The applied value of basic research is underestimated

An analysis of genomics publications shows that citation-based indicators miss up to 86% of a study's impact on U.S. innovation

M. Tranchero,[1*] C. Guthmann[2]

[1]The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA

[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

[*]To whom correspondence should be addressed: mtranc@wharton.upenn.edu.

Main text

Governments increasingly assess the return on investment (ROI) in basic research by estimating its impact on downstream technologies, counting the new innovations that arise from publicly funded science (*1*). The most used metrics—citations from corporate patents to publications that acknowledge government support (*2, 3*)—hinge on imperfect norms about what "deserves" a citation (*4*). Citation tallies are least informative when knowledge diffuses through informal channels and produces spillovers that leave no clear citation trail, both features typical of basic science (*5*). Here, we propose an alternative approach to track how discoveries move through knowledge spaces (*6*), using natural language processing to measure innovation spillovers from scientific findings. In the context of genome-wide association studies (GWAS), a much-debated class of genomics studies (*7, 8*), our method shows that standard citation-based metrics capture only a minority of their contributions to patented technologies and marketed therapies. Our results suggest a systematically understated picture of the role that basic science plays in the nation's innovative capacity.

Since the postwar era, federal support for basic research has been grounded in Vannevar

Bush's view that autonomous scientific inquiry generates broad societal benefits, even when its effects unfold slowly and are difficult to quantify. Recent trends, however, have heightened attention to how these effects are measured (*9*). The political environment is increasingly shaping federal research budgets (*10*), and proposed reductions to major science agencies have intensified scrutiny of the ROI in basic research (*11*). Yet, assessing the value of basic science is challenging because the ways in which it shapes innovation extend far beyond formal acknowledgments. Survey evidence shows that industrial R&D managers rely on conferences, informal exchanges, and consulting to access research results (*5*). Because many of these channels leave no paper trail, conventional metrics capture only a limited share of how science is used in downstream innovation. More comprehensive measures of these technological spillovers are crucial for informing debates about public research funding.

We illustrate and quantify the unmeasured impact of basic science by focusing on GWAS, large-scale case-control studies that compare the genomes of individuals to identify genotype–phenotype associations. These studies have expanded the set of genes and pathways implicated in human disease, yet their impact remains contested (*8*). Most associations explain only a modest share of heritability and seldom reveal causal genes directly, complicating efforts to translate individual findings into therapeutic advances (*7*). The uncertainty about the value of GWAS persists despite substantial public investment: by our estimates, at least US$3.3 billion was invested in these studies between 2005 and 2020, with 59% of them acknowledging support from the U.S. National Institutes of Health (NIH) [see supplementary materials (SM) for calculations]. Taken together, these factors make GWAS a compelling case for examining the extent to which citation-based metrics appropriately capture the innovation returns to research.

A spatial approach to measuring spillovers from science

We measure how basic research spills over into innovation by starting from the knowledge space studied rather than from citation links between documents. In many fields, research is organized around a set of underlying targets, components, or topics that define where innovators direct their effort. Examples include genes involved in a disease pathway, materials used in a battery, or classes of algorithms used in machine learning. Together, these elements form knowl-

edge spaces whose structure reflects functional or conceptual relationships between individual elements (*6*). Such spaces are increasingly formalized through machine-readable knowledge graphs (*12*). Because the components of a knowledge space are related, a scientific finding about one element can also convey information about nearby ones, generating innovation externalities that leave no direct citation trail. In medicine, for instance, a discovery about one genetic target can reshape expectations about other genes involved in the same biological pathway. GWAS produce this type of discovery, providing a setting in which to observe spillovers that propagate through the knowledge space without being directly acknowledged.

To implement this approach, we represent the biomedical landscape at the gene–disease pair level, treating each pair as a potential focus of inventive effort. Time-stamped discoveries come from the manually curated GWAS Catalog (*13*), which records when a study first reports that a specific gene is associated with a disease [see SM for data details]. We then place these pairs in a biological network based on KEGG pathway data, linking genes that participate in the same biological function and defining the distance between pairs by the shortest path between their genes. Next, we use BioBERT-based named-entity recognition to identify genes and diseases mentioned in U.S. pharmaceutical patents and map them to standardized ontologies [data and code are publicly available at (*14*)]. Combining this text-mined patent information with the pathway network provides, for each year, both the volume of patenting that refers to each gene-disease pair and whether that pair (or its biological neighbors) has recently been the subject of a GWAS discovery. These data allow us to track how inventive activity changes after a new finding and to measure impacts that conventional patent-to-paper citation measures miss.

We provide an illustration with a case study of Inflammatory Bowel Disease (IBD; MeSH ID: D015212). A landmark GWAS published in 2006 identified a strong association between IBD and mutations in the gene IL23R. This gene is embedded in the Th17 cell differentiation pathway, a dense set of functionally connected genes involved in immunological responses [see SM for visualization and details]. We track how IBD-related inventive activity evolved after the discovery not only for IL23R but also for its neighboring genes. Visualizations of the Th17 pathway before and after the GWAS show a marked increase in patent mentions to IL23R and to genes connected to it by one to three biological steps (Panels (a) and (b) of Figure 1). Aggregated over time, these changes reveal inventive effort spreading across the genetic space,

as the original discovery also illuminates therapeutic opportunities for related genes. Panel (c) of Figure 1 shows that the yearly U.S. patents referencing genes in the Th17 pathway rose from an average of 8.17 before the GWAS to 23.14 afterward, a change far larger than what would be inferred by examining the IL23R–IBD link in isolation.

Regression estimates pooling all GWAS discoveries made between 2005 and 2020 show that the IBD case study reflects a broader pattern in the data. Figure 2 reports event study estimates that compare gene–disease pairs that are featured in a GWAS (treatment group) with pairs that are not (control group). Conditional on stringent controls, there are no statistically significant differences in patenting between these groups before the GWAS publication. However, after the study is published, patenting on the newly identified gene–disease pair rises sharply and remains elevated. A GWAS discovery leads to 0.176 additional patents per year focused on that gene–disease pair, a 52% increase over the sample mean (Table SM2). Patenting also increases for genes that share a pathway with the focal target, with responses that decline smoothly as the number of steps between genes grows: immediate pathway neighbors gain, on average, 0.0843 patents per year, genes two steps away gain 0.0387, and genes three steps away gain 0.0168. These spillover effects are 48%, 22%, and 9.6% as large as the direct effect on the focal pair, respectively, indicating that much of the innovation response accrues in the surrounding biological network rather than solely at the discovery site. The same pattern holds for FDA Orange Book patents, linking GWAS spillovers to marketed small-molecule therapies (Table SM3).

The innovation returns to funding basic research

Quantifying these spillovers is essential for assessing the ROI of basic research. Following established literature (*1–3*), the benefits for the U.S. innovation ecosystem can be expressed as additional patented inventions that would not have occurred in the absence of GWAS discoveries. Using the estimated coefficients above and applying them to the set of gene–disease associations first identified by GWAS, we estimate that these studies are causally associated with 2,071 additional U.S. corporate patents that focus on those pairs each year [see SM for calculations]. Accounting for spillovers to related genetic targets further magnifies these figures. For instance, extending the same calculation to genes within two steps of the pathway

network yields an additional 12,377 yearly U.S. corporate patents that focus on these neighboring genes due to GWAS. These numbers suggest that up to 86% of the innovative response to genomics research accrues to related targets in the surrounding knowledge space rather than being concentrated only on the gene–disease pairs reported.

We next consider the costs of the 3,695 distinct GWAS in our sample, excluding meta-studies that reuse existing data. Conservative calculations using NIH RePORTER indicate that direct NIH funding exceeds US$63 million per year. Because many GWAS acknowledge multiple funders, we complement this figure by directly estimating total study costs. For each GWAS, we multiply its reported sample size by the average genotyping cost per person in the year of publication, which yields total direct genotyping expenditures of about US$205 million per year between 2005 and 2020. Since these figures omit salaries, infrastructure, and overhead, they provide a lower bound on total public investment in GWAS. In the SM, we show that combining these cost estimates with the patent responses above implies that an additional US$10 million invested in GWAS yields 6.5 additional private-sector patents focused on the newly discovered targets or closely related genes. This return is more than twice the leading estimate that biomedical research yields approximately 2.7 additional corporate patents per US$10 million of NIH grant funding (*2*). In the context of genomics, incorporating spillovers therefore more than doubles the implied cost-effectiveness of investments in basic science.

