A Knowledge-Space Approach to Estimating Spillovers from Basic Research: Evidence from Genomics*

Matteo Tranchero† Charlie Guthmann‡

October 27, 2026

Abstract

Scientific discoveries can create innovation opportunities beyond their immediate applications, often through spillovers that leave no observable link to the original research. We develop a knowledge-space approach to trace how these effects extend across related areas of knowledge. We apply it to human genomics, studying 11,785 gene-disease associations first reported in 1,402 genome-wide association studies (GWAS). Exploiting variation in the timing of genomic discoveries, we show that GWAS stimulate innovation both for the gene-disease relationships they report and among biologically related targets for the same disease. The response to a discovery attenuates with biological distance but remains visible as development progresses from patenting to clinical testing and marketed therapies. Although the effect is strongest for the reported associations, we estimate that 77% and 95% of the aggregate increases in corporate patenting and new drugs, respectively, occur among gene-disease relationships not reported by the original GWAS. These findings show that most of the downstream impact of basic research operates through spillovers, a margin that metrics based on direct applications or citations can miss.

Science is a fundamental source of technological progress and economic growth [4, 34, 35]. A central reason is that new scientific knowledge expands the opportunities for subsequent innovation [37, 45]. A discovery made for one purpose can reveal other unforeseen promising directions, guiding innovation efforts toward areas or domains that were previously overlooked [16]. These dynamic spillovers are especially important in basic research, where

*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 at the Wharton School for funding support. All errors remain with the authors. The authors declare no competing interest.

†The Wharton School, University of Pennsylvania, Philadelphia, PA 19104. Corresponding author. E-mail: mtranc@wharton.upenn.edu.

‡National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894.

downstream applications are difficult to anticipate and often emerge only after years of follow-on investments [3, 5]. By then, much of the resulting impact may be hard to quantify and attribute to the original finding. As a result, we know surprisingly little about how large the spillovers from basic research are or how far they extend through the knowledge space [18, 19, 32].

This measurement problem creates a structural asymmetry in the evaluation of public investments in science. Research costs are visible upfront in objective budgets, whereas benefits must often be inferred from the downstream outcomes that can be traced back to the original research. Empirical measures of research impact, therefore, tend to rely on citations and other observable links tying funded discoveries to downstream patents and technologies [17, 28, 30, 33]. Such measures capture important direct effects, but can miss precisely the spillovers that shape subsequent innovation without leaving a paper trail to the originating research [32]. Commonly used metrics are thus likely to understate the downstream effects of basic research [10, 23], making it challenging to weigh its broader benefits against its readily observed costs.

In this paper, we develop a knowledge-space approach to measuring spillovers from basic research. We begin with the observation that many scientific fields are organized around stable underlying entities, such as genes, materials, or topics [12]. Connections among these entities define a conceptual space in which scientific discoveries and subsequent innovations can be located [1, 15]. The relationships that structure this space provide a natural measure of "distance" from an original finding, allowing us to trace how downstream innovation changes across related areas of knowledge [36]. We can then distinguish applications that build directly on the original finding from those that reflect spillovers to related opportunities.

We apply this novel approach to human genomics, where basic scientific discoveries play a key role in understanding disease and guiding the search for new therapies. We focus on genome-wide association studies (GWAS), which identify associations between genes and human diseases [39, 42, 44]. Genes, and the proteins they encode, can serve as targets for therapeutic intervention. Therefore, GWAS findings can identify potential targets for subsequent drug development [40], even if their translational value remains debated [8]. We represent each finding as a location in a gene-disease knowledge space. To define proximity around that location, we use biological pathways, which describe how genes and their products interact within biological processes [24]. We then examine whether a GWAS discovery changes subsequent innovation at the gene-disease relationship it identifies and among biologically related genes for the same disease.

Empirically, we study new gene-disease associations discovered between 2005 and 2020 [9] and measure subsequent innovation using U.S. patents, clinical trials, and marketed therapies [14, 27, 46]. Our design exploits the staggered timing of GWAS [40], comparing changes in downstream activity around newly discovered gene-disease pairs with contemporaneous changes at otherwise similar pairs. We find that GWAS discoveries stimulate innovation involving the gene-disease relationships they report, but also generate large spillovers to biologically related targets. These effects decline with biological distance but persist as drug development progresses to clinical testing and marketed therapies. In aggregate, 77% of the increase in patenting and 95% of the increases in both clinical trials and therapies attributable to GWAS occur beyond the associations identified by such studies.

Our contribution is to introduce a strategy for measuring spillovers from basic science

when downstream innovation cannot be directly linked to the research that generated it. Prior work has measured research and development (R&D) spillovers across firms, geographic areas, and technologies [5, 7, 21, 22, 32]. Yet defining where these effects should appear and how far they extend remains a longstanding empirical challenge [18]. We show that relationships among the underlying entities of scientific and technological activity can define meaningful proximity, allowing spillovers to be measured across related opportunities rather than only through observable links to the original research. Although our application focuses on genomics, the same logic applies whenever discoveries and downstream innovation can be organized around common underlying entities or topics. In such settings, a knowledge space design provides a general framework for measuring spillovers.

1 Quantifying Science Spillovers in Knowledge Space

We begin by representing innovation as a combinatorial search problem over a large knowledge space [2, 15, 26, 45]. Each location corresponds to a combination of knowledge entities, and innovators allocate effort toward those they expect to be most promising [40]. Against this backdrop, scientific knowledge can serve as a map of the landscape, helping innovators focus their search [16]. When different locations involve functionally related entities, a discovery concerning one combination can also change expectations about others and stimulate or reduce efforts toward them [36]. For instance, consider the GWAS that first linked variants in the IL23R gene to inflammatory bowel disease (IBD) in 2006 [13]. Because IL23R is involved in a broader set of immune processes, the discovery indirectly revealed information about related biological targets. By 2019, at least 18 therapies targeting genes related to IL23R had been approved or were in clinical development for chronic inflammatory diseases, yet none targeted IL23R itself [6]. The example illustrates how the downstream consequences of a discovery can, in large part, arise through spillovers to related entities.

