

THE STREETLIGHT EFFECT IN DATA-DRIVEN EXPLORATION *

Johannes Hoelzemann
University of Vienna

Gustavo Manso
UC Berkeley

Abhishek Nagaraj
UC Berkeley & NBER

Matteo Tranchero
The Wharton School

December 16, 2025

Abstract

We study exploration under uncertainty and show how access to data on past attempts can paradoxically hinder breakthrough discovery. We develop a model of the “streetlight effect” demonstrating that when data highlights attractive but ultimately suboptimal projects, it can narrow exploration and suppress innovation. In a laboratory experiment, we find that revealing the value of an enticing project lowers payoffs and reduces breakthrough discoveries. This drop stems from increased free-riding behavior, which crowds out the generation of new data. We then apply our theory in the context of scientific research into the genetic origins of human diseases, focusing on the drivers of limited exploration. To identify the causal impact of past data, we use an instrumental variable that leverages exogenous genetic overlaps between humans and laboratory mice, which reduces research costs for specific genes and leads to prioritized data collection about them. We find that diseases with early evidence of promising genetic targets are 16 percentage points less likely to yield breakthroughs than those where early efforts failed. While competition attenuates the streetlight effect, it does not eliminate it. Our paper provides the first analysis of this phenomenon, outlining the conditions under which data leads agents to look under the lamppost rather than engage in socially beneficial exploration.

JEL Classification: C73, C92, D81, D83.

Keywords: Streetlight effect; Data; Exploration and exploitation; Genetics research; Dynamic public-good problem; Laboratory experiment.

*The experiment was conducted online via the Vienna Center for Experimental Economics in December 2024. The experiment received approval from the University of Vienna and the University of Toronto Research Ethics Board (approval #00038482), and was pre-registered at <https://osf.io/zs2mu/>. We are grateful to participants at the SIOE 2025 conference, the SIE workshop, the MOM workshop at HBS, the MAD conference at Columbia, the Paris Bounded Rationality Workshop at PSE, the 2025 Esade Workshop in Behavioral and Experimental Economics, the SMS Special Conference at Bocconi, the Strategy Science conference, as well as seminars at Berkeley, BU, CEU, GeorgiaTech, Masaryk, Michigan, Minnesota, Montreal, NBER, Purdue, Rotman, UCLA, Vienna, Wharton, WashU, and WU Vienna for their suggestions. We thank Adit Jain for his outstanding help in designing the experimental platform. Eva Chang and especially Cecil-Francis Brenninkmeijer provided excellent research assistance. Authors are listed in alphabetical order. Corresponding author: manso@berkeley.edu.

1 Introduction

A central challenge in medical research is identifying the genetic drivers of human disease from over 19,000 potential gene candidates. Puzzlingly, more than two decades after the Human Genome Project mapped all human genes, the genetic landscape remains relatively underexplored (Edwards et al., 2011). Fewer than 10% of genes have been targeted by approved drugs, despite recognition that many less-studied genes may offer better therapeutic opportunities (Stoeger et al., 2018; Gates et al., 2021). While this issue has been noted in policy discussions (Rodgers et al., 2018), it poses a puzzle: why is collective exploration so limited, neglecting many potentially valuable options, when scientists should have strong incentives to search broadly? Understanding the drivers of scientific underexploration is critical in an era of apparent diminishing returns to research effort (Gordon, 2016; Bloom et al., 2020). Whether these trends reflect intrinsic limits to innovation or a narrowly focused search shaped by innovators' incentive structures remains an open question (Bhattacharya and Packalen, 2020).

To shed light on this issue, we start by observing that innovative search rarely begins on a blank slate. For instance, a scientist studying a disease typically draws on data from past research before selecting a genetic target. We develop a framework to understand how such data shapes the direction of future exploration. Our thinking is motivated by the parable of the *streetlight effect*, where agents disproportionately focus their search in areas with readily available data rather than allocating effort based on scientific theory, market potential, or policy relevance. In our simple model, we show how information on past discoveries can narrow search and, paradoxically, reduce both individual and social returns. This runs counter to the conventional view that accurate data should improve outcomes by reducing uncertainty and making exploration more efficient. Our paper reconciles these perspectives by studying how the streetlight effect can emerge in exploratory search and identifying the conditions under which greater data availability may hinder rather than help innovation.

We start from a simplified multi-armed bandit model amenable to experimental and empirical testing. Agents choose among risky projects that can be of low, medium, or high value, but whose quality is revealed only through exploration over time. In each period, the decision-maker uses existing information to choose between investing in a previously explored project or taking a risk by exploring a new one. Exploration costs are borne privately, but the resulting data become publicly available. Within this setup, we examine how providing data on the value of one opportunity influences exploration choices. Our central result is that the impact of data hinges on the type of project illuminated:

information about a medium-value project can *reduce* both individual and group payoffs relative to having data on a low-value project or even no data at all.

The intuition behind our result is that when the medium-value project exceeds the expected return from exploring riskier alternatives, it becomes individually rational for agents to pursue the option highlighted by the data. Since this logic applies to all agents, it induces herding behavior: data reduces uncertainty but also narrows the direction of follow-on investment, collectively suppressing exploration that would result in new data generation. As a result, even rational agents may underexplore due to free riding on the informational externalities of others. Our baseline model, where followers receive the same payoff as initial innovators, reflects settings like science and technology, where knowledge is partially non-excludable (Aghion et al., 2008; Krieger, 2021). When we introduce competition by reducing the rewards for follow-on innovators, the effect persists under moderate rivalry but weakens as competition intensifies. Thus, while competitive pressures can undermine innovation quality through racing dynamics (Hill and Stein, 2025a), our simple setup suggest that their absence may similarly hurt innovation due to the streetlight effect.

Next, we implement an online laboratory experiment to test whether our theoretical predictions hold with human participants. Groups of players take part in a two-period game involving strategic exploration. In the baseline condition, players sequentially select from five unknown options randomly drawn from a known payoff distribution. In the first period, they choose one project without immediate feedback; in the second, they observe all first-round payoffs before selecting again. Payoffs are non-rival and cumulative. We then run the same game but provide players with information on one project—either low, medium, or high in value. The results align with our theory: revealing data on the medium-value project reduces group payoffs by 5% and the likelihood of finding the best outcome by 56%, relative to the no-data baseline. Information on low-value projects has no significant effects, while data on high-value projects improve outcomes. We also vary the degree of payoff rivalry and find that the streetlight effect persists under moderate rivalry but diminishes in magnitude.

Our theoretical and laboratory results provide an intriguing explanation for the persistence of under-exploration equilibria in scientific research. In the final part of the paper, we return to our motivating example of scientists looking for disease-related genes to employ as drug targets. Searching for the genetic roots of human diseases closely mirrors our theoretical setup: researchers face over 19,000 protein-coding genes, and pinpointing the right targets involves individually risky exploration that can yield large payoffs for drug development. It is also a collective endeavor, with scientists learning

from one another and drawing on data from publications. For instance, consider Tangier disease, a rare condition characterized by extremely low levels of HDL cholesterol in the blood. Decades of research had focused on genes that early data suggested as moderately promising, but unlikely to lead to therapeutic breakthroughs—until a scientist definitely linked the disease to mutations in the ABCA1 gene. We leverage these parallels to examine whether dynamics akin to the streetlight effect might be similarly steering scientists away from discoveries.

We leverage data from DisGeNET, a bibliographic database that links scientific publications to the specific diseases and genes they investigate. DisGeNET assigns a normalized score to rank the scientific value of each gene-disease combination, which we use to classify them into low, medium, and high categories. Our dataset covers genetic discoveries for 3,864 diseases between 1980 and 2019. We use this data to examine how the scientific promise of early discoveries shapes subsequent innovation at the disease level. The main analysis cross-sectionally explores the implications of our model using careful controls for disease type and total research effort received. Since the distribution of past data is unlikely to be random, we also employ a complementary identification strategy. We use an instrumental variable (IV) approach that exploits variation in genetic similarity between human and mouse genes. Research on a human gene is less costly when scientists can study the same gene in laboratory mice, so genes shared across species tend to be explored earlier (Stoeger et al., 2018). However, diseases differ in the likelihood that such shared genes are of high scientific value. At the disease level, this variation creates quasi-random differences in early data, which we use to instrument for the promise of initial discovery and estimate causal effects.

Our results show that disease areas with promising but suboptimal genes are 16 percentage points less likely to report a major breakthrough afterward, compared to diseases where all earlier data unveiled low promise targets. In practical terms, discovering a medium-value genetic target delays a breakthrough by an average of 2.8 years, roughly 14% longer than the sample mean of 20.2 years. These findings are confirmed by our IV framework. Event study estimates further show a sharp decline in the number of new genes explored following a medium-value discovery, with no evidence of pre-trends. Consistent with our theory, the mechanism seems to be that early medium-value discoveries reduce the diversity of follow-on research, narrowing exploration and lowering the likelihood of identifying high-impact gene-disease associations. Also in line with our model, we find that the streetlight effect is muted in disease areas with greater competition. Taken together, our simple theoretical framework provides a useful lens for understanding the mechanisms that can give rise to underexploration in genetic research.

Our three-part study relates to several strands of research. First, we add to a growing literature on how data is generated and how it shapes economic outcomes (Bergemann and Bonatti, 2019; Bessen et al., 2022; Farboodi and Veldkamp, 2020; Jones and Tonetti, 2020). Rather than treating data as a homogeneous commodity, we show that the nature of the data itself (specifically what it illuminates or omits) shapes agents' exploration choices. While our findings resonate with the literature that studies data as a public good (e.g., Nagaraj and Tranchero, 2024), they apply more generally to information that offers signals about the payoffs of uncertain projects. Note also that we operationalized data as instrumental information, i.e., unbiased and directly payoff-relevant, so our results do not depend on biased or uninformative data (Henrich et al., 2010; Cao et al., 2024). Rather, our paper proposes a novel mechanism by which data can hinder exploration: by leading agents to implicitly coordinate on dominated projects, thus crowding out new data generation to the detriment of collective outcomes.

Second, we contribute to the literature on experimentation and social learning (Bolton and Harris, 1999; Keller et al., 2005; Klein and Rady, 2011; Hörner et al., 2022). We build on recent experimental work examining behavior under strategic interdependence and informational externalities (Boyce et al., 2016; Hoelzemann and Klein, 2021, 2025). Relative to the commonly studied single-agent bandit problem (Bergemann and Valimaki, 2008), we show how informational spillovers in collective experimentation can create free-rider problems that endogenously limit aggregate data generation and dynamically lower payoffs.¹ We further demonstrate that this mechanism aligns with empirical patterns in scientific research on disease-causing genes (Gates et al., 2021; Edwards et al., 2011; Haynes et al., 2018), illustrating how our framework helps explain real-world search dynamics.

Finally, we contribute to the innovation search literature that examines what drives risky exploration among innovators (March, 1991; Levinthal, 1997; Azoulay et al., 2011; Ederer and Manso, 2013; Henry et al., 2022). Our paper highlights the role of the information environment in driving underexploration. We also build on research exploring how different types of data influence experimentation under technological uncertainty (Ewens et al., 2018; Krieger, 2021). In particular, we show how data might have counterintuitive effects in search, offering a less sanguine outlook for how innovation will be shaped in the age of big data and AI (Agrawal et al., 2024; Cockburn et al., 2019). Our evidence on disease-relevant genetic discovery adds to prior work examining how databases shape scientific productivity in the biomedical field (Kao, 2024; Williams, 2013; Tranchero, 2025).