Although our analysis focuses on GWAS and biomedical pathways, the underlying logic is not specific to genomics (*6*). Most domains are organized around identifiable entities and relationships that shape where firms search for innovations. In such settings, the main consequence of a basic discovery is to reallocate attention across a neighborhood of related elements rather than simply initiating applied projects that cite the original study. Our implementation relies on current KEGG pathway coverage and text-mining procedures that map patents to genes and diseases; therefore, it is likely to miss spillovers that propagate through incompletely mapped relationships or through more qualitative shifts in understanding. Because we hold disease fixed, the estimates also underestimate spillovers that move across related diseases. Even with these limitations, we demonstrate how transitioning from citation counts to knowledge spaces can recover portions of the innovative response that standard indicators overlook. In our data, for example, we identify 1,781 U.S. patents that explicitly cite a GWAS publication. Even un-

der the strong assumption that each citation reflects an indispensable input to the invention, this total would still be lower than our estimates.

For policymakers, our findings imply that citation-based ROI estimates are best viewed as lower-bound proxies rather than comprehensive measures of the innovative returns to basic research. The large spillovers we document are consistent with other evidence that direct, traceable links between public science and commercial technologies underestimate industry's dependence on basic research. Fewer than 10% of drugs approved since 1985 list government-funded patents in the Orange Book (*15*), yet surveys of laboratory managers indicate that public sector science informs more than 40% of industrial drug R&D projects (*5*). More broadly, our results highlight a structural asymmetry in the science policy debate. The costs of basic research are paid up front and recorded in objective budgets, whereas many of the benefits diffuse through knowledge spaces, accumulate gradually, and are only partially registered in citations. Narrowing this measurement gap is crucial if research funding decisions are to accurately reflect the contribution that basic science makes to the innovative capacity of the U.S. economy.

Acknowledgments

This research was supported in part by the National Center for Biotechnology Information of the National Library of Medicine, National Institutes of Health (NIH). The contributions of the NIH author(s) are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services. We are grateful to Heidi Sofia and Mike Feolo for their support and feedback. Bikash Kumar Panda provided excellent research assistance in building the data used in this analysis. M.T. thanks the NBER i3 Fellowship and the Mack Institute for Innovation Management at the Wharton School for funding support. All errors remain with the authors.

Authors' Contributions

M.T.: Conceptualization, Methodology, Validation, Writing – original draft, Writing – review and editing, Supervision, Project administration, Funding acquisition. C.G.: Data curation,

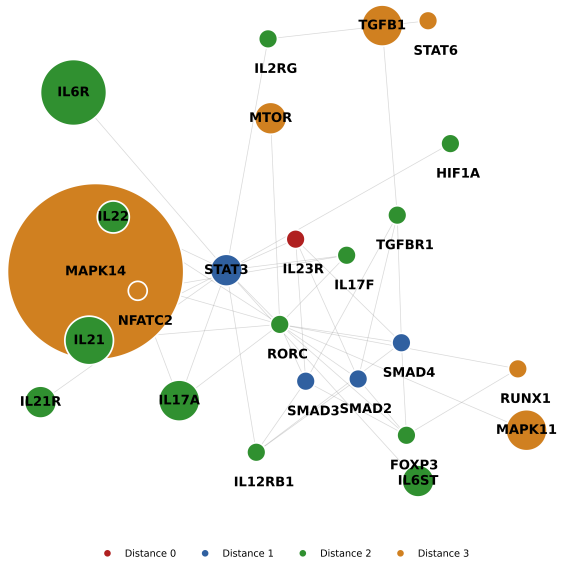
Formal analysis, Visualization, Writing – review and editing.

Conflict of Interest Statements

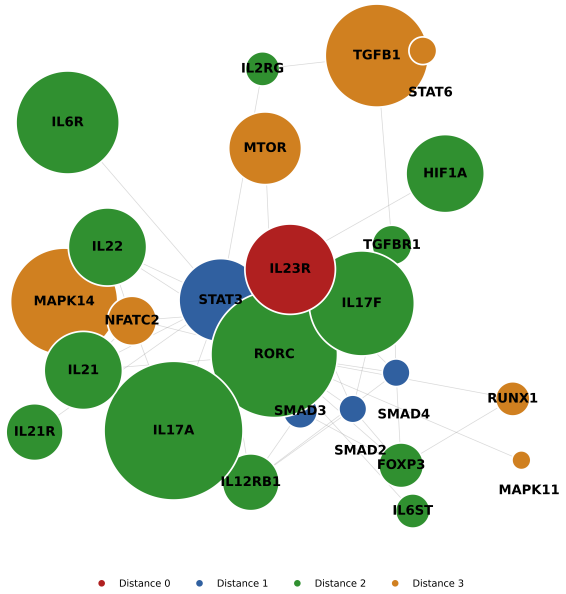
The authors declare no competing interests.

References and Notes

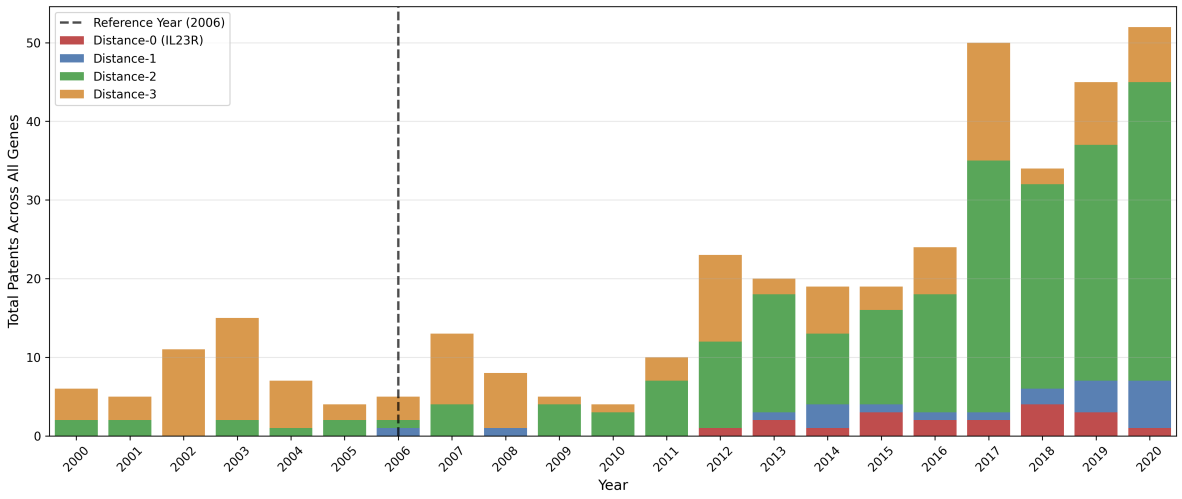
1. D. Li, P. Azoulay, B. N. Sampat, *Science* **356**, 78 (2017).
2. P. Azoulay, J. S. Graff Zivin, D. Li, B. N. Sampat, *The Review of Economic Studies* **86**, 117 (2019).
3. L. Fleming, H. Greene, G. Li, M. Marx, D. Yao, *Science* **364**, 1139 (2019).
4. S. Kwon, *Science* **385**, 936 (2024).
5. W. M. Cohen, R. R. Nelson, J. P. Walsh, *Management Science* **48**, 1 (2002).
6. P. Aceves, J. A. Evans, *Organization Science* **35**, 788 (2024).
7. E. Callaway, *Nature* **546**, 463 (2017).
8. V. Tam, *et al.*, *Nature Reviews Genetics* **20**, 467 (2019).
9. C. Macilwain, *Nature* **465** (2010).
10. A. C. Furnas, N. Fishman, L. Rosenstiel, D. Wang, *Science* **389**, 1195 (2025).
11. P. Azoulay, M. Clancy, D. Li, B. N. Sampat, *Science* **389**, 1303 (2025).
12. C. Peng, F. Xia, M. Naseriparsa, F. Osborne, *Artificial Intelligence Review* **56**, 13071 (2023).
13. M. Cerezo, *et al.*, *Nucleic Acids Research* **53**, D998 (2025).
14. M. Tranchero, C. Guthmann, B. K. Panda, *Zenodo* (2026).
15. L. L. Ouellette, B. N. Sampat, *JAMA Health Forum* **5**, e243775 (2024).



