The Streetlight Effect in Data-Driven Exploration

Online Appendix: Additional Results

Johannes Hoelzemann
University of Vienna

Gustavo Manso
UC Berkeley-Haas

Abhishek Nagaraj
UC Berkeley-Haas & NBER

Matteo Tranchero
UC Berkeley-Haas

April 21, 2024

A Logistics of the Experiment

Figure A.1 summarizes how our experimental sessions unfolded. When participants join, they are assigned either to a data or to a no-data condition.[20] The experiment begins when a total of ten players are assigned to the same experimental set. Then, from each of these experimental sets, two groups of five people are randomly drawn to play the first five rounds (what we labeled as "block"). At the end of the block, the composition of the two groups is randomly reshuffled, and a second block of five rounds is played. This procedure is repeated a total of four times so that each player ends up playing exactly twenty rounds. The order of blocks seen by participants in different experimental sessions is random. The payoff structure changes each round according to a pre-recorded script generated stochastically so that the actual payoffs of each round appear random for the player. Similarly, the specific order in which specific gems are revealed in the treatment condition is generated by a random script before the experiment begins.

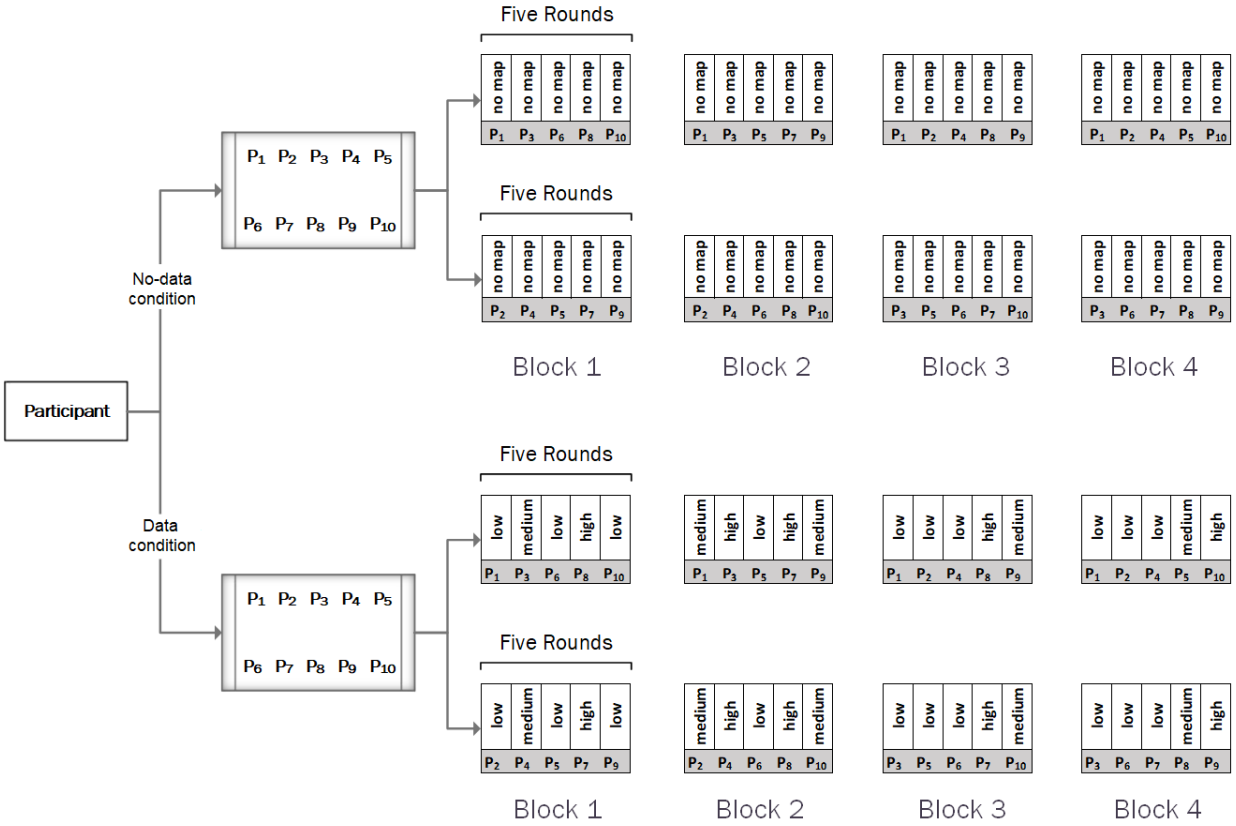


Figure A.1: Flowchart of the experimental setup.

Note: This figure provides an overview of the experiment for one actual session that took place in September 2022.

[20] Not in every experimental session there was a no data condition, in which case the players would be randomly split (and then reshuffled) across two distinct data conditions.

Table A.1 presents the main descriptive statistics for the experimental data, shown separately by treatment condition. Overall, the table already shows our main results in terms of payoffs and discovery while also reassuring that player characteristics were well-balanced across conditions.

Table A.1: Descriptive statistics of the experimental data.

	N	Mean	SD	Median	Min	Max
Individual Payoff (Share)						
Low	2870	72.24	18.63	62.50	8	100
Median	1270	57.78	19.01	58.33	8	100
High	1260	98.14	10.91	100.00	8	100
No Data	1600	68.05	18.93	62.50	8	100
I(Individual found max)						
Low	2870	0.98	0.15	1.00	0	1
Median	1270	0.30	0.46	0.00	0	1
High	1260	0.98	0.13	1.00	0	1
No Data	1600	0.94	0.24	1.00	0	1
I(Group found max)						
Low	574	1.00	0.00	1.00	1	1
Median	254	0.46	0.50	0.00	0	1
High	252	1.00	0.00	1.00	1	1
No Data	320	0.99	0.08	1.00	0	1
Players Age						
Low	2870	22.94	4.43	22.00	18	65
Median	1270	22.65	4.13	22.00	18	57
High	1260	22.50	4.12	21.00	18	59
No Data	1600	24.71	6.18	23.00	18	55
Female Players						
Low	2870	0.65	0.48	1.00	0	1
Median	1270	0.66	0.47	1.00	0	1
High	1260	0.65	0.48	1.00	0	1
No Data	1600	0.71	0.45	1.00	0	1

Note: The table presents descriptive statistics on the 7000 participants in the 1400 rounds of the experiment. *Individual Payoff Share*= individual round payoffs as a share of the maximum achievable; *I(Individual found max)*:0/1=1 if the location of the maximum was found by the participant; *I(Group found max)*:0/1=1 if the location of the maximum was found by at least one participant in the round; *Players Age*= age of the participant at the time of the experiment, in years; *Female Players*= share of participants who voluntarily reported to identify as female.