¹A related literature in computer science examines rule-based bandit learning, where a single decision-maker follows fixed decision rules (Vermorel and Mohri, 2005). In contrast, the welfare losses we document arise from incentive misalignment between individually and socially optimal behavior, rather than from bounded rationality.

The remainder of the paper proceeds as follows. Section 2 provides an overview of the theoretical framework. Section 3 describes the laboratory experiment. Sections 4 and 5 present the empirical analysis in the context of genetic research. Section 6 concludes.

2 Theoretical Framework

Setup. There are N agents engaged in a search to maximize their individual payoffs, choosing from A projects of initially unknown value, with $N \geq A$. Project payoffs are independent and fall into one of three categories: with probability p_L , a project yields a low payoff (L); with probability p_M , a medium payoff (M); and with probability p_H , a high payoff (H), where $0 \leq L < M < H$ and $p_L + p_M + p_H = 1$. While agents know this distribution in advance, they have no prior information about the specific payoff of any given project. Each agent lives for two periods, is risk-neutral, and discounts future payoffs at zero. Agents cannot communicate directly. This setup reflects real-world environments in which individuals face a set of unknown opportunities, where valuable projects are rare but highly rewarding (Kerr et al., 2014; Manso, 2016).

Dynamics. In each period, the N agents choose projects sequentially in a random order. While they can observe the choices made by earlier movers, they do not yet see the payoffs associated with those choices. Once all agents have selected a project, the payoffs of the chosen projects are revealed to everyone, marking the end of period 1. In period 2, the process repeats with the same order. This time, agents know the payoffs of previously explored projects and can choose either a known project or an unexplored one, whose payoff will again be revealed at the end of the period. Payoffs are cumulative across the two periods, so agents earn the sum of the values of the projects they choose. Importantly, payoffs are non-rival, so if multiple agents select the same project, each receives its full value. Unlike classic payoff externalities in public goods problems, here an agent is affected by others only through the data their choices produce over time (Hoelzemann and Klein, 2021, 2025). This setup mimics competitive markets where organizations conduct parallel R&D. Although projects do not directly compete, the information they generate is valuable to all participants (Krieger, 2021).

Equilibrium without Data. We begin by considering the equilibrium in a setting where no data about project payoffs is available before the game begins. At the start of period 1, all projects offer the same expected payoff based on the known probability distribution. The sequential structure of the game leads agents to choose different projects to generate more data that can guide decisions in period 2. Since $N \geq A$, agents can implicitly coordinate to explore all projects, so the highest payoff

is identified before the second period begins. This means that each agent earns the expected value of a random draw in period 1, followed by the highest available payoff in period 2. The probability that the best discovered project has payoff L is p_L^A , payoff M is $(1 - p_H)^A - p_L^A$, and payoff H is $1 - (1 - p_H)^A$. The collective expected payoff and the likelihood of discovering a high-value project are as follows:

$$\text{[Group Payoff]} \quad N[(p_L + p_L^A)L + (p_M + (1 - p_H)^A - p_L^A)M + (p_H + 1 - (1 - p_H)^A)H] \quad (1)$$

$$\text{[Group Breakthrough]} \quad 1 - (1 - p_H)^A \quad (2)$$

Equilibrium with Data on L or H Projects. We now compare the setup above to a scenario where the payoff of one project is publicly revealed at the start of the game. The effect of this data depends on the value of the disclosed project. If the revealed project has a payoff of H , all agents immediately coordinate on it, each earning $2H$, and the group achieves the maximum total payoff of $2H \cdot N$. The probability of a breakthrough is 1, showing how data can lead directly to the best possible outcome by eliminating uncertainty. If, instead, the revealed project has a payoff of L , agents simply avoid that option, and they are back to the original setup with one fewer low-value project. In this case, the group's expected payoff is $N[(p_L + p_L^{(A-1)})L + (p_M + (1 - p_H)^{(A-1)} - p_L^{(A-1)})M + (p_H + 1 - (1 - p_H)^{(A-1)})H]$, and the probability of a breakthrough is $1 - (1 - p_H)^{(A-1)}$. These outcomes are similar to the no-data case and converge to it as $A \rightarrow \infty$. In other words, when a low-payoff project is revealed and the search space is large enough, there is still dispersed exploration.

Equilibrium with Data on M project. What is arguably more interesting, and so far understudied, is the intermediate case where a medium-value project is revealed. Here, a non-empty parameter space exists in which data can be detrimental due to the streetlight effect. This arises when the payoff from choosing M is attractive enough relative to exploring other, unknown-value projects. If the loss from exploration, given by $M - (p_L L + p_M M + p_H H)$, exceeds the potential gain from exploration, $p_H(H - M)$, then all agents choose the medium project in equilibrium.² This leads to the following condition:

Assumption 1 (“Medium Project is Good Enough”).

$$M > \frac{p_L L + p_H 2H}{2 - p_L - 2p_M} \quad (3)$$

Assumption 1 ensures that selecting the medium project dominates searching for a high-value one.

²Suppose there was an equilibrium where some agents selected other projects. By backward induction, the last such agent would strictly prefer the medium project under this condition.

Rational agents choose it in both periods, yielding an expected group payoff of $2M \cdot N$. However, when M is not too large relative to L and H , we can show—perhaps counterintuitively—that payoffs with data are actually lower than those with no data. More formally, we introduce:

Proposition 1 (“Group Payoff with Data on Medium Project”). *Under Assumption 1 and if*

$$M < \frac{(p_L + p_L^A)L + (p_H + 1 - (1 - p_H)^A)H}{2 - (1 - p_H)^A + p_L^A - p_M} \quad (4)$$

the group payoff without data is higher than when a medium project is revealed.

Proof. We need to show that the expected group payoff without data exceeds the expected group payoff whenever an M project is revealed upfront. This is true if $N[(p_L + p_L^A)L + (p_M + (1 - p_H)^A - p_L^A)M + (p_H + 1 - (1 - p_H)^A)H] > 2M \cdot N$, which is equivalent to the condition in the proposition. ■

The fact that the medium option offers high individual payoffs does not guarantee it is socially optimal. On the contrary, it can lure agents into avoiding exploration. The known option is tempting when the individual odds of finding the high-value project are low, but at the cost of hurting collective welfare. What rational agents fail to account for are the information externalities created by their own experimentation, even when unsuccessful. This leads to the following two results:

Proposition 2 (“Exploration with Data on Medium Project”). *If $\mu|i$ is defined as the unmapped share of projects chosen in period 1 given data i , then under Assumption 1, the following weak inequalities hold: $\mu|H \leq \mu|M \leq \mu|L \leq \mu|\emptyset$*

Proof. The proof directly derives from our preceding discussion. If H is revealed, agents will choose that project, so $\mu = 0$. If M is appealing enough, agents forfeit exploration and only choose the revealed project, so $\mu = 0$. If no data is provided or L is revealed, then agents explore all remaining unknown options in period 1, so $\mu = 1$. ■

Proposition 3 (“Breakthrough with Data on Medium Project”). *If $P(H|i)$ is defined as the conditional probability of discovering H given data i , then under Assumption 1, the following strict inequality holds: $P(H|M) < P(H|i)$ where $i \in \{\emptyset, L, H\}$*

Proof. If M is appealing enough, agents never achieve a breakthrough, i.e., never discover H , so $P(H|M) = 0$. If no data is provided or L is revealed ex ante, then agents explore all remaining unknown options in period 1, and the probability that H is discovered at all is $(1 - (1 - p_H)^A)$ and $(1 - (1 - p_H)^{A-1})$ respectively, which are both strictly greater than 0. The statement is trivially true whenever H is revealed. ■

The streetlight effect arises when the medium payoff is tempting enough for the individual, yet exploration still holds social value—that is, when Assumption 1 and Equation (4) both hold. This

requires a skewed payoff distribution. If the distribution of payoffs was symmetric, the expected value of an unknown draw would equal M , making exploration risk-free with a potential upside of $p_H(H - M)$. In that case, Assumption 1 would be violated, and the streetlight equilibrium would not emerge. However, the effect also vanishes under extreme payoff skewness. If the breakthrough is too rare (very small p_H), the expected social value of exploration falls below M , violating Equation (4). If the breakthrough is too common (very large p_H), the private upside $p_H(H - M)$ becomes very attractive, breaking Assumption 1. Thus, the streetlight equilibrium appears only under moderate skewness of the payoff distribution.³

The Role of Competition. Our theoretical framework assumes non-rivalry in payoffs, meaning that agents still earn the full reward even if they were not the first to choose a project. While a simplification, this assumption fits reasonably well in fields like scientific research. For instance, Hill and Stein (2025b) find that follow-on projects receive about 79% as many citations as similar projects that were first to the finding, even if they appear on less prestigious outlets. Scientific innovation is often non-rivalrous because early discoveries generate new opportunities for others in the same domain. Still, in many other settings, one agent’s choice can largely diminish the value of that option for others. To capture this, we now introduce rivalry into the model. Specifically, we assume that a project’s payoff falls to zero when the number N of agents already selecting that project is greater than \bar{N} . This adjustment makes individual payoffs sensitive to competition, with smaller \bar{N} reflecting stronger payoff rivalry. The rest of the model remains unchanged. The results below show how this affects exploration and discovery:

Proposition 4 (“Exploration under Rivalry”). *If payoff rivalry is not extreme (i.e., $\bar{N} > N - A + 1$), then the original weak inequalities still hold under Assumption 1: $\mu|H \leq \mu|M \leq \mu|L \leq \mu|\emptyset$. Moreover, exploration is increasing in rivalry.*

Proof. Without any data, agents still explore all projects in the first period, so $\mu|\emptyset = 1$. If L is revealed ex ante, then agents will explore all the unknown projects in the first period so that $\mu|L = 1$. If H is revealed ex ante, then \bar{N} agents will select the mapped project. The remaining $N - \bar{N}$ agents will randomly select as many as the remaining $A - 1$ projects as possible. Therefore, since $N - \bar{N} < A - 1$, $\mu|H = \frac{N - \bar{N}}{A - 1}$. Similarly, if M is revealed ex ante, and Assumption 1 holds, then \bar{N} agents will select the mapped project. The remaining $N - \bar{N}$ agents will randomly select as many as the remaining $A - 1$ projects as possible. Since $N - \bar{N} < A - 1$, then $\mu|M = \frac{N - \bar{N}}{A - 1}$. Now, suppose we increase rivalry to $\bar{N} - 1$. Since $N - \bar{N} < A - 1$, when a medium project is revealed, an additional unknown project is explored, and $\mu|M$ increases. In the extreme case, when $\bar{N} = 1$, the revelation of an M

³For example, the following parameters satisfy Assumption 1 and Equation (4), and thus lead to the outcome where revealing information about a medium project reduces social welfare and lowers the probability of a breakthrough: $L = 0, M = 6, H = 15, p_L = 7/10, p_M = 1/10, p_H = 2/10, A = 5, N = 5$.

project has no impact on exploration as agents will explore all remaining unknown options in the first period. Note this result can be analogously stated in terms of N (instead of \bar{N}). Holding \bar{N} constant, if we increase the number of agents to $N + 1$, then if $N - \bar{N} < A - 1$, an additional unknown project is explored, and $\mu|M$ increases. ■

Proposition 5 (“Breakthroughs under Rivalry”). *If payoff rivalry is not extreme (i.e., $\bar{N} > N - A + 1$), then the original strict inequality still holds under Assumption 1: $P(H|M) < P(H|i)$ where $i \in \{\emptyset, L, H\}$. Moreover, breakthrough discoveries are increasing in rivalry.*

Proof. The proof follows directly from the analysis above. If no data is provided, then the probability that H is discovered is still $(1 - (1 - p_H)^A)$. If L is revealed ex ante, then the probability that H is discovered is still $(1 - (1 - p_H)^{A-1})$. If H is revealed ex ante, then $P(H|H)$ is still trivially 1. If M is revealed ex ante, the probability that H is discovered is $(1 - (1 - p_H)^{N-\bar{N}})$ and, since $A - 1 > N - \bar{N}$, $P(H|M) < P(H|L) < P(H|\emptyset) < P(H|H)$. Now, suppose we increase rivalry to $\bar{N} - 1$. Since $N - \bar{N} < A - 1$, then an additional unknown project is explored, and $P(H|M)$ increases. Similar to before, this result can also be stated in terms of N . Suppose we increase the number of agents to $N + 1$. Since $N - \bar{N} < A - 1$, then an additional unknown project is explored, and $P(H|M)$ increases. ■

Our key finding is that the streetlight effect persists under modest levels of rivalry but weakens as rivalry increases. Competition pushes agents to explore more, increasing the likelihood of discovering a high-value project. This highlights payoff rivalry as a boundary condition for the streetlight effect.