(a) Genetic targets in U.S. corporate patents on IBD before 2006

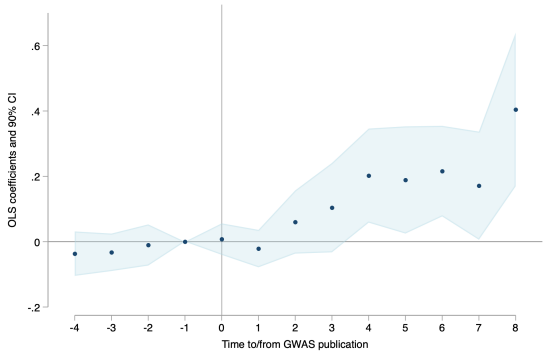


(b) Genetic targets in U.S. corporate patents on IBD after 2006

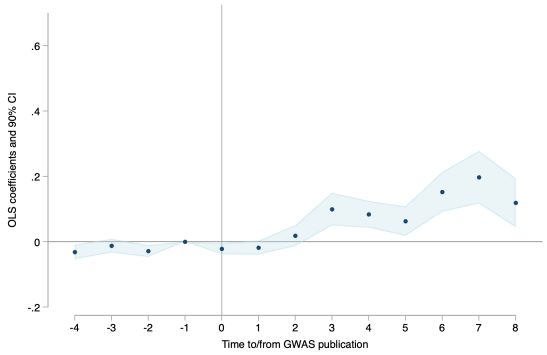


(c) Annual counts of U.S. corporate patents related to IBD, by network distance from IL23R in the Th17 pathway

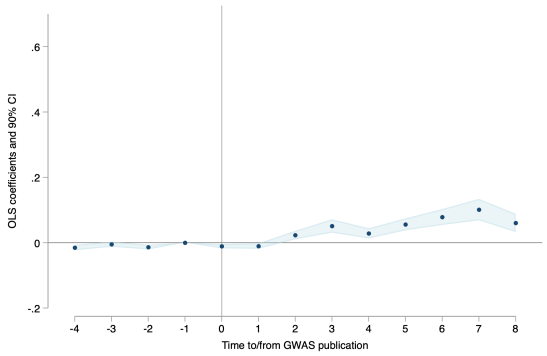
Figure 1: Corporate innovation in IBD shifts toward genes near IL23R after a GWAS discovery. Panels (a) and (b) show average yearly U.S. patents on genetic targets for IBD, before and after the publication of the GWAS first linking it to the gene IL23R. Nodes represent Th17 pathway genes and edges represent activation or inhibition relationships [see SM for a representation of the Th17 pathway]. The network layout is held fixed across panels and node size is proportional to the number of patents that reference each gene. Panel (c) shows yearly counts of U.S. corporate patents on IBD that reference Th17 genes as targets, separated by each gene's network distance from IL23R.



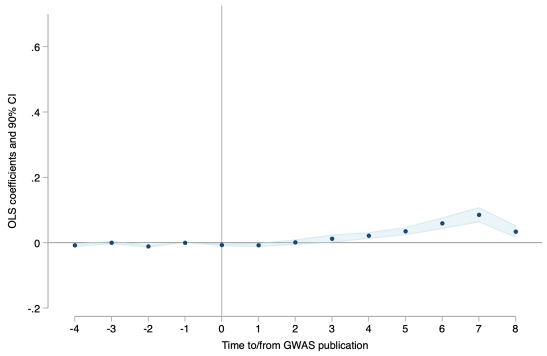
(a) Focal gene–disease pair (distance 0)



(b) Gene–disease pairs at distance 1 in the pathway



(c) Gene–disease pairs at distance 2 in the pathway



(d) Gene–disease pairs at distance 3 in the pathway

Figure 2: GWAS discoveries cause significant and sustained increases in U.S. corporate patents on focal and nearby gene–disease pairs. Each panel reports coefficients from an event study OLS regression that tracks the cumulative change in U.S. patents mentioning gene–disease pairs at different network distances from a GWAS discovery, relative to the year before the discovery. Shaded areas indicate 90% confidence intervals. All models include gene–disease, gene–year, and disease–year fixed effects [see SM for regression tables and specification details]. Panel (a) shows the patenting increase for the gene–disease association reported in the GWAS. Panels (b), (c), and (d) show the increase in patenting for gene–disease pairs that are not reported in the GWAS but lie in the same biological pathways at increasing network distances from the focal gene.

The applied value of basic research is underestimated

Supplementary Information

M. Tranchero,[1*] C. Guthmann[2]

[1]The Wharton School, University of Pennsylvania
[2]National Center for Biotechnology Information, NLM, NIH

[*]To whom correspondence should be addressed: mtranc@wharton.upenn.edu.

The PDF file includes:

Materials and Methods

Figures SM1 to SM6

Tables SM1 to SM7

Materials and Methods

Detailed information on materials and methods is provided in this supplement and organized as follows:

1) Data sources
2) Extracting bio-entities from patent text
3) Data cleaning procedures
4) Descriptive statistics
5) Case study: Inflammatory Bowel Disease (IBD)
6) Estimating corporate patents generated by GWAS studies (2005–2020)
7) Estimating total investment in GWAS studies (2005–2020)
8) Estimating ROI of GWAS studies (2005–2020)

1) Data sources

Our analysis integrates data from the following primary sources:

- **The GWAS Catalog**: To identify scientific discoveries, we use the NHGRI-EBI GWAS Catalog (version 1.0.2), a manually curated repository of genome-wide association studies (GWAS) and the resulting associations.[1] We focus on GWAS published between 2005 and 2020. The start date corresponds to the emergence of GWAS as a research design, while the endpoint aligns the discovery data with our patent panel. For our analysis, we consider all studies that introduce at least one novel association between a disease and a protein-coding gene. The Catalog reports genes using standardized NCBI Gene identifiers and traits using Experimental Factor Ontology (EFO) terms, which we later harmonize to MeSH in order to link GWAS findings to patent data. We keep in our sample only GWAS related to a disease, syndrome, or abnormality with health implications.

[1] https://www.ebi.ac.uk/gwas/docs/file-downloads

- **Google Patents**: To observe downstream private-sector innovation, we use the full text of U.S. patent documents accessed through Google Patents.[2] This source provides searchable full-text patent specifications and associated bibliographic information for the corpus of granted U.S. patents. We use patents granted between 2000 and 2020: the early years provide a pre-GWAS baseline, while the 2020 endpoint limits truncation bias from patent grant delays. We use these documents to measure pharmaceutical innovation and, based on the language of the patent itself, infer which genes and diseases the underlying inventions target. Data on patent citations to scientific publications are from the Reliance on Science database.[3]

- **KEGG HSA Pathway Database**: To measure biological proximity across genetic targets, we construct a gene-gene network from the Kyoto Encyclopedia of Genes and Genomes (KEGG) human pathway database.[4] We first retrieve all human pathway identifiers and then download pathway-specific KGML files. These XML files report gene entries and their pairwise relations. After parsing all gene-gene relations and treating the resulting graph as an undirected network, we obtain 51,228 unique gene pairs. This network provides the biological space used to define distances between a GWAS discovery and other genetic targets.

- **NIH RePORTER**: To estimate public funding associated with GWAS research, we query NIH RePORTER using the grants acknowledged by the GWAS papers in our sample.[5] We use these data to recover project-level funding records, link grants to publications, and estimate the extent of NIH investment in GWAS studies.

- **Orange Book**: To identify patents linked to marketed pharmaceuticals, we use the FDA's *Approved Drug Products with Therapeutic Equivalence Evaluations*, commonly known as the Orange Book.[6] The Orange Book links approved small-molecule drugs to the

[2] https://patents.google.com/

[3] M. Marx, and A. Fuegi (2020). "Reliance on Science: Worldwide Front-Page Patent Citations to Scientific Articles", Strategic Management Journal, 41(9):1572-1594.

[4] https://rest.kegg.jp

[5] https://reporter.nih.gov/

[6] https://www.fda.gov/drugs/drug-approvals-and-databases/orange-book-data-files

patents and regulatory exclusivities reported for them. We use these data as a complementary source to identify the subset of patents in our sample that protect approved drug products.