B Searching for the Genetic Roots of Human Diseases: Additional Details

B.1 Scientific Background

Genetics is the branch of biology that studies genes, heredity, and variation in living organisms. Genes are segments of DNA (deoxyribonucleic acid) that contain the information necessary for living organisms' development, functioning, and reproduction. In practice, each gene is a portion of DNA that contains instructions for building one or more proteins, which are the fundamental constituents of an organism. Genes often acquire mutations (or variants) in their sequence, most of which are harmless. However, some mutations can lead the gene to alter its behavior and affect phenotypic traits, sometimes with significant consequences and the emergence of severe health conditions. Discovering which mutations are responsible for specific human diseases is thus a first-order priority since genes associated with a condition can often be used as drug targets (Nelson et al., 2015). When a drug molecule binds to its genetic target, it can modify its functioning, favorably affecting the outcome of a disease. Therefore, knowing the genetic roots of diseases has important practical consequences in the design of pharmaceutical drugs.

Diseases caused by single gene mutations are called Mendelian disorders, but such diseases are typically rare. Most common human diseases have a polygenic nature, meaning they are not due to a single genetic factor but rather by many genes. This class of diseases is called complex and genetic mutations can increase the risk of presenting the condition even without being neither necessary nor sufficient. Despite often clustering in families, polygenic disorders do not have a predictable inheritance pattern because convoluted interactions between genes and environmental factors determine them. This means that scientists need to search through the over 19,000 protein-coding genes to find the mutations involved in each of the thousands of polygenic diseases (Tranchero, 2024).

Researchers have noted that even after the completion of the Human Genome Project, most scientists continue to investigate the same small number of genes (Stoeger et al., 2018). Gates et al. (2021) report that 1% of genes still receive 22% of all gene-related publications, helping to explain why current treatments exploit only around 10% of the potentially druggable targets. This situation is probably suboptimal since our chances of finding a cure for polygenic diseases would benefit from exploring a larger number of genes (Edwards et al., 2011) and several understudied genes showing high promise have been identified (Nguyen et al., 2017; Stoeger et al., 2018). Interestingly, despite much debate on this extreme concentration of attention on a small number of theoretically well-known genes, we still

lack an explanation for its drivers. Some scholars have attributed it to scientists' preference for genes with past data that permit the formulation of functional hypotheses (Haynes et al., 2018), akin to what we characterized as a streetlight effect in this paper.

B.2 Data Description

DisGeNET. Our main data source is DisGeNET (v7.0), which is considered a complete repository of scientific results linking human diseases to their genetic causes (Piñero et al., 2020). This database aggregates all novel gene-disease associations studied by publications indexed in PubMed. This information is harvested from specialized sources, including curated datasets such as ClinVar, UniProt, and Orphanet.[21] In addition, DisGeNET complements these data with information extracted from the scientific literature indexed in PubMed using text-mining approaches. Our resulting data are at the gene-disease-paper level, because for each association we observe both the publication that introduced it and the list of all follow-up articles that investigated it.

Genes. Each gene in the database is identified by a unique identifier derived from Entrez Gene, a gene-centric database curated by the National Center for Biotechnology Information (NCBI). Entrez Gene provides tracked, unique gene identifiers that are integer and species-specific. In other words, the integer assigned to a given human gene differs from that of the homolog gene in any other species. DisGeNET compiles the Entrez Gene ID of each gene studied by papers in PubMed. We then limit our sample to protein-coding genes given their prominence in the drug discovery process (Nelson et al., 2015).

Diseases. Disease entries in DisGeNET are annotated using vocabulary from the Unified Medical Language System (UMLS), a set of crosswalks that bring together many health and biomedical vocabularies and standards to enable interoperability between databases. DisGeNET compiles the UMLS ID of each disease studied by papers in PubMed. Since we focus on human diseases, we keep any entries that map to the following UMLS semantic types: disease or syndrome; neoplastic process; acquired abnormality; anatomical abnormality; congenital abnormality; and mental or behavioral dysfunction. Using the UMLS ID, we also obtain disease relations from Kehoe and Torvik (2019), which contains all pairwise relationships in the Medical Subject Headings vocabulary (MeSH) hierarchy.

Gene-Disease Pair Score. DisGeNET provides a stable score for each gene-disease association it records. The score ranges from 0 to 1 and takes into account the number and type of sources supporting the association. In practice, it is a sum of the number of publications studying the

[21] For the complete list of sources aggregated by DisGeNET, see https://www.disgenet.org/dbinfo.

association, weighted by their level of curation and reliability. This synthetic metric reflects how well-established is a particular association based on current knowledge and provides a parsimonious way to assess the scientific value of any given gene-disease pair (Piñero et al., 2020; Tranchero, 2024). We then consider any score below the 60th percentile as a low payoff, between the 60th and 90th percentile as a medium payoff, and above the 90th percentile as a high payoff.[22]

Descriptive Statistics. We report the main descriptive statistics of our dataset in Table B.1. Panel A summarizes the data at the disease level. Around 58% of the 4,369 diseases in our sample achieved a breakthrough by 2019, which is the last year of our data. On average, it takes 22.3 years and the exploration of 131 genes to find a high-value genetic target for a disease. Panel B summarizes the data at the disease-year level. In any given year, the average disease receives 5.8 publications exploring its genetic roots, usually entailing the exploration of 3.3 new genetic associations.

Table B.1: Descriptive statistics of the DisGeNET database.

Panel A: Disease level

	Mean	Median	Sd	Min	Max	N
Max Found: Low (0/1)	0.11	0.00	0.32	0	1	4369
Max Found: Medium (0/1)	0.31	0.00	0.46	0	1	4369
Max Found: High (0/1)	0.58	1.00	0.49	0	1	4369
Year Reached 10% Papers	2001	2001	5.55	1981	2017	4369
Max Gene Score During Exploration	84.79	92.00	24.27	0	100	4369
Year of First Low Score	1991	1991	7.08	1980	2016	1588
Year of First Medium Score	1995	1994	8.27	1980	2018	2208
Year of First High Score	1996	1995	8.24	1980	2019	3164
Delay (Years since 1980)	22.26	20.00	12.48	0	39	4369
Total Publications	229.84	72.00	526.05	23	5312	4369
Total Genes Explored	131.02	50.00	248.22	1	2555	4369
New Genes Per Paper (Post-Exploration)	0.69	0.68	0.44	0	6	4369

Panel B: Disease-year level

	Mean	Median	Sd	Min	Max	N
Maximum Gene Score In Year	44.06	1.00	45.81	0	100	174760
Yearly Count of Publications	5.75	1.00	24.07	0	836	174760
Yearly Count of Genes Explored	3.28	0.00	11.36	0	646	174760

Note: Panel A presents descriptive statistics on papers that introduce new gene-disease associations after 2005. Panel B presents descriptive statistics of the panel dataset at the disease-year level that we used for the event-study analysis shown in Figure 7 and Appendix Figure C.6.

[22] As already noted in the main text, all our results are robust to the adoption of alternative threshold values (Appendix Tables C.6 and C.7).

