, 5, e10068.

Felin, T., A. Gambardella, and T. Zenger (2024): "Theory-Based Decisions: Foundations and Introduction," *Strategy Science*, 9, 297–310.

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

Furman, J. L. and F. Teodoridis (2020): "Automation, research technology, and researchers' trajectories: Evidence from computer science and electrical engineering," *Organization Science*, 31, 330–354.

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

Goldsmith-Pinkham, P., I. Sorkin, and H. Swift (2020): "Bartik instruments: What, when, why, and how," *American Economic Review*, 110, 2586–2624.

Hakonarson, H., S. F. Grant, J. P. Bradfield, L. Marchand, C. E. Kim, J. T. Glessner, R. Grabs, et al. (2007): "A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene," *Nature*, 448, 591–594.

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

Hermosilla, M. and J. Lemus (2019): "Therapeutic translation of genomic science," in *Economic dimensions of personalized and precision medicine*, University of Chicago Press.

Hoelzemann, J., G. Manso, A. Nagaraj, and M. Tranchero (2025): "The streetlight effect in data-driven exploration," *NBER wp 32401*.

Jayaraj, S. and M. Gittelman (2018): "Scientific maps and innovation: Impact of the Human Genome on drug discovery," *DRUID Society Conference Paper*.

Kang, S. (2025): "From Outward to Inward: Reframing Search with New Mapping Criteria," in *Academy of Management Proceedings*, Academy of Management Valhalla, NY 10595, vol. 2025, 10129.

Kao, J. (2024): "Charted Territory: Mapping the Cancer Genome and R&D Decisions in the Pharmaceutical Industry," *UCLA Anderson*.

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

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.

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

Kuhn, T. S. (1962): *The structure of scientific revolutions*, Chicago University Press.

Leonelli, S. (2014): "What difference does quantity make? On the epistemology of Big Data in biology," *Big Data & Society*, 1, 2053951714534395.

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

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

Loos, R. J. (2020): "15 years of genome-wide association studies and no signs of slowing down," *Nature Communications*, 11, 5900.

Ludwig, J. and S. Mullainathan (2024): "Machine learning as a tool for hypothesis generation," *The Quarterly Journal of Economics*, 139, 751–827.

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

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

Maggitti, P. G., K. G. Smith, and R. Katila (2013): "The complex search process of invention," *Research Policy*, 42, 90–100.

March, J. G. (1991): "Exploration and exploitation in organizational learning," *Organization Science*, 2, 71–87.

Mazzocchi, F. (2015): "Could Big Data be the end of theory in science?" *EMBO reports*, 16, 1250–1255.

Merchant, A., S. Batzner, S. S. Schoenholz, et al. (2023): "Scaling deep learning for materials discovery," *Nature*, 624, 80–85.

Merton, R. K. (1948): "The bearing of empirical research upon the development of social theory," *American Sociological Review*, 13, 505–515.

Mishra, A., J. Diesner, and V. I. Torvik (2023): "A probabilistic model of 'Hype' in scientific abstracts," *International Society of Scientometrics and Informetrics Conference 2023 (ISSI)*.

Mokyr, J. (2002): *The gifts of Athena: Historical origins of the knowledge economy*, Princeton: Princeton University Press.

Mullainathan, S. and A. Rambachan (2024): "From predictive algorithms to automatic generation of anomalies," *NBER wp 32422*.

Mullainathan, S. and J. Spiess (2017): "Machine learning: An applied econometric approach," *Journal of Economic Perspectives*, 31, 87–106.

Murray, F., P. Aghion, M. Dewatripont, J. Kolev, and S. Stern (2016): "Of mice and academics: Examining the effect of openness on innovation," *American Economic Journal: Economic Policy*, 8, 212–52.

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

Nagaraj, A. and M. Tranchero (2024): "How Does Data Access Shape Science? The Impact of Federal Statistical Research Data Centers on Economics Research," *NBER Working Paper 31372*.

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

Ott, T. E. and D. P. Hannah (2024): "On the origin of entrepreneurial theories: How entrepreneurs craft complex causal models with theorizing and data," *Strategy Science*, 9, 461–482.

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.

Piñero, J., J. M. Ramírez-Anguita, J. Saüch-Pitarch, Ronzano, et al. (2020): "The DisGeNET knowledge platform for disease genomics: 2019 update," *Nucleic Acids Research*, 48, D845–D855.

Popper, K. (1963): *Conjectures and refutations: The growth of scientific knowledge*, London: Routledge.

Price, D. (1983): "Of sealing wax and string: A philosophy of the experimenter's craft and its role in the genesis of high technology," in *Little science, big science and beyond*, Columbia University Press, 237–53.

Qu, H.-Q., L. Marchand, R. Grabs, and C. Polychronakos (2007): "The IRF5 polymorphism in type 1 diabetes," *Journal of Medical Genetics*, 44, 670–672.

Rosenberg, N. (1982): *Inside the black box: Technology and economics*, Cambridge University Press.

Schilling, M. A. and E. Green (2011): "Recombinant search and breakthrough idea generation: An analysis of high impact papers in the social sciences," *Research Policy*, 40, 1321–1331.

Schliesmann, D. (2025): "The Where of Search," *The Wharton School*.

Shi, F. and J. Evans (2023): "Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines," *Nature Communications*, 14, 1641.

Shmueli, G. (2010): "To explain or to predict?" *Statistical Science*, 289–310.

Shrestha, Y. R., V. F. He, P. Puranam, and G. von Krogh (2021): "Algorithm supported induction for building theory: How can we use prediction models to theorize?" *Organization Science*, 32, 856–880.

Sink, R., S. Gobec, S. Pecar, and A. Zega (2010): "False positives in the early stages of drug discovery," *Current Medicinal Chemistry*, 17, 4231–4255.

Sorenson, O. (2024): "Theory, search, and learning," *Strategy Science*, 9, 372–381.

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.

Stokes, J. M., K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, et al. (2020): "A deep learning approach to antibiotic discovery," *Cell*, 180, 688–702.

Torvik, V. I. and N. R. Smalheiser (2021): "Author-ity 2018 - PubMed author name disambiguated dataset," *University of Illinois at Urbana-Champaign*.

Tranchero, M. (2024): "Finding diamonds in the rough: Data-driven opportunities and pharmaceutical innovation," *The Wharton School*.

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.

Wei, C.-H., A. Allot, P.-T. Lai, R. Leaman, S. Tian, L. Luo, Q. Jin, Z. Wang, Q. Chen, and Z. Lu (2024): "PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge," *Nucleic Acids Research*, 52, W540–W546.

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

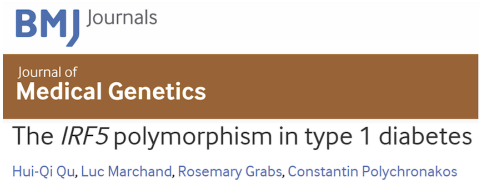
York, D. G., J. Adelman, J. E. Anderson Jr, S. F. Anderson, et al. (2000): "The Sloan Digital Sky Survey: Technical summary," *The Astronomical Journal*, 120, 1579.

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

# 8 Figures and Tables

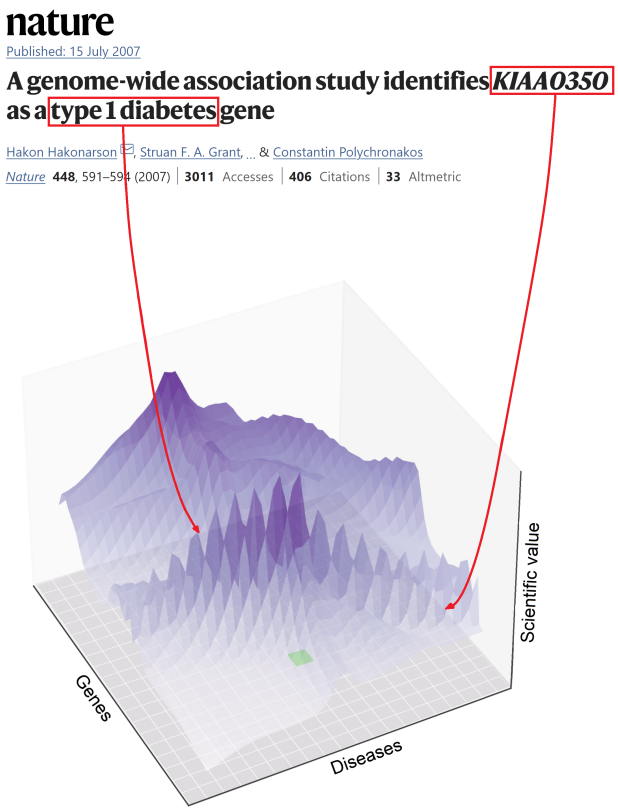
Figure 1: Scientists search for new gene-disease associations with candidate gene studies or with GWAS.

(a) *Example of candidate gene study*




(b) *Example of GWAS*

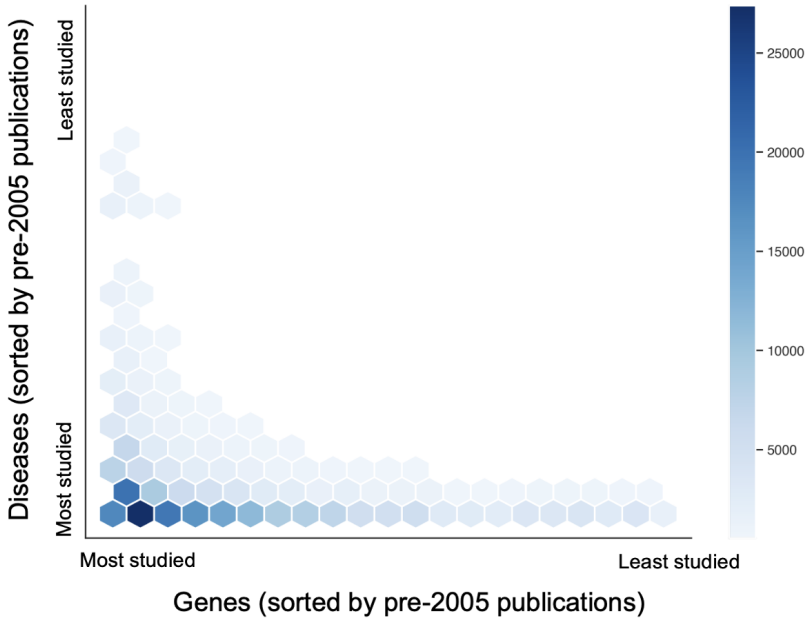

(c) *GWAS as data-driven search on a landscape of gene-disease combinations*



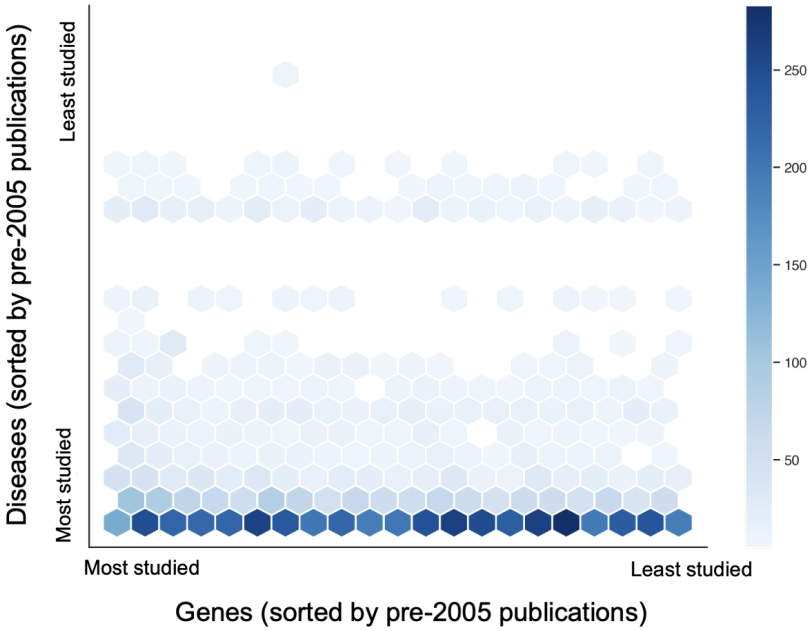

*Note:* The figure compares candidate gene studies with GWAS. Panel (a) shows the title of the candidate gene study by Qu et al. (2007) published in *The Journal of Medical Genetics* in 2007. Panel (b) shows the title of the GWAS by Hakonarson et al. (2007) published in *Nature* in 2007. Both studies searched for the genetic roots of type 1 diabetes and were carried out by the same principal investigator (PI) in the same year. More detail on these studies is provided in the case study in Appendix B. Panel (c) depicts how a typical GWAS introduces a new gene-disease association in the combinatorial landscape of all possible gene-disease pairs. Each combination of gene and disease has a specific scientific value, ideally captured by the elevation at that location. See text for details.

Figure 2: Conditional on the choice of disease, GWAS introduce new gene-disease associations spanning a larger portion of the genetic landscape relative to candidate gene studies.

(a) *Gene-disease associations discovered by candidate gene studies*



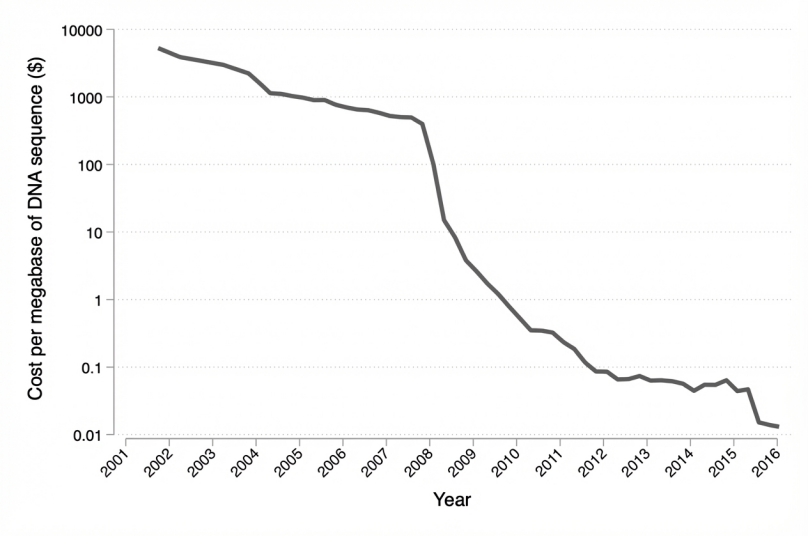

(b) *Gene-disease associations discovered by GWAS*



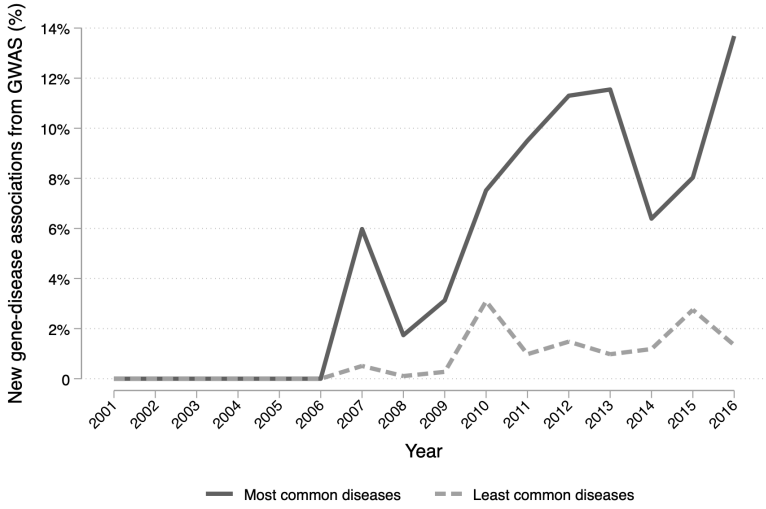

*Note:* The figure graphically shows the breadth of discovery for different search strategies. Panel (a) shows a heatmap of new gene-disease associations discovered after 2005 by candidate gene studies. Panel (b) shows a heatmap of new gene-disease associations discovered after 2005 by genome-wide association studies. Both panels have 14,136 genes on the X axis, sorted from the most to the least studied in the pre-GWAS era, and 9,863 narrowly defined disease categories on the Y axis, sorted from the most to the least studied in the pre-GWAS era. Darker blue bins correspond to a higher number of associations involving those genes. See text for details.

Figure 3: The reduction in the cost of microarray-based DNA sequencing led to the diffusion of GWAS, especially for the most common human diseases.

(a) *Cost of DNA sequencing*



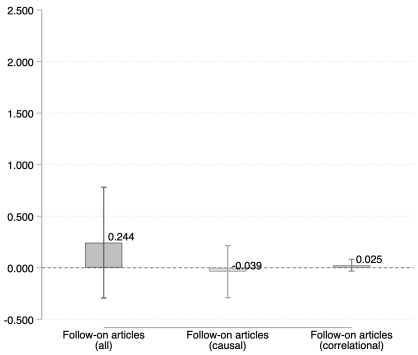

(b) *Share of new gene-disease associations discovered by GWAS*



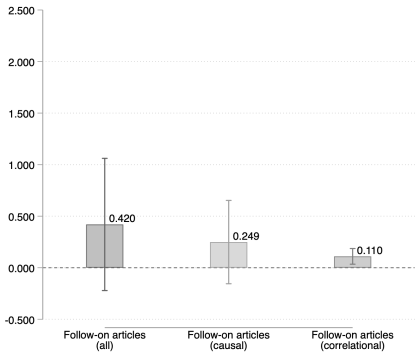

*Note:* The figure illustrates the intuition behind the shift-share instrumental variable (IV) strategy. Panel (a) shows the average cost of sequencing one megabase (i.e., one million bases) of DNA over time, highlighting a sharp decline that led to the diffusion of GWAS. Data are available at: `www.genome.gov/sequencingcostsdata`. Panel (b) plots the share of new gene-disease associations (GDAs) recorded in DisGeNET that were discovered by GWAS over time. The figure is shown separately for diseases in the top and bottom quartiles of prevalence (i.e., the number of people affected) in the pre-GWAS era. The figure shows that while declining genotyping costs increased the use of GWAS overall, the increase was larger for the most common diseases, due to ease in assembling study samples and the greater availability of research funding. More detail on the IV is provided in Appendix F. See text for details.

Figure 4: Mechanism analysis: GWAS generate more follow-on research than candidate-gene studies only when their discoveries are empirical anomalies that challenge existing theory.