To make such spillovers observable, we shift the unit of analysis from documents to the knowledge entities they contain. A publication records a discovery, but the underlying contribution concerns a particular combination of entities. In the case of genomics, for instance, a location in the knowledge space corresponds to a particular gene-disease pair. A GWAS that first associates gene g with disease d introduces new information at location (g, d). Patents, clinical trials, and marketed therapies can then be mapped to the same space based on the genes and diseases they target. Downstream outputs targeting gene g for disease d are located at the focal pair (g, d), while outputs targeting another gene h for the same disease are located at (h, d). This representation reveals *where* innovation follows a discovery, regardless of whether later outputs cite the original study.

Relationships among the underlying entities provide a measure of distance in this space. In genomics, an example is given by biological pathways, which connect genes involved in shared biological processes [24] [see SI Appendix for details]. For a GWAS linking gene g to disease d, we can define the distance to another location (h, d) as the shortest path connecting genes g and h in the pathway network. For a given representation of this network, we hold these relationships fixed over time, allowing us to trace changes in downstream innovation over a common knowledge landscape. Returning to IL23R, this measure orders potential targets from IL23R itself to genes at progressively greater biological distance. If

the information revealed by a discovery is less informative about more distant biological targets, its effect on downstream innovation should attenuate with distance [18, 32, 36].

Finally, estimating these spillovers requires a counterfactual for how downstream innovation would have evolved in the absence of a discovery. We construct this counterfactual using a difference-in-differences design, comparing changes in innovation at locations exposed to a discovery with contemporaneous changes at otherwise comparable locations not exposed. In our setting, GWAS discoveries occur at different times, allowing us to estimate how innovation changes after a discovery at the focal gene-disease pair and among genes at increasing pathway distances. The aggregate importance of spillovers, however, depends not only on the size of this response but also on the number of related opportunities over which it extends. Even when effects attenuate as the neighborhood expands, spillovers can be large in aggregate if they reach a sufficiently broad set of nearby entities.

2 An Application to Genomic Research

We begin the empirical analysis by returning to the discovery linking IL23R to IBD. The NHGRI-EBI GWAS Catalog [9] identifies the 2006 study by Duerr and colleagues [13] as the first GWAS to report this association. To locate the discovery in biological space, we use human pathway data from the Kyoto Encyclopedia of Genes and Genomes (KEGG), which encode functional relationships among genes [24]. IL23R is embedded in the Th17 cell differentiation pathway, a network of functionally related genes involved in immune and inflammatory processes. We measure proximity by the shortest pathway connecting each gene to IL23R. Extending this neighborhood three steps from the focal gene yields 24 related genes, comprising 4 direct neighbors, 13 genes two steps away, and 7 three steps away.

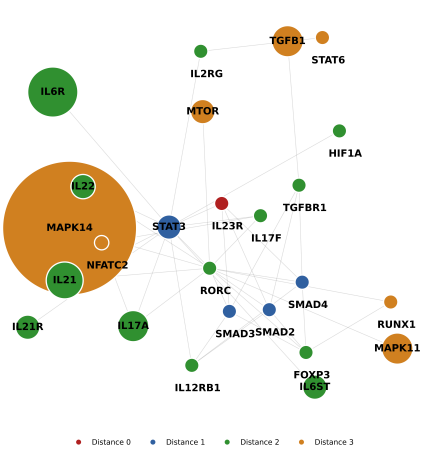
We measure inventive activity using U.S. corporate patents [29]. Patent titles, abstracts, and claims describe the technical content of an invention, allowing us to identify the genetic targets and diseases it addresses. We use BERN, a biomedical named-entity recognition and normalization framework built on BioBERT [27] to identify and normalize gene and disease mentions in patent text. We then assign each patent to the gene-disease pair(s) it contains. A patent covering an IBD therapy that targets IL23R is assigned to the same gene-disease pair identified by the GWAS, whereas a patent with a different drug target is assigned to the corresponding gene-disease pair. Aggregating patents by gene-disease pair and year yields an annual measure of inventive activity at each location in the genetic space.

Figure 1 shows how IBD-related inventive activity changes across this biological neighborhood before and after the 2006 GWAS. Panels A and B hold the structure of the Th17 network fixed and compare patenting involving each gene, with node size proportional to inventive activity. Before the discovery, patenting for therapies addressing IBD is limited in this area of the biological space. Afterward, activity is substantially higher both for IL23R and for biologically related genes. Panel C shows that annual patenting involving Th17 targets for IBD rises from an average of 8.17 patents before the discovery to 23.14 afterward, with most of the increase occurring for genes other than IL23R. The broader pattern is only partially visible through citations to the original study. Between 2006 and 2020, 85 U.S. patents directly cite the GWAS, whereas our content-based measure identifies 386 patents

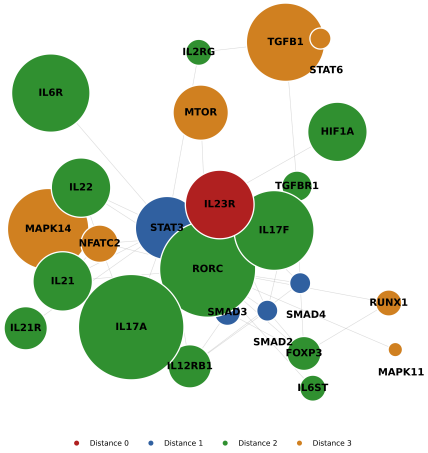
about IBD that target IL23R or related genes. The comparison illustrates the measurement gap motivating our approach: substantial inventive activity can occur in the biological neighborhood of a discovery without a citation to the original publication.