B.3 Case Study: Gardner's Syndrome and Tangier's Disease

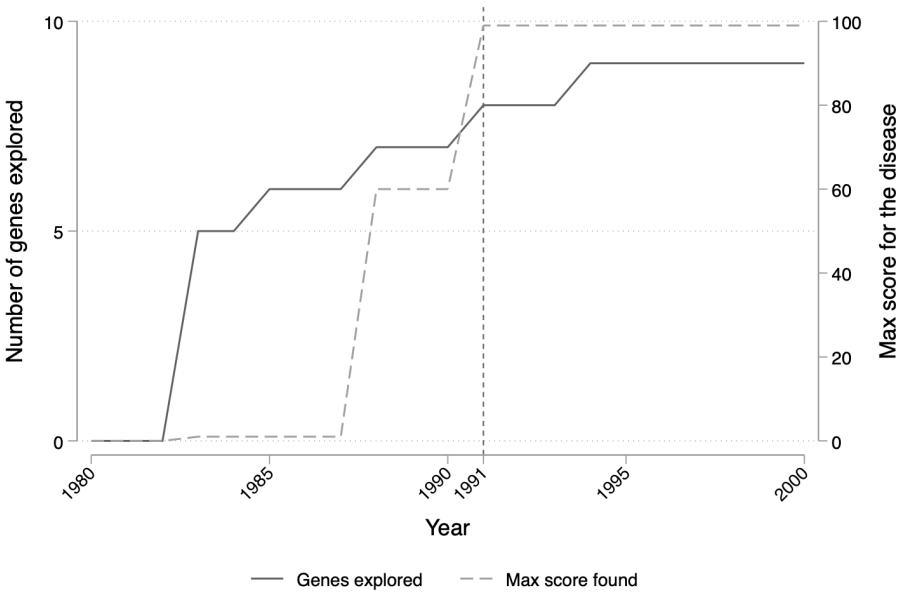
To exemplify our empirical application to genetic research, consider the following two genetic diseases. Gardner syndrome (MeSH ID: D005736) is a rare disorder that falls under the umbrella of familial adenomatous polyposis. It is characterized by the development of numerous polyps, particularly in the colon and rectum. These polyps have the potential to become cancerous if left untreated. In addition to gastrointestinal manifestations, individuals with Gardner syndrome may exhibit extra-colonic features, such as the development of osteomas (benign bone tumors), particularly in the skull and jaw. Importantly, Gardner syndrome is associated with mutations in the APC gene. This tumor suppressor gene is responsible for regulating cell growth and preventing cells from dividing and multiplying too quickly.

Tangier disease (MeSH ID: D013631) is a rare disorder characterized by a deficiency of high-density lipoprotein cholesterol (HDL-C) in the blood. HDL-C is responsible for transporting cholesterol away from tissues and back to the liver, playing a crucial role in cholesterol metabolism. Individuals with Tangier disease typically experience enlarged and dysfunctional tonsils that exhibit a characteristic orange discoloration. Additionally, patients may experience an increased risk of atherosclerosis and cardiovascular disease due to the decreased ability to remove cholesterol from the bloodstream. Tangier disease is an inherited genetic disease due to mutations in the ABCA1 gene. When this gene is abnormal, a problem with its instructions makes the body unable to transport lipids onto apolipoproteins, leading to a significant reduction in functional HDL-C particles.

Figure B.1 compares the history of genetic discoveries for both diseases. In the case of Gardner syndrome (Panel A), early attempts did not find any promising genes, leading to a prolonged period of exploration which culminated in the discovery of mutations in the APC gene (Nishisho et al., 1991). Instead, Tangier disease (Panel B) saw the immediate discovery of a promising association with the gene APOA1 in 1982. Such discovery stifled the exploration of new genes and led to resources being poured on a target similar to a medium-value finding in our theoretical framework (APOA1 turned out to have a DisGeNET score in the 60^{th} percentile). The gene responsible for Tangier disease, ABCA1, was only discovered in 1999 by Brooks-Wilson et al. (1999).[23] This case study highlights the unfolding of the streetlight effect in a real-world example. Somewhat paradoxically, the disease for which earlier inroads were made is also the one that reached the breakthrough later. Instead, the lack of early discoveries for Gardner syndrome led to more exploration, resulting in the responsible gene being discovered 8 years earlier.

[23] Both these papers are very influential: Nishisho et al. (1991) and Brooks-Wilson et al. (1999) received over 2,400 and 2,100 cites in Google Scholar as of the year 2023, respectively.

Panel A: Gardner's syndrome



Panel B: Tangier's disease

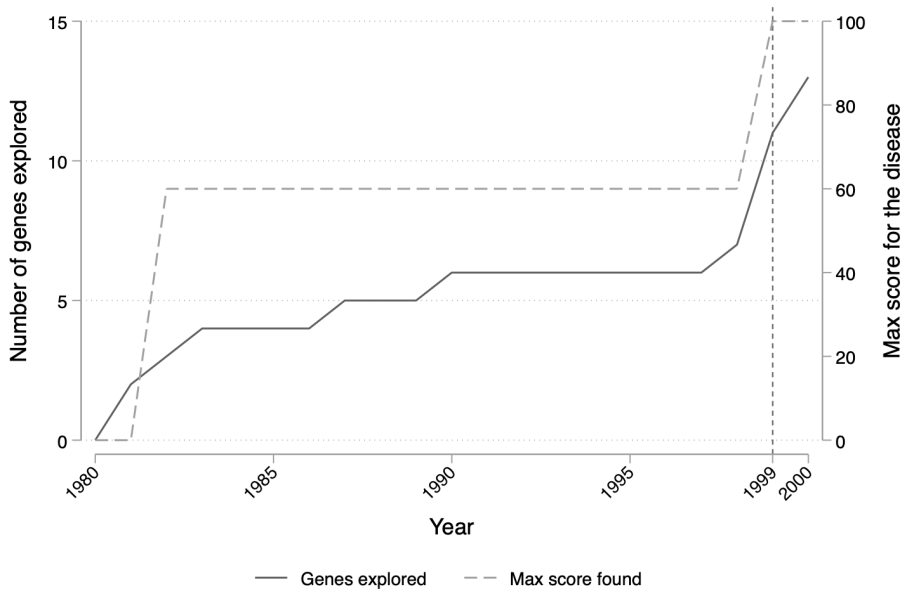
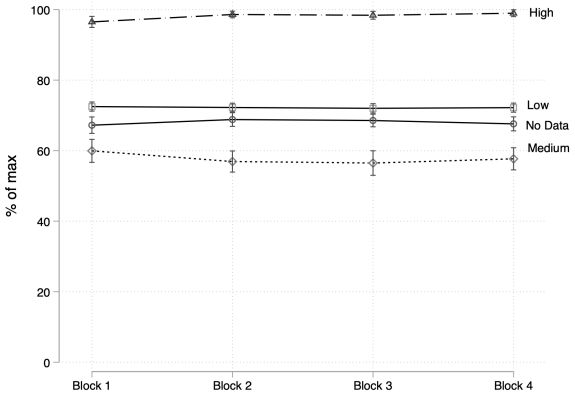


Figure B.1: Two case examples in search for the genetic origins of diseases.