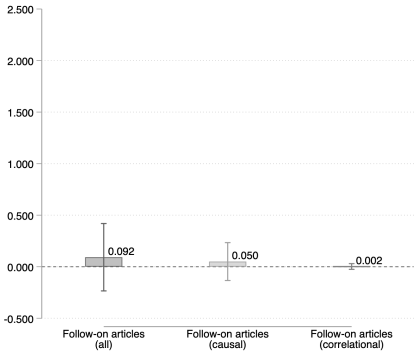
*(a) Follow-on research on low-anomaly GWAS discoveries in low-theory disease areas*



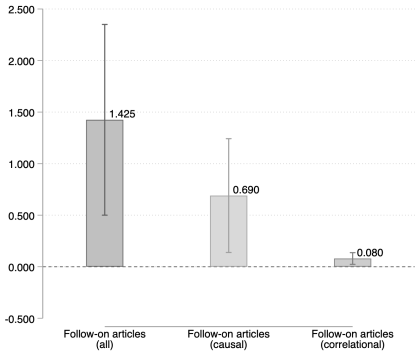

*(b) Follow-on research on low-anomaly GWAS discoveries in high-theory disease areas*




*(c) Follow-on research on high-anomaly GWAS discoveries in low-theory disease areas*




*(d) Follow-on research on high-anomaly GWAS discoveries in high-theory disease areas*

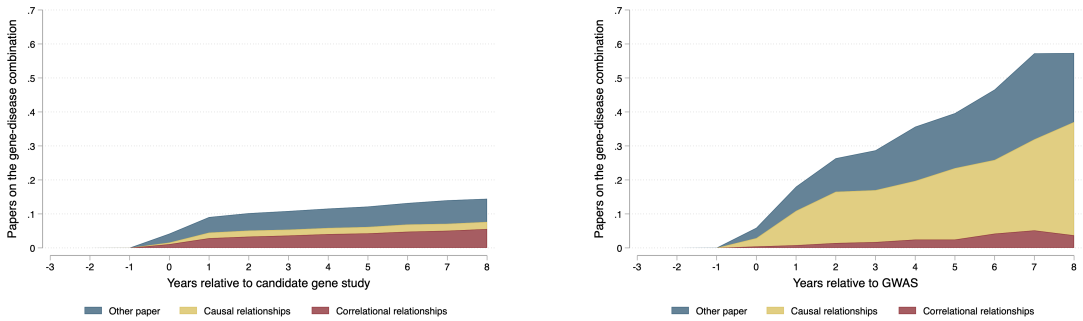



*Note:* This figure examines heterogeneity in the GWAS effect on follow-on research by splitting newly introduced gene-disease associations along two dimensions: the degree to which the discovery is anomalous relative to prior knowledge and the amount of pre-existing genetic knowledge in the focal disease area. Each panel presents the same estimates of Table 3, panel (b), but separately for associations with a different degree of anomaly relative to existing theory. Panel (a) reports estimates for low-anomaly GDAs (i.e., with Anomaly Score below median) in low-theory disease areas (i.e., with a below median number of previous genetic associations). Panel (b) reports estimates for low-anomaly GDAs in high-theory disease areas (i.e., with an above median number of previous genetic associations). Panel (c) reports estimates for high-anomaly GDAs (i.e., with Anomaly Score above median) in low-theory disease areas. Panel (d) reports estimates for high-anomaly GDAs in high-theory disease areas. Each panel plots the coefficient $\beta$ from the cross-sectional specification estimated separately within the corresponding subsample: $(Follow\text{-}on\ articles)_{i,j} = \alpha + \beta\, GWAS_{i,j} + \gamma\, Disease_i + \omega\, Scientist_j + \delta\, Year_{i,j} + \epsilon_{i,j}$. The charts plot values of $\beta$ for different dependent variables. *Follow-on articles (all)*: count of all subsequent studies investigating the gene-disease association; *Follow-on articles (causal)*: count of subsequent studies investigating causal relationships between the gene and the disease; *Follow-on articles (correlational)*: count of subsequent studies investigating correlational relationships between the gene and the disease. Follow-on articles are papers directly working on the gene-disease combinations, regardless of whether they cite the study that first introduced it. Information on which papers explore causal or correlational relationships comes from the AI engine of PubTator3 (Wei et al., 2024). See text for details.

Figure 5: Relative to candidate gene studies, GWAS introduce new gene-disease associations that attract more follow-on research, especially studies investigating causal mechanisms.

*(a) Follow-on articles on gene-disease associations discovered by candidate gene studies*

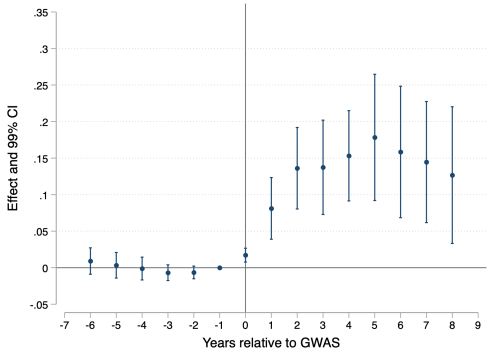
*(b) Follow-on articles on gene-disease associations discovered by GWAS*



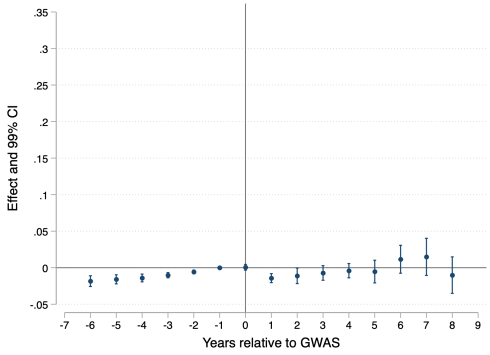

*Note:* Panel (a) shows the average yearly number of papers exploring gene-disease associations first discovered by a candidate gene study. Panel (b) shows the average yearly number of papers exploring gene-disease associations first discovered by a GWAS. Note that follow-on papers are those directly studying the same gene-disease pair, regardless of whether they cite the study that introduced it. In both figures, the count of papers is split out by whether the follow-on papers explore causal or correlational relationships between the gene and the disease. Information on which papers explore causal or correlational relationships comes from PubTator3 (Wei et al., 2024). See text for details.

Figure 6: Event studies show that GWAS discoveries sustain a durable increase in theorizing relative to candidate gene studies.

*(a) Increase in articles investigating causal relationships about the gene-disease combination*




*(b) Increase in articles investigating correlational relationships about the gene-disease combination*




*Note:* This figure shows the event study version of the difference-in-differences regressions evaluating the increase in follow-on publications for new gene-disease associations discovered by GWAS relative to new gene-disease associations introduced by candidate gene studies. Each panel shows the event study coefficients estimated from the following specification: $Papers_{i,j,t} = \alpha + \sum_z \gamma_z Post_z \times Publication_{i,j} \times 1(z) + \sum_z \beta_z Post_z \times Publication \times GWAS_{i,j} \times 1(z) + \lambda GD_{i,j} + \delta_t \times Gene_i + \omega_t \times Disease_j + \epsilon_{i,j,t}$. The charts plot values of the triple-difference coefficient $\beta_z$ for different lags $z$ before and after the gene-disease association is first reported by a GWAS with 99% confidence intervals. Standard errors are clustered at the disease level. The dependent variables are the yearly count of papers investigating causal relationships about the gene-disease pair (panel (a)) and the yearly count of papers investigating correlational relationships about the gene-disease pair (panel (b)). Information on which papers explore causal or correlational relationships comes from PubTator3 (Wei et al., 2024). See text for details.

Table 1: Descriptive statistics.

| | Panel A: paper-level descriptives (cross sectional) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Candidate gene studies | | | | | | GWAS | | | | | |
| | mean | median | st d | min | max | N | mean | median | st d | min | max | N |
| Associations per paper | 2.588 | 2 | 5.158 | 1 | 916 | 141,395 | 6.934 | 3 | 14.089 | 1 | 244 | 1,250 |
| Genes per paper | 1.531 | 1 | 2.378 | 1 | 641 | 141,395 | 4.508 | 2 | 7.722 | 1 | 88 | 1,250 |
| Year of publication | 2011.07 | 2011 | 3.394 | 2005 | 2016 | 141,395 | 2012.28 | 2012 | 2.527 | 2005 | 2016 | 1,250 |

| | Panel B: gene-disease level descriptives (cross sectional) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Candidate gene studies | | | | | | GWAS | | | | | |
| | mean | median | st d | min | max | N | mean | median | st d | min | max | N |
| With never associated genes (%) | 0.121 | 0 | 0.326 | 0 | 1 | 360,647 | 0.404 | 0 | 0.491 | 0 | 1 | 8,655 |
| With recently discovered gene (%) | 0.098 | 0 | 0.298 | 0 | 1 | 360,647 | 0.256 | 0 | 0.437 | 0 | 1 | 8,655 |
| Average DisGeNET Score | 0.056 | 0.01 | 0.100 | 0.01 | 1 | 360,647 | 0.069 | 0.01 | 0.122 | 0.01 | 1 | 8,655 |
| In bottom 50% DisGeNET Score (%) | 0.500 | 0 | 0.500 | 0 | 1 | 360,647 | 0.576 | 0 | 0.494 | 0 | 1 | 8,655 |
| In top 10% DisGeNET Score (%) | 0.094 | 0 | 0.291 | 0 | 1 | 360,647 | 0.150 | 0 | 0.357 | 0 | 1 | 8,655 |
| Follow-on articles (all) | 0.570 | 1 | 3.199 | 0 | 447 | 360,647 | 1.054 | 0 | 8.904 | 0 | 459 | 8,655 |
| Follow-on articles (causal) | 0.198 | 0 | 1.469 | 0 | 276 | 360,647 | 0.513 | 0 | 5.020 | 0 | 271 | 8,655 |
| Follow-on articles (correlational) | 0.129 | 0 | 1.047 | 0 | 235 | 360,647 | 0.021 | 0 | 0.378 | 0 | 23 | 8,655 |
| Anomaly Index | 0.024 | 0 | 0.135 | 0 | 1 | 357,840 | 0.105 | 0 | 0.291 | 0 | 1 | 8,617 |
| Year of the association | 2011.04 | 2011 | 3.385 | 2005 | 2016 | 360,647 | 2012.64 | 2013 | 2.614 | 2005 | 2016 | 8,655 |

*Note:* Panel A presents descriptive statistics on papers that introduce new gene-disease associations (GDAs) after 2005. *Associations per paper*= number of new GDAs introduced by the focal article; *Genes per paper*= number of genes associated with a disease by the focal article; *Year of publication*= year in which the focal article is published. Panel B presents descriptive statistics on new gene-disease associations introduced after 2005. *With never associated genes (%)*= share of GDAs that include a gene never associated with a disease before 2005; *With recently discovered genes (%)*= share of GDAs that include a gene discovered after the year 2000 (i.e., after the Human Genome Project); *Average DisGeNET Score*= average DisGeNET Score of GDAs' scientific quality; *In bottom 50% DisGeNET Score (%)*= share of GDAs that fall below the median of DisGeNET Score; *In top 10% DisGeNET Score (%)*= share of GDAs that fall in the top $90^{th}$ percentile of DisGeNET Score; *Follow-on articles (all)*: count of all subsequent studies investigating the gene-disease association; *Follow-on articles (causal)*: count of subsequent studies investigating causal relationships between the gene and the disease; *Follow-on articles (correlational)*: count of subsequent studies investigating correlational relationships between the gene and the disease; *Anomaly Index*= a synthetic measure of how unlikely a new gene–disease association is relative to prior genetic knowledge (more details in Appendix D); *Year of the association*= year in which the article introducing the GDA is published. The table reports only data that are effectively used in the empirical estimates, i.e., excluding observations that are dropped by the inclusion of fixed effects. See text for details.

Table 2: GWAS are more likely to discover gene-disease associations involving less-studied genes relative to candidate gene studies.

| | I(GDA with never associated gene>0) | | I(GDA with recently discovered gene>0) | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| GWAS (0/1) | 0.261*** | 0.188*** | 0.144*** | 0.108*** |
| | (0.0129) | (0.0152) | (0.0110) | (0.0131) |
| Disease FE | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES |
| Observations | 369,302 | 348,921 | 369,302 | 348,921 |
| Number of diseases | 9,863 | 9,504 | 9,863 | 9,504 |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the gene and disease level. All models include dummies controlling for disease and year fixed effects; Columns (2) and (4) also include dummies controlling for the principal investigator (PI) of the articles introducing a GDA. *I(GDA with never associated gene>0)*: 0/1 = 1 if the new gene-disease association encompasses a gene never associated with a disease before 2005 (the year of the first GWAS); *I(GDA with recently discovered gene>0)*: 0/1 = 1 if the new gene-disease association encompasses a gene discovered after the year 2000 (the year of the Human Genome Project's first draft completion); *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. See text for details and Appendix Table G.1 for the corresponding results using continuous versions of the dependent variables.

Table 3: GWAS discover gene-disease associations of higher scientific quality and that receive more follow-on research relative to candidate gene studies.

**Panel A: Effects of GWAS on the quality of discovery**

| | DisGeNET Score of the GDA | | I(GDA in bottom 50% DisGeNET Score>0) | | I(GDA in top 10% DisGeNET Score>0) | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| GWAS (0/1) | 0.0117* | 0.0111* | 0.0625*** | 0.0369* | 0.0411** | 0.0355* |
| | (0.00454) | (0.00476) | (0.0148) | (0.0170) | (0.0140) | (0.0149) |
| Disease FE | YES | YES | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES | NO | YES |
| Observations | 369,302 | 348,921 | 369,302 | 348,921 | 369,302 | 348,921 |
| Number of diseases | 9,863 | 9,504 | 9,863 | 9,504 | 9,863 | 9,504 |

**Panel B: Effects of GWAS on follow-on science**

| | Follow-on articles on the GDA (all) | | Follow-on articles on the GDA (causal) | | Follow-on articles on the GDA (correlational) | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| GWAS (0/1) | 0.816*** | 0.543** | 0.450*** | 0.288** | -0.00253 | 0.0611 |
| | (0.146) | (0.188) | (0.0748) | (0.109) | (0.0114) | (0.0406) |
| Disease FE | YES | YES | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES | NO | YES |
| Observations | 369,302 | 348,921 | 369,302 | 348,921 | 369,302 | 348,921 |
| Number of diseases | 9,863 | 9,504 | 9,863 | 9,504 | 9,863 | 9,504 |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the gene and disease level. All models include dummies controlling for disease and year fixed effects; Columns (2), (4) and (6) also include dummies controlling for the principal investigator (PI) of the articles introducing a GDA. *DisGeNET Score of the GDA* : synthetic measure of scientific value of the discovery provided by DisGeNET; *I(GDA in bottom 50% DisGeNET Score>0)*: 0/1 = 1 if the gene-disease association has a DisGeNET Score below the sample median; *I(GDA in top 10% DisGeNET Score>0)*: 0/1 = 1 if the gene-disease association has a DisGeNET Score in the top sample decile; *Follow-on articles on the GDA (all)*: count of all subsequent studies investigating the gene-disease association; *Follow-on articles on the GDA (causal)*: count of subsequent studies investigating causal relationships between the gene and the disease; *Follow-on articles on the GDA (correlational)*: count of subsequent studies investigating correlational relationships between the gene and the disease; *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. Follow-on articles are papers directly working on the gene-disease combinations, regardless of whether they cite the study that first introduced it. See text for details.

Table 4: Instrumental variable estimates confirm that GWAS introduce gene-disease associations with higher scientific quality and that generate more follow-on research relative to candidate gene studies.

**Panel A: Causal effects of GWAS on the quality of discovery**

| | First Stage | Second Stage (2SLS) | | |
|---|---|---|---|---|
| | GWAS (0/1) | DisGeNET Score of the GDA | I(GDA in bottom 50% DisGeNET Score>0) | I(GDA in top 10% DisGeNET Score>0) |
| | (1) | (2) | (3) | (4) |
| IV: Genotyping cost shift-share | -0.0104** | | | |
| | (0.00235) | | | |
| GWAS (0/1) | | 0.155*** | 0.0374 | 0.429* |
| | | (0.0186) | (0.0280) | (0.154) |
| **F-Statistic (First Stage)** | 19.5 | | | |
| Disease FE | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES |
| Observations | 88,581 | 88,581 | 88,581 | 88,581 |
| Number of diseases | 1,724 | 1,724 | 1,724 | 1,724 |

**Panel B: Causal effects of GWAS on follow-on science**

| | First Stage | Second Stage (2SLS) | | |
|---|---|---|---|---|
| | GWAS (0/1) | Follow-on articles on the GDA (all) | Follow-on articles on the GDA (causal) | Follow-on articles on the GDA (correlational) |
| | (1) | (2) | (3) | (4) |
| IV: Genotyping cost shift-share | -0.0104** | | | |
| | (0.00235) | | | |
| GWAS (0/1) | | 3.131** | 1.736** | 2.448 |
| | | (0.652) | (0.422) | (2.000) |
| **F-Statistic (First Stage)** | 19.5 | | | |
| Disease FE | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES |
| Observations | 88,581 | 88,581 | 88,581 | 88,581 |
| Number of diseases | 1,724 | 1,724 | 1,724 | 1,724 |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the gene and disease class level. All models include dummies controlling for disease and year fixed effects. Note that the number of observations is smaller because some diseases lack information on their prevalence in the pre-GWAS period. The first stage regresses the *GWAS* indicator on the instrument *Genotyping cost shift-share*. The second stage uses the fitted values of *Genotyping cost shift-share* as the regressor. *DisGeNET Score*: synthetic measure of scientific value of the discovery provided by DisGeNET; *I(GDA in bottom 50% DisGeNET Score>0)*: 0/1 = 1 if the gene-disease association has a DisGeNET Score below the sample median; *I(GDA in top 10% DisGeNET Score>0)*: 0/1 = 1 if the gene-disease association has a DisGeNET Score in the top sample decile; *Follow-on articles on the GDA (all)*: count of all subsequent studies investigating the gene-disease association; *Follow-on articles on the GDA (causal)*: count of subsequent studies investigating causal relationships between the gene and the disease; *Follow-on articles on the GDA (correlational)*: count of subsequent studies investigating correlational relationships between the gene and the disease. Follow-on articles are papers directly working on the gene-disease combinations, regardless of whether they cite the study that first introduced it. See text and Appendix F for more details.

# Data-Driven Search and the Birth of Theory: Evidence from Genome-Wide Association Studies

## Appendix

# A Additional Details on Genomic Research

## A.1 Scientific Background

Genomics is the branch of biological science focused on the study of genomes—that is, the complete set of an organism's genes. Genes are sequences of DNA bases that encode the "instructions" for synthesizing gene products, most notably proteins. Genes play a fundamental role in the functioning of the human body, but their sequences can sometimes acquire mutations. When this happens, they may alter their behavior, sometimes with serious consequences and the emergence of health conditions. At the same time, the role of genes in the etiology of disease creates opportunities for therapeutic intervention: genes causally linked to a condition can often serve as drug targets (Nelson et al., 2015). When a drug molecule binds to its target, it can modify the gene's function, potentially improving the condition. As a result, understanding the genetic roots of disease has direct implications for the design of pharmaceutical drugs.

Genetic research has historically been very effective at identifying individual genes responsible for specific disease conditions. These are known as Mendelian disorders, usually visible from birth and traceable through family history. Understanding the causes of Mendelian disorders was one of the earliest and most important successes of genetic research (Bush and Moore, 2012). Take, for example, the case of cystic fibrosis. This rare disorder is caused by DNA mutations that tend to cluster in a specific region of the genome: the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene. The approach to make this discovery involved genotyping families affected by cystic fibrosis. Despite small sample sizes, due to the rarity of the disease, the strong effect and localization of the responsible mutations in a single gene made it possible to unambiguously identify the gene as causally linked to the condition.