2) Extracting bio-entities from patent text

Our objective is to measure how basic scientific discoveries reshape corporate innovation. The standard approach in the literature is to infer science-to-technology linkages using explicit citations from patents to publications. However, patent-to-paper citations are, at best, an imperfect proxy for the technological use of scientific discoveries. They are affected by errors of omission, because patents may build on a scientific finding without citing the focal paper, and by errors of commission, because patents may cite scientific articles as general background even when the underlying invention is not substantively based on the article's finding. These limitations are especially severe for basic research, whose downstream effects often unfold through follow-on work and become embedded in later inventions without leaving a clean citation trail.

For this reason, we adopt an entity-based measurement strategy. Rather than tracing the impact of science through citation links between documents, we recover the substantive content of inventions from the entities discussed in patent text. This approach enables us to measure whether a scientific discovery redirects inventive effort toward particular parts of the underlying knowledge space. Entities mentioned in a patent's text provide a closer proxy for that allocation of effort, since they reveal the targets, phenomena, or problem areas targeted by corporate inventive efforts. This approach also captures patents whose technical content is substantively related to a scientific advance, even when the corresponding paper is not cited, while excluding patents that cite a publication solely as background and do not operate within the same scientific domain.

We implement this approach in our context by assembling a comprehensive dataset of U.S. utility patents granted between 2000 and 2020. The analysis period begins five years before the first GWAS discoveries in our sample, allowing us to observe baseline inventive activity before the scientific shock, and extends through 2020 to capture downstream responses while

limiting truncation from patent grant delays. To focus on technological domains most relevant to biomedical and pharmaceutical innovation, we restrict attention to patents classified under the IPC/CPC codes C09, C08, A61, G01, A62, A24, C10, C07, A01, C05, C12, C11, A23, C13, C06, C14, and G01N. For each patent, we retrieve detailed textual information from Google Patents, including the title, abstract, and claims section. These text fields provide the basis for our natural language processing pipeline.

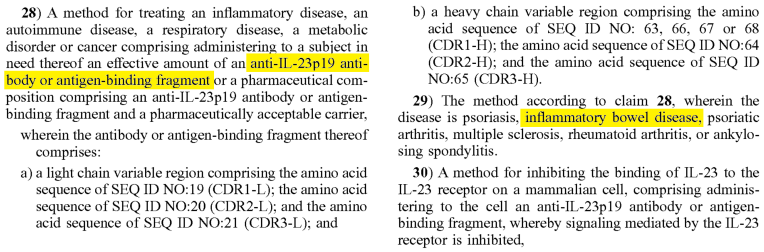
To extract biomedical entities from patent text, we use BERN (Biomedical Entity Recognition and Normalization), a neural network framework for biomedical named entity recognition and normalization built on BioBERT.[7] BERN is well-suited to our setting because it performs two tasks. First, it identifies entity mentions in raw text. Second, it maps these mentions into standardized identifiers and taxonomies. This normalization step is essential because the same underlying biological concept often appears under multiple synonyms or naming conventions. In practice, we submit patent text to the BERN RESTful API and collect the returned annotations in structured JSON format. For each detected mention, the output reports the entity, its semantic class, the corresponding normalized identifier, and a model-based confidence score. For the purposes of this paper, we retain only the two entity classes central to our analysis, namely genes[8] and diseases. Gene mentions are normalized to NCBI Gene IDs, while disease mentions are normalized to MeSH identifiers.

This measurement strategy is especially useful when scientific discoveries reshape downstream research without leaving a direct citation trail in patents. To illustrate, Figure SM1 reports selected claims from patent US 2016/0333091 A1, titled "Anti-IL-23 Antibodies," filed by Boehringer Ingelheim. As the claims make clear, the invention is substantively grounded in IL23R's role in inflammatory bowel disease. That association was first reported in a 2006 GWAS by Duerr and colleagues.[9] Nonetheless, the patent does not directly cite Duerr et al.

[7] https://github.com/dmis-lab/bern

[8] We group genes and proteins together because, in pharmaceutical and patent text, references to a target gene and to its encoded protein are often used interchangeably to denote the same biological target. In practice, the nomenclature is usually shared or closely overlapping. For example, references to *IL23R* may denote either the gene or the corresponding protein, depending on context, with both pointing to the same therapeutic target.

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



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

wherein the antibody or antigen-binding fragment thereof comprises:

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

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

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

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

Figure SM1: Claims 28, 29, and 30 of patent application US 2016/0333091 A1

Note: This figure shows selected claims from the patent application titled "Anti-IL-23 Antibodies," filed by Boehringer Ingelheim. The patent does not cite the GWAS by Duerr et al. (2006), even though it builds on the IL23R-IBD association first reported in that study. Instead, it cites later papers that themselves cite Duerr et al. (2006), consistent with the patent being part of the downstream research trajectory opened by the original finding.

(2006). Instead, it cites later papers that elaborate on the biological mechanism underlying the IL23R-IBD association and that, in turn, cite the original GWAS finding. This example clarifies why our measurement strategy focuses on biological content rather than citation links: what matters for our purposes is whether an invention operates in the scientific space first opened by a discovery, not whether it formally acknowledges the paper in which that discovery first appeared.

3) Data cleaning and panel construction

Following entity extraction, we retain, for each U.S. patent, mentions of genes and diseases, mapped to their respective NCBI Gene IDs and MeSH codes. To reduce measurement noise, we take a conservative approach and keep only entities extracted with confidence greater than or equal to 0.99. We then aggregate these data to the gene-disease-year level by counting, for each year, the number of patents mentioning a given gene-disease pair. The analysis is restricted to the 2000-2020 period. This window provides sufficient pre-period before the emergence of GWAS and sufficient post-period to limit truncation bias in patenting outcomes.

We clean the GWAS data so that scientific discoveries can be merged with the patent panel. While genes are reported by Gene ID, the GWAS Catalog reports traits using the EFO disease ontology. We therefore translate GWAS traits into MeSH using the official EFO-MeSH cross-

walk; for traits lacking a direct match, we supplement with exact and fuzzy string matching, requiring a similarity score of at least 90%. Once mapped to MeSH, we standardize all disease concepts to MeSH level 4. This level provides a middle ground between categories that are too broad to be therapeutically meaningful and categories that are too narrow to be consistently observed in patent text.[10] For each gene-disease pair discovered by GWAS, we retain the earliest publication year in which it is reported, and we restrict attention to pairs first discovered by 2020. The resulting file identifies the set of gene-disease associations first reported by a GWAS, along with the year each association entered the scientific literature. After these harmonization steps, we are left with 92,611 USPTO patents whose title, abstract, or claims mention both a genetic target and a human disease.

We next use KEGG pathways to classify biologically related genes around each GWAS discovery. We define distance over genes using KEGG's gene-gene network, while holding the disease constant. Starting from the cleaned GWAS gene-disease associations, we classify as "distance 1" all pairs that involve the same disease and a gene located one KEGG step away from the GWAS gene. We then iterate the same procedure to define pairs at distances 2 and 3. To ensure mutually exclusive treatment categories, direct GWAS pairs are excluded from the spillover sets, and any pair that qualifies at multiple distances is assigned only once, at the shortest distance from a GWAS discovery. This procedure yields a biologically grounded way to study how scientific advances propagate across nearby regions of biological space.

Note that our baseline design deliberately holds the disease fixed when measuring spillovers across genes. This choice isolates the most direct form of spillovers in our setting: whether a GWAS discovery for gene g and disease d redirects inventive activity toward nearby genes for the same disease. It does not assign treatment to other diseases that are clinically comorbid, phenotypically related, or connected through shared etiology. This restriction means our results are conservative, since GWAS discoveries may also extend into disease space. Patents that reflect this type of disease-space spillover enter our estimates only if they also mention the focal MeSH disease; otherwise, they are excluded from our spillover estimates.