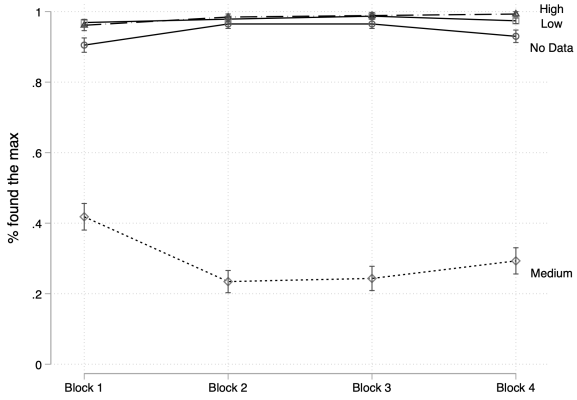
Note: The solid black line represents the cumulative number of gene candidates discovered for the disease up to that year. The dashed line represents the maximum DisGeNET score observed for the genes associated with that disease up to that year. Panel A presents the data for Gardner's syndrome, and the vertical line indicates the year when the association with the APC gene was discovered (DisGeNET score in the 99th percentile). Panel B presents the data for Tangier's disease, and the vertical line indicates the year when the association with the ABCA1 gene was discovered (DisGeNET score in the 100th percentile).

C Additional Figures and Tables

(i) Average round payoffs by block



(ii) Likelihood of individual breakthrough by block



(iii) Likelihood of group breakthrough by block

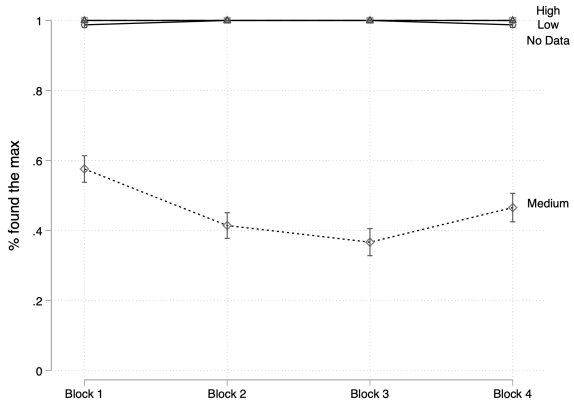
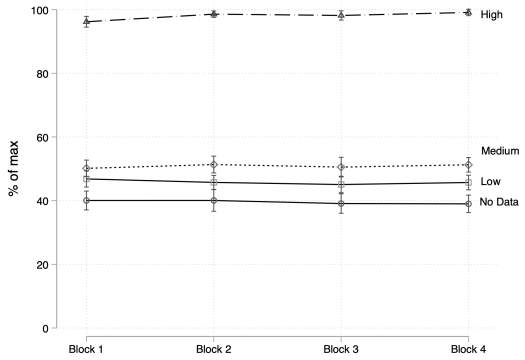


Figure C.1: Robustness of the main results over time.

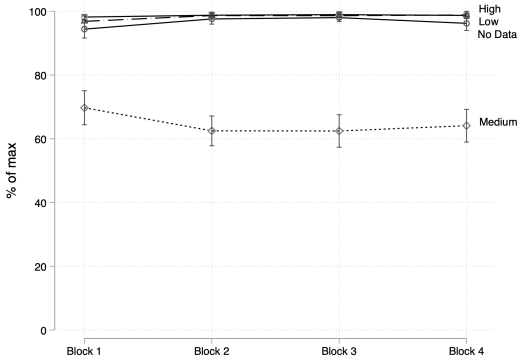
Note: The figures depict the impact of data on group outcomes as the experimental session progresses. Figure (i) shows for each block of 5 rounds the average group payoffs divided by experimental condition. Payoffs are reported as a share of the maximum available in each round. Figure (ii) shows the share of participants who found the location of the maximum divided by experimental condition for each block. Figure (iii) shows the share of rounds for each block where the maximum was found divided by experimental condition.

Panel A: Payoffs

(i) Payoffs in period 1

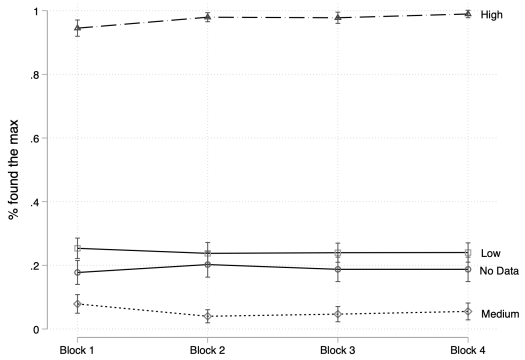


(ii) Payoffs in period 2



Panel B: Breakthroughs

(iii) Likelihood of breakthrough in period 1



(iv) Likelihood of breakthrough in period 2

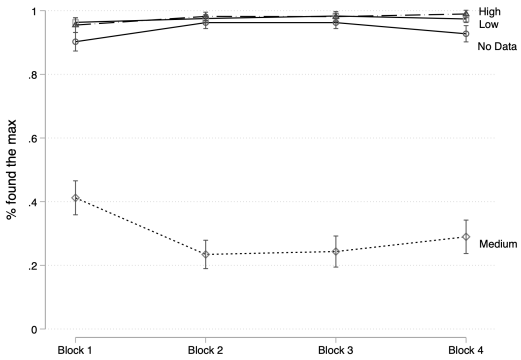


Figure C.2: Outcomes over time and by period of the game.

Note: Panel A reports the experimental results on the period payoffs computed as a share of the maximum possible in each period. Figure (i) shows the average collective payoffs achieved in period 1 by experimental condition and over time. Figure (ii) shows the average collective payoffs achieved in period 2 by experimental condition and over time. Panel B reports the experimental results on the likelihood of an individual breakthrough in each round. Figure (iii) shows the share of participants that found the maximum in period 1 by experimental condition and over time. Figure (iv) shows the share of participants that found the maximum in period 2 by experimental condition and over time.

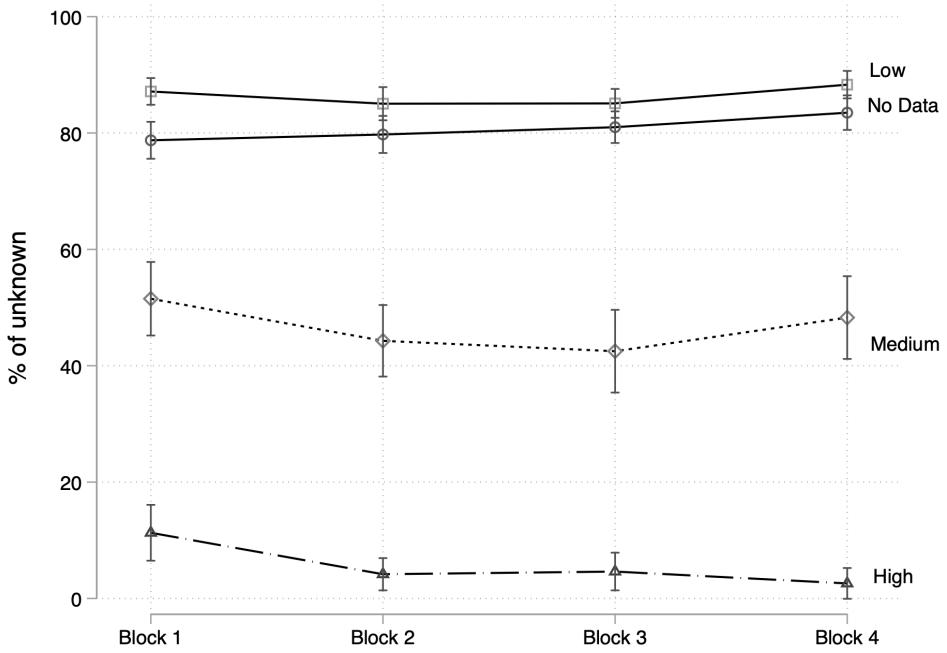
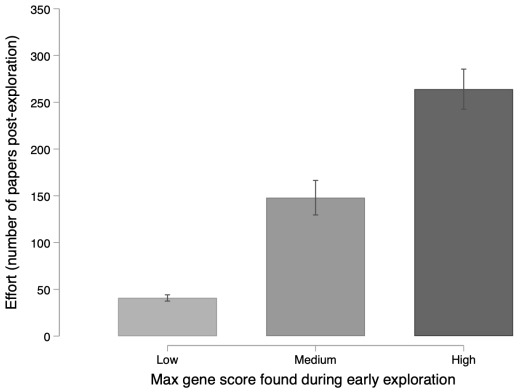