However, Mendelian diseases are rare, as their severity renders them susceptible to negative evolutionary selection. Much more common are complex diseases, which are caused by mutations in multiple genes. In these cases, any mutation may increase disease risk without being either necessary or sufficient, meaning that it usually accounts for only a small fraction of the condition's heritability. Although complex disorders also cluster within families, they do not follow predictable inheritance patterns, as they are shaped by intricate interactions between genetic and environmental factors. For this reason, family-based studies have proven less effective when applied to more common polygenic diseases.

## A.2 A Scientific Primer on GWAS

Towards the end of the last century, researchers increasingly embraced the idea that common disorders are likely influenced by genetic mutations that are also common in the population (Reich and Lander, 2001).

Rather than searching for individual genes with strong effects, the field has shifted toward studying common variants that individually have only a small impact on disease risk (Bush and Moore, 2012). But what exactly is a variant? At the most basic level, two genomes differ at a specific genetic location if they contain different single nucleotides (adenine, thymine, cytosine, or guanine) at that position. When such a difference occurs in at least 1% of the population, it is referred to as a single-nucleotide polymorphism (SNP). Connecting SNPs with diseases is based on the idea that a causative variant should appear more frequently in individuals with the disease than in those without it. In practice, researchers search for statistical associations between variants and diseases in large population samples.

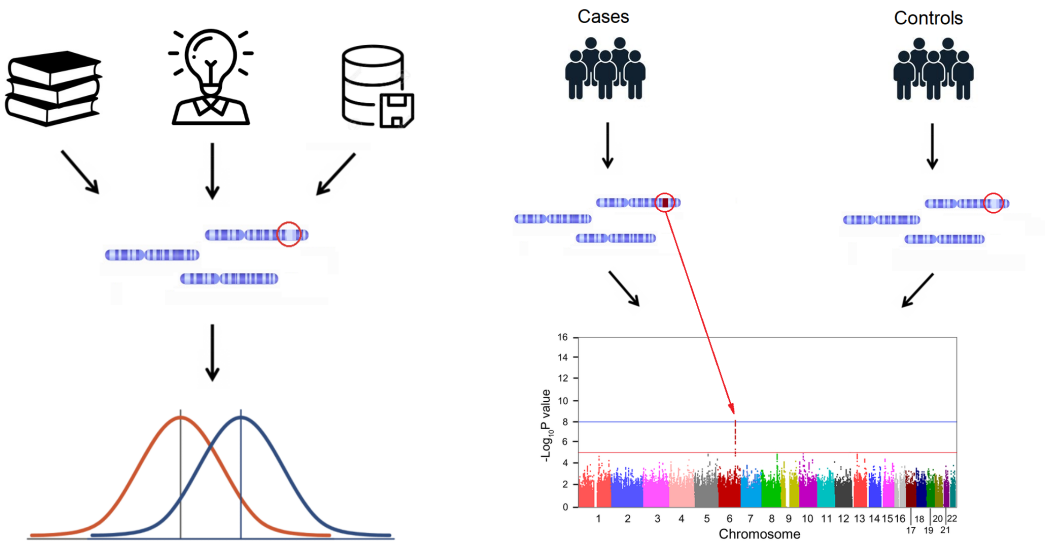
Building on this logic, candidate gene studies are a hypothesis-driven approach used in genetic research to identify associations between specific genes and diseases. Panel (a) of Figure A.1 shows a stylized representation. Researchers begin by selecting one or more genes based on prior biological knowledge, typically genes believed to be involved in the physiological pathways relevant to the disease under investigation. The process involves genotyping individuals in a case-control design to detect whether genetic variants in the candidate genes occur more frequently in individuals with the disease than in those without it. Statistical tests are then used to determine whether these variants are significantly associated with the presence or severity of the disease. Note that this approach genotypes individuals only at the specific genetic locations hypothesized as important. Candidate gene studies are thus limited by their reliance on existing theory, which can bias discovery toward well-known genes and overlook novel or unexpected genetic contributors.

In contrast, genome-wide association studies (GWAS) are hypothesis-free methods for identifying gene-disease associations (Visscher et al., 2017). Panel (b) of Figure A.1 shows a representation. Like candidate gene studies, GWAS are case-control studies: researchers collect DNA from patients with the disease under study and individuals without it. However, in this case, researchers rely on high-throughput microarrays to collect data on millions of genetic variants across the entire genome of the individual. Compared to candidate gene studies, GWAS typically require much larger sample sizes to achieve statistical significance, given the large number of multiple hypotheses tested, thereby increasing their costs. It is important to note that array-based GWAS do not sequence DNA base by base, since they only detect the presence or absence of specific SNPs. While microarrays can genotype millions of SNPs, they still cover less than 0.1% of the genome. However, because the co-occurrence of nearby genetic variants is not random (a phenomenon known as linkage), researchers can use reference genomes to efficiently infer the characteristics of the broader genome from the smaller subset of SNPs directly genotyped (Bush and Moore, 2012; Uffelmann et al., 2021).[14]

[14] GWAS collect data only on selected genetic loci whose variation can be extrapolated to represent their surrounding genetic regions. As a result, researchers can efficiently infer the broader structure of the genome by sampling only a tiny fraction of DNA bases

Figure A.1: Stylized comparison of alternative types of genetic studies.

(a) *Schema of a typical candidate gene study* (b) *Schema of a typical GWAS*




*Note:* Panel (a) shows a schema of how a candidate gene study unfolds. First, researchers select the disease of interest and decide which gene(s) to investigate based on previous literature, existing evidence, or their own intuition. Then, the genome of people with and without the condition is genotyped *at that specific genetic location* in search of differences. Finally, statistical methods are used to test the association between mutations in the targeted genes and the disease of interest. Panel (b) shows a schema of how a GWAS unfolds. First, researchers select the disease of interest, but not the gene(s). Then, *the entire genome* of people with and without the condition is genotyped in search of differences at any genetic location. Finally, statistical methods are used to test the association between mutations in each gene and the disease of interest. The results are often graphically represented as a "Manhattan plot," which shows the p-value of multiple statistical tests comparing DNAs between the case and control groups. The y-axis reports $-log_{10}(p-value)$, and hence higher values correspond to stronger associations.

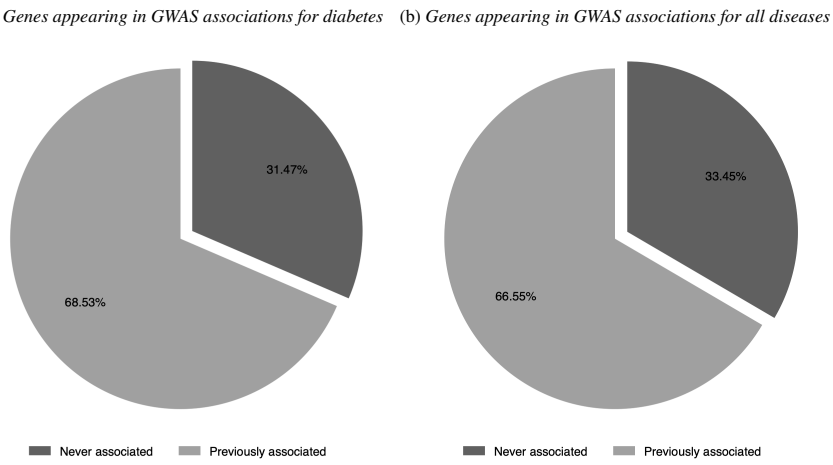
GWAS have led to many important scientific discoveries, but have also come under scrutiny for several limitations. First, genome-wide scans detect common variants that only serve as markers for broader genomic regions likely to contain causal mutations. However, GWAS cannot pinpoint the exact causal SNPs with certainty, so additional analyses or follow-up studies are typically required to narrow down the association region. Second, even when the relevant mutation is identified, GWAS are unable to illuminate the biological mechanism behind its role in human health, which is essential for drug discovery (Nelson et al., 2015). Third, most complex diseases are influenced by a large number of genes, meaning the role of any single variant is usually small. This limits the therapeutic value of GWAS findings, as variants with small effects may not be actionable drug targets (Goldstein, 2009). Finally, critics have noted that GWAS tend to overlook more complex genetic architectures, since they are primarily designed to test pairwise gene-disease associations (Boyle et al., 2017). As a result, their cost-effectiveness remains debated (Callaway, 2017).

(Visscher et al., 2017). This approach differs from more recent whole-genome sequencing, which reads every DNA base, and has superseded GWAS in the past few years.

## A.3 GWAS as "Unbiased" Data-Driven Search

A central feature of GWAS is that they are designed to screen genetic variation across the entire genome, without restricting attention to genes selected based on prior theory. Major reviews emphasize that *"GWASs are unbiased with respect to prior biological knowledge (or prior beliefs) and with respect to genome location"* (Visscher et al., 2012). This design choice is often summarized by describing GWAS as a *"hypothesis-free method"* for identifying associations between genes and diseases.[15] While this hypothesis-free property enhances the potential for discovery when biological understanding is incomplete (Pearson and Manolio, 2008), the design of GWAS also limits investigators' ability to steer the genetic search toward particular gene-disease links, since the same association-testing procedure is applied to the entire genome. Reporting norms further limit selective disclosure, because GWAS publications conventionally present the full set of statistically significant associations emerging from the scan rather than a hand-picked subset of "novel" genes. This convention is reflected in the GWAS Catalog, whose curators manually extract and record all the reported associations from GWAS publications.

Figure A.2: Rediscovery of known genes in GWAS reported associations.




*Note:* Each panel reports the share of genes that appear in GWAS reported associations that were already linked to the corresponding disease prior to 2005 ("known") versus genes that were not previously linked ("novel"). Panel (a) restricts to GWAS on diabetes, while panel (b) aggregates across all diseases in the sample.

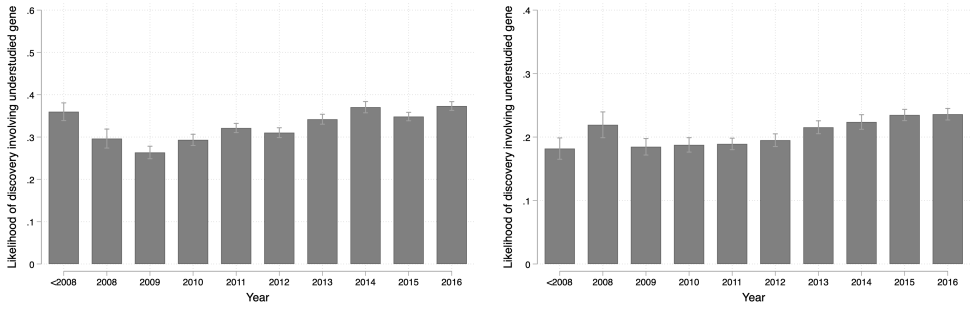
Several falsification tests support interpreting GWAS as an *unbiased* data-driven search procedure. First, because GWAS implement a genome-wide screen rather than targeted testing, studies should not only surface novel loci but also re-identify previously established associations. This is visible in the motivating example of Hakonarson et al. (2007): Figure B.1 shows that alongside a novel association involving *KIAA0350*, the GWAS also re-discovers known associations at four genes (*PTPN22*, *INS*, *COL1A2*, *LPHN2*). The

[15] Source: https://www.ebi.ac.uk/gwas-catalogue/what-are-genome-wide-association-studies/.

same pattern appears in the aggregate: across all GWAS in my sample, 66.5% of genes reported as GWAS discoveries were already known prior to 2005, as summarized in Figure A.2. Second, I examine whether the observable characteristics of GWAS discoveries display secular time trends, which would be inconsistent with a stable genome-wide screening procedure. Figure A.3 shows that the main discovery features remain stable over time, indicating that GWAS generate similar kinds of discoveries across years.

Figure A.3: The novelty yield of GWAS is stable over time.

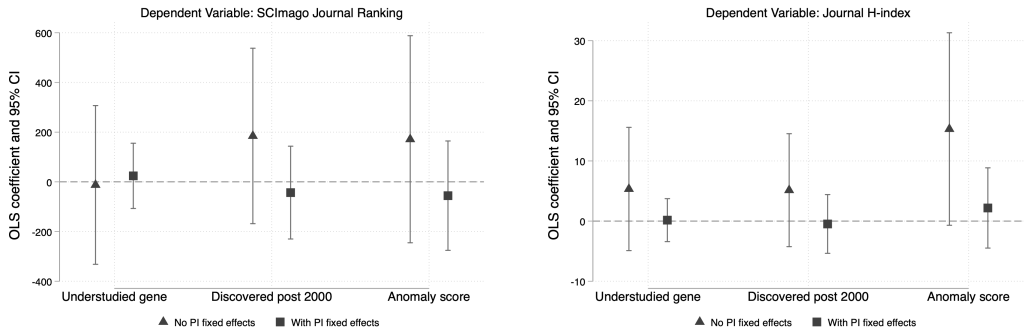
(a) *I(GDA with never associated gene>0)* (b) *I(GDA with recently discovered gene>0)*




Finally, I test whether novelty is rewarded in the GWAS publication process. If GWAS researchers had strong incentives to pursue novel findings, then studies reporting more novel or more counterintuitive gene-disease associations should systematically appear in more prestigious outlets. Figure A.4 shows no such relationship: the novelty or degree of anomaly of GWAS discoveries is not associated with better publication outcomes. The absence of a novelty premium suggests that GWAS researchers face limited incentives to strategically emphasize novelty in what they report or publish, reinforcing the interpretation that the novelty content of GWAS findings is largely a product of genome-wide screening.

Figure A.4: The novelty of GWAS discoveries is not related to better publication outcomes.

(a) *Journal prestige (SCImago rank)* (b) *Journal impact (journal h-index)*




*Note:* Each panel relates the novelty or anomaly of GWAS-reported associations to publication outcomes. Panel (a) measures journal prestige using the SCImago Journal Rank, and panel (b) measures journal citation impact using the journal h-index.

# B Detailed Case Study: Qu et al. (2007) vs. Hakonarson et al. (2007)

## B.1 Genetic Research on type 1 diabetes

It is estimated that 10% of Americans (around 37.3 million people) have diabetes.[16] The two main forms of diabetes are type 1 and type 2, with type 1 accounting for approximately 5–10% of cases. Type 1 diabetes is a chronic condition that typically begins in childhood. More specifically, it is an autoimmune disease caused by the immune system's destruction of pancreatic $\beta$-cells (DiMeglio et al., 2018). As a result, the pancreas is unable to produce sufficient insulin, the hormone essential for enabling sugar to enter cells, produce energy, and regulate blood glucose levels. To date, there is no known way to prevent type 1 diabetes, and lifelong insulin therapy remains necessary for patient survival.

Genetics plays a significant role in the onset of type 1 diabetes: children with a parent affected by the condition face a relative risk of 1–9% of developing it themselves (DiMeglio et al., 2018). Early candidate gene studies identified several genetic determinants, primarily within a tightly linked group of genes known as the major histocompatibility complex (MHC). MHC genes encode cell surface proteins that are essential for initiating and directing the immune response. When these genes malfunction, they can trigger autoimmune activity, such as the immune system attacking the body's own $\beta$-cells in type 1 diabetes. However, MHC genes account for only slightly more than half of the genetic risk associated with the disease, suggesting that other, still unidentified, loci are involved. A systematic effort to detect these remaining genes could reveal alternative therapeutic targets and shed light on the deeper causes of type 1 diabetes.

## B.2 A Tale of Two Studies

In June 2007, Qu et al. (2007) published a candidate gene study based on a sample of 947 nuclear family trios with type 1 diabetes (i.e., one affected child and both parents). Using a targeted sequencing assay, the researchers tested the association between type 1 diabetes and a specific gene: interferon regulatory factor 5 (IRF5). The choice of IRF5 was informed by prior findings linking it to autoimmune diseases such as systemic lupus erythematosus. Based on these similarities, the authors hypothesized that IRF5 might also play a role in type 1 diabetes, another autoimmune condition with overlapping features. As they put it, the *"[...] association of IRF5 with other autoimmune diseases, such as T1D, has a high prior probability"* based on existing evidence and genetic understanding. However, despite being grounded in theoretical reasoning and prior knowledge, the hypothesis turned out to be incorrect. The results showed no significant association between mutations in IRF5 and type 1 diabetes risk, leading Qu et al. (2007) to conclude that their proposed

[16]The figure is taken from: https://www.cdc.gov/diabetes

gene-disease combination had limited therapeutic potential.

Two months later, in August 2007, Hakonarson et al. (2007) published a GWAS based on a study population of 563 patients with type 1 diabetes and 1,146 controls.[17] The analysis was conducted using a microarray capable of genotyping 550,000 single nucleotide polymorphisms (SNPs) across the whole genome. The study identified several SNPs significantly associated with type 1 diabetes. Figure B.1 presents the key results from Hakonarson et al. (2007). Some of the significant SNPs were found in genes already known to be related to diabetes (for example, the insulin gene INS), but three were located in KIAA0350. It is now understood that this gene helps regulate $\beta$-cell function and thus plays a role in preventing diabetes. However, KIAA0350 was among the least studied human genes at the time, and its function was largely unknown (Soleimanpour et al., 2014). The association between KIAA0350 and type 1 diabetes proved robust and has since been investigated in multiple follow-up studies, fundamentally contributing to a deeper genetic understanding of the disease (Gingerich et al., 2020).

## B.3 The Effect of GWAS on Follow-on Research

The studies by Hakonarson et al. (2007) and Qu et al. (2007) offer an unusually clean comparison because both were conducted in the same year, on the same disease, and by the same principal investigator. Figure B.2 summarizes the quantity of follow-on research each gene-disease combination received, based on DisGeNET.

[17] The study also replicated its main analysis in a separate sample of 483 nuclear family trios, leveraging genetic differences between affected children and their parents.

Figure B.1: Main results from the GWAS analysis of Hakonarson et al. (2007).