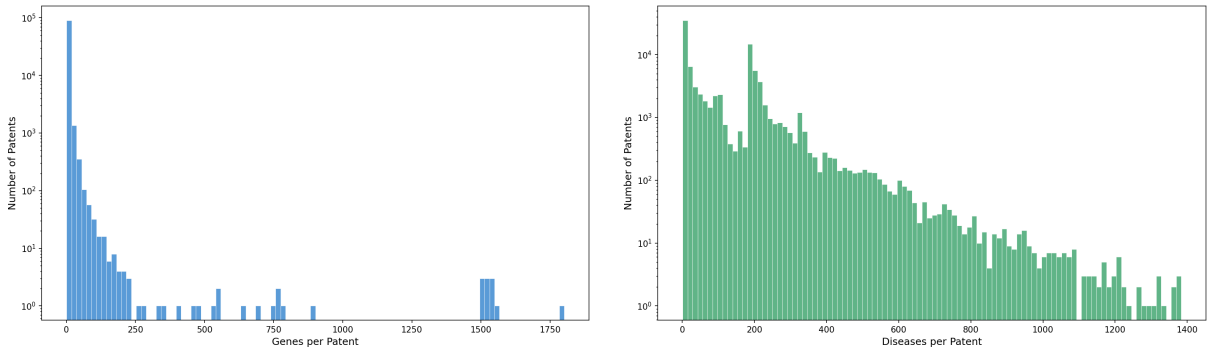
[10] More specifically, level-5 traits are aggregated to their level-4 parent, level-3 traits are expanded to their level-4 descendants, and traits at levels 1–2 or 6 and above are dropped.

Finally, we construct the analysis dataset as a balanced gene-disease-year panel for all pairs observed at least once in the patent data between 2000 and 2020. We assign a value of 0 to years with no patenting activity, then merge the GWAS treatment timing and KEGG pathway distance from the closest GWAS finding, as described above. The resulting panel distinguishes two objects central to the empirical design: whether a gene-disease pair was directly reported by a GWAS or lies near one in the biological network, and when each type of exposure first occurs.

4) Descriptive statistics

Our data identify 92,611 USPTO-granted patents whose titles, abstracts, or claims mention both a protein-coding human gene and a human disease. These patents reference 14,368 protein-coding genes and 2,435 MeSH diseases. The distribution of biological breadth across patents is skewed: the median patent mentions 2 genetic targets and 51 diseases, although some patents mention as many as 1,801 genes and 1,383 diseases. Figure SM2 summarizes this heterogeneity. This pattern is consistent with pharmaceutical patents being comparatively narrow in target space but broader in stated clinical scope. Combining the gene and disease mentions within each patent yields 7,053,555 distinct gene-disease pairs that appear at least once in patent text. The final balanced panel contains 148,124,655 pair-year observations spanning these 7,053,555 unique gene-disease pairs.

The GWAS Catalog contains information on 4,491 GWAS published between 2005 and 2020. Many of these studies, however, do not introduce new gene-disease associations that enter our analysis. After applying the cleaning and harmonization procedures described above, our sample contains 11,785 new gene-disease associations first reported in 1,402 GWAS papers. Yet the relevant biological neighborhood around them is much larger. As Table SM1 shows, the KEGG-based spillover sets contain 84,001 unique gene-disease pairs at distance 1, 277,655 at distance 2, and 333,698 at distance 3. Altogether, 707,139 unique pairs, about 10% of the full panel, lie within three biological steps of a GWAS discovery. Figure SM3 visualizes this structure. Panel (a) shows that direct GWAS discoveries are greatly outnumbered by nearby exposed



(a) Genes mentioned per U.S. patent

(b) Diseases mentioned per U.S. patent

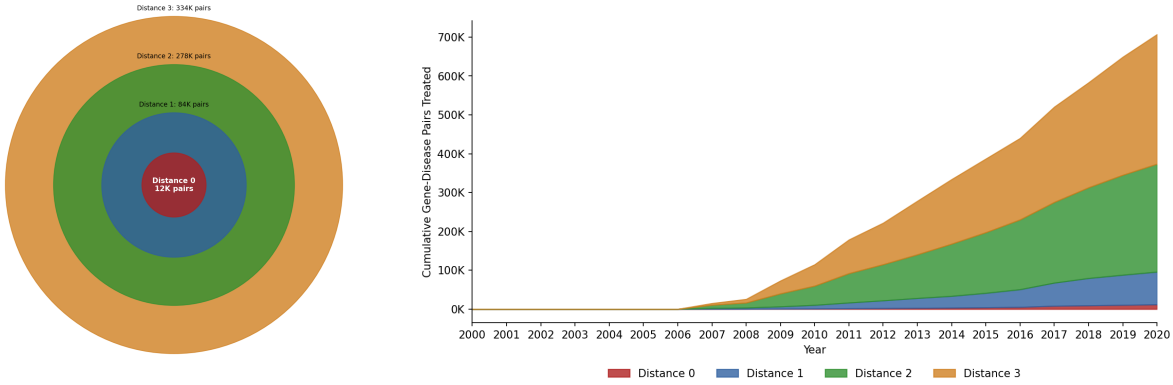
Figure SM2: Distribution of gene and disease mentions across patents in the analysis sample. Panel (a) reports the distribution of the number of distinct genes mentioned in USPTO granted patents. Panel (b) reports the distribution of the number of distinct diseases mentioned in USPTO granted patents. In both panels, the vertical axis is shown on a logarithmic scale.

pairs. Panel (b) shows that, over time, the expansion of the treated search space occurs mainly through these indirect layers. If one were to evaluate the influence of GWAS only at the focal pairs first reported in the scientific article, one would miss most of the biologically adjacent search space potentially affected by the discovery.

Table SM1: Summary of the panel by biological distance from a GWAS discovery

KEGG Distance from GWAS	Gene-Disease-Year Observations	Share of Full Panel	Unique Gene-Disease Pairs	Patents per Gene-Disease-Year	Total Patents per Gene-Disease Pair
0	247,485	0.17%	11,785	0.67	14.15
1	1,764,021	1.19%	84,001	0.76	15.93
2	5,830,755	3.94%	277,655	0.53	11.16
3	7,007,658	4.73%	333,698	0.45	9.39
Full Dataset	**148,124,655**	**100%**	**7,053,555**	**0.34**	**7.05**

Notes: The unit of observation is a gene-disease-year. Distance over genes is defined using the KEGG pathway network, holding the disease fixed. Distance 0 denotes gene-disease pairs first reported by a GWAS; distances 1, 2, and 3 denote pairs involving the same disease and genes located one, two, or three KEGG steps away from a GWAS gene. Each pair is assigned to the closest qualifying distance, making the categories mutually exclusive. The rows for distances 0–3 cover only the subset of the panel exposed, directly or indirectly, to a GWAS discovery; the full dataset also includes pairs never assigned to any GWAS-based distance category. "Patents per Gene-Disease-Year" reports average patents for each gene-disease-year, and "Total Patents per Gene-Disease Pair" reports total patents for gene-disease pairs over 2000–2020. Note that distances are defined over genes while holding the disease fixed; the classification does not incorporate disease–disease links, comorbidities, or spillovers across related disease categories.



(a) Unique treated pairs by distance

(b) Cumulative treated pairs over time by distance

Figure SM3: GWAS discoveries expose a large surrounding set of biologically related gene-disease pairs. Panel (a) reports the number of unique gene-disease pairs assigned to each GWAS-based distance category. Distance 0 denotes gene-disease pairs first reported by a GWAS, while distances 1, 2, and 3 denote pairs involving the same disease and genes located one, two, or three KEGG pathway steps away from a GWAS discovery. Panel (b) reports the cumulative number of treated gene-disease pairs over time by distance from the closest GWAS finding.

Across the full panel, the average gene-disease pair receives 0.34 patents per year and 7.05 patents cumulatively over the 2000–2020 period (Table SM1). These averages are higher for gene-disease pairs linked, directly or indirectly, to GWAS discoveries. Direct GWAS pairs receive 0.67 patents per year on average, while pairs at distances 1, 2, and 3 receive 0.76, 0.53, and 0.45 patents per year, respectively. In cumulative terms, the corresponding totals are 14.15, 15.93, 11.16, and 9.39 patents per pair. Even at the descriptive level, inventive activity appears to extend beyond the associations reported in the scientific article and into the surrounding biologically related targets.

5) Case study: Inflammatory Bowel Disease (IBD)

We use a specific scientific discovery and its surrounding biological neighborhood to illustrate the spatial logic of our measurement strategy. The empirical setting is inflammatory bowel disease (IBD; MeSH ID: D015212), a disease area in which genetic discoveries have been

important for downstream therapeutic search. We focus on the 2006 GWAS by Duerr and colleagues, which first reported an association between IL23R (interleukin-23 receptor, NCBI Gene ID: 149233) and IBD.