Figure C.3: Average level of exploration over time by experimental condition.

Note: The figure shows for each block of five rounds the impact of data on exploration choices divided by experimental condition. The number of mountains explored is reported as a share of the unknown mountains in each round to account for the fact that rounds without data have one more unknown option.

(i) Average search effort for diseases



(ii) Average exploration of new genes for diseases

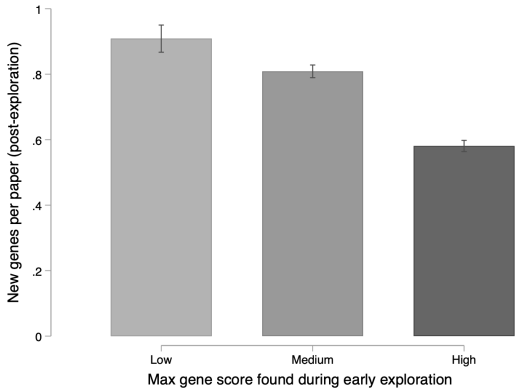
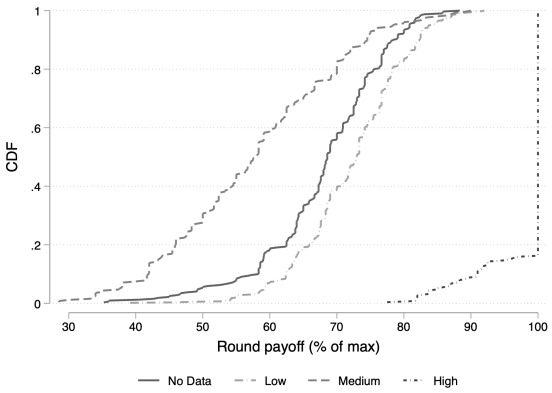


Figure C.4: Impact of early discoveries on search effort and exploration for the genetic origins of diseases.

Note: For each human disease, we compute the highest DisGeNET score identified in genetic publications linked to the disease during the early search phase (defined as the first 10% of publications on the disease). We classify maximum 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. Panel (i) displays the mean number of publications about the genetic roots of a disease following the early search window. Panel (ii) displays the average number of new genes explored per paper about a disease following the early search window. Error bars represent 95% confidence intervals. See text for more details.

(i) CDFs of round payoffs in the experiment



(ii) CDFs of delay to breakthrough in genetics

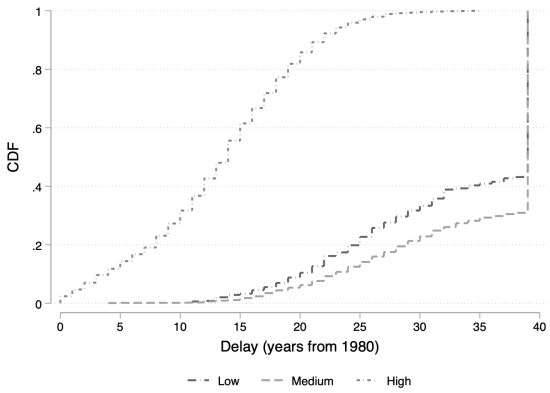
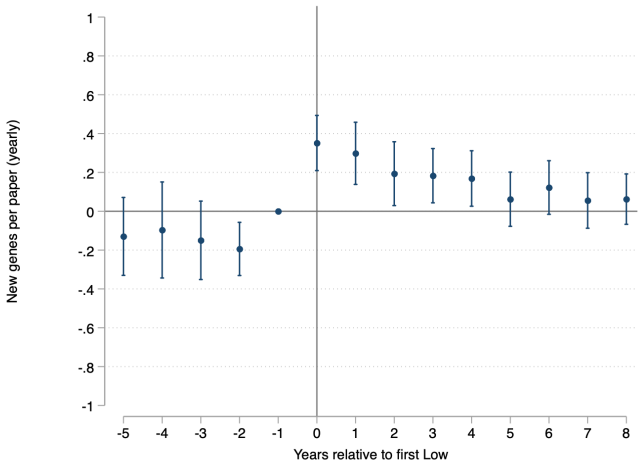


Figure C.5: Suggestive comparison between experimental findings and patterns of genetic discovery.

Note: Panel (i) plots the cumulative density function of round payoffs by experimental condition. Panel (ii) displays the CDF of the number of years it takes to discover the first "high" score after 1980 (the first sample year) for each of the three groups. For each human disease, we compute the highest DisGeNET score identified in genetic publications linked to the disease during the early search phase (defined as the first 10% of publications on the disease). We classify maximum scores below the 60^{th} percentile as a "low" gene discovery, scores between the 60^{th} and 90^{th} percentile as a "medium" gene discovery, and scores above the 90^{th} percentile as a "high" (or breakthrough) gene discovery.

Panel A: Discovery of a low-value genetic association



Panel B: Discovery of a high-value genetic association

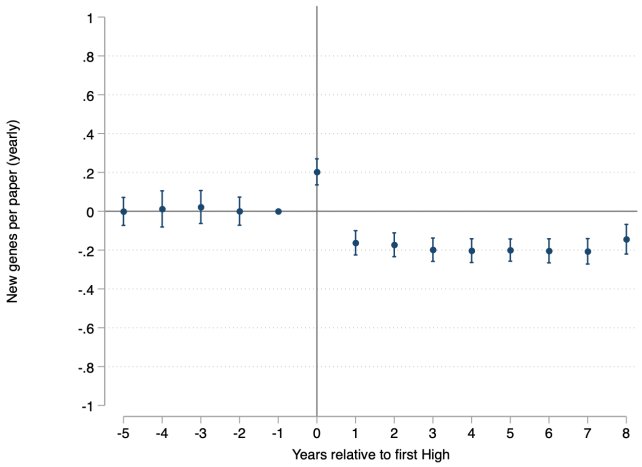
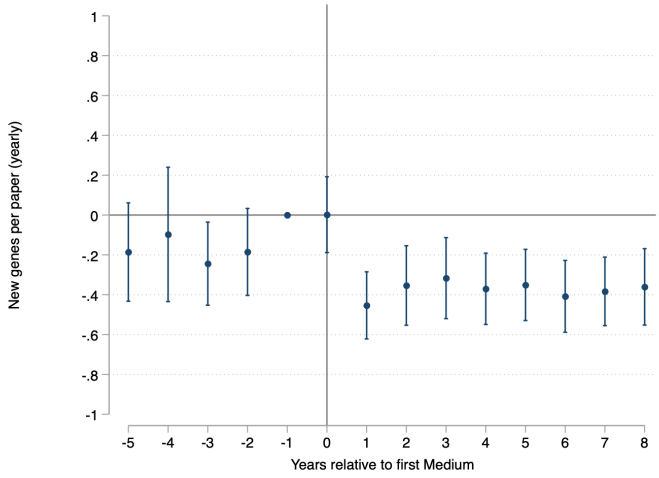