3 Laboratory Experiment: Design and Results

While our simple theoretical framework illustrate how the streetlight effect can emerge, it remains an open question whether it accurately reflects how agents behave in practice. To explore this, we conducted an online experiment mirroring exactly the structure of our simplified model.

3.1 Experimental Procedure and Logistics

Participants logged into the experimental platform remotely and were assigned to either the data or no-data condition in groups of ten. Upon joining, they received detailed written instructions and watched a mandatory seven-minute video that reiterated the rules and introduced the platform.⁴ Participants were then required to complete a short quiz as an attention and comprehension test. They also had continuous access to the instructions and could contact an experimenter via cell phone or Zoom for support. The experiment consisted of independent “rounds,” each following the structure of our theoretical framework. Each round had two periods over which payoffs were calculated. Participants

⁴The videos shown to participants are available upon request.

were randomly assigned to groups of five, with groups reshuffled every five rounds. In total, each participant played 20 rounds. At the end of the experiment, we collected demographic information and measured risk preferences using a monetarily incentivized, upscaled version of the Holt and Laury task (Holt and Laury, 2002). Final payments included earnings from one randomly selected round, a show-up fee, and the outcome of the risk elicitation task.

The experiment was programmed using the open-source platform oTree (Chen et al., 2016) and conducted at the Vienna Center for Experimental Economics (VCEE). Participants were recruited from VCEE’s subject pool via ORSEE (Greiner, 2015), targeting undergraduate and master’s students who had previously participated in no more than five experiments. Participation was voluntary, and individuals could withdraw at any time. We ran 18 sessions with a total of 180 participants, ensuring that no one took part in more than one session. Participants ranged in age from 18 to 52, with an average age of 24.7 years and a standard deviation of 4.7. All sessions were conducted in December 2024. The experimental task lasted approximately 50 minutes, with additional 10 minutes allocated for reading instructions, watching the explanatory video, and completing the attention quiz. Average participant earnings were €15.4, with a standard deviation of €4.6.

3.2 Task Description and Implementation

Participants took on the role of individuals searching for precious gems (Panel A of Figure 1). In each round, they faced five mountains, each hiding one type of gem that could only be revealed through exploration. There were three types of gems, differing in rarity and value: topazes (L), rubies (M), and diamonds (H). While the exact monetary values varied across rounds, diamonds were always more valuable than rubies, and rubies were always more valuable than topazes. Participants were informed that topazes appeared with a 60% probability, rubies with 20%, and diamonds with 20%, though they were not told which gem was hidden behind which mountain. The goal of the game was to find the most valuable gems, as their value directly determined participants’ earnings.

In addition to displaying the values and distributions of the gems, the interface tracks the current period and the round number as participants progress through the experiment.⁵ Each group of five players remains anonymous, and participants cannot interact or communicate directly with one another.⁶

⁵The interface also shows the “block” number, which indicates when participant groups are reshuffled. A new block begins every five rounds, after which players remain in the same group for the next five rounds.

⁶Although participants are aware that their co-players change every five rounds, they are never able to identify who they are playing with. When a player selects a mountain, the others see a message such as “one player selected this mountain,” but never learn who made the choice. See Figure 1 for an illustration.

Within each round, players take turns selecting a mountain to explore in a randomly determined order that changes every round. A dynamic indicator on the screen highlights when it is their turn to choose. At the start of each round, no player has private information about the locations of the gems, which are randomly reassigned each round (but remain fixed between the two periods of a given round). While waiting for their turn, players can observe which mountains have already been selected. When it is their turn, they are free to choose the same mountain as someone else or a different one.

In the no-data condition, participants begin by selecting one of five mountains to explore in period 1. Once all players have made their choices, the gems hidden in the selected mountains are revealed to everyone, and each player earns the value of the gem from their chosen mountain. In period 2, players again choose from the same mountains, in the same random order, with gem locations unchanged. Now, however, they can see the gems uncovered in period 1 and can either stick with their previous choice or switch to a different mountain. The newly selected mountains are revealed, and their gem values are added to each player's payoff. The data condition follows the same structure, with one key difference: at the start of each round, one mountain is "mapped," and its gem is revealed to all participants. This is the only information available at the outset. Panel B of Figure 1 illustrates this setup. Figure (i) shows the no-data condition, where all mountains are hidden, while Figures (ii), (iii), and (iv) depict the three possible data scenarios, where the revealed mountain contains a low-, medium-, or high-value gem. The revealed mountain is selected by a script using a random sequence.

We collected data from a total of 720 rounds. In 120 of these, participants received data revealing a low-value gem; in 240 rounds, they saw data on a medium-value gem; and in another 120 rounds, the revealed gem was high-value. In the remaining 240 rounds, no initial data about gem locations was provided. We determined the proportion of rounds assigned to each treatment condition based on power calculations. Across the experiment, we used five different combinations of payoff parameters. Specifically, the values for low, medium, and high-value gems were set to one of the following: $(L, M, H) = \{(1, 6, 11), (1, 6, 11.5), (2, 6, 11), (2, 6, 11.5), (3, 7, 12)\}$.

3.3 Results

Group Payoffs. We begin by examining group-level earnings. For each round, we calculate the maximum possible group payoff and express realized group earnings as a percentage of this value. This allows us to compare outcomes across rounds, despite variation in the values and distributions of the low-, medium-, and high-value gems. Panel (i) of Figure 2 plots the average group payoff by condition, comparing the three data treatments to the no-data baseline. Strikingly, revealing data on a

medium-value project leads to lower group payoffs than all other conditions, including the case where no data is provided. To quantify these differences, we estimate the following OLS specification:

$$Group\ Payoff_{j,k} = \alpha + \beta Initial\ Data_k + \gamma \mathbf{X}_k + \epsilon_{j,k}, \quad (5)$$

where $Group\ Payoff_{j,k}$ denotes the payoff for group j in round k , $Initial\ Data_k$ is a categorical variable indicating the type of project revealed at the start of the round, and \mathbf{X}_k is a vector of fixed effects that accounts for the session, the specific payoff structure, and the round's position in the session. Standard errors are clustered at the session level. Column 1 of Table 1 presents the results. We find that revealing data on a medium-value mountain reduces group payoffs by €3.133, or approximately 4% relative to the no-data condition, consistent with Proposition 1. Providing data on a high-value mountain increases payoffs by 44.5 percentage points. In contrast, revealing a low-value mountain has no statistically significant effect on group performance.

Group Exploration. Our theoretical framework suggests that partial data on project value can discourage exploration, effectively crowding out data generation. To test this, our next outcome of interest is the share of unmapped mountains explored in a round. Panel (ii) of Figure 2 shows that revealing the location of a medium-value gem significantly reduces exploration. We quantify this using an OLS specification similar to Equation (5), with the dependent variable defined as the share of unmapped mountains explored by the group across both periods.⁷ The results in Table 1 show that revealing a high-value gem eliminates the need for exploration, while revealing a low-value gem has no measurable effect. Most notably, revealing a medium-value gem decreases the share of mountains explored by 38.6 percentage points relative to the no-data condition (Column 2). This provides a clear demonstration of the streetlight effect: data can shift the balance from exploration to exploitation, ultimately reducing social welfare by leaving participants stuck on a suboptimal outcome.

Group Breakthroughs. The final outcome of interest is the likelihood that participants discover the high-value option. Panel (iii) of Figure 2 shows that revealing the location of a medium-value gem significantly lowers the chances of a breakthrough. We quantify this effect using a linear probability model based on the specification in equation (5), with the dependent variable indicating whether a group discovers a high-value gem. Since not all rounds contain a diamond, we limit the analysis to rounds where at least one high-value gem is present. As shown in Column 3 of Table 1, revealing a medium-value mountain reduces the likelihood of discovering the maximum by 56% compared to the no-data condition. In contrast, we find no such reduction when the revealed data points to a low-

⁷Note that there are four unmapped mountains in each of the three data conditions and five in the no-data condition.

or high-value gem. Taken together, these results support the predictions of Proposition 3: while data can increase payoffs when it points to the best option, it can also impose substantial societal costs depending on the underlying payoff structure.

The Impact of Competition. Our theory shows that the presence of payoff competition can reduce the intensity of the streetlight effect. We test this experimentally by varying \bar{N} , the number of players who can choose a mountain before payoffs fall to zero. The results from the baseline case, presented earlier, implicitly correspond to $\bar{N} = 5$, where players can choose without penalty. We then examine two more conditions: intermediate rivalry ($\bar{N} = 3$) and extreme rivalry ($\bar{N} = 1$). In the intermediate case (Panel A, Table 2), the streetlight effect weakens but does not disappear. Revealing the medium option no longer affects payoffs, but still reduces exploration by roughly 20 percentage points and the likelihood of a breakthrough by 24.5 percentage points (significant at the 10% level). Revealing the high option still increases payoffs and reduces exploration, while revealing a low option continues to have no effect. Under extreme rivalry (Panel B, Table 2), initial data has no significant impact on payoffs, exploration, or breakthroughs. Consistent with Propositions 4 and 5, the streetlight effect declines with increased rivalry and disappears when payoff competition is strongest.

4 Empirical Application: The Genetic Roots of Human Diseases

The preceding sections formalized and tested how the streetlight effect can emerge in lab-based search tasks. We now turn back to our motivating example, showing how our framework helps explain real-world patterns in scientific research.