(a) *Results from main analysis*

| Case-control cohort | | | | |
|---|---|---|---|---|
| Chr. | SNP | OR (95% CI) | *P*-value | Locus |
| 1 | rs2476601 | 1.80 (1.44, 2.24) | $1.32 \times 10^{-7}$ | *PTPN22* |
| 11 | rs1004446 | 0.62 (0.53, 0.73) | $4.38 \times 10^{-9}$ | *INS* |
| 16 | rs2903692 | 0.65 (0.56, 0.76) | $4.77 \times 10^{-8}$ | *KIAA0350* |
| 11 | rs6356 | 1.52 (1.31, 1.76) | $1.78 \times 10^{-8}$ | *INS* |
| 16 | rs725613 | 0.67 (0.58, 0.78) | $3.24 \times 10^{-7}$ | *KIAA0350* |
| 7 | rs10255021 | 0.58 (0.44, 0.77) | $1.16 \times 10^{-4}$ | *COL1A2* |
| 11 | rs10770141 | 0.65 (0.56, 0.76) | $7.20 \times 10^{-8}$ | *INS* |
| 1 | rs672797 | 1.54 (1.29, 1.85) | $2.67 \times 10^{-6}$ | *LPHN2* |
| 16 | rs17673553 | 0.66 (0.55, 0.78) | $1.30 \times 10^{-6}$ | *KIAA0350* |
| 11 | rs7111341 | 0.63 (0.53, 0.76) | $3.77 \times 10^{-7}$ | *INS* |
| 11 | rs10743152 | 0.67 (0.57, 0.78) | $4.73 \times 10^{-7}$ | *INS* |

(b) *Text excerpt on KIAA0350 gene*

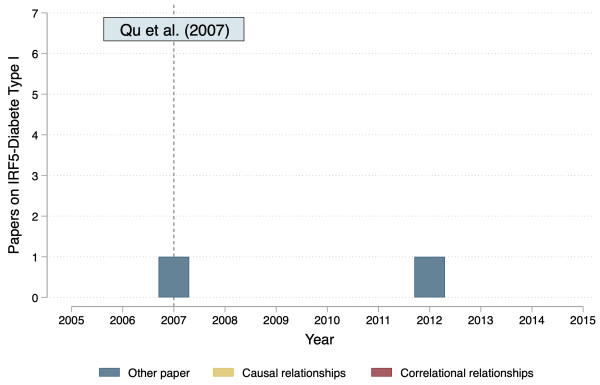
This locus resides in a 233-kb block of LD that contains only *KIAA0350* and no other genes, making this gene a prime candidate for harbouring the causative variant. *KIAA0350* encodes a protein of unknown function and its genomic location is next to the suppressor of cytokine signalling 1 (*SOCS1*) gene. The almost exclusive expression specificity of *KIAA0350* in immune cells (http://symatlas.gnf.org/SymAtlas), including dendritic cells, B lymphocytes and natural killer (NK) cells, all of which are pivotal in the pathogenesis of T1D[27,28], indicates that the variant probably contributes to the disease by modulating immunity.

*Note:* Panel (a) shows an excerpt from Table 1 of Hakonarson et al. (2007). The genetic location of the single nucleotide polymorphisms significantly associated with type 1 diabetes is shown in the rightmost column. Highlighted in red are the three genetic variants located in the KIAA0350 gene. Panel (b) shows the passage of Hakonarson et al. (2007) describing the inferred role of KIAA0350. T1D stands for type 1 diabetes. Highlighted in red is the description of the GWAS' key findings.

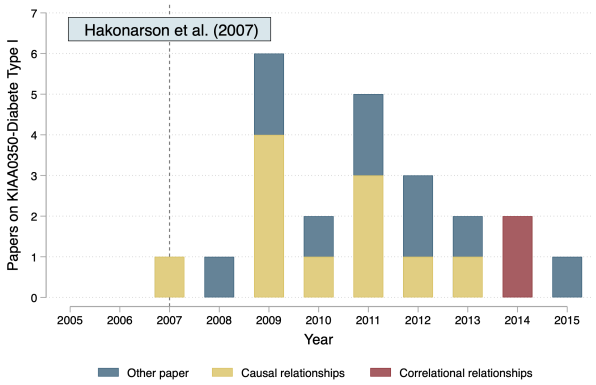
Beyond this difference in volume, the contrast is also qualitative. The follow-on research triggered by the GWAS association with KIAA0350 (now renamed as *CLEC16A*) is disproportionately organized around understanding if and how it is causally tied to diabetes: it takes the statistical association as a puzzle, then iteratively translates it into a chain of biological claims that can be tested and refined.

Figure B.2: Follow-on papers investigating the association between type 1 diabetes and the IRF5 and KIAA0350 genes, respectively.

(a) *Papers on IRF5-T1D after Qu et al. (2007)*




(b) *Papers on KIAA0350-T1D after Hakonarson et al. (2007)*




*Note:* Panel (a) shows the yearly count of publications exploring the relationship between IRF5 and type 1 diabetes following the candidate gene study by Qu et al. (2007). Panel (b) shows the yearly count of publications exploring the relationship between KIAA0350 and type 1 diabetes following the GWAS by Hakonarson et al. (2007). Data on publications and the gene-diseases studied come from DisGeNET, while information on the type of relationship studied is from PubTator3. In both panels, the focal paper introducing the gene-disease combination is excluded.

A qualitative reading of these paper suggests how the GWAS finding influenced subsequent research. A first set of follow-on papers treats the GWAS association as an *identification problem* and asks what, exactly, in the genetic locus of KIAA0350 could be responsible for disease risk. Because the original association spans a broader region, subsequent work concentrates on fine mapping and functional annotation, re-sequencing the region and testing whether risk alleles act through gene regulation. In so doing, studies building on Hakonarson et al. (2007) do not merely replicate the association, but they aim to specify how mutations in the KIAA0350/CLEC16A region could plausibly increase type 1 diabetes risk. In this sense, the anomaly created an impetus to reconcile genotype-level evidence with a biologically interpretable target.

A second set of follow-on papers treats the locus as a *mechanistic puzzle* and asks what the implicated gene does in disease-relevant tissue, and through what biological pathway. This work moves from association to theory-building by proposing testable hypotheses, identifying molecular interaction pathways, and using experimental methods to establish directionality. For example, later work characterizes CLEC16A as an endosomal protein and links it to mitochondrial quality control through mitophagy in pancreatic $\beta$-

cells, with downstream effects on insulin secretion and glucose homeostasis (Soleimanpour et al., 2014). This exemplifies the core dynamics highlighted by my results: an anomalous GWAS discovery motivates theorizing by forcing researchers to build and test a causal narrative that connects an unexpected gene to a specific physiological function. Empirically, this shows up as a shift toward studies that make mechanism-oriented claims, such as demonstrating that altering gene function changes disease-relevant phenotypes.

In contrast, the candidate-gene study of Qu et al. (2007) generated little follow-on work (Panel (a)), maybe not surprisingly given the weak findings. Importantly, this difference also reflects how the two approaches structure subsequent research. In the IRF5 case, prior knowledge guided attention to a familiar and already intensively studied gene, so the study's main contribution was an assessment of an existing theory-driven conjecture, rather than uncovering of a new puzzle. In the case of Hakonarson et al. (2007), the GWAS result pointed to an anomaly relative to prior expectations and, in turn, a clearer opportunity for follow-on researchers to generate new understanding from the discovery.

# C DisGeNET Data

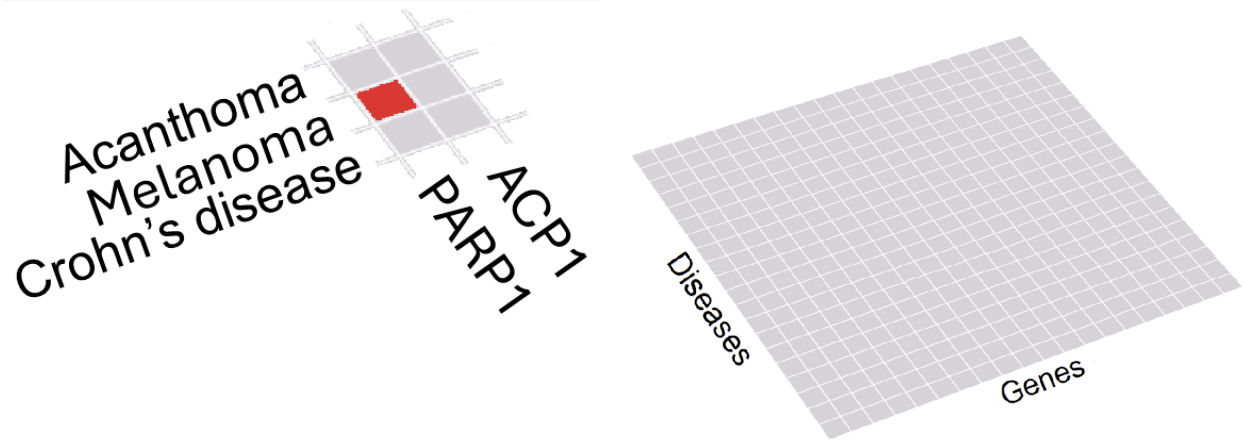
This appendix describes the empirical approach used in the paper and then provides additional details on the DisGeNET data.

## C.1 The Gene-Disease Landscape

One advantage of focusing on a specific search problem, such as identifying the genetic roots of human diseases, is that the search landscape is well-defined. The relevant components in this combinatorial problem are genes and diseases. In principle, any gene could contain mutations that contribute to a given condition. This means the search task can, in first approximation, be reduced to identifying gene–disease pairs $< i, j >$ in which mutations in gene $i$ are causally linked to the emergence of disease $j$ (Panel (a) of Figure C.1). Another benefit of this setting is that all the relevant components of the problem have already been mapped and codified in exhaustive taxonomies. The Human Genome Project cataloged approximately 19,000 protein-coding genes and assigned them unique names and identifiers. Similarly, thousands of human diseases have been classified in systems such as the Unified Medical Language System (UMLS), which also provides unique disease identifiers. The Cartesian product of these entities defines the full combinatorial landscape in which search takes place, as illustrated in Panel (b) of Figure C.1.

Figure C.1: Empirically tracing the combinatorial landscape to study search.

(a) *Example of gene-disease combinations* (b) *The gene-disease combinatorial landscape*




Note: Panel (a) shows an example of gene-disease pairwise combinations. Genes and diseases can be sorted based on similarity or relatedness, such as melanoma and acanthoma, which are both skin tumors. The figure identifies in red the combination of PARP1 and melanoma, which was introduced by a GWAS. Panel (b) shows the idealized combinatorial landscape of all possible gene-disease pairs.

In terms of empirical design, the combinatorial landscape represents the "ground truth," that is, the space of entities over which researchers conduct their search activities. Mapping this knowledge landscape allows

for the empirical study of how actors search across it (Aharonson and Schilling, 2016). Specifically, one can locate research outputs (e.g., scientific publications) on the landscape by extracting the relevant entities (e.g., genes and diseases), thereby characterizing search in spatial terms. This approach offers two main advantages. First, it enables researchers to record the features of each combination relative to its position on the landscape. For example, Figure 2 illustrates how the spatial distribution of new combinations can be used to assess whether a search is "local" or "distant." Second, treating search as a landscape-based activity removes the need for citations to track knowledge evolution (Arts et al., 2025). Instead, one can observe changes in follow-on research that directly engages with the same entities before and after a given study. For instance, Figure 5 plots the yearly number of papers focused on a specific gene–disease combination, regardless of whether they cite the original study that introduced it.

This empirical approach extends beyond the specific application in this paper. In practice, it can be used in any context where the relevant search landscape can be defined ex ante. Advances in science and technology have increasingly made this possible. From genetic atlases (Kang, 2025; Kao, 2024) to satellite imagery (Nagaraj, 2022), large-scale mapping efforts are turning a growing number of search problems into well-defined landscapes (Tranchero, 2024).

### C.2 The DisGeNET Score

The main source of data for this paper is DisGeNET (v7.0), a platform that integrates information from multiple sources to create a comprehensive repository of scientific findings linking human diseases to their genetic causes (Hermosilla and Lemus, 2019; Piñero et al., 2020). The database compiles gene-disease associations (GDAs) from manually curated datasets, experimental studies on animal models, and literature mining from PubMed. The version used in this paper covers publications through 2016 and includes over 628,000 gene–disease pairs involving 17,549 genes and 24,166 diseases. Genes are identified using their NCBI Gene ID (formerly EntrezGene ID), and diseases are coded using UMLS concept unique identifiers. The empirical analyses focus on GDAs first introduced between 2005 and 2016. The start date reflects the emergence of GWAS as a research design in 2005, which is necessary to compare genome-wide and targeted approaches during the same time window. Instead, the end date is determined by data availability, since the DisGeNET version I use covers publications through 2016, and later versions are not freely available to researchers.

DisGeNET is designed to help researchers in both academia and industry prioritize promising genetic targets. To support this goal, it provides a synthetic DisGeNET Score for each gene–disease combination.

The Score ranges from 0 to 1, with higher scores indicating associations that are scientifically more robust and therapeutically more promising. In that sense, it is closer to a "ground truth" assessment of whether an association ultimately proves scientifically consequential, rather than a mere measure of how much attention the discovery happens to attract. In the version used in this paper (v7.0), the Score is defined as follows:

$$\text{DisGeNET Score of gene-disease combination} < \text{i, j} > = C_{i,j} + M_{i,j} + I_{i,j} + L_{i,j}$$

The first component $C_{i,j}$ summarizes the evidence from curated sources reporting gene-disease combination $< i, j >$:

$$C_{i,j} = \begin{cases} 0.6 & \text{if } N_{sources_c} > 2 \\ 0.5 & \text{if } N_{sources_c} = 2 \\ 0.3 & \text{if } N_{sources_c} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{C.1}$$

where $N_{sources_c}$ is the number of curated sources supporting a gene-disease association, including CGI, ClinGen, Genomics England, CTD, PsyGeNET, Orphanet, and UniProt.

The second component $M_{i,j}$ summarizes the evidence from experiments using animal models reporting gene-disease combination $< i, j >$:

$$M_{i,j} = \begin{cases} 0.2 & \text{if } N_{sources_m} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{C.2}$$

where $N_{sources_m}$ is the number of sources using the lab rat or lab mouse from RGD, MGD, and CTD.

The third component $I_{i,j}$ summarizes the evidence inferred from experiments on gene-disease combination $< i, j >$:

$$I_{i,j} = \begin{cases} 0.1 & \text{if } N_{sources_i} > 1 \\ 0 & \text{otherwise} \end{cases} \tag{C.3}$$

where $N_{sources_i}$ is the number of sources from HPO and CLINVAR.[18]

Finally, the component $L_{i,j}$ summarizes the evidence mined from the literature about gene-disease combi-

[18] In the original version of the Score, papers listed in the GWAS Catalog where also included in this count. However, this would generate a mechanical upward bias in the DisGeNET Score of gene-disease pairs introduced by GWAS. Therefore, the current paper excludes those sources from the calculation of the DisGeNET Score, ensuring that the results in Table 3 are not due to a different weighting of GWAS papers.

nation $< i, j >$:

$$L_{i,j} = \begin{cases} 0.1 & \text{if } N_{publications} > 9 \\ \text{N}_{publications}\text{*0.01} & \text{if } N_{publications} <= 9 \end{cases} \tag{C.4}$$

where $N_{publications}$ is the number of publications supporting a gene-disease association as mined by LHGDN and BEFREE.

The DisGeNET Score has strong face validity and has been thoroughly validated (Piñero et al., 2020). Returning to the earlier example of type 1 diabetes, the data show that the IRF5–type 1 diabetes combination has a low DisGeNET Score of 0.03, compared to a much higher score of 0.46 for the KIAA0350–type 1 diabetes pair. These scores align well with the relative scientific and therapeutic impact of the two combinations (Gingerich et al., 2020). Table C.1 shows that publications introducing novel gene-disease combinations with higher DisGeNET Score receive a much larger number of citations. The result is robust to controlling for scientist fixed effects. As an additional validation, Table C.2 compares the DisGeNET Score of each gene–disease combination with real-world innovation and therapeutic outcomes at the gene-disease level. If the Score is a good proxy for the underlying quality of a combination, it should correlate with successful downstream pharmaceutical developments. This is exactly what the data show: combinations with higher DisGeNET Scores are associated with significantly more follow-on applications in patents, clinical trials, and FDA-approved drugs. Importantly, none of these downstream outcomes are used in calculating the Score itself.

Table C.1: The DisGeNET Score of a given gene-disease combination is strongly associated with the citations accrued to the paper first introducing it.

| | Citations to the paper introducing the gene-disease pair | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| DisGeNET Score of the gene-disease pair | 117.4*** | 106.1*** | 65.79*** |
| | (8.564) | (8.802) | (5.731) |
| Year of discovery FE | YES | YES | YES |
| Disease FE | NO | YES | YES |
| Principal investigator FE | NO | NO | YES |
| Observations | 374,596 | 369,305 | 347,311 |
| Number of diseases | 15,155 | 9,864 | 9,507 |
| Mean of the DV | 45.18 | 45.18 | 45.18 |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the gene and disease level. All models include dummies controlling for year fixed effects. Column 2 adds disease fixed effects, and Column 3 adds scientist fixed effects. *Citations to the paper introducing the gene-disease pair*: count of scientific citations received by the focal article introducing the new gene-disease combination (data from NIH's iCite); *DisGeNET Score of the gene-disease pair*: a synthetic measure of scientific reliability of the gene-disease association provided by DisGeNET.See text for details.

Table C.2: The DisGeNET Score of a given gene-disease combination is strongly associated with more USPTO granted patents, clinical impact, and FDA-approved drugs on the same combination.

| | Patents granted | Clinical citations | Drugs approved |
|---|---|---|---|
| | (1) | (2) | (3) |
| DisGeNET Score of the gene-disease pair | 1.313*** | 27.31*** | 0.0440*** |
| | (0.207) | (3.381) | (0.0128) |
| Disease FE | YES | YES | YES |
| Year of discovery FE | YES | YES | YES |
| Observations | 369,291 | 369,291 | 369,291 |
| Number of diseases | 9,862 | 9,862 | 9,862 |
| Mean of the DV | 0.193 | 2.128 | 0.008 |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the gene and disease level. All models include dummies controlling for disease and year fixed effects. *Patents granted*: count of USPTO patents granted from 2015 to 2023 that mention in their text a specific gene-disease combination. Data from EBI's SureChEMBL Workshop 2024 is available here: https://ftp.ebi.ac.uk/databases/SureChEMBL/SureChEMBLWorkshop2024/. *Clinical citations*: count of clinical articles up to 2024 citing the paper introducing a specific gene-disease combination. Data from NIH's iCite is available here: https://icite.od.nih.gov/. *Drugs approved*: count of FDA-approved drugs up to 2023 targeting a specific gene-disease combination. Data from DrugCentral 2023 is available here: https://drugcentral.org/. *DisGeNET Score of the gene-disease pair*: a synthetic measure of scientific reliability of the gene-disease association provided by DisGeNET. See text for details.

# D Anomaly Index

This appendix provides details on the construction of the Anomaly Index for each novel gene–disease association. It begins with a stylized example to illustrate the logic of the measure, which is computed based on established patterns of genetic co-occurrence, followed by a more detailed explanation and a set of robustness checks.

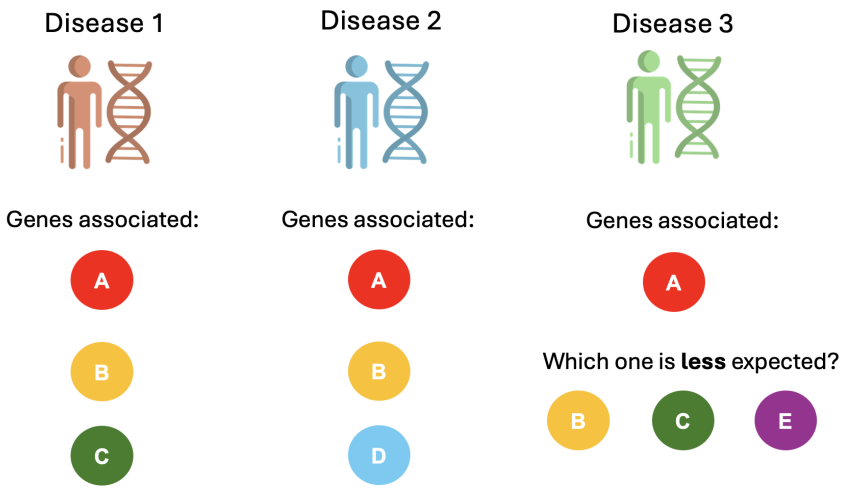
## D.1 Construction and Stylized Example

Consider the following example. There are three diseases:

- Disease 1, which has been associated with genes $\{A, B, C\}$ at time $t$;
- Disease 2, which has been associated with genes $\{A, B, D\}$ at time $t$;
- Disease 3, which has been associated with gene $\{A\}$ at time $t$.