Figure C.6: Dynamic effects of the discovery of a low or high-value genetic association on the exploration of new genes.

Note: For each human disease, we compute the highest DisGeNET score identified in genetic publications linked to the disease during the early search phase (defined as the first 10% of publications on the disease). We classify maximum scores below the 60th percentile as a "low" gene discovery and scores above the 90th percentile as a "high" (or breakthrough) gene discovery. Panel (i) plots OLS estimates and 95% confidence intervals from an event study design that explores how genetic exploration in each disease evolves in the years before and after the discovery of the first low-value genetic association. Panel (ii) plots OLS estimates and 95% confidence intervals from an event study design that explores how genetic exploration in each disease evolves in the years before and after the discovery of the first high-value genetic association. Standard errors are clustered at the disease class level. See text for more details.

Panel A: Keeping sibling and parent diseases



Panel B: Keeping only sibling diseases

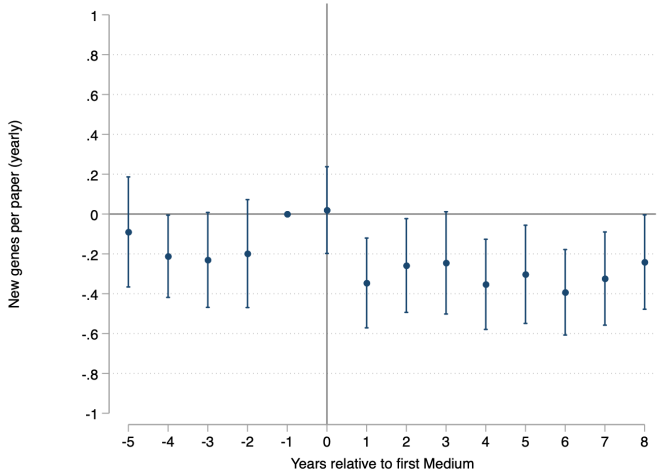


Figure C.7: Considering only diseases genetically related to a disease with a breakthrough.

Note: This figure replicates our baseline event study but only considers diseases in the sample that are genetically related to a disease with a known breakthrough (genetic discoveries with scores above the 90th percentile of DisGeNET score). We obtain genetic relations from the Medical Subject Headings vocabulary (MeSH). In Panel A, we keep both sibling diseases (i.e., diseases sharing the same parent MeSH code) and parent diseases (i.e., diseases one level up in the MeSH tree) of diseases with a breakthrough. In Panel B we keep only sibling diseases (i.e., diseases sharing the same parent MeSH code) of diseases with a breakthrough. This figure plots OLS estimates and 95% confidence intervals from an event study design that explores how genetic exploration in each disease evolves in the years before and after the discovery of the first medium-value genetic association. Standard errors are clustered at the disease class level. See text for more details.

Table C.1: Effects of revealing a medium-value project for different parameter values.

	Exploration with parameter set 1	Exploration with parameter set 2
	(1)	(2)
	Round	Round
High	-75.318***	-75.246***
	(2.968)	(2.622)
Low	6.473*	4.433**
	(2.370)	(1.469)
Medium	-38.886***	-26.168***
	(4.028)	(3.123)
Constant	86.290***	79.152***
	(3.035)	(2.067)
Round Order FE	No	No
Block order FE	Yes	Yes
Payoff structure FE	Yes	Yes
Observations	800	600

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

Estimates from OLS models. The sample in both columns is at the group-round level. Column 1 shows the results for rounds where the parameters used satisfy Assumption 1 and the condition in equation (5). Column 2 shows the results for rounds where the parameters used satisfy Assumption 1 and the condition in equation (4). The excluded category captured by the constant is the condition without data.

Table C.2: Risk aversion and decision not to choose the known outcome in period 1 when medium is revealed.

	I(Exploration if M is revealed)		
Risk aversion	-0.003		
	(0.018)		
Top quartile risk aversion		0.000	
		(0.034)	
Bottom quartile risk aversion			-0.052
			(0.030)
Constant	0.329***	0.328***	0.343***
	(0.054)	(0.051)	(0.059)
Round Order FE	Yes	Yes	Yes
Block order FE	Yes	Yes	Yes
Payoff structure FE	Yes	Yes	Yes
Observations	1270	1270	1270

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

Round-participant level observations, estimates from OLS models. The sample includes all the individual observations for the 254 rounds where the medium value was revealed. *I(Exploration if M is revealed)*:0/1=1 if the player did not choose the medium value in period 1. *Risk aversion* = standardized measure of individual risk aversion (Holt and Laury, 2002); *Top quartile risk aversion*:0/1=1 if the participant is in the top quartile of the risk aversion distribution in our sample; *Bottom quartile risk aversion*:0/1=1 if the participant is in the bottom quartile of the risk aversion distribution in our sample.

Table C.3: Correlates of the decision not to choose the known outcome in period 1 when medium is revealed.

	I(Exploration if M is revealed)			
English native	0.012			
	(0.038)			
Wrong quizzes		0.043*		
		(0.017)		
Round number			-0.012	
			(0.008)	
Order of choice				0.013
				(0.011)
Constant	0.322***	0.324***	0.369***	0.316***
	(0.067)	(0.054)	(0.055)	(0.056)
Round Order FE	Yes	Yes	Yes	No
Block order FE	Yes	Yes	Yes	Yes
Payoff structure FE	Yes	Yes	Yes	Yes
Observations	1270	1270	1270	1270

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

Round-player level observations, estimates from OLS models. The sample includes all the individual observations for the 254 rounds where the medium value was revealed. *I(Exploration if M is revealed)*:0/1=1 if the player did not choose the medium value in period 1. *English native*:0/1=1 if the participant is a native English speaker based on her reported nationality; *Wrong quizzes* = standardized number of wrong answers to the initial comprehension test; *Round number* = progressive order in which the rounds were played in the experimental session; *Order of choice* = random sequential order in which the player chose in that round.