4.1 Setting

A key endeavor in biomedical research is identifying the genetic mutations that cause human diseases (see Appendix B for details). Genes carrying causal mutations can serve as drug targets, substantially improving the chances of developing effective treatments (Nelson et al., 2015). However, finding breakthrough drug targets is a complex search problem among the over 19,000 protein-coding human genes. In practice, scientists must choose between further investigating known genetic targets or exploring novel candidates. Despite individual incentives to establish priority in new areas (Bobtcheff et al., 2017; Hill and Stein, 2025a), exploration across the genetic space has remained surprisingly limited (Edwards et al., 2011). Research continues to focus on a subset of human genes, a puzzling pattern given widespread recognition that promising drug targets may lie among less-studied genes (Rodgers et al., 2018; Stoeger et al., 2018). One explanation, echoing the streetlight effect, is that

earlier data on promising—but ultimately unproductive—genes has focused scientists’ efforts away from exploring more valuable alternatives (Haynes et al., 2018).

To illustrate this, consider two examples of genetic disorders described in Figure 3. As noted in the introduction, research on Tangier disease followed a revealing trajectory. A 1982 study identified a moderate link to the APOA1 gene, which attracted subsequent attention and diverted exploration away from alternative candidates. However, Tangier disease is actually caused by mutations in the ABCA1 gene, which impair the production of functional HDL-C particles. This genetic target was only discovered in 1999. In contrast, the search for the cause of Gardner syndrome, a genetic colon polyposis, unfolded differently. Early investigations yielded only weak associations, prompting a broader search effort. This eventually led to the discovery of mutations in the APC gene, a tumor suppressor that plays a central role in controlling cell growth and is strongly linked to the condition. The APC discovery happened in 1991, eight years before the key breakthrough in Tangier disease, despite both diseases receiving a similar number of publications. These contrasting case studies highlight how the disease that initially showed clearer research progress reached its breakthrough much later due to the streetlight effect.

Building on these cases, we turn to a systematic empirical investigation. Our central proposition is that early discoveries of moderate promise can narrow scientific focus and slow the identification of true genetic drivers. In contrast, weaker early findings tend to promote broader exploration and accelerate discovery. The setting offers suggestive parallels with our theoretical framework: just as agents in our model search for valuable projects or participants in the lab look for gems hidden in mountains, scientists navigate a vast genetic landscape in pursuit of scientific breakthroughs.

4.2 Data

DisGeNET Database. We compile a dataset of genetics research from 1980 to 2019 using DisGeNET (v7.0), a comprehensive database of gene–disease links drawn from curated sources and PubMed-indexed publications (Piñero et al., 2020; Tranchero, 2025). Because DisGeNET does not include author information, we supplement it with disambiguated data from Author-ity 2018 (Torvik and Smalheiser, 2021). Additional details on both data sources are provided in Appendix B. Our analysis focuses on articles investigating associations between protein-coding genes and diseases, syndromes, or abnormalities with clear health relevance. For each disease, we record the number of publications along with information on the novel genetic candidates identified each year. To filter out conditions unlikely to have a genetic basis, we restrict the sample to diseases with at least 10 publications over

the study period, but results are fully robust to different cut-offs. The final dataset captures the search and discovery trajectories of 5,519 diseases over a 40-year span.

Measuring the Scientific Value of Genetic Discoveries. Scientists aim to identify genes of high scientific value for each disease. Mirroring our theoretical setup, we classify genetic candidates for a disease into three categories: weak targets (L), middle-value leads (M), and breakthroughs (H). We rely on the score provided by DisGeNET for each gene–disease pair, which ranges from 0 to 1 and summarizes the strength of its experimental evidence. DisGeNET builds this score as a tool to prioritize R&D in pharmaceutical firms towards the most promising drug targets. We provide extensive details on the DisGeNET score and its features in Appendix B.3. For interpretability, we express a gene–disease pair’s scientific value as its percentile within the overall score distribution. Genes below the 60th percentile are classified as low value, those between the 60th and 90th percentiles as medium value, and those above the 90th percentile as high value. These categories closely align with real-world indicators of therapeutic relevance: clinical citations, approved patents, and granted drugs all increase monotonically with our score categories (Appendix Figure B.1).

Data Generated by Past Exploration. Our objective is to assess how the scientific value of known gene candidates shapes subsequent exploration patterns for a given disease. We build on the idea that early exploration provides data that scientists can choose to exploit through repeated studies, rather than search for new candidates. We define the early search window as the period from 1980 to 2000, which marks the first half of our sample and accounts for just 10% of all publications, during which scientists began identifying potential gene targets.⁸ For each disease, we record the highest-scoring gene candidate identified during this period, classifying it as low (L), medium (M), or high (H) based on the categories described above. This captures the genetic targets known to researchers as of 2000. We then examine how the nature of this early data shapes research activity in the second half of the sample period (2000–2019). Figure C.1 provides a stylized overview of this empirical setup.

Dependent Variables. We construct a dataset at the disease level to examine how cross-sectional differences in early data shape subsequent exploration. Our first dependent variable captures whether scientists identified a gene-disease pair with a high DisGeNET score, corresponding to a breakthrough discovery at the group level in our experimental setup. The second dependent variable measures the number of new gene candidates explored after the early search window, allowing us to assess how the scientific promise of early data affects the diversity of follow-on research. To account for variation

⁸However, our results are fully robust to choosing alternative windows (Appendix Tables C.10 and C.11).

in research intensity across diseases, we divide the number of new genes explored by the number of publications. In practice, this variable captures changes in the propensity to explore new genes for a given disease. The third dependent variable measures the number of years required to reach a breakthrough, defined as the number of years since 1980 (the start of our sample period). This group-level delay offers a concrete indication of the societal cost imposed by the streetlight effect. We include the total number of publications focused on each disease as a control to account for variation in research effort. In addition, DisGeNET assigns each condition a set of disease classes based on the MeSH vocabulary, and our data include 536 unique disease class combinations. Disease class captures features such as whether a condition is congenital or acquired. We include disease-class fixed effects to control for unobserved characteristics shared by similar diseases, and cluster standard errors at the disease-class level to account for correlations across related conditions.

Summary Statistics. Table 3 reports descriptive statistics for the 5,519 diseases in our sample. By the year 2000, 10% of diseases show early data pointing to an L target, 32% to an M target, and the remainder to an H target. The H category is less informative for our purposes, as a breakthrough has already occurred in the early exploration window.⁹ On average, it takes 21.8 years from the first publication on a disease to identify a high-value genetic target. Each disease is linked to approximately 295 publications, involving 186 unique principal investigators (PIs), and associated with the discovery of about 130 genes.

5 Empirical Results

5.1 Cross-Sectional Evidence

We begin by examining how the likelihood of a breakthrough varies with early scientific data, comparing outcomes for diseases with information only on low-value genes to those with data on genes of medium or high value. To do this, we estimate the following cross-sectional OLS specification at the disease level:

$$Breakthrough (0/1)_i = \alpha + \beta(Max Found : X_i) + \gamma \mathbf{X}_i + \epsilon_i, \quad (6)$$

where $Breakthrough (0/1)_i$ equals 1 if at least one publication discovers a genetic target with a high DisGeNET score for disease i , and 0 otherwise. The variable $(Max Found : X_i)$ is a categorical indicator for the highest DisGeNET score identified in the early search window, classified as L , M , or H . \mathbf{X}_i is a vector of controls that includes the number of publications on the specific disease,

⁹Note that we do not include a “no data” condition here, as most diseases had seen some level of investment before 2000.

as a proxy for search efforts, and fixed effects for disease class, taking into account broader genetic similarities between related conditions. The results are reported in Panel A of Table 4. While early data on a high-value genetic target mechanically increases the likelihood of a breakthrough, the more interesting comparison lies with medium-value targets. Diseases with early data on medium-value genes are 11 percentage points less likely to experience a breakthrough than those with only low-value initial findings.

One possible explanation for this counterintuitive finding is that the early discovery of a promising—but ultimately suboptimal—genetic target diverts attention from the search for a true breakthrough. Column 2 of Table 4 presents evidence consistent with this mechanism. Using the same specification as in Equation (6), we find that early data on a medium-value target reduces exploration of new genes by almost 20 percentage points. Notably, this drop is nearly half as large as the effect of an early breakthrough itself. Column 3 quantifies the real-world cost of reduced exploration. Identifying a medium-value target early on delays the eventual breakthrough by 1.7 years, which corresponds to an increase of 8% relative to the sample mean of 21.7 years.

Our theoretical framework and behavioral experiments suggest that the streetlight effect weakens with increasing rivalry. Does this prediction hold in our empirical setting? In our experiments, we can directly manipulate the threshold \bar{N} , which represents the number of individuals who can benefit from a project before the payoff erodes. Here, we take \bar{N} as fixed, assuming it is broadly similar across diseases due to common scientific norms of credit allocation. Still, our comparative statics in Propositions 4 and 5 show that even with a constant \bar{N} , increasing N (i.e., the number of competitors) should reduce the streetlight effect. This insight allows us to proxy rivalry empirically using the number of scientists who have studied each disease, echoing recent work on competition in science (Hill and Stein, 2025a). We classify diseases as more or less competitive based on the number of active PIs. Using our cross-sectional specification from Equation (6), we run split-sample regressions for diseases in the top and bottom quartiles of this distribution.

The results of this analysis are presented in Panel B of Table 4. In Column 1, where we restrict the sample to diseases with fewer scientists, we find that early data on a medium-value target reduces the likelihood of a breakthrough by 17 percentage points. By contrast, for diseases with more competition (Column 2), there is no significant change in breakthrough likelihood. A similar pattern holds for exploration activity: early data reduces exploration of new genes by 20% relative to the sample mean in diseases with fewer scientists (Column 3), but this effect disappears when more PIs are engaged

in the search for genetic roots (Column 4). Finally, early data on a tempting genetic target increases the time to breakthrough in diseases with fewer researchers involved (Column 5), while the effect is not significant in more competitive settings (Column 6). Taken together, these results suggest that competition helps offset the streetlight effect that uneven data availability might create.

5.2 Instrumental Variable (IV) Estimates

The empirical patterns are consistent with our theorization and echo concerns raised by scientists about the lack of exploration in this field (Haynes et al., 2018; Stoeger et al., 2018). However, issues about causality remain, since the generation of early data reflects scientists' endogenous exploration choices. As a first step, we note that including fixed effects for disease classes helps control for unobserved characteristics shared across related diseases. Yet, certain disease-specific features could still correlate with the nature and volume of early scientific data, potentially driving our results.

To bolster the causal interpretation of our findings, we leverage the fact that many human genes have orthologous counterparts—that is, genes in other species that share a common ancestor gene and thus retain similar biological sequences and functions. Scientists frequently use animals as models to experimentally study human orthologs at lower cost and with fewer ethical constraints (Li et al., 2017). In particular, genes with orthologs in the commonly used laboratory mice tend to receive more attention from scientists out of sheer convenience (Stoeger et al., 2018). We retrieve information on gene orthology from the National Center for Biotechnology Information (NCBI). In our data, human genes with mouse orthologs appear 2.6 years earlier in scientific publications and are about 27% more likely to have been explored before the year 2000 (Figure 4, Panel A). This confirms that researchers prioritize these genes, and within a disease, information about them emerges earlier (Appendix Table C.1). Yet, nothing ensures that orthologs are equally relevant for every disease. Since the strength of associations between these mouse-overlapping genes and a given disease is effectively unforeseen by scientists, delays stemming from focusing on medium-value orthologs can be attributed to convenience rather than to unobserved disease characteristics.