GWAS Discoveries Increase Patenting Across Biological Space

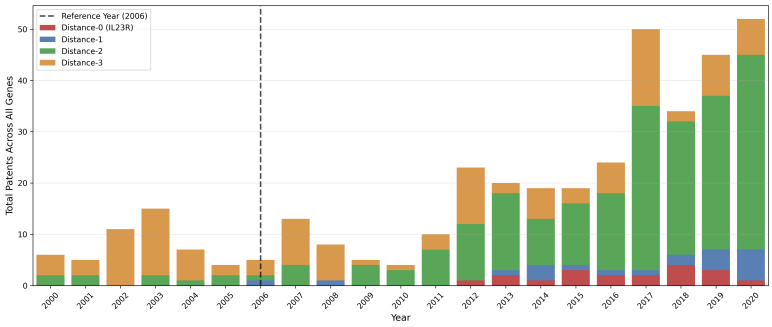
To distinguish the effect of a GWAS discovery from potential confounds, we extend the analysis to 11,785 gene-disease associations first reported in 1,402 distinct GWAS. Our difference-in-differences design compares changes in patenting after a gene-disease pair is first reported in a GWAS with contemporaneous changes at otherwise similar pairs. GWAS



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

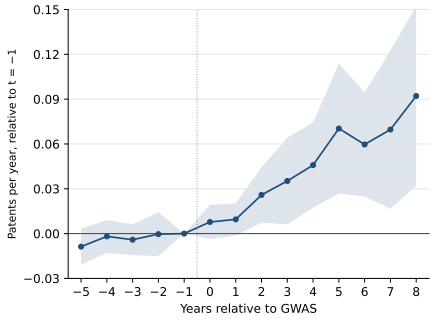


(b) Genetic targets in U.S. corporate patents on IBD after 2006

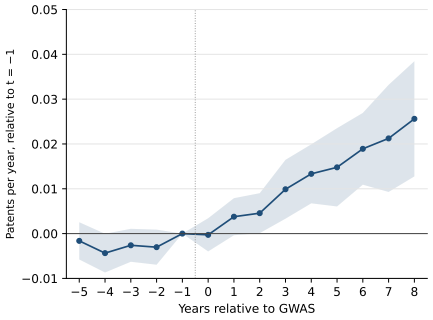


(c) Annual counts of U.S. corporate patents related to IBD, by network distance from IL23R in the Th17 pathway

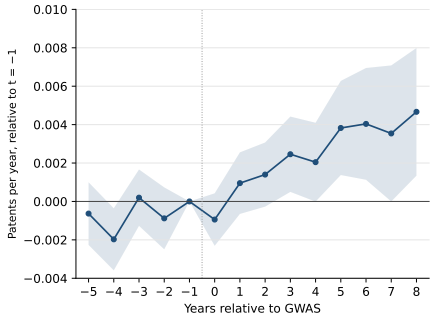
Figure 1: Corporate innovation in IBD shifts toward genes near IL23R after a GWAS discovery. Panels (a) and (b) show average yearly U.S. patents on genetic targets for IBD, before and after the publication of the GWAS first linking it to the gene IL23R. Nodes represent Th17 pathway genes and edges represent activation or inhibition relationships [see SI Figure S4 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.



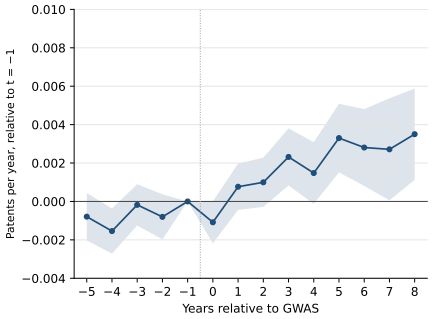
(a) Focal gene–disease pair (distance 0)



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



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



(d) Gene–disease pairs within 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 changes 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) reports the response at the gene-disease association identified by the GWAS. Panels (b), (c), and (d) progressively expand the treated neighborhood to include gene-disease pairs involving the same disease and genes located within one, two, or three pathway steps of a GWAS gene, respectively. See SI Appendix Section 6 for specification details and corresponding difference-in-differences estimates.

scan the genome broadly rather than targeting particular genes ex ante [41, 42]. Conditional on the disease being studied, this makes the timing of a particular gene-disease association less likely to be driven by prior inventive activity at that pair [40, 44]. Gene-disease fixed effects absorb persistent differences across pairs, while gene-year and disease-year fixed effects control for trends affecting the same gene or disease more broadly. The resulting coefficient is identified from the pair-specific change in patenting when an association is first reported, net of these broader shifts.

Figure 2 shows that the pattern observed for IL23R generalizes across GWAS discoveries. Patenting is relatively stable in the years preceding a discovery and rises afterward, both at the focal gene-disease pair and within broader biological neighborhoods. At the focal pair, the