Suppose now that Disease 3 can potentially be associated with genes $B$, $C$, and $E$ in time $t + 1$. How anomalous would be each potential combination between these genes and Disease 3, based on pre-existing association patterns?

Figure D.1: Example of how the Anomaly Index for each new gene-disease association is computed.




*Note:* See text for the step-by-step computation based on this simple example.

The first step involves computing the frequency of gene co-occurrences from past discoveries. In this example, we have the pairs of genes $\{(A, B), (A, C), (B, C)\}$ from Disease 1 and $\{(A, B), (A, D), (B, D)\}$ from Disease 2. Counting in how many diseases each pair occurs gives us the following frequencies:

- $\text{freq}(A, B) = 2$ (appears in Diseases 1 and 2)

- $\text{freq}(A, C) = 1$ (appears only in Disease 1)
- $\text{freq}(B, C) = 1$ (appears only in Disease 1)
- $\text{freq}(A, D) = 1$ (appears only in Disease 2)
- $\text{freq}(B, D) = 1$ (appears only in Disease 2)
- Any other potential pair (e.g. $(C, D), (A, E)$) has frequency 0.

Against this backdrop, Disease 3 has only been associated with the gene $\{A\}$ up to time $t$. Based on the pattern of genetic co-occurrences in Diseases 1 and 2, the association between Disease 3 and gene $\{A\}$ allows me to quantify how strongly prior genetic knowledge would anticipate each potential new gene link for Disease 3 at time $t + 1$. Accordingly, the *theory-consistency score* of linking a generic gene $g$ to Disease 3, given its existing gene set $\{A\}$, is:

$\text{Theory Consistency}(g, \text{Disease 3}) = \sum_{x \in \{A\}} \text{freq}(g, x) = \text{freq}(g, A)$

Which then leads to the following:

- $\text{freq}(B, A) = 2 \implies \text{Theory Consistency}(B, \text{Disease 3}) = 2$
- $\text{freq}(C, A) = 1 \implies \text{Theory Consistency}(C, \text{Disease 3}) = 1$
- $\text{freq}(E, A) = 0 \implies \text{Theory Consistency}(E, \text{Disease 3}) = 0$

From this, I define the Anomaly Index as:

$\text{Anomaly Index}(g) = \frac{1}{1+\text{Theory Consistency}(g)}$,

thus leading to the following calculations:

$$\text{Anomaly Index}(B, \text{Disease 3}) = \frac{1}{1+2} = \frac{1}{3} \approx 0.33,$$
$$\text{Anomaly Index}(C, \text{Disease 3}) = \frac{1}{1+1} = \frac{1}{2} = 0.50,$$
$$\text{Anomaly Index}(E, \text{Disease 3}) = \frac{1}{1+0} = 1.0.$$

Intuitively, the Anomaly Index ranges from 0 to 1, with higher values indicating that a given gene–disease pair is more discordant with prior discovery patterns observed in other diseases. The index is undefined for the first gene ever linked to a disease. It also equals 1 by construction when the gene has never been previously associated with any disease, as with gene $E$ in the example above.

### D.2 Validation of the Anomaly Index

Following the procedure outlined above, I computed the Anomaly Index for every gene–disease combination in DisGeNET introduced after 2005. In total, I was able to calculate the Index for 367,586 gene–disease

pairs (98.1% of the sample); the remaining 6,040 pairs involve the first gene ever associated with a given disease, for which the Index is not defined. The Anomaly Index takes the maximum value of 1 for the 11,289 combinations that include a gene never previously linked to any disease. Consider again the studies by Hakonarson et al. (2007) and Qu et al. (2007), both published in 2007 and focused on type 1 diabetes—a disease that, by the end of 2006, had already been associated with 1,020 genes. The Anomaly Index for the KIAA0350–type 1 diabetes combination is 1, since it marked the first time KIAA0350 had been linked to any disease. In contrast, the Anomaly Index for the IRF5–type 1 diabetes combination is 0.00077, reflecting the fact that IRF5 had frequently co-occurred with genes already associated with diabetes up to that point.

Table D.1: The Anomaly Index is higher for gene-disease combinations described as "novel" in the paper abstract, but not for other words generically denoting hype

| | Anomaly Index | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (46 |
| I(Results described as "Novel">0) | 0.0123*** | 0.0105*** | | | | |
| | (0.00144) | (0.00159) | | | | |
| I(Results described as "Major">0) | | | -0.000851 | -0.000852 | | |
| | | | (0.00124) | (0.00178) | | |
| I(Results described as "Critical">0) | | | | | 0.000171 | -0.000130 |
| | | | | | (0.00125) | (0.00170) |
| Disease FE | YES | YES | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES | NO | YES |
| Observations | 365,340 | 345,040 | 365,340 | 345,040 | 365,340 | 345,040 |
| Number of diseases | 8,746 | 8,450 | 8,746 | 8,450 | 8,746 | 8,450 |
| Mean of the DV | 0.027 | 0.027 | 0.027 | 0.027 | 0.027 | 0.027 |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the gene and disease level. All models include dummies controlling for disease and year fixed effects; Columns (2), (4) and (6) also include dummies controlling for the principal investigator (PI) of the articles introducing a GDA. *Anomaly Index*: a synthetic measure of how unlikely a given gene was to be associated with a disease given the pattern of genes previously associated with that disease up to the year before; *I(Results described as "Novel">0)*: 0/1 = 1 if the gene-disease association is described as "novel" in the abstract of the paper reporting it, according to the data by Mishra et al. (2023); *I(Results described as "Major">0)*: 0/1 = 1 if the gene-disease association is described as "major" in the abstract of the paper reporting it, according to the data by Mishra et al. (2023); *I(Results described as "Critical">0)*: 0/1 = 1 if the gene-disease association is described as "critical" in the abstract of the paper reporting it, according to the data by Mishra et al. (2023). See text for details.

Several additional analyses support the face validity of the Anomaly Index. First, I link the Index to how authors describe their findings in the abstract of the originating publication. Using data from Mishra et al. (2023), who apply a probabilistic model to classify adjectives in abstracts, I show that papers that label their discovery as "novel" introduce gene–disease pairs with an Anomaly Index that is 39–46% higher on average

(Table D.1). By contrast, I find no comparable relationship for broader evaluative terms such as "major" or "critical," which suggests that the result is not driven by generic hype in the abstract but by the Index capturing departures from established discovery patterns.

Second, I benchmark the Anomaly Index against expert assessments from the Faculty Opinions platform, following Shi and Evans (2023). Faculty Opinions provides post-publication evaluations in which domain experts annotate papers using standardized tags, including whether a study reports a "new finding" or a "technical advance." These annotations are available for 11,465 of the papers in my sample that introduce new gene–disease associations. Table D.2 shows that, in this subset, the Anomaly Index is positively and significantly associated with papers tagged as a "new finding," but not with tags capturing other forms of novelty such as the use of a new technique. This pattern supports the interpretation of the Index as capturing conceptual surprise in the content of the discovery.

Table D.2: The Anomaly Index is higher for gene-disease combinations rated as "new findings", but not for other type of ratings by scientists on Faculty Opinions.

| | Anomaly Index | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| I(Rated as "New Finding">0) | 0.0877** | | | |
| | (0.0267) | | | |
| I(Rated as "New Drug Target">0) | | 0.0209 | | |
| | | (0.0148) | | |
| I(Rated as "Controversial">0) | | | -0.0163 | |
| | | | (0.0266) | |
| I(Rated as "New Technique">0) | | | | -0.0119 |
| | | | | (0.0215) |
| Disease FE | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES |
| Principal investigator FE | YES | YES | YES | YES |
| Observations | 11,465 | 11,465 | 11,465 | 11,465 |
| Number of diseases | 1,471 | 1,471 | 1,471 | 1,471 |
| Mean of the DV | 0.057 | 0.057 | 0.057 | 0.057 |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the gene and disease level. All models include dummies controlling for disease, year, and principal investigator (PI) fixed effects. *Anomaly Index*: a synthetic measure of how unlikely a given gene was to be associated with a disease given the pattern of genes previously associated with that disease up to the year before; *I(Results rated as "New Finding">0)*: 0/1 = 1 if the gene-disease association is rated as a "new finding" by scientists in the Faculty Opinions platform; *I(Results rated as "New Drug Target">0)*: 0/1 = 1 if the gene-disease association is rated as constituting a "new drug target" by scientists in the Faculty Opinions platform; *I(Results rated as "Controversial">0)*: 0/1 = 1 if the gene-disease association is rated as being "controversial" by scientists in the Faculty Opinions platform; *I(Results rated as "New Technique">0)*: 0/1 = 1 if the gene-disease association is rated as employing a "new technique" by scientists in the Faculty Opinions platform. See text for details.

# E PubTator3 Data

This appendix provides additional details on PubTator3 and includes examples illustrating how its AI engine captures the epistemic nature of the gene–disease relationships investigated in scientific articles.

## E.1 Details on PubTator3

PubTator3 is an advanced text-mining tool developed by the NIH's National Center for Biotechnology Information (NCBI) to help researchers extract biomedical concepts from scientific literature (Wei et al., 2024). It integrates natural language processing and state-of-the-art AI techniques to automatically recognize, annotate, and normalize bio-entities (such as genes and diseases) as well as the relationships between them (e.g., a gene causing a disease vs just co-occurring with it). PubTator3 currently provides entity and relation annotations across roughly 36 million PubMed abstracts and 6 million full-text articles from the PubMed Central open-access subset. The tool is freely accessible to the research community at: https://www.ncbi.nlm.nih.gov/research/pubtator3/ (see Wei et al. 2024, for details).

PubTator3 employs a high-performance entity search engine that normalizes different forms of the same biological entity into standardized names, ensuring consistent retrieval of relevant articles regardless of terminology variation. The system uses AIONER, an AI annotation tool, to identify six key bio-entities: genes, diseases, chemicals, variants, species, and cell lines. Each entity type is normalized using specialized natural language processing modules tailored to existing terminology standards, such as GNormPlus for genes (mapped to NCBI Gene identifiers) and TaggerOne for diseases (mapped to MeSH terms). This normalization process ensures accurate and consistent identification across the literature. In addition, PubTator3 integrates BioREx, a transformer-based relation extraction model capable of identifying multiple types of relationships between different bio-entities, including gene–disease interactions. BioREx significantly enhances relation extraction performance, achieving an F-score of 79.6% on standard test sets (Wei et al., 2024).

## E.2 Entity Relationships from PubTator3

PubTator3 extracts a total of twelve possible types of relationships among bio-entities using the transformer-based relation extraction method known as BioREx. For the purposes of this paper, I classify the following PubTator3 relationship types as reflecting interactions of a *causal* nature:

- Gene A *causes* or *stimulates* disease B: when the status of one entity increase (or decrease) as the other increase (or decreases);

- Gene A *inhibits* or *prevents* disease B: when the status of one entity increase (or decrease) as the other decreases (or increase);
- Gene A *treats* or *co-treats* disease B: if a chemical or drug treats a disease, alone or in isolation, through a given gene.

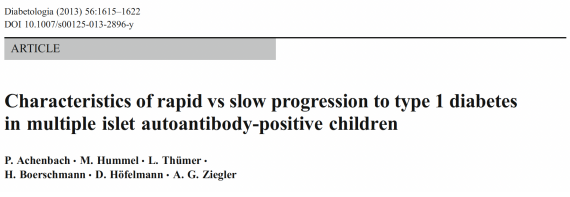
Note that all results remain robust when restricting the analysis to the strict relationship in which Gene A *causes* Disease B. Similarly, I classify the following PubTator3 relationship types as reflecting interactions of a *correlational* nature:

- Gene A *negatively correlate* with disease B: when the status of the two entities tends to be opposite;
- Gene A *positively correlate* with disease B: when the status of one entity tends to increase (or decrease) as the other increase (or decreases);

I use the unique PubMed ID of each paper in DisGeNET to merge it with the corresponding relationships extracted by PubTator3. I retain only those relationships that involve either a causal or correlational link between the relevant genes and diseases. As a result, any paper in DisGeNET can be classified as presenting either causal or correlational relationships, based on how PubTator3 categorizes the gene–disease interactions it contains. Out of the 201,548 unique articles in my dataset, 74,634 include at least one causal relationship, 46,233 include at least one correlational relationship, and 29,432 include both (but involving different gene–disease pairs).

Figure E.1: Example of publications investigating correlational and causal relationships between the gene KIAA0350 and type 1 diabetes.

(a) *Papers on correlational relationship between KIAA0350-T1D*




(b) *Papers on causal relationship between KIAA0350-T1D*


*Note:* Panel (a) reports the abstract from Achenbach et al. (2013). Panel (b) reports the abstract from Soleimanpour et al. (2014). Clec16a is an alternative name for the KIAA0350 gene.

As an example, consider the GWAS by Hakonarson et al. (2007). PubTator3 allows for a more detailed characterization of the follow-on studies investigating the KIAA0350–type 1 diabetes relationship. Panel (a) presents the title of a study by Achenbach et al. (2013), an observational study examining factors that influence

the rate at which children with multiple islet autoantibodies develop type 1 diabetes. Among 1,650 children followed, 23 progressed to diabetes within 3 years (rapid progressors), while 24 remained non-diabetic for over 10 years (slow progressors). The study found that the presence or absence of several mutations, including in the gene KIAA0350, was correlated with the speed of disease progression. Accordingly, this paper is tagged in PubTator3 with the relationship "gene KIAA0350 associates with disease diabetes mellitus".[19]

Panel (b) shows the title of the study by Soleimanpour et al. (2014), which investigates the mechanisms through which KIAA0350 influences the diseases it has been associated with, including type 1 diabetes. The researchers discovered that KIAA0350 encodes a protein that interacts with an enzyme called Nrdp1, protecting it from degradation. This interaction is critical, as Nrdp1 regulates another key protein involved in the clearance of damaged mitochondria. These findings reveal KIAA0350's role in preserving mitochondrial health via the regulation of mitophagy, shedding light on the biological mechanisms through which this gene could emerge as a therapeutic target for diabetes. In PubTator3, this paper is tagged with the relationship "gene KIAA0350 inhibits disease diabetes mellitus." Notably, one of the authors of Soleimanpour et al. (2014) also co-authored the earlier GWAS by Hakonarson et al. (2007).

I use a case–control approach to further validate PubTator3's ability to capture the nature of gene–disease relationships studied in scientific articles. Specifically, one would expect clinical trials to be more likely to investigate causal relationships, given their objective of causally testing the effect of therapies. To test this idea, I draw on data from NIH's iCite, which classifies an article as clinical if it is tagged with MeSH terms such as "Clinical Trial" or "Randomized Controlled Trial." Table E.1 confirms this intuition: articles reporting clinical trial results are significantly more likely to be classified as studying causal relationships in the PubTator3 data, and significantly less likely to present correlational evidence. This provides additional support for the validity of the PubTator3 classification.

[19] Interestingly, Achenbach et al. (2013) does not cite Hakonarson et al. (2007), even if it investigates the KIAA0350-type 1 diabetes first introduced by it. This is one example of how my landscape-based approach can better capture the impact of combinations by not relying on paper-to-paper citations (Arts et al., 2025).

Table E.1: Clinical articles are more likely to explore causal relationships between a gene and a disease, and less likely to explore correlational relationships.

| | Causal relationship between gene and disease (0/1) | | Correlational relationship between gene and disease (0/1) | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Clinical article (0/1) | 0.0684*** | 0.0477** | -0.0443*** | -0.0145 |
| | (0.0129) | (0.0150) | (0.00999) | (0.0110) |
| Disease FE | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES |
| Observations | 369,305 | 347,776 | 369,305 | 347,776 |
| Number of diseases | 9,864 | 9,494 | 9,864 | 9,494 |
| Mean of the DV | 0.386 | 0.386 | 0.246 | 0.246 |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the gene and disease level. All models include dummies controlling for disease and year fixed effects; Columns (2) and (4) also include dummies controlling for the principal investigator (PI) of the articles introducing a GDA. *Causal relationship between gene and disease (0/1)*: 0/1 = 1 if the article introduces at least one causal relationship about a gene-disease pair according to the AI engine of PubTator3 (Wei et al., 2024); *Correlational relationship between gene and disease (0/1)*: 0/1 = 1 if the article introduces at least one correlational relationship about a gene-disease pair according to the AI engine of PubTator3 (Wei et al., 2024); *Clinical article (0/1)*: 0/1 = 1 for new gene-disease associations introduced by an article reporting results of a clinical trial (data on clinical articles from NIH's iCite). See text for details.

# F Details on the Instrumental Variable Strategy

## F.1 Endogeneity Concerns

The main findings indicate that a GWAS approach is associated with systematically different outcomes than pursuing candidate gene studies. However, since GWAS adoption is a researcher's choice, it may be correlated with unobserved determinants of discovery outcomes. Principal investigator (PI) fixed effects strengthen identification by comparing outcomes for the same scientist before and after switching search strategies, but bias could persist if researchers with higher or lower latent abilities to generate breakthroughs are more likely to adopt GWAS. Yet, the sign of this bias is unclear ex ante. If scientists adopt GWAS after struggling to discover valuable gene-disease associations through theory-driven approaches, estimates would be biased downward. If, instead, the most capable scientists are early adopters, estimates would be biased upward. Appendix Figure G.2 helps address this concern by showing that, prior to their first GWAS paper, eventual adopters do not systematically introduce gene-disease combinations of different quality.

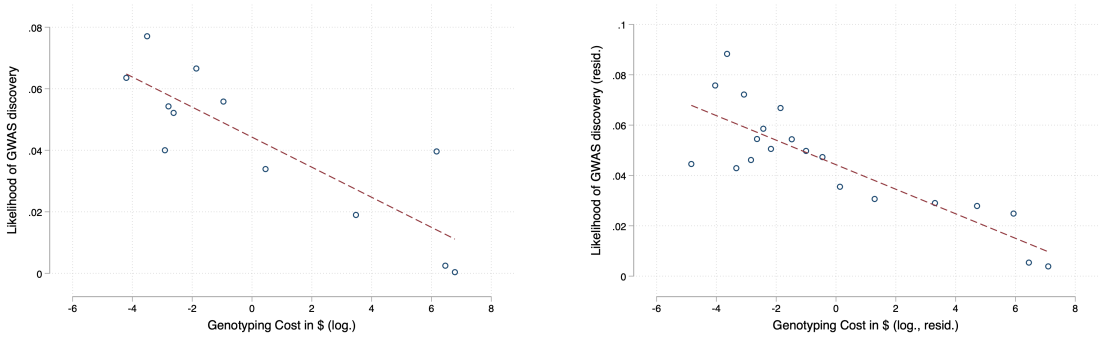
A more serious concern regarding identification is that the timing of GWAS adoption may be influenced by unobserved, time-varying factors that also impact outcomes. At the PI level, shifts in research resources or scientific standing could increase the likelihood of adopting GWAS while simultaneously shaping the quality and impact of subsequent discoveries. At the disease level, adoption incentives may depend on the state of prior knowledge, since diseases with well-developed theory may offer less apparent value from data-driven search. Even with disease and disease-by-year fixed effects, OLS estimates may still be exposed to disease-specific shocks that jointly influence the adoption of a given search strategy and its outcomes. To address these concerns, the remainder of this appendix introduces and validates an instrumental variable strategy (IV) that isolates plausibly exogenous variation in GWAS adoption and provides a complementary test of the baseline estimates.