Building on this intuition, we construct an instrumental variable based on the distribution of orthologous gene candidates. For each disease, we measure the share of orthologous genes classified as medium-value (M) candidates (see Appendix Figure C.2 for a stylized visualization). If the orthologous gene pool contains more medium-value targets for a particular disease, scientists should be more likely to encounter a medium-value discovery early in their exploration. Indeed, our instrument exogenously shifts the probability of identifying a medium-value gene, as confirmed by a strong first-stage

regression (Figure 4, Panel B).¹⁰ For example, Tangier Disease has 27 gene candidates with mouse orthologs, 7 of which are medium-value ($M \text{ share}_{Tangier} = 26\%$); APA01, a medium-value ortholog, was indeed discovered early, in 1982. In contrast, only 2 of Gardner Syndrome’s 23 orthologous genes are of medium-value ($M \text{ share}_{Gardner} = 8\%$), leading scientists to identify the causal APC gene without distraction by medium-value discoveries. Figure C.3 supports this logic across our broader sample of diseases, showing reduced-form evidence linking the M share of ortholog genes to breakthrough likelihood, exploration extent, and discovery delay.

The results of our IV analysis are presented in Table 5. We replicate the same three cross-sectional specifications as before, but now instrument for ($MaxFound : M$) using the share of ortholog genes corresponding to an M target. In this setup, the 2SLS coefficients can be interpreted as a local average treatment effect (LATE), capturing the effect in the subset of diseases for which the instrument shifts the likelihood of identifying an M target. Column 1 presents the first stage of our IV, showing that the M share of ortholog genes is strongly associated with the highest-scoring gene association being M , with an F-statistic of 154. Columns 2 through 4 report the second-stage results for our three main outcomes. Consistent with our earlier findings, we find that early data on a medium-value target reduces the likelihood of a breakthrough (Column 2), decreases the number of new genes explored (Column 3), and increases the delay to a breakthrough (Column 4). This analysis provides additional evidence that the patterns observed in genetics stem from the streetlight effect created by early data (Haynes et al., 2018).

5.3 Exploration Dynamics

In this section, we dig deeper into the exploration dynamic to bolster confidence in our proposed mechanism. If early data on medium-value genetic targets indeed crowds out exploration, we should see a drop in research efforts aimed at discovering new genes in the years following a medium-value gene. To test this idea, we construct a panel at the disease-year level. While our earlier analyses relied on cross-sectional estimates at the disease level, this alternative approach allows us to track how exploration patterns change over time. For each disease, we count the number of new genes investigated in a given year, along with the total number of publications as a proxy for research effort. Appendix

¹⁰The exclusion restriction requires that, conditional on disease class and overall research intensity, the instrument influences breakthrough timing and exploration only through its effect on the probability that early work identifies a medium value target, rather than through intrinsic features of the disease. While this remains an identifying assumption, Appendix Table C.2 shows that diseases that receive more publications, patents, or approved drugs do not systematically exhibit higher or lower values of the instrument, which reduces the concern that it is simply proxying for underlying therapeutic potential.

Table C.3 presents descriptive statistics at the disease-year level. On average, each disease receives 7.4 publications per year focused on its genetic underpinnings, typically leading to the exploration of 3.3 new genes each year.

We then estimate the following event study specification using OLS:

$$Group\ Exploration_{i,t} = \alpha + \sum_z \beta_t Medium\ Gene_i \times 1(z) + \gamma \mathbf{X}_{i,t} + \epsilon_{i,t}, \quad (7)$$

where $Group\ Exploration_{i,t}$ denotes the number of new genes explored for disease i in year t , normalized by the number of articles published. $Medium\ Gene_i \times 1(z)$ the number of years that have elapsed since a medium-value association was first discovered for disease i , and $\mathbf{X}_{i,t}$ is a vector of controls that include disease fixed effects, year fixed effects, and the number of papers published each year. For the small number of diseases with multiple medium-value genes, we define the time lags relative to the discovery of the first one. To account for the mechanical uptick in exploration during the year of discovery, we exclude the focal gene and its corresponding publication from our calculations, but our results are robust to their inclusion.

Panel A of Figure 5 plots the regression coefficients. The results show an immediate, significant, and persistent drop in exploration following the discovery of a medium-value genetic target. Reassuringly, there is no evidence of pre-trends, suggesting that the observed decline is indeed driven by the discovery itself. In Panel B, we re-estimate the specification in Equation (7) and find a similar pattern: research efforts on new genes also decline after the discovery of a high-value target. Appendix Table C.4 reports the corresponding estimates from a difference-in-differences specification. We find that yearly exploration of new genes drops by 24% relative to the sample mean after a medium-value target is identified. The effect is even larger following the discovery of a high-value target, with exploration falling by around 35%. Results are consistent using alternative difference-in-differences estimators (Appendix Table C.5). Taken together with the IV results, these estimates offer additional support for the predictions of our theoretical framework.

5.4 Robustness

We assess the robustness of our results by relaxing several key choices in the main specification. First, while we excluded diseases with very few publications to focus on those more likely to have genetic roots, our results hold under alternative sample cut-offs (Appendix Table C.6). Similarly, excluding the top 1% of most-studied diseases does not change the findings (Appendix Table C.7). Second, we defined payoffs based on percentiles of the DisGeNET score. While our definitions map into

real-world outcomes (Appendix Figure B.1), changing the percentiles used to define an M genetic discovery does not affect our results (Appendix Tables C.8 and C.9). Third, redefining the early search window yields similar results (Appendix Table C.10). Appendix Table C.11 shows robustness to a disease-specific definition, where “early” refers to the years before the first 10% of publications for each disease. Finally, the findings remain stable under alternative windows for tracking exploration dynamics following an M discovery (Appendix Table C.12).

One potential concern is that focusing on a medium-value gene could be a rational choice when there is ambiguity about whether a high-value target exists at all. This might partly explain the drop in exploration following the discovery of an M . To address this, we draw on the genetic relationships between diseases. The MeSH vocabulary defines hierarchical linkages between diseases based on shared etiology, biological mechanisms, and other biomedical features. Using this classification, we restrict the analysis to diseases closely related to conditions where a breakthrough (H) has already occurred. In these cases, the existence of valuable targets is less ambiguous, as related diseases often share underlying biological processes.¹¹ Re-estimating the event study specification in Equation (7), we find consistent results: as shown in Appendix Figure C.4 and Table C.13, data on a medium-value gene still dampens exploration, even within this subset of diseases.

Relatedly, it is possible that our cross-sectional results reflect the absence of valuable genetic targets, rather than suboptimal exploration behavior. To test this, we narrow our analysis to diseases that had a high-value genetic association identified by 2019. These results are presented in Appendix Table C.14. While this restriction prevents us from estimating effects on group breakthroughs, we still observe longer delays for diseases where early data pointed to an M -value target. We find no change in exploration over the full sample period, likely because all diseases in this sample eventually saw a breakthrough and, by definition, received some level of exploration. Still, we detect a significant decline in exploration activity in the years immediately following the early M discovery.

Finally, one reason scientists might continue to focus on M candidates is the prospect of positive spillovers that could benefit research on related diseases. These spillovers could come from complementary insights, such as new methods or a deeper understanding of protein function and genetics. If genes classified as M in one disease often end up as H candidates in related diseases, then continued

¹¹For instance, Ulcerative Colitis [MeSH tree code: C06.405.469.432.249] and Crohn’s Disease [MeSH tree code: C06.405.469.432.500] are “sibling” sub-branches of the “parent” disease Inflammatory Bowel Diseases [MeSH tree code: C06.405.469.432]. Once a gene is identified as a high-value target for Crohn’s there is a higher chance that a breakthrough exists for Ulcerative Colitis (which could either be the same gene or another one).

focus on them might be rational. To evaluate this possibility, we test whether a gene is more likely to be a breakthrough for a given disease when it is classified as an M in a related disease. The results are presented in Panel A of Appendix Table C.15. The effect is small and only slightly larger than when the gene is classified as L , and much smaller than when it is classified as H in a sibling disease. By contrast, as Panel B shows, genes classified as M for one disease are likely to remain M in related diseases. This suggests that spillovers are limited in scope and are unlikely to justify continued attention to M candidates.

6 Conclusion

In this paper, we examine the paradoxical role of data provision in shaping innovative search, a dynamic we refer to as the “streetlight effect.” Our theoretical model shows that access to partial data on past successes can narrow the search space and trigger free-riding, ultimately reducing the diversity of exploration and hampering breakthrough discoveries. This prediction is supported by our empirical findings. In our lab experiments, revealing data on a medium-value project lowered group payoffs by 5% and reduced the likelihood of a breakthrough by 56% compared to the no-data condition. We apply this insight to help explain the patterns observed in genetic research. Our approach includes multiple research designs, including an IV strategy based on exogenous genetic overlaps between human and mouse genes. The results show that diseases with early data on a middle-value target are, on average, 16 percentage points less likely to yield breakthroughs, with discoveries delayed by nearly three years due to reduced exploration. We also find that competition moderates these effects by lowering the attractiveness of known options and breaking the cycle of low data generation. Taken together, our theoretical, experimental, and empirical evidence highlights how the streetlight effect shapes the direction of innovative search.

Our findings challenge the conventional belief that more data is always better for innovation. When data is incomplete and narrowly focused, as in our setting, it can unintentionally steer researchers toward suboptimal projects. Our evidence from genetics highlights how this pattern can emerge endogenously in decentralized and parallel exploration endeavors such as scientific research. This has important implications for policymakers and funding agencies involved in data creation and dissemination, whose goal should be to provide broad “floodlights” that illuminate the entire search space. Our findings reinforce the value of publicly funded, comprehensive mapping initiatives such as the Human Genome Project (Williams, 2013) and Landsat satellite imagery (Nagaraj, 2022), which serve as shared data infrastructure for scientific discovery. They also highlight the importance of strengthening institutions

such as the U.S. Census Bureau’s FSRDCs (Nagaraj and Tranchero, 2024), which enable research access to existing large-scale datasets at relatively low public cost.

More broadly, our evidence suggests that in environments where data availability is uneven, setting aside existing information can promote breakthrough innovation. Our findings lend support to corporate practices like skunkworks, where firms intentionally restrict the internal diffusion of early R&D results. They also underscore the value of delaying the release of intermediate project information unless there is strong evidence that the project represents a high-value lead (Boudreau and Lakhani, 2015). As innovation and decision-making become increasingly data-driven, it is important to recognize that technologies like AI are often trained on uneven historical data. This can inadvertently narrow the scope of exploration by reproducing the streetlight effect (Kim, 2023). While existing work has focused on the risk of false positives in AI predictions (Tranchero, 2024), our evidence suggests that the risk of false negatives in data-driven innovation may be even greater. At the same time, AI enables initiatives like AlphaFold, which provide unfiltered predictions supporting discovery beyond the bounds of known data. Understanding the nuanced implications of AI for innovation is an exciting direction for future research.

While our study draws strength from combining theoretical modeling, laboratory experimentation, and empirical analysis, there remain several opportunities for further improvement. One direction would be to extend the current two-period framework into a continuous learning model, which would better capture the iterative and dynamic nature of innovation. Our model could also be extended to explore how control rights in organizations might help coordinate search efforts and prevent herding (Aghion et al., 2008). Another promising avenue lies in broadening our definition of data to include dimensions such as precision, informativeness, and bias, all of which are likely to shape search behavior in meaningful ways. The observational analysis, while strengthened by an IV approach, could also be complemented by research designs that introduce direct experimental variation in the data provided. Expanding the analysis to consider a broader set of innovation outcomes across diverse domains would further enhance the generalizability of our findings beyond the setting studied.