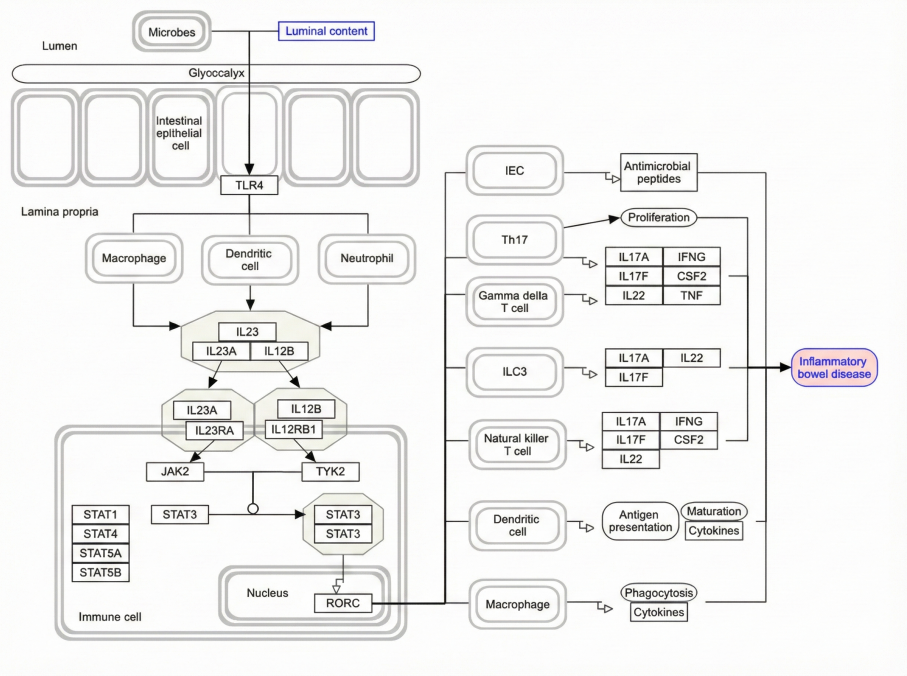


Figure SM4: Schematic representation of the IL23R-related branch of the Th17 cell differentiation pathway. The figure shows the local biological neighborhood used in the IBD case study. IL23R is the focal gene first linked to IBD by Duerr et al. (2006). Figure from WikiPathways.

This discovery is useful for our purposes not only because it was among the earliest GWAS findings, but also because its implications extended beyond the focal gene-disease pair. IL23R is embedded in the Th17 cell differentiation pathway, a biological pathway involving cytokines, receptors, and transcription factors involved in mucosal inflammation (Figure SM4). Taking IL23R as the focal target, we trace its local neighborhood in biological space by following pathway connections up to three steps away, both upstream and downstream. Intersecting this set with the patent panel yields a sample of 25 genes: the focal gene itself, 4 genes at distance 1, 13 at distance 2, and 7 at distance 3.

Between 2005 and 2020, 85 unique U.S. patents explicitly cite the GWAS by Duerr and colleagues. However, over the same period, 386 U.S. patents were granted for inventions and

therapies targeting genes in the Th17 cell differentiation pathway for IBD. As shown plastically in Figure 1, much of the innovative response therefore takes place in the surrounding pathway rather than at the focal association. The case study illustrates the logic of the full analysis: the downstream effects of a scientific discovery are not confined to the exact gene-disease pair first reported in the article, but propagate across nearby biological targets linked to the same mechanism.

6) Estimating corporate patents generated by GWAS studies (2005–2020)

Table SM2 reports the OLS estimates for U.S. patents mentioning gene-disease pairs that received GWAS associations. The coefficients are economically large and decline smoothly with biological distance from the focal GWAS discovery. For the gene-disease association first reported by a GWAS, the estimated effect is 0.176 additional patents per year, corresponding to a 52% increase relative to the sample mean of 0.336. The spillover effects remain substantial as one moves away from the focal target: 0.084 patents at distance 1, 0.039 at distance 2, and 0.017 at distance 3, which correspond to roughly 25%, 12%, and 5% of the sample mean, respectively. The estimates indicate that GWAS discoveries stimulate patenting not only on the focal association itself, but also on gene-disease pairs in the surrounding biological neighborhood. Citation-based measures centered on the original publication miss much of this downstream impact.

Table SM2: OLS Regression results: GWAS discoveries are followed by more U.S. corporate patents on focal and nearby gene–disease pairs.

	Number of U.S. patents targeting a gene-disease pair			
	(1)	(2)	(3)	(4)
GWAS (Focal gene)	0.1758**			
	(0.0840)			
GWAS (Distance 1)		0.0843***		
		(0.0212)		
GWAS (Distance 2)			0.0387***	
			(0.0070)	
GWAS (Distance 3)				0.0168***
				(0.0049)
Gene-Disease FE	YES	YES	YES	YES
Gene-Year FE	YES	YES	YES	YES
Disease-Year FE	YES	YES	YES	YES
Observations	148,119,888	148,119,888	148,119,888	148,119,888
Sample Mean	0.3356	0.3356	0.3356	0.3356

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered two-way by gene and disease in parentheses. Each column reports coefficients from a separate OLS panel regression: $Y_{i,j,t} = \alpha + \beta\, Post_t \times GWAS_{i,j,d} + \gamma GD_{i,j} + \delta_t \times Gene_i + \omega_t \times Disease_j + \epsilon_{i,j,t}$, where the dependent variable $Y_{i,j,t}$ is the yearly count of U.S. pharmaceutical patents that mention gene i and disease j in year t and $Post_t \times GWAS_{i,j,d}$ equals one after the first GWAS reports a gene–disease association at a distance d. All specifications include gene–disease fixed effects ($GD_{i,j}$), gene–year fixed effects ($\delta_t \times Gene_i$), and disease–year fixed effects ($\omega_t \times Disease_j$). In column (1), the main regressor is an indicator that equals one in years after the first GWAS that links the focal gene–disease pair (distance 0) and zero otherwise. Columns (2)–(4) report cumulative spillover effects for genes at increasing KEGG pathway distances from a GWAS gene associated with the same disease. Distance is measured as the number of steps between genes in the KEGG pathway network.

Table SM3 shows that this result is replicated, and even stronger, when the outcome is restricted to the narrower set of U.S. patents listed in the FDA Orange Book of approved products. In absolute terms, the coefficients are smaller because this outcome captures patents linked to marketed small-molecule drugs rather than the broader universe of corporate patenting. However, the magnitudes become even larger relative to the sample mean of 0.0041 Orange Book patents per gene-disease-year. The direct effect is 0.0082, or about 200% of the mean. The corresponding spillover effects are 0.0063 at distance 1, 0.0022 at distance 2, and 0.0014 at distance 3,

which correspond to approximately 154%, 54%, and 34% of the mean, respectively. This result indicates that GWAS discoveries not only stimulate patenting activity, but also the introduction of pharmaceutical products.

Table SM3: OLS Regression results: GWAS discoveries are followed by more U.S. Orange Book's patents on focal and nearby gene–disease pairs.

	Number of U.S. Orange Book patents targeting a gene-disease pair			
	(1)	(2)	(3)	(4)
GWAS (Direct)	0.0082***			
	(0.0010)			
GWAS (Distance 1)		0.0063***		
		(0.0009)		
GWAS (Distance 2)			0.0022***	
			(0.0006)	
GWAS (Distance 3)				0.0014***
				(0.0005)
Gene-Disease FE	YES	YES	YES	YES
Gene-Year FE	YES	YES	YES	YES
Disease-Year FE	YES	YES	YES	YES
Observations	148,119,888	148,119,888	148,119,888	148,119,888
Sample Mean	0.0041	0.0041	0.0041	0.0041

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered two-way by gene and disease in parentheses. Each column reports coefficients from a separate OLS panel regression: $Y_{i,j,t} = \alpha + \beta\, Post_t \times GWAS_{i,j,d} + \gamma GD_{i,j} + \delta_t \times Gene_i + \omega_t \times Disease_j + \epsilon_{i,j,t}$, where the dependent variable $Y_{i,j,t}$ is the yearly count of U.S. pharmaceutical patents in the FDA's Orange Book that mention gene i and disease j in year t and $Post_t \times GWAS_{i,j,d}$ equals one after the first GWAS reports a gene–disease association at a distance d. All specifications include gene–disease fixed effects ($GD_{i,j}$), gene–year fixed effects ($\delta_t \times Gene_i$), and disease–year fixed effects ($\omega_t \times Disease_j$). In column (1), the main regressor is an indicator that equals one in years after the first GWAS that links the focal gene–disease pair (distance 0) and zero otherwise. Columns (2)–(4) report cumulative spillover effects for genes at increasing KEGG pathway distances from a GWAS gene associated with the same disease. Distance is measured as the number of steps between genes in the KEGG pathway network.