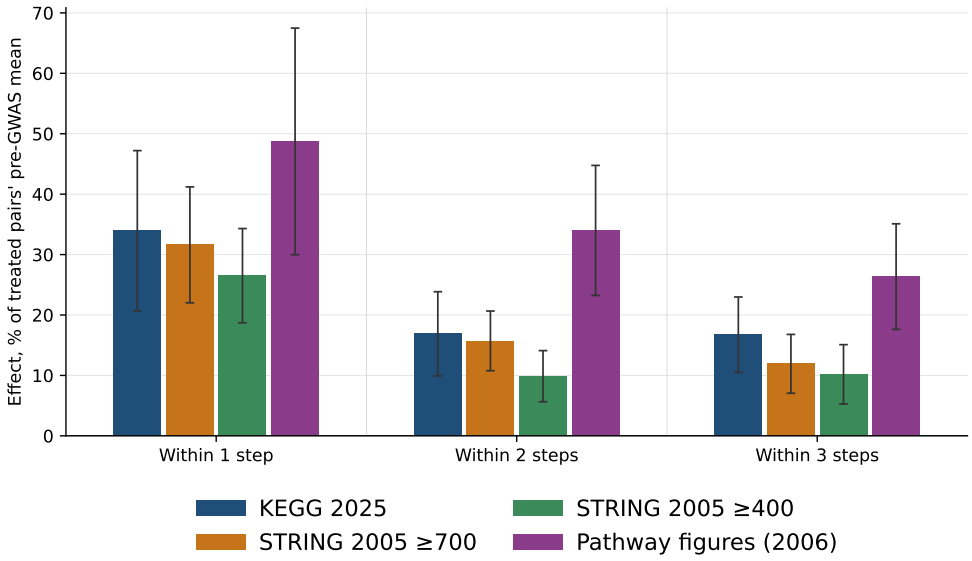


Figure 3: The figure reports difference-in-differences estimates at increasing biological distances from a GWAS discovery, expressed as percentage increase over pre-GWAS patenting among treated pairs. Biological distance is defined using the 2025 KEGG network (our baseline estimate), the 2005 STRING network at two confidence thresholds, and gene relationships extracted from pathway diagrams published before 2006. Error bars report 90% confidence intervals. The estimate for the gene-disease pairs reported by GWAS is not reported because it does not depend on the gene network.

difference-in-differences estimate implies an increase of 0.038 patents per year, or 69% relative to the pre-GWAS mean (SI Appendix, Table S2). There are substantial spillovers as we expand the neighborhood beyond the reported association. When we include genes directly connected to the GWAS gene, patenting increases by 0.013 patents per year, equivalent to a 34% increase. Including more biologically distant genes, the effects decline but remain sizable at roughly 17% of the sample mean. All estimates are statistically significant at the 1% level. The estimates reveal a clear spatial gradient, with the largest response at the focal association and progressively smaller effects as the biological neighborhood expands.

A challenge for our approach is that the network used to measure biological distance may itself contain knowledge generated after the discoveries we study. The version of KEGG that we use is a contemporary representation of gene relationships. If GWAS contributed to establishing some links later recorded in KEGG, the network could partly incorporate the response we seek to measure. We address this concern using alternative measures of the biological space predating the GWAS era. Figure 3 reports estimates using the 2005 release of STRING, which records protein-protein associations [38], and gene relationships extracted from 827 pathway diagrams published before 2006 [20]. These historical networks differ from contemporary KEGG in both coverage and density, yet recover the same spatial findings. For the closest set of genes, the estimated increase in patenting ranges from 26% to 49% across the historical networks and remains positive and statistically significant as the neighborhood expands (SI Appendix, Tables S8 and S9). The corresponding event studies

show the same post-discovery timing under these alternative representations (SI Appendix, Fig. S6), making it unlikely that the findings are an artifact of the specific representation of the biological space used.

Additional analyses further support the empirical design. First, GWAS research is systematically concentrated in diseases with more pre-existing inventive activity, but this variation is absorbed by our disease-year fixed effects [40]. Within a disease, the relationship between prior patenting on a particular gene-disease pair and the timing of its discovery is not significant (SI Appendix, Figs. S7 and S8). Together with the absence of differential pre-trends in Figure 2, this evidence makes it difficult to explain our estimates by inventive activity already shifting toward the same targets before discovery. Second, broad patent language can map a single invention to large numbers of gene-disease pairs. Our baseline excludes patents mentioning more than 100 pairs, but the results remain qualitatively unchanged when we relax this constraint or fractionally weight each patent across the pairs it mentions (SI Appendix, Tables S3 and S4 and Fig. S5).

Spillovers Persist Through Clinical Development

Patents reveal where firms direct their inventive effort, but they capture only an early stage of the innovation process. To assess whether these spillovers extend into therapeutic development, we collect data on candidate drugs that have entered clinical testing. From ClinicalTrials.gov [46], we identify 69,449 trials with their starting date, and map each tested intervention to its genetic target and disease indication. This places clinical trials in the same gene-disease space as patents and allows us to examine whether the spatial response to GWAS discoveries translates to the development of new drugs.

Figure 4 shows that the spillovers generated by GWAS discoveries extend into clinical development. At the focal gene-disease pair, a GWAS discovery increases clinical activity by 0.0074 trials per year, equivalent to 19% of the pre-GWAS mean. For genes directly connected to the GWAS gene, the estimated increase is 0.0064 trials per gene-disease pair and year, or about a 20% increase. The response attenuates with biological distance but remains positive and statistically significant within two and three pathway steps, corresponding to average increases of 19% and 17%, respectively (SI Appendix, Table S5). We next separate trials by their highest registered phase. The spillovers remain visible even among Phase 3/4 trials, showing that the response extends into late-stage clinical development (SI Appendix, Table S6).

We next ask whether these effects shape which therapies reach the market. We use the FDA Orange Book [14] to identify patents protecting marketed drugs and DrugBank [25] to recover their genetic targets. At the focal gene-disease pair, marketed therapies increase by 0.0022 per year. Expanding the neighborhood to genes directly connected to the GWAS target yields a corresponding estimate of 0.0020. The response remains positive and statistically significant as the neighborhood expands to more biologically distant genes (SI Appendix, Table S7). We recover a similar gradient when genetic targets are identified directly from the patents associated with these therapies.