References

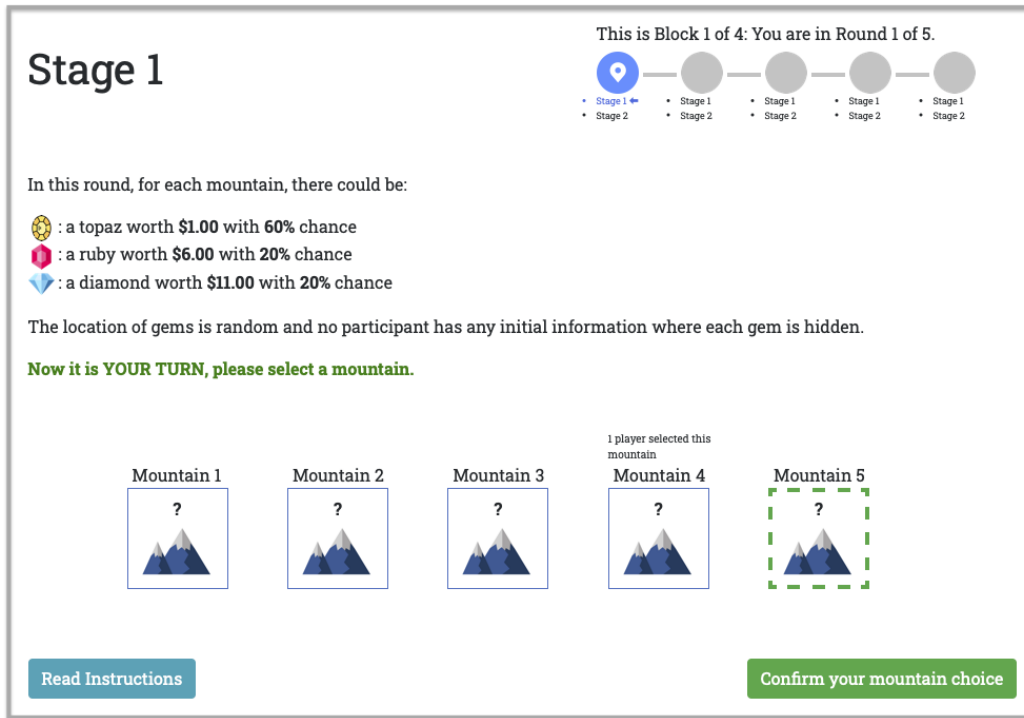
- AGHION, P., M. DEWATRIPONT, AND J. C. STEIN (2008): “Academic freedom, private-sector focus, and the process of innovation,” *The RAND Journal of Economics*, 39, 617–635.
- AGRAWAL, A., J. McHALE, AND A. OETTL (2024): “Artificial intelligence and scientific discovery: A model of prioritized search,” *Research Policy*, 53, 104989.
- AZOULAY, P., J. S. GRAFF ZIVIN, AND G. MANSO (2011): “Incentives and creativity: Evidence from the academic life sciences,” *The RAND Journal of Economics*, 42, 527–554.
- BERGEMANN, D. AND A. BONATTI (2019): “Markets for information: An introduction,” *Annual Review of Economics*, 11, 85–107.
- BERGEMANN, D. AND J. VALIMAKI (2008): “Bandit problems,” in *The New Palgrave Dictionary of Economics*, 2nd ed., Macmillan Press, 336–340.
- BESSEN, J., S. M. IMPINK, L. REICHENSBERGER, AND R. SEAMANS (2022): “The role of data for AI startup growth,” *Research Policy*, 51, 104513.
- BHATTACHARYA, J. AND M. PACKALEN (2020): “Stagnation and scientific incentives,” *NBER Working Paper 26752*.
- BLOOM, N., C. I. JONES, J. VAN REENEN, AND M. WEBB (2020): “Are ideas getting harder to find?” *American Economic Review*, 110, 1104–1144.
- BOBTSCHEFF, C., J. BOLTE, AND T. MARIOTTI (2017): “Researcher’s dilemma,” *The Review of Economic Studies*, 84, 969–1014.
- BOLTON, P. AND C. HARRIS (1999): “Strategic experimentation,” *Econometrica*, 67, 349–374.
- BOUDREAU, K. J. AND K. R. LAKHANI (2015): “‘Open’ disclosure of innovations, incentives and follow-on reuse: Theory on processes of cumulative innovation and a field experiment in computational biology,” *Research Policy*, 44, 4–19.
- BOYCE, J. R., D. M. BRUNER, AND M. MCKEE (2016): “Strategic experimentation in the lab,” *Managerial and Decision Economics*, 37, 375–391.
- CAO, R., R. KONING, AND R. NANDA (2024): “Sampling bias in entrepreneurial experiments,” *Management Science*, 70, 7283–7307.
- CHEN, D. L., M. SCHONGER, AND C. WICKENS (2016): “oTree—An open-source platform for laboratory, online, and field experiments,” *Journal of Behavioral and Experimental Finance*, 9, 88–97.
- COCKBURN, I. M., R. HENDERSON, AND S. STERN (2019): “The impact of artificial intelligence on innovation,” *The Economics of Artificial Intelligence: An Agenda*, 115–152.
- EDERER, F. AND G. MANSO (2013): “Is pay for performance detrimental to innovation?” *Management Science*, 59, 1496–1513.
- EDWARDS, A. M., R. ISSERLIN, G. D. BADER, S. V. FRYE, T. M. WILLSON, AND F. H. YU (2011): “Too many roads not taken,” *Nature*, 470, 163–165.
- EWENS, M., R. NANDA, AND M. RHODES-KROPP (2018): “Cost of experimentation and the evolution of venture capital,” *Journal of Financial Economics*, 128, 422–442.
- FARBOODI, M. AND L. VELDKAMP (2020): “Long-run growth of financial data technology,” *American Economic Review*, 110, 2485–2523.
- GATES, A. J., D. M. GYSI, M. KELLIS, AND A.-L. BARABÁSI (2021): “A wealth of discovery built on the Human Genome Project—by the numbers,” *Nature*, 590, 212–215.

- GORDON, R. (2016): *The rise and fall of American growth: The US standard of living since the civil war*, Princeton University Press.
- GREINER, B. (2015): “Subject pool recruitment procedures: organizing experiments with ORSEE,” *Journal of the Economic Science Association*, 1, 114–125.
- HAYNES, W. A., A. TOMCZAK, AND P. KHATRI (2018): “Gene annotation bias impedes biomedical research,” *Scientific Reports*, 8, 1362.
- HENRICH, J., S. J. HEINE, AND A. NORENZAYAN (2010): “Most people are not WEIRD,” *Nature*, 466, 29–29.
- HENRY, E., M. LOSETO, AND M. OTTAVIANI (2022): “Regulation with experimentation: Ex ante approval, ex post withdrawal, and liability,” *Management Science*, 68, 5330–5347.
- HILL, R. AND C. STEIN (2025a): “Race to the bottom: Competition and quality in science,” *The Quarterly Journal of Economics*, 140, 1111–1185.
- (2025b): “Scooped! Estimating rewards for priority in science,” *Journal of Political Economy*, 133.
- HOELZEMANN, J. AND N. KLEIN (2021): “Bandits in the lab,” *Quantitative Economics*, 12, 1021–1051.
- (2025): “Breakdowns in the lab,” *University of Vienna and University of Montreal*.
- HOLT, C. A. AND S. K. LAURY (2002): “Risk aversion and incentive effects,” *American Economic Review*, 92, 1644–1655.
- HÖRNER, J., N. KLEIN, AND S. RADY (2022): “Overcoming free-riding in bandit games,” *The Review of Economic Studies*, 89, 1948–1992.
- JONES, C. I. AND C. TONETTI (2020): “Non-rivalry and the economics of data,” *American Economic Review*, 110, 2819–2858.
- KAO, J. (2024): “Charted territory: Mapping the cancer genome and R&D decisions in the pharmaceutical industry,” *UCLA Anderson*.
- KEHOE, A. AND V. TORVIK (2019): “Predicting controlled vocabulary based on text and citations: Case studies in medical subject headings in MEDLINE and patents,” *University of Illinois at Urbana-Champaign*.
- KELLER, G., S. RADY, AND M. CRIPPS (2005): “Strategic experimentation with exponential bandits,” *Econometrica*, 73, 39–68.
- KERR, W. R., R. NANDA, AND M. RHODES-KROPF (2014): “Entrepreneurship as experimentation,” *Journal of Economic Perspectives*, 28, 25–48.
- KIM, S. (2023): “Shortcuts to innovation: The use of analogies in knowledge production,” *Columbia Business School*.
- KLEIN, N. AND S. RADY (2011): “Negatively correlated bandits,” *The Review of Economic Studies*, 78, 693–732.
- KRIEGER, J. L. (2021): “Trials and terminations: Learning from competitors’ R&D failures,” *Management Science*, 67, 5525–5548.
- LEVINTHAL, D. A. (1997): “Adaptation on rugged landscapes,” *Management Science*, 43, 934–950.
- LI, D., P. AZOULAY, AND B. N. SAMPAT (2017): “The applied value of public investments in biomedical research,” *Science*, 356, 78–81.
- MANSO, G. (2016): “Experimentation and the returns to entrepreneurship,” *The Review of Financial Studies*, 29, 2319–2340.
- MARCH, J. G. (1991): “Exploration and exploitation in organizational learning,” *Organization Science*, 2, 71–87.

- 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, M. R., H. TIPNEY, J. L. PAINTER, J. SHEN, P. NICOLETTI, Y. SHEN, A. FLORATOS, P. C. SHAM, M. J. LI, J. WANG, ET AL. (2015): “The support of human genetic evidence for approved drug indications,” *Nature Genetics*, 47, 856–860.
- NGUYEN, D.-T., S. MATHIAS, C. BOLOGA, S. BRUNAK, N. FERNANDEZ, A. GAULTON, A. HERSEY, J. HOLMES, L. J. JENSEN, A. KARLSSON, ET AL. (2017): “Pharos: Collating protein information to shed light on the druggable genome,” *Nucleic Acids Research*, 45, D995–D1002.
- OLEA, J. L. M. AND C. PFLUEGER (2013): “A robust test for weak instruments,” *Journal of Business & Economic Statistics*, 31, 358–369.
- 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.
- RODGERS, G., C. AUSTIN, J. ANDERSON, A. PAWLYK, C. COLVIS, R. MARGOLIS, AND J. BAKER (2018): “Glimmers in illuminating the druggable genome,” *Nature Reviews Drug Discovery*, 17, 301–302.
- 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.
- TORVIK, V. AND N. SMALHEISER (2021): “Authority 2018—PubMed author name disambiguated dataset,” *University of Illinois Urbana-Champaign*.
- TRANCHERO, M. (2024): “Finding diamonds in the rough: Data-driven opportunities and pharmaceutical innovation,” *The Wharton School*.
- (2025): “Data-driven search and the birth of theory: Evidence from genome-wide association studies,” *The Wharton School*.
- VERMOREL, J. AND M. MOHRI (2005): “Multi-armed bandit algorithms and empirical evaluation,” in *Machine Learning: ECML 2005*, ed. by J. Gama, R. Camacho, P. B. Brazdil, A. M. Jorge, and L. Torgo, Berlin, Heidelberg: Springer Berlin Heidelberg, 437–448.
- WILLIAMS, H. L. (2013): “Intellectual property rights and innovation: Evidence from the human genome,” *Journal of Political Economy*, 121, 1–27.