Table C.4: Sensitivity to definition of marginally explored diseases

Panel A: Delay in breakthroughs

	Delay (Years From 1980)			
	>0 Pubs	>10 Pubs	>20 Pubs	>30 Pubs
	(1)	(2)	(3)	(4)
Max Found: Medium	3.265***	2.768***	1.691***	1.362*
	(0.366)	(0.392)	(0.452)	(0.602)
Max Found: High	-13.305***	-14.671***	-16.301***	-16.931***
	(0.546)	(0.578)	(0.625)	(0.723)
Final Exploration Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes
N	14208	5529	3828	2965

Panel B: Diversity of follow-on research

	New Genes Per Paper			
	>0 Pubs	>10 Pubs	>20 Pubs	>30 Pubs
	(1)	(2)	(3)	(4)
Max Found: Medium	-0.127***	-0.129***	-0.125***	-0.101***
	(0.025)	(0.021)	(0.021)	(0.026)
Max Found: High	-0.062	-0.125**	-0.151***	-0.136***
	(0.042)	(0.042)	(0.035)	(0.039)
Final Exploration Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes
N	11345	5529	3828	2965

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease class level in parentheses. This table replicates our baseline specification (which removes diseases with less than 25 publications over the sample window) and shows robustness when we keep only diseases with nonzero publications (1), more than 10 publications (2), more than 20 publications (3), and more than 30 publications (4). For each human disease, we compute the highest DisGeNET score identified in the genetic publications linked to the disease during the early search phase (defined as the first 10% of publications on the disease). We classify maximum 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. Panel A shows the impact of early discoveries on the delay in discovering a breakthrough for a given disease, defined as years elapsed from 1980 (the first year of our panel). Panel B shows the impact of early discoveries on the number of new genes explored for a given disease, normalized by the total number of publications in the years following the exploration window. In both cases, diseases that found only low-value genes during the early search period constitute the excluded category. See text for more details.

Table C.5: Sensitivity to the inclusion of outlier diseases.

Panel A: Delay in breakthroughs

	Delay (Years From 1980)			
	(1)	(2)	(3)	(4)
Max Found: Medium	1.959***	1.701***	1.350**	1.611**
	(0.501)	(0.457)	(0.482)	(0.529)
Max Found: High	-19.043***	-18.636***	-18.733***	-16.494***
	(0.647)	(0.627)	(0.694)	(0.726)
Final Exploration Year FE	No	Yes	Yes	Yes
Disease Class FE	No	No	Yes	Yes
Count of Publications	No	No	No	Yes
N	4010	4009	3779	3337

Panel B: Diversity of follow-on research

	New Genes Per Paper			
	(1)	(2)	(3)	(4)
Max Found: Medium	-0.077**	-0.089**	-0.142***	-0.115***
	(0.029)	(0.028)	(0.024)	(0.025)
Max Found: High	-0.311***	-0.316***	-0.239***	-0.150***
	(0.028)	(0.028)	(0.029)	(0.035)
Final Exploration Year FE	No	Yes	Yes	Yes
Disease Class FE	No	No	Yes	Yes
Count of Publications	No	No	No	Yes
N	4010	4009	3779	3337

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease class level in parentheses. This table replicates our baseline specification including also outlier diseases (i.e., those in the top 1% by publications over the sample period). For each human disease, we compute the highest DisGeNET score identified in the genetic publications linked to the disease during the early search phase (defined as the first 10% of publications on the disease). We classify maximum 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. Panel A shows the impact of early discoveries on the delay in discovering a breakthrough for a given disease, defined as years elapsed from 1980 (the first year of our panel). Panel B shows the impact of early discoveries on the number of new genes explored for a given disease, normalized by the total number of publications in the years following the exploration window. In both cases, diseases that found only low-value genes during the early search period constitute the excluded category. See text for more details.

Table C.6: Alternative definitions of low and medium-value genes.

Panel A: Delay in breakthroughs

	Delay (Years From 1980)			
	50^{th} P	60^{th} P	70^{th} P	80^{th} P
	(1)	(2)	(3)	(4)
Max Found: Medium	1.924**	1.611**	1.611**	1.568**
	(0.685)	(0.529)	(0.529)	(0.557)
Max Found: High	-16.044***	-16.494***	-16.494***	-16.858***
	(0.899)	(0.726)	(0.726)	(0.626)
Final Exploration Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes
N	3337	3337	3337	3337

Panel B: Diversity of follow-on research

	New Genes Per Paper			
	50^{th} P	60^{th} P	70^{th} P	80^{th} P
	(1)	(2)	(3)	(4)
Max Found: Medium	-0.055	-0.115***	-0.115***	-0.183***
	(0.031)	(0.025)	(0.025)	(0.025)
Max Found: High	-0.113***	-0.150***	-0.150***	-0.159***
	(0.030)	(0.035)	(0.035)	(0.032)
Final Exploration Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes
N	3337	3337	3337	3337

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease class level in parentheses. This table replicates our baseline specification but varies the cutoff between a low and medium-value genetic association. In our baseline, we adopt the 60th percentile to separate medium and high scores. We test the 50th percentile (1), the baseline (2), the 70th percentile (3), and the 80th percentile (4) instead. For each regression, we hold the cutoff between a medum gene score and high gene score fixed at the 90th percentile (our baseline). Panel A shows the impact of early discoveries on the delay in discovering a breakthrough for a given disease, defined as years elapsed from 1980 (the first year of our panel). Panel B shows the impact of early discoveries on the number of new genes explored for a given disease, normalized by the total number of publications in the years following the exploration window. In both cases, diseases that found only low-value genes during the early search period constitute the excluded category. See text for more details.

Table C.7: Alternative definitions of medium and high-value genes.

Panel A: Delay in breakthroughs

	Delay (Years From 1980)			
	90^{th} P	95^{th} P	98^{th} P	99^{th} P
	(1)	(2)	(3)	(4)
Max Found: Medium	1.611**	0.371	0.494	0.355
	(0.529)	(0.529)	(0.541)	(0.512)
Max Found: High	-16.494***	-17.936***	-18.299***	-18.954***
	(0.726)	(0.793)	(0.803)	(0.821)
Final Exploration Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes
N	3337	3337	3337	3337

Panel B: Diversity of follow-on research

	New Genes Per Paper			
	90^{th} P	95^{th} P	98^{th} P	99^{th} P
	(1)	(2)	(3)	(4)
Max Found: Medium	-0.115***	-0.083***	-0.071**	-0.072**
	(0.025)	(0.024)	(0.024)	(0.025)
Max Found: High	-0.150***	-0.220***	-0.249***	-0.275***
	(0.035)	(0.039)	(0.040)	(0.039)
Final Exploration Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes
N	3337	3337	3337	3337

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease class level in parentheses. This table replicates our baseline specification but varies the cutoff between a medium and high-value genetic association. In our baseline, we adopt the 90th percentile to separate medium and high scores. We test the baseline (1), the 95th percentile (2), the 98th percentile (3), and the 99th percentile (4) instead. For each regression, we hold the cutoff between a low gene score and medium gene score fixed at the 60th percentile (our baseline). Panel A shows the impact of early discoveries on the delay in discovering a breakthrough for a given disease, defined as years elapsed from 1980 (the first year of our panel). Panel B shows the impact of early discoveries on the number of new genes explored for a given disease, normalized by the total number of publications in the years following the exploration window. In both cases, diseases that found only low-value genes during the early search period constitute the excluded category. See text for more details.