Additional analyses indicate that the downstream increase in pharmaceutical innovation does not simply reflect greater activity by researchers and firms already working on these targets. GWAS discoveries are followed by increases in the number of new researchers and

companies working on focal and nearby gene-disease relationships (SI Appendix, Tables S10 and S11). At the disease level, cumulative GWAS discoveries are associated with higher total scientific and clinical activity (SI Appendix, Table S12). Taken together, these patterns are consistent with GWAS expanding the set of opportunities that researchers and firms pursue, rather than simply reallocating a fixed amount of effort across genetic targets.

3 The Aggregate Importance of Research Spillovers

Our estimates characterize how the effect of a discovery varies across individual locations in the knowledge space. To quantify aggregate spillovers, we combine the estimates within each biological radius with the number of pairs it contains. Although the coefficient attenuates with distance, the number of potentially affected genes expands. Focusing only on the gene-disease pairs directly identified by GWAS, our estimates imply that these focal pairs appear in 451 additional patents per year. However, including genes within three pathway steps raises the yearly pair-level patenting response to 1,990, implying that 77% of GWAS' impact occurs beyond the focal associations (SI Appendix, Table S13). At later stages of

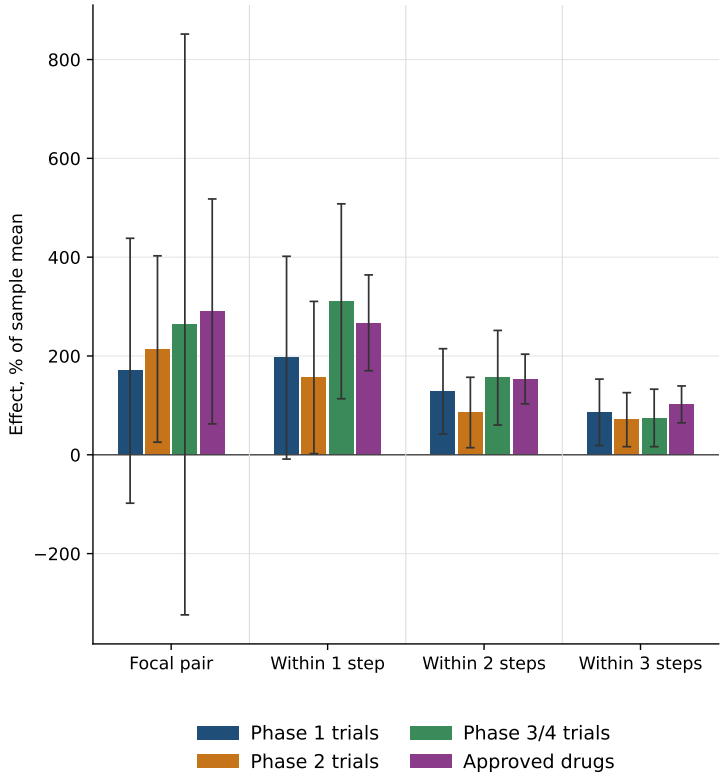


Figure 4: Spillovers from basic science persist through clinical development and shape marketed therapies. Bars report difference-in-differences estimates at increasing biological distances from a GWAS discovery, expressed as a percentage of the pre-GWAS mean among treated pairs. Outcomes include new Phase 1, Phase 2, and Phase 3/4 clinical trials and marketed therapies, identified from FDA Orange Book patents. Error bars report 90% confidence intervals.

development, biologically related targets account for 95% of the increase in both clinical trials and marketed therapies. This pattern echoes the IL23R case, where downstream therapeutic development was not concentrated on the original discovery itself [6].

Accounting for spillovers materially changes the measured research productivity of GWAS. Because our aggregate estimates are constructed at the gene-disease-pair level, a single patent can contribute to multiple pairs. For a back-of-the-envelope comparison with existing estimates, we convert these pair-level responses into patent-equivalent units (SI Appendix, Section 14). Restricting the calculation to focal associations implies that GWAS generated 38 additional corporate patents per year, compared with 166 when genes within three pathway steps are included. To benchmark these magnitudes, we divide the estimated increase in patenting by our estimate of total genotyping expenditures between 2005 and 2020 (totaling US$3.28 billion). This calculation yields 1.8 additional patents per $10 million invested when only focal associations are counted and 8.1 when spillovers are included. By comparison, previous research exploiting NIH funding thresholds estimates 2.7 additional corporate patents per $10 million in NIH funding [5]. Thus, a calculation based only on direct effects falls below this benchmark, whereas including spillovers raises the estimate to roughly three times its magnitude.

A separate comparison suggestively illustrates how much of the estimated impact may be missed by measures that require an observable link to the originating research. Across the full GWAS Catalog, only 534 studies, or 11.9%, are cited by at least one U.S. patent, yielding 1,781 citing patents between 2005 and 2020. Our estimates instead imply approximately 2,653 additional patents over the same period, once spillovers are accounted for. Although these quantities are not directly comparable, the magnitude of the gap is consistent with survey evidence that public science contributes to industrial R&D through a broader set of channels than formal acknowledgments reveal [10]. By locating downstream innovation in the same knowledge space as the original discovery, our approach recovers part of this otherwise unobserved contribution of basic science to innovation.

4 Discussion

Scientific discoveries reshape innovation in ways that go beyond the opportunities they directly identify. In genomics, we document how GWAS findings increase inventive activity across biologically related genes. These effects attenuate with biological distance but remain visible as development progresses from patenting to clinical testing and marketed therapies. In our setting, spillovers are the main channel through which research contributes to technological progress.

Our results qualify how the applied value of basic science should be interpreted. A discovery need not identify the technological target that ultimately succeeds to foster subsequent innovation. This is particularly relevant for GWAS, whose translational value remains debated because reported associations often do not reveal the causal gene or underlying biological mechanism [8, 42, 44]. Our findings show that GWAS can nevertheless redirect subsequent search toward biologically related opportunities that generate substantial downstream innovation. More broadly, much of the value of basic science may come from guiding search across related opportunities rather than from specifying the application that ulti-

mately emerges [16].