7 Tables and Figures

Panel A: User Interface



Panel B: Examples of No-Data Condition and Data Conditions

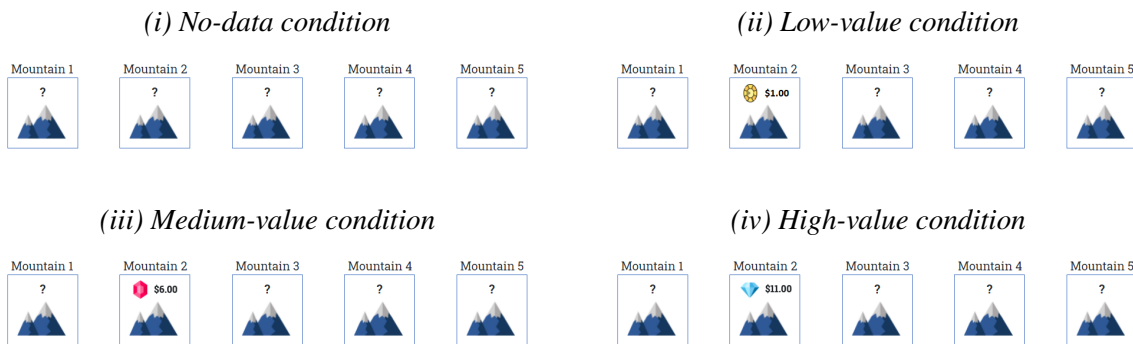


Figure 1: Experimental Platform.

Note: This figure shows the interface participants saw during our online experiment. Panel A illustrates the platform as it appeared in the no-data condition. In this example, Mountain 4 was selected by one other participant, while the user chose Mountain 5. Note that the dollar value of the gems changes in every round and is displayed on the left. Panel B presents the four experimental conditions. In the data condition, participants are shown the value of the gem hidden behind one randomly selected mountain—this could be the medium, the high, or one of the low outcomes. .

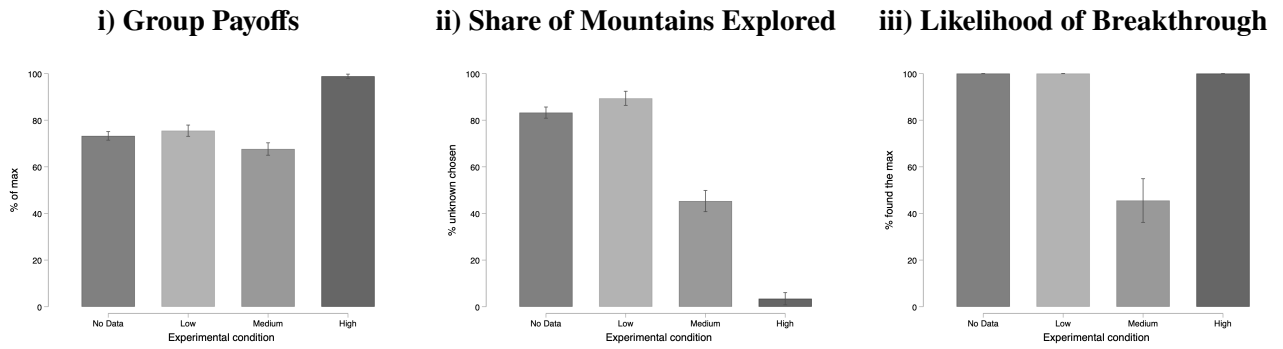


Figure 2: Round Outcomes by Experimental Condition.

Note: Figure (i) displays the average group payoffs per round, by experimental condition. Payoffs are calculated as a share of the maximum possible payoff possible in each round. Figure (ii) shows the average share of unmapped mountains selected per round, by experimental condition. Figure (iii) reports the proportion of rounds in which the maximum payoff was uncovered, by experimental condition. Error bars indicate 95% confidence intervals. See text for more details.

Table 1: Round-Level Experimental Outcomes.

	Group Payoff	Group Exploration	Group Breakthrough
	(1) Group Earnings (€)	(2) Options Explored (%)	(3) Found Maximum (0/1)
High	44.520*** (0.894)	-81.330*** (3.049)	-1.500 (4.277)
Low	1.545 (1.228)	4.598 (3.120)	-2.045 (3.631)
Medium	-3.133*** (0.670)	-38.634*** (2.577)	-56.338*** (5.057)
Session FE	Yes	Yes	Yes
Block order FE	Yes	Yes	Yes
Payoff structure FE	Yes	Yes	Yes
Observations	480	480	364

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the session level in parentheses.

Estimates from OLS models. The unit of analysis is the group-round level (480 rounds in total). Column 2 includes only the rounds in which at least one diamond was present (364 rounds). In all models, payoffs are non-rival if multiple agents choose the same project. *Group Earnings*= sum of payoffs in a group-round (in Euros); *Options Explored*= share of unknown mountains explored in the round; *Found Maximum:0/1=1* if the location of the maximum was found by any participant. The excluded category is the control condition without data. See text for more details.

Table 2: Round-Level Outcomes of the Experiment with Payoff Rivalry.

Panel A: Intermediate Payoff Rivalry

	Group Payoff	Group Exploration	Group Breakthrough
	(1) Group Earnings (€)	(2) Options Explored (%)	(3) Found Maximum (0/1)
High	19.048* (2.253)	-15.520* (1.918)	-3.681 (1.712)
Low	-7.212 (2.803)	8.759 [†] (2.939)	-4.162 (5.394)
Medium	-1.118 (1.714)	-19.870*** (0.524)	-24.470 [†] (5.726)
Session FE	Yes	Yes	Yes
Block order FE	Yes	Yes	Yes
Payoff structure FE	Yes	Yes	Yes
Observations	120	120	90

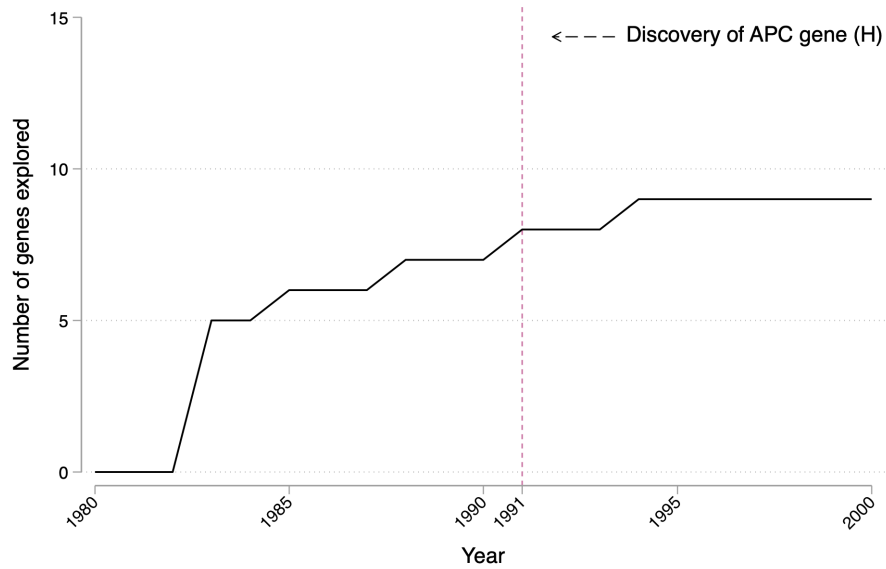
Panel B: Extreme Payoff Rivalry

	Group Payoff	Group Exploration	Group Breakthrough
	(1) Group Earnings (€)	(2) Options Explored (%)	(3) Found Maximum (0/1)
High	3.698 (2.400)	2.74e-14 (2.67e-14)	-2.50e-14 (3.70e-14)
Low	-6.888 [†] (1.753)	3.10e-14 (2.83e-14)	-1.15e-14 (3.75e-14)
Medium	3.345* (0.364)	1.02e-14 (2.65e-14)	-2.53e-14 (3.41e-14)
Session FE	Yes	Yes	Yes
Block order FE	Yes	Yes	Yes
Payoff structure FE	Yes	Yes	Yes
Observations	120	120	90

Note: [†] $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Estimates from OLS models. The unit of analysis is the group-round level. In Panel A, payoffs exhibit intermediate rivalry ($\bar{N} = 3$), meaning that once three agents have selected the same mountain in a given period, any additional agents choosing that mountain will receive a payoff of zero. In Panel B, payoffs exhibit extreme rivalry ($\bar{N} = 1$), where only the first agent selecting a mountain earns a positive payoff, while any subsequent agents choosing the same mountain receive a payoff of zero. *Group Earnings*= sum of payoffs in a group-round (in Euros); *Options Explored*= share of unknown mountains explored in the round; *Found Maximum:0/1*=1 if the location of the maximum was found by any participant. The excluded category is the control condition without data. Note that coefficients in columns (2) and (3) of Panel B are extremely small because the experimental conditions achieve results barely distinguishable from the excluded category of no data. See text for more details.

Panel A: Gardner's Syndrome



Panel B: Tangier's Disease

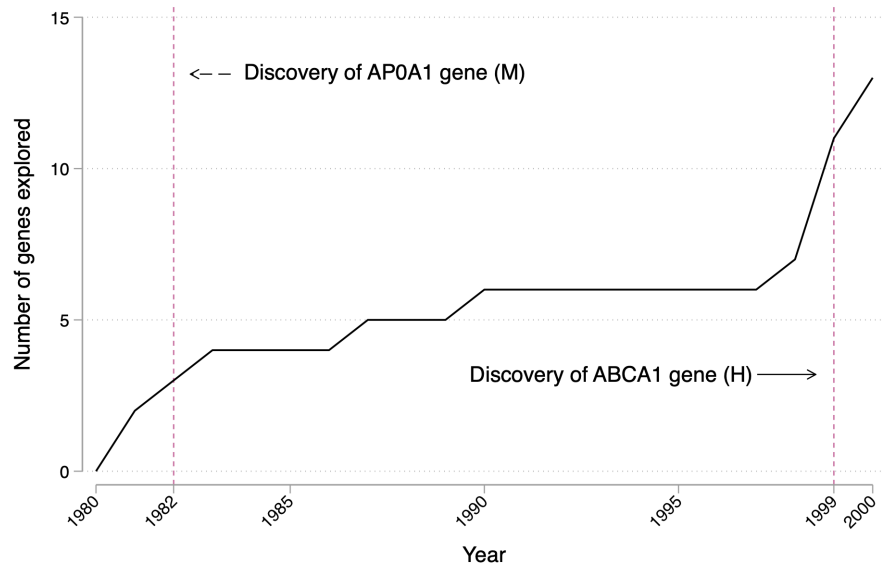


Figure 3: Two Case Studies in Search for the Genetic Origins of Human Diseases.

Note: The solid black line shows the cumulative number of gene candidates explored for the disease up to each year. Panel A displays data for Gardner's syndrome, with the vertical line marking the year the association with the APC gene was discovered (DisGeNET score in the 99th percentile). Panel B shows data for Tangier's disease. The first vertical line marks the discovery of the APOA1 association (DisGeNET score in the 60th percentile), while the second marks the discovery of the ABCA1 association (DisGeNET score in the 99th percentile). All other genes explored were below the 60th percentile of the DisGeNET score.