Table C.8: Different thresholds of publication share to define the early search period.

	Delay (Years From 1980)			
	5%	10%	15%	20%
	(1)	(2)	(3)	(4)
Max Found: Medium	2.598***	1.611**	0.661	1.008
	(0.436)	(0.529)	(0.550)	(0.548)
Max Found: High	-13.657***	-16.494***	-18.465***	-18.532***
	(0.680)	(0.726)	(0.686)	(0.688)
Final Exploration Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes
N	3325	3337	3369	3391

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease class level in parentheses. This table replicates our baseline specification using alternative definitions of "early search period". We test varying thresholds from 5% (1) up to 20% (4) in increments of 5%. This table replicates our baseline specification using alternative windows to define the period of early search. We report the results employing fixed windows, including all years before 1990 (1), before 1995 (2), before 2000 (3), and before 2005 (4). For each human disease, we compute the highest DisGeNET score identified in the genetic publications linked to the disease during the early search phase. We classify maximum 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. See text for more details.

Table C.9: Fixed windows of years to define the early search period.

	Delay (Years From 1980)			
	<1990	<1995	<2000	<2005
	(1)	(2)	(3)	(4)
Max Found: Medium	1.939**	1.811**	2.531***	1.560**
	(0.724)	(0.556)	(0.544)	(0.521)
Max Found: High	-18.706***	-17.826***	-18.596***	-20.288***
	(0.958)	(0.778)	(0.653)	(0.581)
Final Exploration Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes
N	1192	2213	2923	3297

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease class level in parentheses. This table replicates our baseline specification using alternative windows to define the period of early search. We report the results employing fixed windows, including all years before 1990 (1), before 1995 (2), before 2000 (3), and before 2005 (4). For each human disease, we compute the highest DisGeNET score identified in the genetic publications linked to the disease during the early search phase. We classify maximum 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. See text for more details.

Table C.10: Alternative windows to examine follow-on explorative research.

	New Genes Per Paper			
	All Years	5 Years	10 Years	Until H
	(1)	(2)	(3)	(4)
Max Found: Medium	-0.115***	-0.098*	-0.105**	-0.095*
	(0.025)	(0.039)	(0.033)	(0.040)
Max Found: High	-0.150***	-0.183***	-0.179***	
	(0.035)	(0.044)	(0.041)	
Disease Class FE	Yes	Yes	Yes	Yes
Final Exploration Year FE	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes
N	3337	3305	3332	1077

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease class level in parentheses. This table replicates our baseline specification using alternative windows to evaluate the evolution of explorative research. We report the results from the baseline (1), the 5 subsequent years after 10% of publications is reached (2), the 10 subsequent years after 10% of publications is reached (3), and until the first high gene score is found (4). For each human disease, we compute the highest DisGeNET score identified in the genetic publications linked to the disease during the early search phase (defined as the first 10% of publications on the disease). We classify maximum 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. See text for more details.

Table C.11: Alternative measures of delays in breakthroughs.

	Delay (Years From 10% Publications)		
	(1)	(2)	(3)
Max Found: Medium	4.063***	3.182**	3.302**
	(1.003)	(0.994)	(0.994)
Max Found: High	-22.939***	-22.216***	-21.194***
	(1.118)	(1.250)	(1.242)
Disease Class FE	No	Yes	Yes
Count of Publications	No	No	Yes
N	3968	3738	3338

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease class level in parentheses. This table replicates our baseline specification using an alternative measure of delay in breakthrough discovery, here defined as years elapsed from the first 10% of publications on the disease. For each human disease, we compute the highest DisGeNET score identified in the genetic publications linked to the disease during the early search phase (defined as the first 10% of publications on the disease). We classify maximum 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. See text for more details.

Table C.12: Difference-in-difference estimates of the effect of early discoveries on subsequent genetic exploration.

	New Genes Paper Paper (Yearly)					
	Low GDA		Medium GDA		High GDA	
	(1)	(2)	(3)	(4)	(5)	(6)
Post (Low)	0.252***	0.225***				
	(0.039)	(0.039)				
Post (Med)			-0.139***	-0.141***		
			(0.028)	(0.029)		
Post (High)					-0.177***	-0.153***
					(0.019)	(0.019)
Disease FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Count of Publications	No	Yes	No	Yes	No	Yes
N	88134	87997	88134	87997	88134	87997

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease class level in parentheses. For each human disease, we compute the highest DisGeNET score identified in the genetic publications linked to the disease during the early search phase (defined as the first 10% of publications on the disease). We classify maximum 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. This table reports OLS estimates from differences-in-differences that explore how genetic exploration in each disease evolves in the years before and after the discovery of the first low, medium, and high-value genetic association. Standard errors are clustered at the disease class level. See text for more details.

Table C.13: Considering only diseases that have a breakthrough by 2019.

Panel A: Delay in breakthroughs

	Delay (Years From 1980)			
	(1)	(2)	(3)	(4)
Max Found: Medium	0.918	0.759	0.641	2.199***
	(0.594)	(0.507)	(0.560)	(0.473)
Max Found: High	-11.591***	-11.800***	-11.832***	-8.531***
	(0.536)	(0.521)	(0.591)	(0.441)
Final Exploration Year FE	No	Yes	Yes	Yes
Disease Class FE	No	No	Yes	Yes
Count of Publications	No	No	No	Yes
N	3053	3051	2861	2477

Panel B: Diversity of follow-on research

	New Genes Per Paper			
	(1)	(2)	(3)	(4)
Max Found: Medium	0.005	0.004	-0.071	-0.067
	(0.044)	(0.044)	(0.043)	(0.042)
Max Found: High	-0.210***	-0.220***	-0.181***	-0.113**
	(0.037)	(0.038)	(0.031)	(0.042)
Final Exploration Year FE	No	Yes	Yes	Yes
Disease Class FE	No	No	Yes	Yes
Count of Publications	No	No	No	Yes
N	3053	3051	2861	2477

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease class level in parentheses. This table replicates our baseline specification removing any diseases without a breakthrough (i.e. a gene with a "high" score) during the sample period. For each human disease, we compute the highest DisGeNET score identified in the genetic publications linked to the disease during the early search phase (defined as the first 10% of publications on the disease). We classify maximum scores below the 60$^{\text{th}}$ percentile as a "low" gene discovery, scores between the 60$^{\text{th}}$ and 90$^{\text{th}}$ percentile as a "medium" gene discovery, and scores above the 90$^{\text{th}}$ percentile as a "high" (or breakthrough) gene discovery. Panel A shows the impact of early discoveries on the delay in discovering a breakthrough for a given disease, defined as years elapsed from 1980 (the first year of our panel). Panel B shows the impact of early discoveries on the number of new genes explored for a given disease, normalized by the total number of publications in the years following the exploration window. In both cases, diseases that found only low-value genes during the early search period constitute the excluded category. See text for more details.