This finding has implications for how the returns to science are measured. Empirical approaches commonly begin with a publication, grant, or discovery and follow observable links to downstream outcomes. Such measures are informative when downstream applications retain a direct connection to their source, but are less well suited to capture effects that operate through spillovers. We show that this distinction is important, as estimates of research productivity change substantially once these effects are included. Additional analyses also indicate that the increase at biologically related targets does not simply reflect reallocation within a disease. At the disease level, GWAS discoveries are associated with increased total scientific, inventive, and clinical activity (SI Appendix, Table S12).

Finally, our approach provides a way to quantify spillovers in settings where scientific discoveries and subsequent innovations can be represented using common underlying entities. Genomics provides a proof of concept because genes can be linked through biological pathways, but the same logic applies in other domains where relationships among targets, components, or topics define meaningful proximity. Our implementation necessarily captures only a subset of these spillovers. It focuses on biologically related genes for the same disease and therefore misses effects that operate through relationships not represented in existing biological networks, extend across diseases, or alter scientific understanding in ways that are difficult to map. These omissions imply that our estimates are likely conservative. Nonetheless, shifting the unit of analysis from documents to the underlying knowledge space makes it possible to recover downstream effects of basic research that would otherwise remain unobserved.

Genomic discoveries and biological distance

We rely on the NHGRI-EBI GWAS Catalog [9], a manually curated repository of genome-wide association studies and the gene-disease associations they report. We retain associations linking protein-coding genes to human diseases and harmonize disease names across the GWAS and downstream innovation data. For each gene-disease association, we record the publication year of the first GWAS reporting it. Our main sample contains 11,785 new associations first reported in 1,402 GWAS between 2005 and 2020. Details on sample construction are provided in the SI Appendix.

We measure biological proximity using human pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [24]. We construct a gene network from the relationships encoded in KEGG pathways and define the distance between two genes as the length of the shortest path connecting them. For a GWAS discovery linking gene g to disease d, distance zero corresponds to the focal pair (g, d). For any other gene h paired with the same disease d, distance is given by the shortest pathway connecting h to g. Our main analysis considers cumulative neighborhoods extending up to three pathway steps from each discovery, progressively expanding the treated set from the focal pair to genes within one, two, and three steps. In the SI Appendix, we assess robustness using alternative measures of biological proximity based on historical STRING networks [38] and pathway diagrams published before the GWAS era [20].

Downstream innovation

Our primary measure of technological innovation is U.S. corporate patenting [29]. We assemble U.S. patents granted between 2000 and 2020 and use their titles, abstracts, and claims to identify the genes and diseases involved in each invention. We extract genes and diseases from patent text using BERN (Biomedical Entity Recognition and Normalization), a biomedical named-entity recognition and normalization framework built on BioBERT [27]. Gene and protein mentions are normalized to NCBI Gene identifiers and disease mentions to MeSH identifiers, allowing the content of patents to be represented using the same gene-disease pairs. We aggregate the resulting data to the gene-disease-year level. To reduce measurement noise from patents containing broad lists of biological entities, the baseline outcome excludes patents mapped to more than 100 gene-disease pairs. Results using all patents and fractional weighting are reported in the SI Appendix.

We complement patenting with two measures of therapeutic development. First, we use ClinicalTrials.gov [46] to identify 69,449 clinical trials and map each intervention to its genetic target and disease indication. Trials are dated by their start year, and we separately examine them by clinical phase. Second, we identify marketed therapies using patents listed in the FDA Orange Book [14] and use DrugBank [25] to recover the molecular targets of the protected products. These outcomes are aggregated to the same gene-disease-year level as the patent data. Additional details on data harmonization and outcome construction are provided in the SI Appendix.

Empirical design

We estimate the effect of GWAS discoveries using a gene-disease-year panel covering 2000 through 2020. For each biological radius r, we estimate:

$$Y_{gdt} = \beta_r \, \text{PostGWAS}^r_{gdt} + \alpha_{gd} + \lambda_{gt} + \mu_{dt} + \varepsilon_{gdt}, \tag{1}$$

where Y_{gdt} is downstream innovation involving gene g and disease d in year t. The indicator PostGWAS^r_{gdt} equals one beginning in the first year in which the pair is exposed to a GWAS discovery within a biological radius r. We estimate separate specifications for the focal gene-disease pair and for cumulative neighborhoods extending $r = \{1, 2, 3\}$ pathway steps from a GWAS gene. The coefficient $\hat{\beta}_r$ measures the average change in downstream innovation among gene-disease pairs exposed within radius r. The gene-disease fixed effects α_{gd} absorb persistent differences in innovative activity across pairs. Gene-by-year fixed effects λ_{gt} absorb contemporaneous changes affecting a genetic target, while disease-by-year fixed effects μ_{dt} absorb changes in research activity affecting a disease. Standard errors are clustered two-way by gene and disease.

Identification of $\hat{\beta}_r$ comes from changes in innovation for a particular gene-disease pair when it becomes exposed to a GWAS discovery, relative to pairs not exposed and net of common shocks affecting the same gene or disease. We assess the validity of the research design using event-study specifications that replace the post-discovery indicator with indicators for years relative to the first exposure. These estimates allow us to examine pre-discovery trends and the dynamic evolution of downstream innovation following a GWAS finding. In the SI Appendix, we present additional robustness checks.