7) Estimating total investment in GWAS studies (2005–2020)

To quantify economic magnitudes, we estimate aggregate investment in GWAS research during our sample period. Doing so is inherently difficult because there are few standardized benchmarks for GWAS costs and because the composition of studies changes over time. We therefore begin with the limited cost evidence available in the literature. Craig et al. (2009) estimate that a 4,000-person GWAS costs approximately US$1.8 million, implying a cost of roughly US$450 per participant in 2009.[11] Similarly, Visscher et al. (2012) estimate an average direct genotyping cost of about US$500 per SNP chip and infer that the first five years of GWAS amounted to a collective investment of roughly US$250 million.[12] Mills and Rahal (2019) further estimate that approximately 47.6 million individuals were genotyped globally between 2007 and 2018.[13] Under the simplifying assumption that each genotyped individual costs US$500, this would imply a global expenditure of roughly US$23.8 billion.

However, two countervailing forces shape the evolution of GWAS costs over time. On the one hand, genotyping technology becomes cheaper, lowering the cost per participant. On the other hand, GWAS sample sizes increase sharply, thereby raising the number of observations collected by each study. As a result, the cost of the average GWAS is plausibly more stable over time than chip prices alone would suggest. To account for this, we employ two complementary approaches. The first is a bottom-up estimate that imputes the direct genotyping cost of each study as a function of its sample size and publication year. The second is an estimate based on NIH grant funding acknowledged by GWAS publications. Throughout this section, we use the full set of GWAS reported in the GWAS Catalog, rather than only the 1,402 studies in our cleaned regression sample, because the goal is to characterize the overall scale of investment in the scientific enterprise from which these discoveries emerged.

[11] Craig, Jamie E., Alex W. Hewitt, Amy E. McMellon, Anjali K. Henders, Lingjun Ma, Leanne Wallace et al. (2009). "Rapid inexpensive genome-wide association using pooled whole blood," Genome Research, 19(11).

[12] Visscher, Peter M., Matthew A. Brown, Mark I. McCarthy, and Jian Yang (2012). "Five years of GWAS discovery," The American Journal of Human Genetics, 90(1).

[13] Mills, M. C., and C. Rahal (2019). "A scientometric review of genome-wide association studies," Communications Biology, 2(1), 9.

Method 1: Genotyping Costs Approach

The GWAS Catalog reports the genotyping chip used in each study. We exploit this information by collecting benchmark prices for the main arrays appearing in our data, reported in Table SM4. Because these prices are available only for a limited set of platforms and years, we use them to estimate the broad time path of genotyping costs rather than to assign an exact price tag to each study. Specifically, we fit a simple linear regression of genotyping costs over time, as shown in Figure SM5. This procedure is intended to capture the sharp decline in chip prices during the GWAS era, despite the absence of complete price information for every platform used in the sample.

Table SM4: Genotyping Array Prices by Chip Type.

Chip Name	**First Year**	**Price (USD)**	**Source**
Illumina HumanHap 550K	2007	330.00	Press Release
Illumina 1M	2008	615.00	Press Release
Affymetrix Axiom (∼1–2M, imputed)	2009	60.00	Frontiers (2018)
Illumina ImmunoChip (∼200K)	2011	39.00	Cortes et al.
Illumina Exome Array (∼250K)	2014	62.50	Grove et al.
Affymetrix Exome Array (∼250K)	2014	109.00	Iowa U. Genomics
Illumina OncoArray (∼500K)	2019	130.00	Iowa U. Genomics

Notes: The First Year column indicates the first year of appearance in our GWAS dataset.

We apply these year-specific unit costs to each GWAS in the Catalog, so the cost calculation explicitly accounts for the decline in array prices over time. For study (s) published in year (t), we compute the direct genotyping cost as Cost$s = Ns \times c_t$, where N_s is the reported sample size and c_t is the estimated per-sample genotyping cost in year t. We estimate c_t from the time trend in observed platform prices. Thus, later GWAS are assigned lower per-person genotyping costs; at the same time, later studies often include substantially larger samples, so total study costs tends to be constant over time.

Table SM5 reports the resulting estimates under several filtering scenarios. In the full sample, 4,491 studies covering 190.8 million participants imply direct genotyping expenditures of US$10.76 billion in nominal dollars ($17.09 billion in 2025 dollars). However, these figures

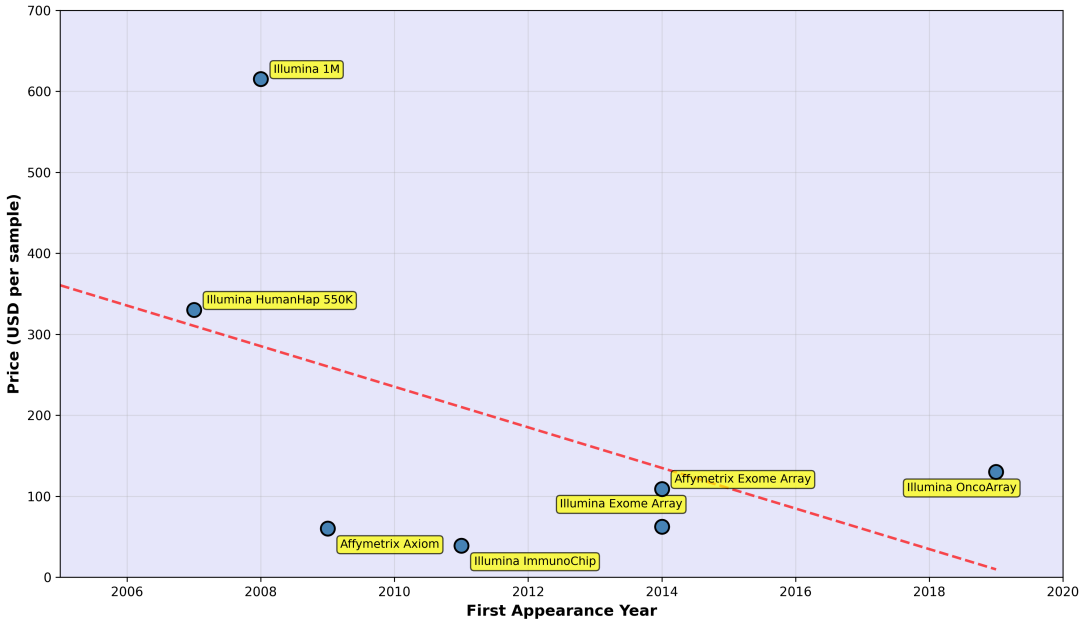


Figure SM5: Estimated per-sample genotyping costs based on observed platform prices. The dashed line shows the fitted linear decline in the cost of genotyping a single individual over time.

Table SM5: GWAS Genotyping Cost Estimates by Filtering Scenario (Full Sample)

Sample	**Studies**	**Samples**	**Cost in Dollars**
All studies	4,491	190,825,050	$10.76B
Excluding meta-analyses	4,014	155,055,236	$8.72B
Excluding large studies	4,099	38,873,361	$4.22B
Excluding meta-analyses and large studies	**3,695**	**30,770,709**	**$3.28B**

Notes: Costs are calculated at the study level as reported sample size multiplied by an estimated publication-year-specific genotyping cost. The estimated per-sample cost declines from US$360 in 2005 by US$25 per year, with a floor of US$30 per sample for 2019–2020. Meta-analyses were identified through keyword matching and by excluding studies with more than 100,000 participants, which were likely drawn from shared biobank data.

likely overstate the cost of creating original data relevant to our setting, because many GWAS reuse previously collected samples. We therefore construct a more conservative benchmark by excluding meta-analyses identified through keyword matching and studies with sample sizes above 100,000, which are more likely to rely on shared biobanks rather than on newly generated data. Under this most restrictive scenario, the sample falls to 3,695 studies and 30.8 million participants, implying direct genotyping expenditures of US$3.28 billion. This constitutes our preferred estimate of the amount invested in GWAS between 2005 and 2020.