## F.2 Exploiting the Decrease in Genotyping Costs

GWAS are carried out using genotyping microarrays (or chips), that is, physical slides with millions of synthetic DNA probes that match with specific known genetic variations (SNPs) in the human genome. When GWAS first emerged around 2005, the cost of conducting a single study was estimated to be on the order of $10 million or more (Visscher et al., 2012). While GWAS were enabled by the broader post-Human Genome Project decline in genetic data costs, genome-wide genotyping became markedly more affordable in the mid-2000s with the diffusion of cheaper microarrays (Panel (a) of Figure 3). I exploit this technology-driven decline as a supply-side shock to the feasibility of GWAS. Specifically, I utilize NIH data

on the average cost of collecting data on one megabase (i.e., one million bases) of DNA in any given year. Consistent with this logic, Figure F.1 shows that the share of new gene–disease associations discovered in year $t$ that are identified via GWAS declines as genotyping costs in year $t$ increase. Lower costs thus generate plausibly exogenous increases in the likelihood of adopting GWAS, which provides the first ingredient of the IV strategy.

Figure F.1: Lower genotyping costs predict greater use of GWAS in new discoveries.




*(a) Raw correlation between genotyping costs and GWAS use*

*(b) Conditional correlation between genotyping costs and GWAS use*

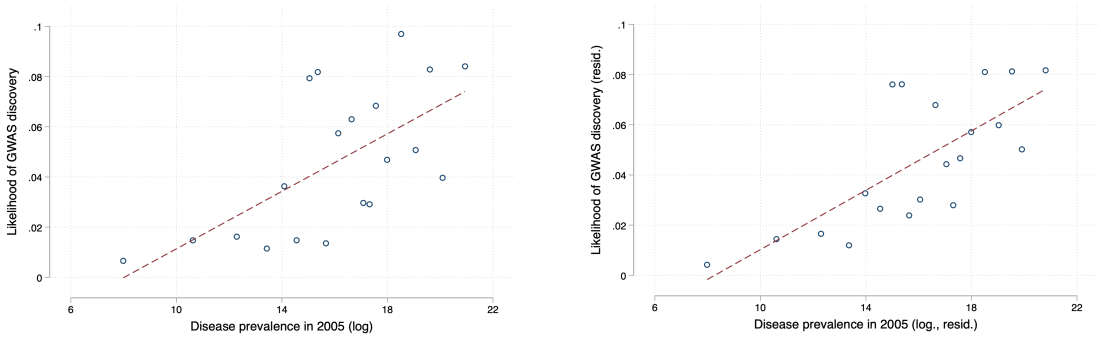
*Note:* The figure plots binned correlations between the genotyping cost in a given year and the likelihood that a gene disease association first reported in that year is introduced via GWAS. Panel (a) reports the unconditional relationship. Panel (b) reports the relationship after residualizing the outcome for disease fixed effects.

While the reduction in genotyping costs was a common trend, the diffusion of GWAS was uneven across diseases. In particular, more common diseases experienced a larger increase in GWAS use (Panel b of Figure 3). Two forces help explain this heterogeneity. On the demand side, funding agencies explicitly prioritized "common complex diseases" with high public health salience (Heeney, 2021). For example, the Wellcome Trust targeted the first GWAS to "seven diseases of major public health importance," such as diabetes and coronary heart disease (WTCCC, 2007). On the supply side, GWAS required assembling large case-control samples, making recruitment more feasible when patient pools were abundant. The Wellcome Trust noted that its early focus was shaped by the fact that "suitable nationally representative sample sets were available" (WTCCC, 2007), which was obviously easier for diseases with higher prevalence in the population. Together, these forces imply that the same cost decline translated into larger effective reductions in the barriers to GWAS for common diseases, a fact exploited by my IV strategy.

To capture how the same cost decline translated into different adoption responses across diseases, I construct a measure of disease prevalence in the pre-GWAS period. Prevalence is the number of individuals living with a given disease at a point in time, reflecting the overall burden of the condition. In the context of my

identification strategy, it proxies both for the prioritization of more important diseases (a pull factor) and for the ease of assembling the large samples required for GWAS (a push factor). I measure prevalence as of 2005, immediately before GWAS diffusion, so it is predetermined with respect to subsequent GWAS adoption and discovery outcomes. I obtain prevalence information from two sources. First, I use the Unified Medical Language System (UMLS) Metathesaurus, drawing on the MRCONSO.RRF file, which provides harmonized data from the main medical vocabularies. Second, I use Orphanet, which reports disease frequency for rare conditions. In total, prevalence information is available for 2,511 diseases, covering 89,368 gene-disease associations discovered after 2005, of which 4,064 are introduced via GWAS.[20] Using these data, I assign each disease a pre-GWAS prevalence measure and hold it fixed at its 2005 value throughout the IV analysis. Figure F.2 shows that GWAS use is higher in disease areas with greater prevalence, even after controlling for year fixed effects. This provides the second ingredient of the IV strategy by capturing cross-disease differences in the feasibility and incentives to adopt GWAS.

Figure F.2: Higher disease prevalence predicts greater use of GWAS in new discoveries.




*(a) Raw correlation between 2005 disease prevalence and GWAS use*

*(b) Conditional correlation between 2005 disease prevalence and GWAS use*

*Note:* The figure plots binned correlations between disease prevalence measured in 2005 and the likelihood that a gene disease association first reported in year $t$ is introduced via GWAS. Panel (a) reports the unconditional relationship. Panel (b) reports the relationship after residualizing the outcome for year fixed effects.

## F.3 Construction of the IV

To build the instrument, I combine the two elements discussed above to capture the idea that the exogenous decrease in genotyping costs had heterogeneous effects across diseases with different baseline prevalence.

[20] However, the estimation sample in Table 4 includes only 1,724 diseases and 88,581 gene–disease associations because the fixed effects absorb diseases with a single observed discovery, which therefore drop out of the estimation.

Following Borusyak et al. (2025), this yields a simplified shift share structure:

$$\text{Genotyping cost shift-share}_{j,t} = \underbrace{\text{Genotyping Cost}_t}_{\text{Shift}} \times \underbrace{\text{Disease Prevalence}_{j,2005}}_{\text{Share}}, \tag{F.1}$$

where the "shift", $Genotyping\ Cost_t$, is a common time series capturing the decline in the cost of collecting genetic data, and the "share", $Disease\ Prevalence_{j,2005}$, is a predetermined measure of disease $j$'s prevalence in the population. Because there is a common time-varying shock, the baseline prevalence can be interpreted as a disease-level exposure, capturing how strongly each disease is predicted to respond to the same exogenous change in genotyping costs. The identifying variation, therefore, comes from differential adoption responses across diseases with different baseline prevalence as costs fall.

The first stage is estimated at the discovery level and relates the probability that discovery $i$ for disease $j$ in year $t$ is introduced via GWAS to the genotyping cost shift-share instrument, controlling for the vector of baseline covariates in $\mathbf{Z}_{i,j,t}$ and fixed effects described in the main text:

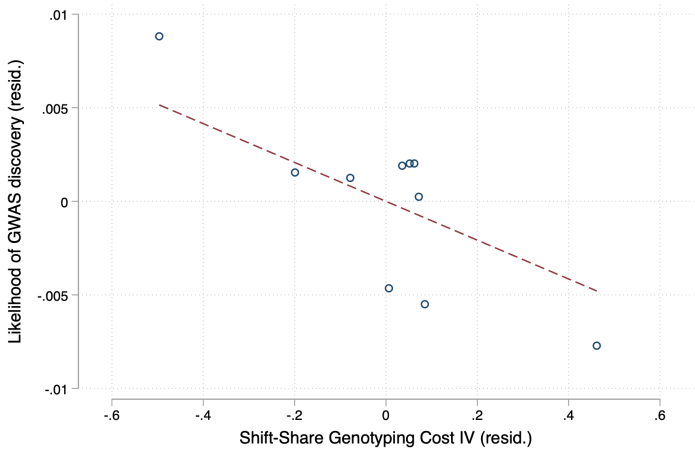
$$\text{GWAS (0/1)}_{i,j,t} = \pi_0 + \pi_1 \text{ Genotyping cost shift-share}_{j,t} + \mathbf{Z}'_{i,j,t}\pi + u_{i,j,t}. \tag{F.2}$$

The instrument varies at the disease-year level, since it depends on the prevalence of disease $j$ measured in 2005 and genotyping costs in year $t$, thereby shifting GWAS adoption through differential exposure to an exogenous cost shock. Empirical estimates show that the first stage is strong: higher values of the instrument, which correspond to years with higher genotyping costs, are associated with a lower likelihood that a discovery is introduced via GWAS ($\hat{\pi}_1 = -0.0104$, s.e. $= 0.00235$, $p = 0.004$). Figure F.3 visualizes the first stage relationship. The instrument is highly relevant, with Cragg Donald Wald $F = 53.6$ and Kleibergen-Paap rk Wald $F = 19.5$.

## F.4 Validation of the IV

In this section, I follow recent econometric work on shift-share instruments to probe the validity of the proposed IV (Goldsmith-Pinkham et al., 2020; Borusyak et al., 2025). Here, I focus on the suitability of the shares, namely disease prevalence measured in 2005 (before GWAS emergence), as a disease-level exposure measure to the common exogenous shock generated by changes in genotyping costs. Identification relies on a "parallel trends type" assumption with respect to the shares: absent the cost decline, high-prevalence and low-prevalence diseases would have exhibited similar trajectories in GWAS adoption and in the discovery outcomes studied, conditional on the fixed effects and controls included in the empirical specification. Below, I follow the checklist proposed by Borusyak et al. (2025) to assess these conditions in my context.

Figure F.3: First stage: higher values of the shift-share genotyping costs IV predict lower GWAS adoption.

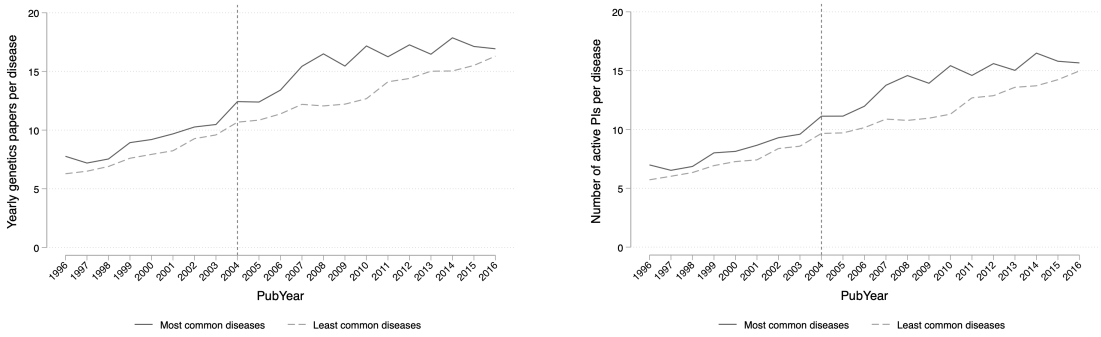



*Note:* The figure plots a binned relationship between the genotyping cost shift-share instrument and the likelihood that a gene disease association first reported in year $t$ is introduced via GWAS. The estimates are residualized by disease, year, and PI fixed effects.

**1. Conceptual suitability:** Disease prevalence measured in 2005 is a suitable exposure share because it is predetermined relative to GWAS diffusion and it captures a disease's differential sensitivity to the common decline in genotyping costs. When genome-wide data becomes cheaper, the adoption response should be stronger precisely where GWAS is most feasible and most valuable: diseases with larger potential case pools for assembling high-powered samples and greater public health salience that attracts early funding (Figure F.2). Figure F.4 provides additional descriptive evidence on this interpretation by plotting long-run trends in scientific activity for diseases with higher versus lower baseline prevalence. Reassuringly, these series display no evidence of differential pre-trends that would be consistent with confounds (such as pre-existing funding changes). The divergence in scientific activity emerges only when GWAS emerge.

**2. Unit-level controls:** The shift-share instrument varies only through the interaction between baseline prevalence and annual genotyping costs. The empirical specifications include two sets of unit-level controls that isolate this interaction from obvious confounds. Year fixed effects absorb any shocks common to all diseases in a given year, including broad changes in the funding environment. Disease fixed effects absorb all time-invariant differences across diseases, such as baseline research intensity or intrinsic difficulty. With these fixed effects, the instrument does not compare high- versus low-prevalence diseases in levels. Instead, identification comes from comparing how outcomes within the same disease evolve as genotyping costs fall, with the predicted adoption response scaled by that disease's prevalence in 2005.

Figure F.4: Scientific activity evolves similarly across high and low prevalence diseases prior to GWAS diffusion.




*(a) Publications on a disease*

*(b) PIs researching a disease*

*Note:* The figure plots descriptive trends in scientific activity for diseases above versus below the median prevalence measured in 2005. Panel (a) reports the yearly number of publications in each group. Panel (b) reports the yearly number of active principal investigators in each group.

**3. Characterizing which shares matter for identification:** With a single exposure share and a single aggregate shock, the shift-share design does not admit the standard Rotemberg weights decomposition (Goldsmith-Pinkham et al., 2020). The relevant diagnostic in this setting is therefore to characterize which diseases contribute most to the identifying covariances in the first stage and reduced form. Let the instrument be $G_{j,t} = Prevalence_{j,2005} \times GenotypingCost_t$ and the endogenous regressor be $D_{i,j,t} = \text{GWAS}_{i,j,t}$. Let $\widetilde{G}_{j,t}$ and $\widetilde{D}_{i,j,t}$ denote the residuals of $G_{j,t}$ and $D_{i,j,t}$ after partialling out the controls and fixed effects used in the main specification. In this context, instrument relevance is governed by the first-stage covariance $\sum_{i,j,t} \widetilde{G}_{j,t}\widetilde{D}_{i,j,t}$, which admits an exact decomposition by disease. Define each disease's contribution as $C_j = \sum_{i,t,j} \widetilde{G}_{j,t}\, \widetilde{D}_{i,j,t}$ and $\sum_j C_j = \sum_{i,j,t} \widetilde{G}_{j,t}\, \widetilde{D}_{i,j,t}$ and the corresponding share as $w_j = \frac{C_j}{\sum_{j'} C_{j'}}$. The shares $w_j$ sum to one and measure how much each disease contributes to the first-stage covariance. Column (2) of Table F.1 reports these quantities. The distribution is not concentrated in a small set of diseases: the top five diseases account for approximately 35% of the total covariance share, substantially less than in several applications discussed by Borusyak et al. (2025). A complementary sensitivity analysis implements a "leave-one-disease-out" procedure. For each disease $j$, I re-estimate the 2SLS specification for the main outcomes after dropping all discoveries associated with $j$ and report the resulting coefficient $\hat{\beta}_{-j}$. Substantial changes in $\hat{\beta}_{-j}$ relative to the full-sample estimate would indicate that the results are driven by a small set of diseases. Reassuringly, Columns (3) through (6) of Table F.1 show that omitting any single disease yields estimates that are very close to the baseline IV coefficients, and when differences arise, they tend to increase the magnitude of the estimated effects, implying that the baseline estimates are conservative.

Table F.1: Robustness check: diseases that contribute most to IV identification and sensitivity to their exclusion.

| ID | Disease Name | GDAs (1) | Weight (2) | $\hat{\beta}_{-j}^{(IV,U)}$ (3) | $\hat{\beta}_{-j}^{(IV,D)}$ (4) | $\hat{\beta}_{-j}^{(IV,S)}$ (5) | $\hat{\beta}_{-j}^{(IV,F)}$ (6) |
|---|---|---|---|---|---|---|---|
| 2774 | Migraine Disorders | 191 | 0.110 | 0.504 | 0.359 | 0.187 | 3.198 |
| 202 | Asthma | 928 | 0.088 | 0.468 | 0.258 | 0.217 | 4.752 |
| 586 | Dental caries | 75 | 0.080 | 0.522 | 0.224 | 0.163 | 1.979 |
| 231 | Barrett Esophagus | 263 | 0.043 | 0.527 | 0.279 | 0.118 | 1.892 |
| 5676 | Common Migraine | 48 | 0.034 | 0.502 | 0.220 | 0.142 | 3.087 |
| 604 | Atopic Dermatitis | 366 | 0.025 | 0.481 | 0.250 | 0.154 | 4.831 |
| 1425 | Systemic Lupus Erythematosus | 836 | 0.022 | 0.460 | 0.246 | 0.147 | 2.685 |
| 2149 | Systemic Scleroderma | 373 | 0.015 | 0.474 | 0.268 | 0.141 | 2.673 |
| 998 | Sensorineural Hearing Loss | 119 | 0.015 | 0.365 | 0.246 | 0.160 | 2.358 |
| 619 | Type 1 Diabetes | 634 | 0.015 | 0.467 | 0.263 | 0.163 | 3.249 |

*Note:* This table reports diagnostics for the genotyping costs shift-share IV. Column (1) reports the number of post-2005 gene-disease associations involving disease $j$. Column (2) reports the disease contribution weight $w_j$ computed from the residualized first stage, as explained in the text above. Columns (3)–(6) report leave-one-disease-out 2SLS coefficients, $\hat{\beta}_{-j}^{IV}$, for the main outcomes: introduction of GDAs including a gene never studied before (U) or recently discovered (D), DisGeNET score of teh GDA (S), and follow-on publications on the GDA (F). The full-sample 2SLS coefficients are reported in the corresponding main IV tables.

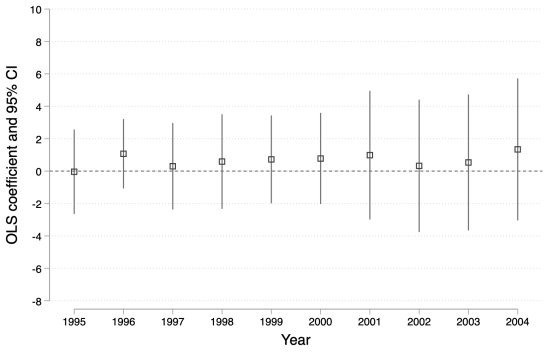
**4. Balance tests of the shares:** Although the exclusion restriction cannot be tested directly, recent work on shift-share designs emphasizes balance tests as a plausibility check (Borusyak et al., 2025). The logic is that, if baseline prevalence is a valid exposure share, then before GWAS emerges and genotyping costs decline, diseases with higher versus lower prevalence should not display systematically different trends in the outcomes studied. In this setting, this corresponds to the parallel-trends component of the identifying assumption: absent the cost decline, discovery outcomes in high-prevalence and low-prevalence diseases would have evolved similarly, conditional on the fixed effects and controls in the empirical specification. Figure F.5 implements this diagnostic by comparing pre-2005 trends in the main dependent variables for diseases in the top versus bottom quartile of prevalence measured in 2005. Across outcomes, there is no evidence of differential pre-trends prior to GWAS diffusion. This pattern supports the interpretation of baseline prevalence as an appropriate exposure share for the common cost shock.