Aggregate spillovers

To quantify the aggregate importance of spillovers, we combine the estimated response within each biological radius with the number of gene-disease pairs exposed within that neighborhood. For radius r, we calculate the total increase in activity as $\hat{\beta}_r \times N_r$, where N_r is the number of gene-disease pairs exposed. We compare this quantity with the corresponding increase at the focal pairs, $\hat{\beta}_0 \times N_0$, to calculate the share of the estimated effect attributable to gene-disease relationships not directly identified by GWAS. We apply the same procedure to patents, clinical trials, and marketed therapies. The SI Appendix describes the additional assumptions used to convert pair-level patent outcomes into patent-equivalent units and to compare these estimates with direct investment in GWAS research.

Processed data underlying the analyses, figures, and tables in this study are available in Zenodo [41]. All raw data are drawn from publicly available sources described in the Materials and Methods and SI Appendix, including the NHGRI-EBI GWAS Catalog, U.S. patent records, KEGG, STRING, ClinicalTrials.gov, the FDA Orange Book, PubMed/PubTator3, and NIH RePORTER. Drug target information was obtained from DrugBank subject to its terms of use.

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, the Mack Institute at the Wharton School, and the Alfred P. Sloan Foundation for funding support. All errors remain with the authors.

References

[1] Pedro Aceves and James A Evans. Mobilizing conceptual spaces: How word embedding models can inform measurement and theory within organization science. *Organization Science*, 35(3):788–814, 2024.

[2] Ajay Agrawal, John McHale, and Alexander Oettl. Artificial intelligence and scientific discovery: A model of prioritized search. *Research Policy*, 53(5):104989, 2024.

[3] Mohammad Ahmadpoor and Benjamin F Jones. The dual frontier: Patented inventions and prior scientific advance. *Science*, 357(6351):583–587, 2017.

[4] Ufuk Akcigit, Douglas Hanley, and Nicolas Serrano-Velarde. Back to basics: Basic research spillovers, innovation policy, and growth. *The Review of Economic Studies*, 88 (1):1–43, 2021.

[5] Pierre Azoulay, Joshua S Graff Zivin, Danielle Li, and Bhaven N Sampat. Public r&d investments and private-sector patenting: evidence from nih funding rules. *The Review of Economic Studies*, 86(1):117–152, 2019.

[6] Elisabetta Bianchi and Lars Rogge. The il-23/il-17 pathway in human chronic inflammatory diseases—new insight from genetics and targeted therapies. *Genes & Immunity*, 20(5):415–425, 2019.

[7] Nicholas Bloom, Mark Schankerman, and John Van Reenen. Identifying technology spillovers and product market rivalry. *Econometrica*, 81(4):1347–1393, 2013.

[8] Ewen Callaway. New concerns raised over value of genome-wide disease studies. *Nature*, 546(7659):463–464, 2017.

[9] Maria Cerezo, Elliot Sollis, Yue Ji, Elizabeth Lewis, Ala Abid, Karatuğ Ozan Bircan, Peggy Hall, James Hayhurst, Sajo John, Abayomi Mosaku, et al. The nhgri-ebi gwas catalog: Standards for reusability, sustainability and diversity. *Nucleic Acids Research*, 53(D1):D998–D1005, 2025.

[10] Wesley M Cohen, Richard R Nelson, and John P Walsh. Links and impacts: the influence of public research on industrial r&d. *Management Science*, 48(1):1–23, 2002.

[11] Jamie E Craig, Alex W Hewitt, Amy E McMellon, Anjali K Henders, Lingjun Ma, Leanne Wallace, Shiwani Sharma, Kathryn P Burdon, Peter M Visscher, Grant W Montgomery, and Stuart MacGregor. Rapid inexpensive genome-wide association using pooled whole blood. *Genome Research*, 19(11):2075–2080, 2009.

[12] Ying Ding, Min Song, Jia Han, Qi Yu, Erjia Yan, Lili Lin, and Tamy Chambers. Entitymetrics: Measuring the impact of entities. *PloS one*, 8(8):e71416, 2013.

[13] Richard H Duerr, Kent D Taylor, Steven R Brant, John D Rioux, Mark S Silverberg, Mark J Daly, A Hillary Steinhart, Clara Abraham, Miguel Regueiro, Anne Griffiths, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*, 314(5804):1461–1463, 2006.

[14] Maya Durvasula, C Scott Hemphill, Lisa Larrimore Ouellette, Bhaven Sampat, and Heidi L Williams. The nber orange book dataset: A user's guide. *Research Policy*, 52 (7):104791, 2023.

[15] Lee Fleming. Recombinant uncertainty in technological search. *Management Science*, 47(1):117–132, 2001.

[16] Lee Fleming and Olav Sorenson. Science as a map in technological search. *Strategic Management Journal*, 25(8-9):909–928, 2004.

[17] Lee Fleming, Hillary Greene, G Li, Matt Marx, and Dennis Yao. Government-funded research increasingly fuels innovation. *Science*, 364(6446):1139–1141, 2019.

[18] Zvi Griliches. The search for R&D spillovers. *The Scandinavian Journal of Economics*, 94(Supplement):S29–S47, 1992.

[19] Bronwyn H Hall, Jacques Mairesse, and Pierre Mohnen. Measuring the returns to r&d. In *Handbook of the Economics of Innovation*, volume 2, pages 1033–1082. Elsevier, 2010.

[20] Kristina Hanspers, Anders Riutta, Martina Summer-Kutmon, and Alexander R Pico. Pathway information extracted from 25 years of pathway figures. *Genome Biology*, 21 (1):273, 2020.

[21] Adam B. Jaffe. Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits, and market value. *American Economic Review*, 76(5):984–1001, 1986.

[22] Adam B Jaffe, Manuel Trajtenberg, and Rebecca Henderson. Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics*, 108(3):577–598, 1993.

[23] Benjamin F Jones and Lawrence H Summers. A calculation of the social returns to innovation. In *Innovation and Public Policy*, pages 13–60. University of Chicago Press, 2022.