Method 2: NIH Funding Analysis

We compare the estimates above with a calculation based on NIH grants acknowledged by GWAS publications. The purpose of this exercise is to estimate the scale of NIH funding devoted to GWAS research. To do so, we query the NIH RePORTER API for all grants acknowledged by the GWAS papers in our sample. Figure SM6 shows that NIH support is pervasive throughout the sample period and is not driven by a small number of unusually funded years. The main difficulty, however, is attributing what share of each grant should be assigned to the specific GWAS in our data. NIH grants often support multiple projects and generate multiple publications, so aggregate funding depends on how each grant is allocated across outputs.

We report two potential attribution rules in Table SM6. The first divides each grant by the number of GWAS papers in our sample that acknowledge it and then sums the resulting fractional allocations. This removes double-counting, but it likely overstates GWAS costs because the same grants also support other lines of research and because GWAS papers account for only an unspecified share of the total outputs associated with these awards. The second divides each grant by the total number of publications it produced, including non-GWAS outputs. This yields a much smaller total of US$1.01 billion and is best viewed as a lower bound, because GWAS are substantially more capital-intensive than the average publication supported by the same grant. In practice, direct NIH investment in GWAS likely ranges from US$1.01 billion to US$32.03 billion. Taken together, these estimates confirm that GWAS represent a research effort of considerable economic magnitude, sustained in large part by public funding. That said, the estimates in Table SM5 are likely closer to the full cost of the underlying research effort,

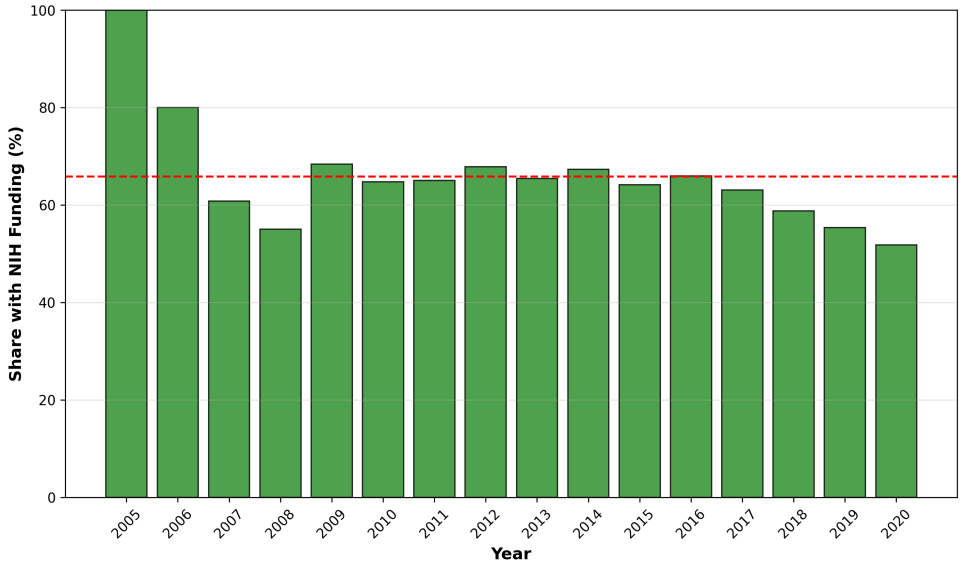


Figure SM6: Share of GWAS papers acknowledging NIH funding, by publication year. The figure shows the proportion of studies acknowledging NIH funding across the full sample of 4,491 GWAS papers. The dashed line indicates the sample-wide average.

since GWAS are funded not only by NIH but also by other public agencies, foundations, and universities. For this reason, we rely on those estimates for the ROI calculations that follow.

Table SM6: NIH Funding Estimates by Filtering Scenario (Full Sample)

Sample	**Studies**	**NIH Funding**	**Divided by GWAS**	**Divided by Total**
All studies	4,491	2,772 (62%)	$41.01B	$1.39B
Excluding meta-analyses	4,014	2,390 (60%)	$34.52B	$1.12B
Excluding large studies	4,099	2,521 (62%)	$37.02B	$1.23B
Excluding meta-analyses and large studies	**3,695**	**2,195 (59%)**	**$32.03B**	**$1.01B**

Notes: Three NIH funding attribution approaches based on NIH RePORTER data. Divided by GWAS divides each grant's value by the number of GWAS papers in the sample that cite it. Divided by Total divides each grant's value by the total number of publications supported by that grant. NIH Funding reports the number and share of GWAS papers in each scenario that acknowledge at least one valid NIH grant. Meta-analyses are identified through keyword matching. Large studies are conservatively defined as those with more than 100,000 participants.

8) Estimating ROI of GWAS studies (2005–2020)

The regressions in this paper are estimated at the gene-disease-year level, so the first step in the ROI calculation is to translate the estimated coefficients into the implied amount of additional patenting generated by GWAS discoveries across biological space. Table SM7 reports this back-of-the-envelope calculation. For each cumulative radius around a GWAS discovery, we multiply the corresponding coefficient by the number of gene-disease pairs exposed at that radius. Focusing only on the exact gene-disease associations first reported by GWAS yields 2,071.2 additional patent mentions building on the discovery. Expanding the radius to include genes within one biological step yields 8,073.6, while extending it to genes within two steps yields 14,448.8.[14] Relative to the 2,071.2 additional patent mentions on the focal pairs themselves, the remaining 12,377.6 accrue to nearby targets within two biological steps. Put differently, once one moves beyond the focal discovery site, roughly 86% of the innovation response to GWAS takes the form of inventive effort on neighboring biological targets.

Table SM7: Back-of-the-Envelope Calculation of GWAS Impact on U.S. Patenting

Cumulative Level	**OLS Coefficient** β	**Gene-Disease Pairs** N	**Effect** $\beta \times N$
0-distance only	0.1757510	11,785	2,071.2
0+1-distance	0.0843076	95,786	8,073.6
0+1+2-distance	0.0386909	373,441	14,448.8
0+1+2+3-distance (All)	0.0168481	707,139	11,914.0

Note: The OLS Coefficient β can be interpreted as the increase in the amount of patents that mention a gene-disease pair at the level in one year.

These quantities are expressed in gene-disease pairs, which are the natural units of the empirical design, but do not map one-for-one to distinct patent documents because a single patent can mention multiple pairs. We convert the pair-level counts to patent-equivalent units by dividing by the median number of gene-disease pairs per patent, which is 108. Under this conservative conversion, the direct effect of GWAS corresponds to approximately 19.2 additional patents

[14] Interestingly, the implied effect declines once the radius is extended to distance 3. This suggests some substitution or crowding-out effects, meaning that the spillover generated by a discovery weakens once one moves far enough away from the focal target in biological space.

per year focused on the exact associations first reported in the scientific literature. Extending the calculation to the neighborhood within two biological steps yields approximately 133.8 additional patents per year, of which 114.6 arise in nearby targets rather than at the focal pair itself.

We then compare these patent-equivalent quantities to the cost estimates from Supplementary Material Section 7. Using the conservative bottom-up estimates, the 3,695 GWAS in the filtered sample imply direct genotyping expenditures of about US$3.28 billion, or roughly US$205 million per year over 2005–2020. Combining this annual cost with the patent response of 133.8 implies that an additional US$10 million invested in GWAS yields about 6.5 additional U.S. patents focused on the newly discovered targets or closely related genes. Restricting attention only to the focal gene-disease pairs would instead imply roughly 0.9 additional patents per US$10 million. For comparison, Azoulay et al. (2019) estimate that US$10 million invested in NIH extramural grants yields approximately 2.7 additional corporate patents, a substantially lower amount.[15]

Citation-based approaches would also miss much of this impact. In the full GWAS Catalog, only 534 studies, or 11.9%, have been cited by at least one U.S. patent, for a total of 1,781 citing patents in our sample period. Even under the strong assumption that each citation reflects an indispensable use of the science, this total remains below 2,140.8 additional patents implied by our results once nearby biological targets are taken into account. In this setting, citation tallies are therefore best interpreted as lower-bound proxies rather than comprehensive measures of the innovation returns to basic research.

[15] Azoulay, P., Graff Zivin, J.S., Li, D. and Sampat, B.N. (2019). "Public R&D investments and private-sector patenting: evidence from NIH funding rules," The Review of Economic Studies, 86(1), 117-152.