Overall, the evidence supports the validity of the IV strategy. The instrument is grounded in a technology-driven cost shock that predicts GWAS adoption, and its heterogeneity is anchored in a predetermined measure of baseline prevalence that captures differential exposure to cost reductions. Consistent with the identifying logic, the balance and sensitivity exercises do not reveal patterns that would suggest systematic confounding, strengthening the interpretation of the 2SLS estimates as reflecting variation in search strategy induced by

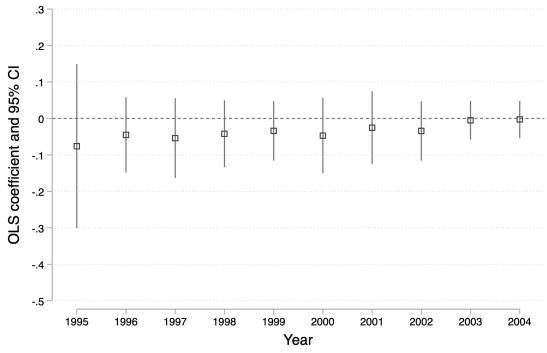
the decline in genotyping costs.

Figure F.5: No evidence of pre-trends in discovery outcomes for high- versus low-prevalence diseases.

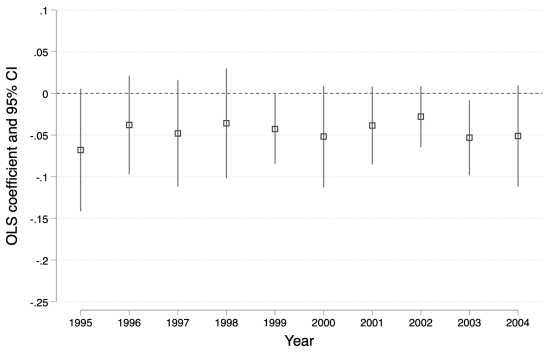
*(a) Amount of past work on the genes discovered*



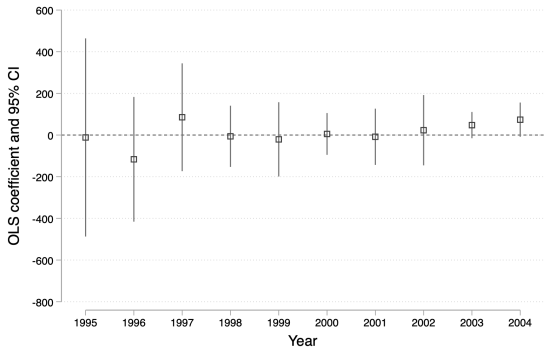

*(b) Anomaly of discoveries*




*(c) DisGeNET score of discoveries*



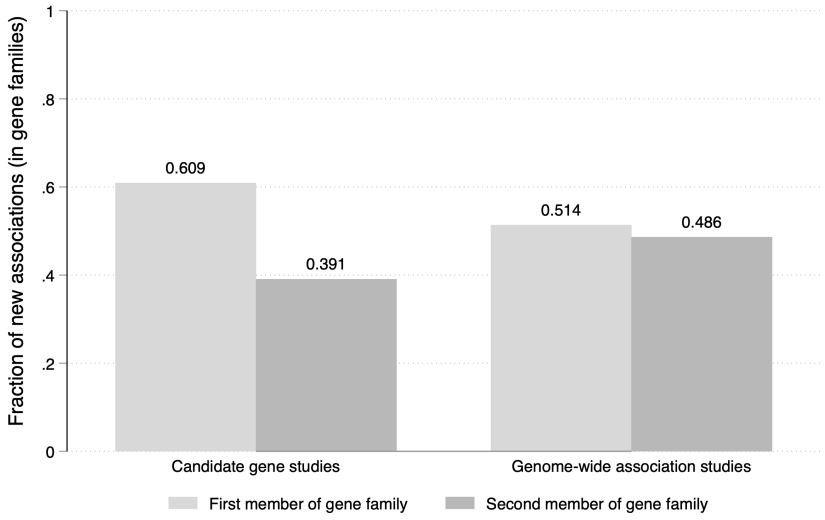

*(d) Follow-on research on discoveries*




*Note:* The figure plots average outcomes by calendar year separately for diseases in the top versus bottom quartile of prevalence measured in 2005. The pre-2005 period provides a balance-style diagnostic for the share-based identifying assumption: absent the decline in genotyping costs and the diffusion of GWAS, outcomes for high- and low-prevalence diseases should display similar trends, conditional on controls.

# G Additional Figures and Tables

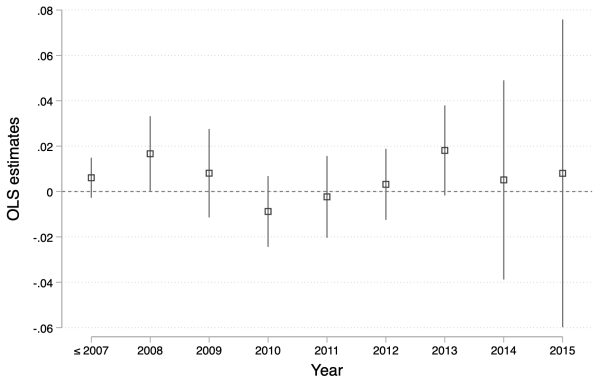
Figure G.1: GWAS are not biased toward the first-discovered member of gene families, unlike candidate gene studies.



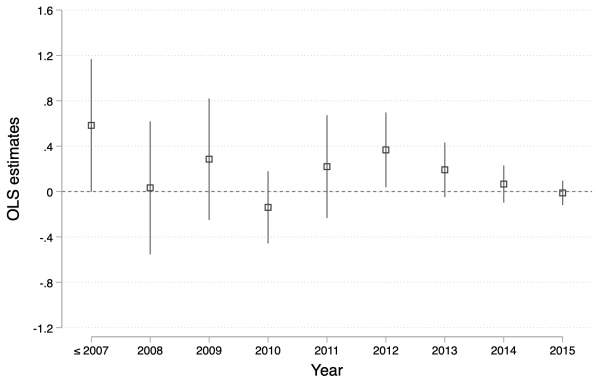

*Note:* The figure shows the share of new gene-disease associations involving the first versus the second member of a gene family, separately by the method used in the paper introducing the association. The corresponding regression estimates are reported in Appendix Table G.2.

Figure G.2: Robustness check: Before their first GWAS, principal investigators that adopt the genome-wide approach are not systematically more likely to introduce gene-disease associations of higher value or that receive more follow-on work.

*(a) Pre-GWAS average DisGeNET Score of discoveries by future GWAS adopters*



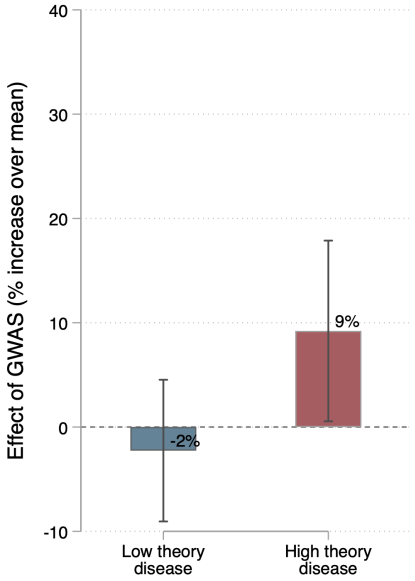

*(b) Pre-GWAS average follow-on studies for discoveries by future GWAS adopters*



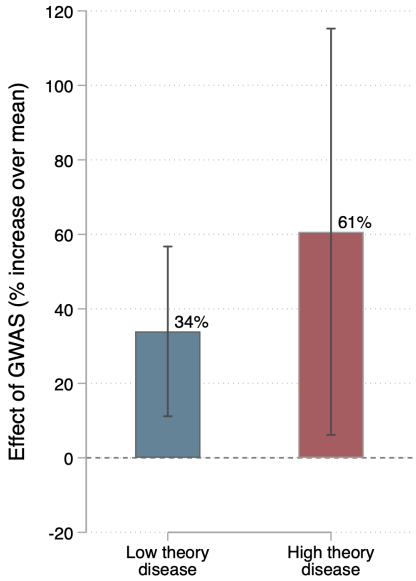

*Note:* This figure investigates if there are systematic differences in the scientific quality of gene-disease combinations introduced by scientists that self-select into publishing GWAS. Every coefficient is estimated from a separate regression for a given year, where I compare principal investigators who never publish a GWAS with principal investigators who will eventually publish one but have not done so already (that is, PIs are excluded from the sample after their first GWAS). In practice, each $\beta$ coefficient is estimated from the following specification: $(Feature\ of\ gene\text{-}disease\ combinations\ in\ year\ t)_{i,j} = \alpha + \beta(PI\ will\ publish\ GWAS)_{i,j,t} + \mathbf{Z}'_{i,j}\gamma + \epsilon_{i,j,t}$. Panel (a) plots the coefficients and confidence intervals from regressing the average DisGeNET score of gene-disease pairs introduced in a given year over a dummy for whether the principal investigator of the study will eventually publish at least one GWAS in my sample period. Panel (b) plots the coefficients and confidence intervals from regressing the average count of follow on papers received by gene-disease pairs introduced in a given year over a dummy for whether the principal investigator of the study will eventually publish at least one GWAS in my sample period. See text for details.

Figure G.3: Heterogeneity in GWAS discovery quality across diseases with more versus less pre-existing genetic knowledge.

*(ii) I(GDA in bottom 50% DisGeNET Score $>$ 0)*

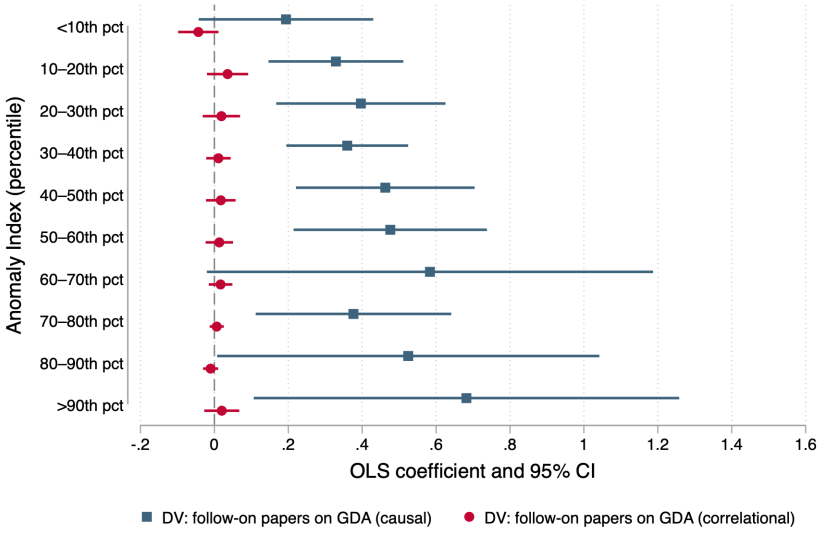



*(iii) I(GDA in top 10% DisGeNET Score $>$ 0)*




*Note:* The figure illustrates how the effect of GWAS on discovery quality varies with the depth of pre-existing genetic knowledge in a disease area. Each bar plots the coefficient on the *GWAS (0/1)* indicator from an OLS regression estimated separately in two subsamples. The subsamples split discoveries by whether the associated disease falls in a low-knowledge (below-median) or high-knowledge (above-median) disease area, measured by the number of genes linked to the disease prior to 2005. The dependent variables are indicators for whether a newly introduced gene–disease association falls in the bottom half or the top decile of the DisGeNET Score distribution. The unit of observation is a gene–disease association, and all specifications include year and disease fixed effects. Standard errors are two-way clustered by gene and disease. See text for details.

Figure G.4: More anomalous discoveries attract disproportionately more causal follow-on research.




*Note:* The figure relates the anomalousness of a newly introduced gene-disease association to the amount and type of follow-on research it attracts. New associations are sorted into deciles of the Anomaly Index measured at the time of discovery. For each decile, the outcome is the number of subsequent papers investigating the same gene-disease pair, reported separately for follow-on work classified as causal and for follow-on work classified as correlational in nature using the PubTator3 data. See text for details.

Table G.1: Robustness check: GWAS introduce associations involving genes that received fewer publications or were discovered later, relative to candidate gene studies.

| | Publications on the gene (pre-2005) | | Years since the discovery of the gene (as of 2005) | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| GWAS (0/1) | -50.84*** | -29.44*** | -2.978*** | -2.126*** |
| | (5.789) | (6.121) | (0.206) | (0.241) |
| Disease FE | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES |
| Observations | 369,302 | 348,921 | 364,661 | 344,416 |
| Number of diseases | 9,863 | 9,504 | 9,786 | 9,422 |
| Mean of the DV | 100.57 | 100.57 | 12.81 | 12.81 |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the gene and disease level. All models include dummies controlling for disease and year fixed effects; Columns (2) and (4) also include dummies controlling for the principal investigator (PI) of the articles introducing a GDA. *Publications on the gene (pre-2005)*: count of articles published before 2005 that study the gene in the newly reported gene–disease association; *Years since the discovery of the gene (as of 2005)*: years between the gene's discovery year and 2005 for the gene in the newly reported association; *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. See text for details.

Table G.2: Mechanisms: GWAS are more likely to associate the second member of gene families relative to candidate gene studies.

| | I(GDA with the second member of a gene family>0) | |
|---|---|---|
| Subsample: | Gene family members | |
| GWAS (0/1) | 0.0873*** | 0.0394† |
| | (0.0125) | (0.0238) |
| Disease FE | YES | YES |
| Year of discovery FE | YES | YES |
| Principal investigator FE | NO | YES |
| Observations | 91,643 | 79,359 |
| Number of diseases | 5,234 | 4,730 |
| Mean of the DV | 0.393 | 0.393 |

*Note:* †, *, **,*** denote significance at the 10%, 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the disease level. All models include dummies controlling for disease and year fixed effects; Column (2) also includes dummies controlling for the principal investigator (PI) of the articles introducing a GDA. The sample is limited to gene-disease associations involving gene family members (data on gene families from Stoeger et al. 2018). *I(GDA with the second member of a gene family>0)*: 0/1 = 1 if the new gene-disease association involves a gene that is the second member of a gene family; *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. See text for details.

Table G.3: Robustness check: GWAS introduce more gene-disease associations in the bottom tail of scientific quality relative to candidate gene studies.

| GDA in... | Bottom 5% DisGeNET Score | | Bottom 10% DisGeNET Score | | Bottom 20% DisGeNET Score | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| GWAS (0/1) | 0.0111*** | 0.0108* | 0.0199*** | 0.0128† | 0.0336*** | 0.0225* |
| | (0.00300) | (0.00440) | (0.00467) | (0.00661) | (0.00744) | (0.0103) |
| Disease FE | YES | YES | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES | NO | YES |
| Observations | 369,302 | 348,921 | 369,302 | 348,921 | 369,302 | 348,921 |
| Number of diseases | 9,863 | 9,504 | 9,863 | 9,504 | 9,863 | 9,504 |
| Mean of the DV | 0.050 | 0.050 | 0.100 | 0.100 | 0.200 | 0.200 |

*Note:* †, *, **,*** denote significance at the 10%, 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the gene and disease level. All models include dummies controlling for disease and year, while columns (2), (4) and (6) also inclide principal investigator (PI) fixed effects. *GDA in bottom 5% DisGeNET Score*: 0/1 = 1 if the gene-disease association has a DisGeNET Score below the 5$^{th}$ percentile of the sample; *GDA in bottom 10% DisGeNET Score*: 0/1 = 1 if the gene-disease association has a DisGeNET Score below the 10$^{th}$ percentile of the sample; *GDA in bottom 20% DisGeNET Score*: 0/1 = 1 if the gene-disease association has a DisGeNET Score below the 20$^{th}$ percentile of the sample; *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. See text for details.

Table G.4: Robustness check: GWAS introduce more gene-disease associations in the top tail of scientific quality relative to candidate gene studies.

| GDA in... | Top 5% DisGeNET Score | | Top 10% DisGeNET Score | | Top 20% DisGeNET Score | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| GWAS (0/1) | 0.0398** | 0.0292* | 0.0411** | 0.0355* | 0.00371* | -0.000461 |
| | (0.0121) | (0.0126) | (0.0140) | (0.0149) | (0.00171) | (0.00260) |
| Disease FE | YES | YES | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES | NO | YES |
| Observations | 369,302 | 348,921 | 369,302 | 348,921 | 369,302 | 348,921 |
| Number of diseases | 9,863 | 9,504 | 9,863 | 9,504 | 9,863 | 9,504 |
| Mean of the DV | 0.050 | 0.050 | 0.100 | 0.100 | 0.200 | 0.200 |

*Note:* †, *, **,*** denote significance at the 10%, 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the disease level. All models include dummies controlling for disease and year, while columns (2), (4) and (6) also inclide principal investigator (PI) fixed effects. *GDA in top 5% DisGeNET Score*: 0/1 = 1 if the gene-disease association has a DisGeNET Score above the 95$^{th}$ percentile of the sample; *GDA in top 10% DisGeNET Score*: 0/1 = 1 if the gene-disease association has a DisGeNET Score above the 90$^{th}$ percentile of the sample; *GDA in top 20% DisGeNET Score*: 0/1 = 1 if the gene-disease association has a DisGeNET Score above the 80$^{th}$ percentile of the sample; *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. See text for details.

Table G.5: Robustness check: Results are robust to the inclusion of more stringent fixed effect structures.