Table 3: Descriptive Statistics of the DisGeNET Database.

	Mean	Median	Sd	Min	Max	N
Max Found: Low (0/1)	0.10	0.00	0.30	0	1	5519
Max Found: Medium (0/1)	0.32	0.00	0.47	0	1	5519
Max Found: High (0/1)	0.58	1.00	0.49	0	1	5519
Year of First Low Score	1991.45	1992.00	5.58	1980	2000	1530
Year of First Medium Score	1993.71	1994.00	7.42	1980	2019	2890
Year of First High Score	1994.98	1995.00	8.13	1980	2019	3964
Delay (Years since 1980)	21.75	18.00	12.82	0	39	5519
Total Publications	294.63	48.00	1983.81	9	94470	5519
Total Genes Discovered	129.95	32.00	394.35	1	8545	5519
New Genes per Paper	0.73	0.72	0.50	0	9	5519
Total PIs on disease	186.34	38.00	1042.30	5	43749	5519

Note: This table presents cross-sectional descriptive statistics for our sample at the disease level. *Max Found: Low:* 0/1=1 if the gene with the highest DisGeNET score found during the early exploration period is classified as *L*. *Max Found: Medium:* 0/1=1 if the gene with the highest DisGeNET score found during the early exploration period is classified as *M*. *Max Found: High:* 0/1=1 if the gene with the highest DisGeNET score found during the early exploration period is classified as *H*. *Year of First Low Score* = the year of the first discovery involving a gene in the *L* category. *Year of First Medium Score:* the year of the first discovery involving a gene in the *M* category. *Year of First High Score:* the year of the first discovery involving a gene in the *H* category. *Delay (Years since 1980)* = the number of years elapsed before any *H* gene is discovered for the disease. *Total Publications* = the number of publications about the disease during the sample period (1980-2019). *Total Genes Discovered* = the number of genes explored for the disease during the sample period (1980-2019). *New Genes per Publication* = the number of new genes explored per scientific publication during the sample period (1980-2019). *Total PIs* = the number of unique principal investigators (PIs) that have studied the disease during the sample period (1980-2019). See text for more details.

Table 4: Disease-Level Outcomes of Genetic Search.

Panel A: Full Sample

	Group Breakthrough		Group Exploration		Group Delay	
	(1) High-Value Gene (0/1)		(2) New Genes/Papers		(3) Years From 1980	
Max Found: M	-0.105** (0.033)		-0.144*** (0.023)		1.743*** (0.519)	
Max Found: H	0.514*** (0.042)		-0.261*** (0.028)		-20.371*** (0.692)	
Disease Class FE	Yes		Yes		Yes	
Count of Publications	Yes		Yes		Yes	
N	4760		4760		4760	

Panel B: Split Samples

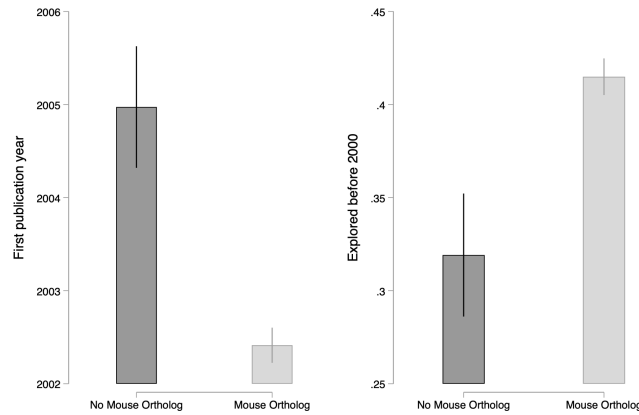
	Group Breakthrough		Group Exploration		Group Delay	
	High-Value Gene (0/1)		New Genes/Papers		Years From 1980	
	(1) No	(2) Yes	(3) No	(4) Yes	(5) No	(6) Yes
High Competition:						
Max Found: M	-0.165*** (0.0433)	-0.0439 (0.154)	-0.144** (0.0510)	0.0779 (0.0972)	2.492*** (0.637)	2.296 (2.973)
Max Found: H	0.516*** (0.0470)	0.541*** (0.159)	-0.132 [†] (0.0790)	-0.00319 (0.0949)	-18.66*** (0.674)	-21.45*** (2.992)
Disease Class FE	Yes	Yes	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes	Yes	Yes
N	1106	1236	1106	1236	1106	1236

Note: [†] $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Standard errors clustered at the disease-class level in parentheses.

Estimates from OLS models. The unit of analysis is the disease level. In Panel A, we report estimates from the full sample of diseases. For each human disease, we identify the highest DisGeNET score among all genes discovered during the exploration period (i.e., before 2000). In Panel B, we report split-sample results that test how our results vary between diseases with more or less competition. Columns (1), (3), and (5) present results for diseases with a bottom-quartile number of principal investigators during the exploration window, while Columns (2), (4), and (6) show results for those with a top-quartile number. We categorize the maximum scores as follows: scores below the 60th percentile are labeled L , those between the 60th and 90th percentiles as M , and those above the 90th percentile as H . *High-Value Gene*: 0/1=1 if any H candidate was discovered for the disease. *New Genes/Papers*= the number of new genes explored per scientific publication in the years following the exploration period. *Years From 1980*= the number of years until the first H candidate is discovered. In all models, diseases in category L serve as the reference group. We include disease-class fixed effects and control for the number of publications post-2000. See text for more details.

Panel A: Genes with a Mouse Orthologs are Explored Earlier



Panel B: First-Stage Evidence for the Instrumental Variable

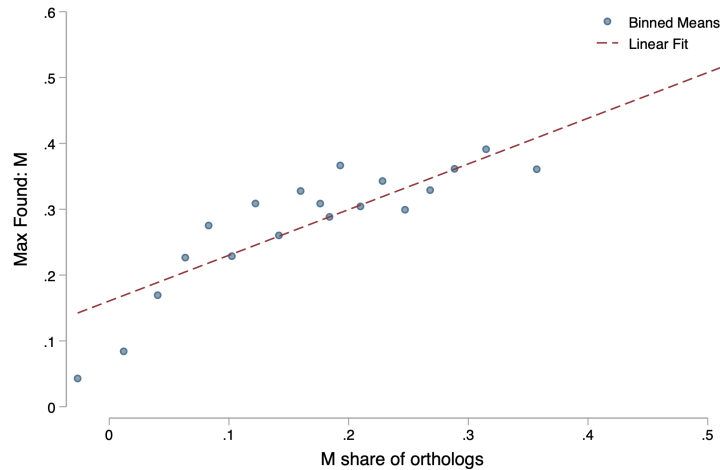


Figure 4: Visual Evidence for Our Instrumental Variable Strategy.

Note: Panel A provides evidence at the gene level that early research tends to focus on genes with mouse orthologs. Each chart shows OLS estimates and 95% confidence intervals estimated from a regression. *First year*= the first year a study exploring a given gene is published. *Explored before 2000*: 0/1=1 if the gene was explored before the year 2000 for at least one disease. Panel B provides a binscatter of the first stage of our disease-level instrumental variable in Table 5. *M Share of Orthologs*: share of orthologous genes (i.e., those with a mouse ortholog) that fall into the *M* category for each disease. (*Max Found: M*): 0/1=1 if the maximum DisGeNET score found during the exploration period is classified as *M*. See text for more details

Table 5: Instrumental Variable Evidence from Human-Mouse Gene Orthologs.

	First Stage	Second Stage		
	Max Found: M (1)	High-Value Gene (0/1) (2)	New Genes/Papers (3)	Years From 1980 (4)
<i>M</i> Share of Orthologs	0.694*** (0.0559)			
Max Found: <i>M</i>		-0.600*** (0.0567)	-0.847*** (0.197)	15.93*** (2.132)
F-Statistic (First Stage)	154.12			
Disease Class FE	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes
N	4757	4757	4757	4757

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease-class level in parentheses. We report the effective first-stage F statistic from Olea and Pflueger (2013).

Estimates from 2SLS models. The sample is at the disease level. For each human disease, we identify the highest DisGeNET score among all genes discovered during the exploration period (i.e., before 2000). To construct our instrument, we calculate the share of each disease's orthologous gene candidates (i.e., those with a mouse ortholog) that fall into the *M* category. We categorize the maximum scores as follows: scores below the 60th percentile are labeled *L*, those between the 60th and 90th percentiles as *M*, and those above the 90th percentile as *H*. *High-Value Gene*: 0/1=1 if any *H* candidate was discovered for the disease. *New Genes/Papers*= the number of new genes explored per scientific publication in the years following the exploration period. *Years From 1980*= the number of years until the first *H* candidate is discovered. In all models, diseases in categories *L* and *H* serve as the reference group. We include disease-class fixed effects and control for the number of publications post-2000. See text for more details.

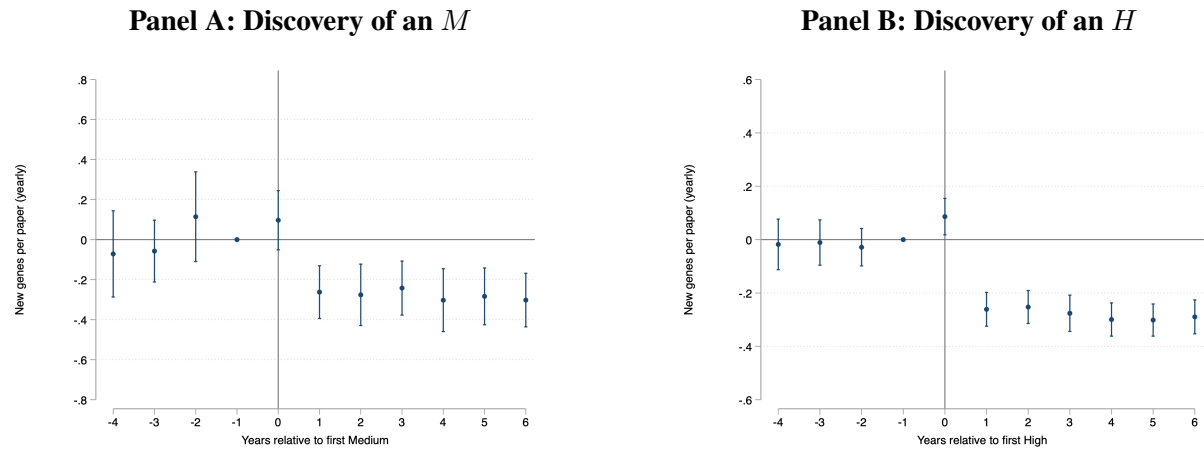


Figure 5: Dynamic Effects of the Discovery of an M or H Genetic Target on Exploration.

Note: Panel A plots OLS estimates and 95% confidence intervals from an event study design that explores how genetic exploration in each disease evolves after the discovery of the first medium-value genetic target. Panel B plots analogous estimates for the discovery of the first high-value genetic target. For each human disease, we classify DisGeNET scores below the 60th percentile as a “low” gene discovery, scores between the 60th and 90th percentile as a “medium” gene discovery, and scores above the 90th percentile as a “high” (or breakthrough) gene discovery. Standard errors are clustered at the disease class level. See text for more details.