**Panel A: Including Disease × Year Fixed Effects**

| | DisGeNET Score of the GDA (1) | I(GDA in bottom 50% DisGeNET Score>0) (2) | I(GDA in top 10% DisGeNET Score>0) (3) | Follow-on articles on the GDA (all) (4) | Follow-on articles on the GDA (causal) (5) | Follow-on articles on the GDA (correlational) (6) |
|---|---|---|---|---|---|---|
| GWAS (0/1) | 0.0174*** | 0.0306† | 0.0477** | 0.743*** | 0.407** | 0.0290 |
| | (0.00465) | (0.0175) | (0.0145) | (0.219) | (0.127) | (0.0177) |
| Disease × year FE | YES | YES | YES | YES | YES | YES |
| Principal investigator FE | YES | YES | YES | YES | YES | YES |
| Observations | 330,378 | 330,378 | 330,378 | 330,378 | 330,378 | 330,378 |
| Number of diseases | 7,891 | 7,891 | 7,891 | 7,891 | 7,891 | 7,891 |
| Mean of the DV | 0.132 | 0.500 | 0.094 | 0.595 | 0.212 | 0.136 |

**Panel B: Including Disease × Year and Disease × PI Fixed Effects**

| | DisGeNET Score of the GDA (1) | I(GDA in bottom 50% DisGeNET Score>0) (2) | I(GDA in top 10% DisGeNET Score>0) (3) | Follow-on articles on the GDA (all) (4) | Follow-on articles on the GDA (causal) (5) | Follow-on articles on the GDA (correlational) (6) |
|---|---|---|---|---|---|---|
| GWAS (0/1) | 0.0215* | 0.0269 | 0.0523* | 1.247*** | 0.708** | 0.0450 |
| | (0.00837) | (0.0278) | (0.0255) | (0.353) | (0.217) | (0.0221) |
| Disease × year FE | YES | YES | YES | YES | YES | YES |
| Disease × PI FE | YES | YES | YES | YES | YES | YES |
| Observations | 206,256 | 206,256 | 206,256 | 206,256 | 206,256 | 206,256 |
| Number of diseases | 7,321 | 7,321 | 7,321 | 7,321 | 7,321 | 7,321 |
| Mean of the DV | 0.126 | 0.513 | 0.104 | 0.557 | 0.201 | 0.130 |

*Note:* †, *, **,*** denote significance at the 10%, 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the gene and disease level. Panel A includes disease × year fixed effects, while Panel B includes disease × year and disease × PI fixed effects. *DisGeNET Score of the GDA*: synthetic measure of scientific value of the discovery provided by DisGeNET; *I(GDA in bottom 50% DisGeNET Score>0)*: 0/1 = 1 if the gene-disease association has a DisGeNET Score below the sample median; *I(GDA in top 10% DisGeNET Score>0)*: 0/1 = 1 if the gene-disease association has a DisGeNET Score in the top sample decile; *Follow-on articles on the GDA (all)*: count of all subsequent studies investigating the gene-disease association; *Follow-on articles on the GDA (causal)*: count of subsequent studies investigating causal relationships between the gene and the disease; *Follow-on articles on the GDA (correlational)*: count of subsequent studies investigating correlational relationships between the gene and the disease; *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. Follow-on articles are papers directly working on the gene-disease combinations, regardless of whether they cite the study that first introduced it. See text for details.

Table G.6: Robustness check: Results are robust to the inclusion of gene fixed effects.

**Including Gene Fixed Effects**

| | DisGeNET Score of the GDA (1) | I(GDA in bottom 50% DisGeNET Score>0) (2) | I(GDA in top 10% DisGeNET Score>0) (3) | Follow-on articles on the GDA (all) (4) | Follow-on articles on the GDA (causal) (5) | Follow-on articles on the GDA (correlational) (6) |
|---|---|---|---|---|---|---|
| GWAS (0/1) | 0.0206*** | -0.0255 | 0.0560*** | 0.978*** | 0.497*** | 0.108*** |
| | (0.00415) | (0.0148) | (0.0126) | (0.240) | (0.133) | (0.0208) |
| Disease FE | YES | YES | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES | YES | YES |
| Principal investigator FE | YES | YES | YES | YES | YES | YES |
| Gene FE | YES | YES | YES | YES | YES | YES |
| Observations | 347,366 | 347,366 | 347,366 | 347,366 | 347,366 | 347,366 |
| Number of diseases | 9,491 | 9,491 | 9,491 | 9,491 | 9,491 | 9,491 |
| Mean of the DV | 0.127 | 0.501 | 0.097 | 0.591 | 0.209 | 0.130 |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the disease level. *DisGeNET Score of the GDA*: synthetic measure of scientific value of the discovery provided by DisGeNET; *I(GDA in bottom 50% DisGeNET Score>0)*: 0/1 = 1 if the gene-disease association has a DisGeNET Score below the sample median; *I(GDA in top 10% DisGeNET Score>0)*: 0/1 = 1 if the gene-disease association has a DisGeNET Score in the top sample decile; *Follow-on articles on the GDA (all)*: count of all subsequent studies investigating the gene-disease association; *Follow-on articles on the GDA (causal)*: count of subsequent studies investigating causal relationships between the gene and the disease; *Follow-on articles on the GDA (correlational)*: count of subsequent studies investigating correlational relationships between the gene and the disease; *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. Follow-on articles are papers directly working on the gene-disease combinations, regardless of whether they cite the study that first introduced it. See text for details.

Table G.7: Instrumental variable estimates: GWAS broaden the novelty and anomalousness of discovery relative to candidate gene studies.

| | First Stage | Second Stage (2SLS) | | |
|---|---|---|---|---|
| | GWAS (0/1) | I(GDA with recently associated gene>0) | I(GDA with recently discovered gene>0) | Anomaly Index |
| | (1) | (2) | (3) | (4) |
| IV: Genotyping cost shift-share | -0.0104** | | | |
| | (0.00235) | | | |
| GWAS (0/1) | | 0.472*** | 0.257*** | 0.355* |
| | | (0.116) | (0.0625) | (0.157) |
| **F-Statistic (First Stage)** | 19.5 | | | |
| Disease FE | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES |
| Observations | 88,581 | 88,581 | 88,581 | 88,032 |
| Number of diseases | 1,724 | 1,724 | 1,724 | 1,571 |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the gene and disease class level. All models include dummies controlling for disease and year fixed effects. Note that the number of observations is smaller because some diseases lack information on their prevalence in the pre-GWAS period. The first stage regresses the *GWAS* indicator on the instrument *Genotyping cost shift-share*. The second stage uses the fitted values of *Genotyping cost shift-share* as the regressor. *I(GDA with never associated gene>0)*: 0/1 = 1 if the new gene-disease association encompasses a gene never associated with a disease before 2005 (the year of the first GWAS); *I(GDA with recently discovered gene>0)*: 0/1 = 1 if the new gene-disease association encompasses a gene discovered after the year 2000 (the year of the Human Genome Project's first draft completion); *Anomaly Index*: a synthetic measure of how unlikely a given gene was to be associated with a disease given the pattern of genes previously associated with that disease up to the year before; *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS.

Table G.8: Anomalous gene-disease associations are more likely to fall at both extremes of scientific quality, but not more likely to generate follow on scientific interest.

**Panel A: Empirical anomalies and the quality of discovery**

| | DisGeNET Score of the GDA | | I(GDA in bottom 50% DisGeNET Score>0) | | I(GDA in top 10% DisGeNET Score>0) | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Anomaly Index | 0.00994*** | -0.000259 | 0.130*** | 0.111*** | 0.0455*** | 0.00952* |
| | (0.00296) | (0.00174) | (0.0129) | (0.00918) | (0.00746) | (0.00436) |
| Disease FE | YES | YES | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES | NO | YES |
| Observations | 369,302 | 348,921 | 369,302 | 348,921 | 369,302 | 348,921 |
| Number of diseases | 9,863 | 9,504 | 9,863 | 9,504 | 9,863 | 9,504 |
| Mean of the DV | 0.059 | 0.059 | 0.500 | 0.500 | 0.100 | 0.100 |

**Panel B: Empirical anomalies and follow-on science**

| | Follow-on articles on the GDA (all) | | Follow-on articles on the GDA (causal) | | Follow-on articles on the GDA (correlational) | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Anomaly Index | -0.149 | -0.148 | -0.0995* | -0.114 | -0.148*** | -0.0977 |
| | (0.154) | (0.158) | (0.0502) | (0.0675) | (0.0308) | (0.0736) |
| Disease FE | YES | YES | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES | NO | YES |
| Observations | 369,302 | 348,921 | 369,302 | 348,921 | 369,302 | 348,921 |
| Number of diseases | 9,863 | 9,504 | 9,863 | 9,504 | 9,863 | 9,504 |
| Mean of the DV | 0.059 | 0.059 | 0.500 | 0.500 | 0.100 | 0.100 |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the gene and disease level. All models include dummies controlling for disease and year fixed effects; Columns (2), (4) and (6) also include dummies controlling for the principal investigator (PI) of the articles introducing a GDA. *DisGeNET Score of the GDA* : synthetic measure of scientific value of the discovery provided by DisGeNET; *I(GDA in bottom 50% DisGeNET Score>0)*: 0/1 = 1 if the gene-disease association has a DisGeNET Score below the sample median; *I(GDA in top 10% DisGeNET Score>0)*: 0/1 = 1 if the gene-disease association has a DisGeNET Score in the top sample decile; *Follow-on articles on the GDA (all)*: count of all subsequent studies investigating the gene-disease association; *Follow-on articles on the GDA (causal)*: count of subsequent studies investigating causal relationships between the gene and the disease; *Follow-on articles on the GDA (correlational)*: count of subsequent studies investigating correlational relationships between the gene and the disease; *Anomaly Index*: a synthetic measure of how unlikely a given gene was to be associated with a disease given the pattern of genes previously associated with that disease up to the year before. See text and Appendix D for details.

Table G.9: GWAS are more likely to introduce anomalous gene-disease associations relative to candidate gene studies.

| | Anomaly Index of the GDA | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| GWAS (0/1) | 0.0805*** | 0.0713*** | 0.00698*** | 0.00757*** |
| | (0.00765) | (0.00794) | (0.000927) | (0.00171) |
| Disease FE | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES |
| Observations | 365,340 | 345,040 | 359,303 | 338,984 |
| Number of diseases | 8,730 | 8,429 | 8,730 | 8,429 |
| Mean of the DV | 0.027 | 0.027 | 0.011 | 0.011 |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the gene and disease level. All models include dummies controlling for disease and year fixed effects; Columns (2) and (4) also include dummies controlling for the principal investigator (PI) of the articles introducing a GDA. Columns (1) and (2) include the whole sample, while Columns (3) and (4) exclude new gene-disease associations involving a gene never studied previously (which mechanically have Unexpectedness Index equal to 1). *Anomaly Index*: a synthetic measure of how unlikely a given gene was to be associated with a disease given the pattern of genes previously associated with that disease up to the year before; *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. See text for details.

Table G.10: Descriptive statistics of the gene-disease-year panel.

| | Candidate gene studies | | | | | | GWAS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mean | median | st d | min | max | N | mean | median | st d | min | max | N |
| Papers about the gene-disease pair | 0.047 | 0 | 0.387 | 0 | 101 | 4,673,131 | 0.088 | 0 | 0.980 | 0 | 88 | 108,940 |
| . . . investigating causal relationships | 0.024 | 0 | 0.219 | 0 | 72 | 4,673,131 | 0.051 | 0 | 0.570 | 0 | 53 | 108,940 |
| . . . investigating correlational relationships | 0.016 | 0 | 0.172 | 0 | 56 | 4,673,131 | 0.006 | 0 | 0.092 | 0 | 7 | 108,940 |
| Articles weighted by scientific citations | 2.211 | 0 | 35.641 | 0 | 11,338 | 4,673,131 | 10.269 | 0 | 106.084 | 0 | 6,983 | 108,940 |
| Articles weighted by clinical citations | 0.035 | 0 | 1.058 | 0 | 767 | 4,673,131 | 0.199 | 0 | 2.627 | 0 | 285 | 108,940 |
| Articles weighted by journal SJR | 3.979 | 0 | 32.567 | 0 | 8,672 | 4,673,131 | 7.650 | 0 | 84.748 | 0 | 7,993 | 108,940 |
| Post publication (0/1) | 0.458 | 0 | 0.498 | 0 | 1 | 4,673,131 | 0.337 | 0 | 0.473 | 0 | 1 | 108,940 |
| Year | 2010 | 2010 | 3.742 | 2004 | 2016 | 4,673,131 | 2010 | 2010 | 3.742 | 2004 | 2016 | 108,940 |

*Note:* This table presents descriptive statistics for the gene-disease-year level panel. *Papers about the gene-disease pair*= count of yearly papers about a specific gene-disease combination; *Papers about the gene-disease pair investigating causal relationships*= count of yearly papers about a specific gene-disease combination that investigate causal relationships; *Papers about the gene-disease pair investigating correlational relationships*= count of yearly papers about a specific gene-disease combination that investigate correlational relationships; *Articles weighted by scientific citations*: yearly count of all scientific papers investigating a gene-disease pair weighted by the scientific citations they received; *Articles weighted by clinical citations*: yearly count of all scientific papers investigating a gene-disease pair weighted by the clinical citations they received; *Articles weighted by journal SJR*: yearly count of all scientific papers investigating a gene-disease pair weighted by the SCImago Journal Rank (SJR) of the journal where they appeared published; *Post publication (0/1)*= 0/1 = 1 in all years after a gene-disease pair is reported by a paper for the first time; *Year*= average year of the observations in the panel. The table reports only data that are effectively used in the empirical estimates, i.e., excluding observations that are dropped by the inclusion of fixed effects. See text for details.

Table G.11: Relative to candidate gene studies, GWAS introduce gene-disease associations that generate more studies focused on the causal mechanisms linking the same gene-disease pair.

| | All papers | Papers about causal relationships | Papers about correlational relationships |
|---|---|---|---|
| | (1) | (2) | (3) |
| Post × Publication (0/1) | 0.0561*** | 0.0250*** | 0.0151*** |
| | (0.00139) | (0.000764) | (0.000700) |
| ... × GWAS (0/1) | 0.153*** | 0.0922*** | 0.00410 |
| | (0.0222) | (0.0138) | (0.00238) |
| Gene-disease FE | YES | YES | YES |
| Disease × year FE | YES | YES | YES |
| Gene × year FE | YES | YES | YES |
| Observations | 4,782,071 | 4,782,071 | 4,782,071 |
| Number of diseases | 9,853 | 9,853 | 9,853 |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Difference-in-differences panel regressions at the gene-disease-year level. Std. err. clustered at the gene and disease level. All models include gene-disease pair, disease × year, and gene × year fixed effects. *All papers*: yearly count of all scientific papers investigating a gene-disease pair; *Papers about causal relationships*: yearly count of all scientific papers investigating causal relationships about a gene-disease pair according to the AI engine of PubTator3 (Wei et al., 2024); *Papers about correlational relationships*: yearly count of all scientific papers investigating correlational relationships about a gene-disease pair according to the AI engine of PubTator3 (Wei et al., 2024); *Post Publication*: 0/1 = 1 in all years after a gene-disease pair is first reported in a study; *GWAS*: 0/1 = 1 for gene-disease associations introduced by a GWAS. All dependent variables count papers directly working on the gene-disease combinations, regardless of whether they cite the study that first introduced it. See text for details.

Table G.12: Relative to candidate gene studies, GWAS introduce gene–disease associations that generate higher-impact follow-on research and are published in more prestigious journals.

| | Articles weighted by scientific citations (1) | Articles weighted by clinical citations (2) | Articles weighted by journal SJR (3) |
|---|---|---|---|
| Post × Publication (0/1) | 4.834*** | 0.0843*** | 5.192*** |
| | (0.280) | (0.00697) | (0.120) |
| ... × GWAS (0/1) | 20.38*** | 0.375*** | 13.33*** |
| | (2.969) | (0.0676) | (1.919) |
| Gene-disease FE | YES | YES | YES |
| Disease × year FE | YES | YES | YES |
| Gene × year FE | YES | YES | YES |
| Observations | 4,782,071 | 4,782,071 | 4,782,071 |
| Number of diseases | 9,853 | 9,853 | 9,853 |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Difference-in-differences panel regressions at the gene-disease-year level. Std. err. clustered at the gene and disease level. All models include gene-disease pair, disease × year, and gene × year fixed effects. *Articles weighted by scientific citations*: yearly count of all scientific papers investigating a gene-disease pair weighted by the scientific citations they received; *Articles weighted by clinical citations*: yearly count of all scientific papers investigating a gene-disease pair weighted by the clinical citations they received; *Articles weighted by journal SJR*: yearly count of all scientific papers investigating a gene-disease pair weighted by the SCImago Journal Rank (SJR) of the journal where they appeared published; *Post* × *Publication*: 0/1 = 1 in all years after a gene-disease pair is first reported in a study; *GWAS*: 0/1 = 1 for gene-disease associations introduced by a GWAS. All dependent variables count papers directly working on the gene-disease combinations, regardless of whether they cite the study that first introduced it. See text for details.

# Appendix References


Achenbach, P., M. Hummel, L. Thümer, H. Boerschmann, D. Höfelmann, and A. Ziegler (2013): "Characteristics of rapid vs slow progression to type 1 diabetes in multiple islet autoantibody-positive children," *Diabetologia*, 56, 1615–1622.

Aharonson, B. S. and M. A. Schilling (2016): "Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution," *Research Policy*, 45, 81–96.

Arts, S., N. Melluso, and R. Veugelers (2025): "Beyond citations: Measuring novel scientific ideas and their impact in publication text," *Review of Economics and Statistics*, 1–33.

Borusyak, K., P. Hull, and X. Jaravel (2025): "A practical guide to shift-share instruments," *Journal of Economic Perspectives*, 39, 181–204.

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.

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

Callaway, E. (2017): "New concerns raised over value of genome-wide disease studies." *Nature*, 546, 463–464.

DiMeglio, L. A., C. Evans-Molina, and R. A. Oram (2018): "Type 1 diabetes," *The Lancet*, 391, 2449–2462.

Gingerich, M. A., V. Sidarala, and S. A. Soleimanpour (2020): "Clarifying the function of genes at the chromosome 16p13 locus in type 1 diabetes: CLEC16A and DEXI," *Genes & Immunity*, 21, 79–82.

Goldsmith-Pinkham, P., I. Sorkin, and H. Swift (2020): "Bartik instruments: What, when, why, and how," *American Economic Review*, 110, 2586–2624.

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

Hakonarson, H., S. F. Grant, J. P. Bradfield, L. Marchand, C. E. Kim, J. T. Glessner, R. Grabs, et al. (2007): "A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene," *Nature*, 448, 591–594.

Heeney, C. (2021): "Problems and promises: How to tell the story of a Genome Wide Association Study?" *Studies in History and Philosophy of Science Part A*, 89, 1–10.

Kang, S. (2025): "From Outward to Inward: Reframing Search with New Mapping Criteria," in *Academy of Management Proceedings*, Academy of Management Valhalla, NY 10595, vol. 2025, 10129.

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

Pearson, T. A. and T. A. Manolio (2008): "How to interpret a genome-wide association study," *Journal of the American Medical Association*, 299, 1335–1344.

Qu, H.-Q., L. Marchand, R. Grabs, and C. Polychronakos (2007): "The IRF5 polymorphism in type 1 diabetes," *Journal of Medical Genetics*, 44, 670–672.

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

Soleimanpour, S. A., A. Gupta, M. Bakay, A. M. Ferrari, D. N. Groff, J. Fadista, L. A. Spruce, J. A. Kushner, L. Groop, S. H. Seeholzer, et al. (2014): "The diabetes susceptibility gene Clec16a regulates mitophagy," *Cell*, 157, 1577–1590.

Tranchero, M. (2024): "Finding diamonds in the rough: Data-driven opportunities and pharmaceutical innovation," *The Wharton School*.

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., M. A. Brown, M. I. McCarthy, and J. Yang (2012): "Five years of GWAS discovery," *American Journal of Human Genetics*, 90, 7–24.

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

Wei, C.-H., A. Allot, P.-T. Lai, R. Leaman, S. Tian, L. Luo, Q. Jin, Z. Wang, Q. Chen, and Z. Lu (2024): "PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge," *Nucleic Acids Research*, 52, W540–W546.

WTCCC (2007): "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, 447, 661–678.