[24] Minoru Kanehisa, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, et al. Kegg for linking genomes to life and the environment. *Nucleic Acids Research*, 36(suppl_1):D480–D484, 2007.

[25] Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, et al. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic Acids Research*, 52(D1):D1265–D1275, 2024.

[26] Pat Langley, Herbert A. Simon, Gary L. Bradshaw, and Jan M. Zytkow. *Scientific Discovery: Computational Explorations of the Creative Processes*. MIT Press, Cambridge, MA, 1987.

[27] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[28] Danielle Li, Pierre Azoulay, and Bhaven N Sampat. The applied value of public investments in biomedical research. *Science*, 356(6333):78–81, 2017.

[29] Alan C Marco, Amanda Myers, Stuart JH Graham, Paul D'Agostino, and Kirsten Apple. The uspto patent assignment dataset: Descriptions and analysis. *USPTO Economic working paper 2015-2*, 2015.

[30] Matt Marx and Aaron Fuegi. Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal*, 41(9):1572–1594, 2020.

[31] Melinda C Mills and Charles Rahal. A scientometric review of genome-wide association studies. *Communications Biology*, 2(1):9, 2019.

[32] Kyle R Myers and Lauren Lanahan. Estimating spillovers from publicly funded r&d: Evidence from the us department of energy. *American Economic Review*, 112(7):2393–2423, 2022.

[33] Francis Narin, Kimberly S Hamilton, and Dominic Olivastro. The increasing linkage between us technology and public science. *Research Policy*, 26(3):317–330, 1997.

[34] Richard R Nelson. The simple economics of basic scientific research. *Journal of Political Economy*, 67(3):297–306, 1959.

[35] Paul M Romer. Endogenous technological change. *Journal of Political Economy*, 98(5, Part 2):S71–S102, 1990.

[36] Daniel Schliesmann and Matteo Tranchero. Learning about roads not taken. In *Academy of Management Proceedings*, 2026.

[37] Suzanne Scotchmer. Standing on the shoulders of giants: cumulative research and the patent law. *Journal of Economic Perspectives*, 5(1):29–41, 1991.

[38] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2023.

[39] Matteo Tranchero. Data-driven search and the birth of theory: Evidence from genome-wide association studies. *University of Pennsylvania*, 2026.

[40] Matteo Tranchero. Finding diamonds in the rough: Data-driven opportunities and pharmaceutical innovation. *University of Pennsylvania*, 2026.

[41] Matteo Tranchero, Charlie Guthmann, and Bikash Kumar Panda. Biological entities in us patents - data, figures and tables. *Zenodo*, 2026.

[42] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R Martin, Hilary C Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):1–21, 2021.

[43] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[44] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[45] Martin L Weitzman. Recombinant growth. *The quarterly journal of economics*, 113(2): 331–360, 1998.

[46] Deborah A Zarin, Tony Tse, Rebecca J Williams, Robert M Califf, and Nicholas C Ide. The clinicaltrials. gov results database—update and key issues. *New England Journal of Medicine*, 364(9):852–860, 2011.

Supplementary Information

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 and panel construction
4) Descriptive statistics
5) Case study: Inflammatory Bowel Disease (IBD)
6) Empirical design and baseline estimates of corporate patents generated by GWAS studies (2005–2020)
7) Robustness to alternative patent measures
8) Other outcomes
9) Robustness to alternative representations of biological space
10) Entry and expansion of research effort after GWAS discoveries
11) Timing of GWAS discoveries and pre-existing inventive activity
12) Aggregate importance of spillovers
13) Estimating investment in GWAS research
14) Spillovers and measured research productivity

1) Data sources

Our analysis integrates data from the following primary sources:

- **The GWAS Catalog**: To identify genomic discoveries, we use the NHGRI-EBI GWAS Catalog (version 1.0.2), a manually curated repository of genome-wide association studies (GWAS) and the associations they report.[1] The Catalog reports genes using standardized NCBI Gene identifiers and traits using Experimental Factor Ontology (EFO) terms, which we harmonize to MeSH to link GWAS discoveries to downstream innovation. We retain associations involving protein-coding genes and human diseases, syndromes, or abnormalities with health implications. 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. Section 3 describes the construction of the analysis sample and the timing of first discoveries.

- **U.S. corporate 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

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

and 2020, with the early years providing a pre-GWAS baseline. 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. We identify patents assigned to firms using the USPTO Patent Assignment Dataset. Data on patent citations to scientific publications are from the Reliance on Science database [30].

- **Biological pathways**: To measure biological proximity across genetic targets, we construct a gene-gene network using human pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG).[3] We first retrieve all human pathway identifiers and then download pathway-specific KGML files as of 2025. 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.

- **Historical gene networks**: Because contemporary pathway databases may incorporate knowledge accumulated after the discoveries we study, we also construct biological distance using two historical sources. First, we use the 2005 release of STRING, a protein-association network available before the emergence of GWAS [36, 38]. We consider both high-confidence associations with a combined score of at least 700 and a broader network using a threshold of 400 as a sensitivity check. Second, we use gene relationships extracted from 827 pathway diagrams published before 2006 [20]. These historical representations differ substantially from contemporary KEGG in coverage and density, providing alternative measures of biological proximity based on knowledge available at the beginning of the GWAS era.

- **Clinical trials records**: To measure therapeutic development, we use ClinicalTrials.gov records for clinical studies whose interventions can be mapped to genetic targets and disease indications.[4] We identify 69,449 trials whose interventions can be mapped to a genetic target and whose indications can be matched to the disease concepts used elsewhere in the analysis. Each trial is assigned to the corresponding gene-disease pair and dated by its reported start year. We also classify trials by their highest registered phase, pooling Phases 3 and 4 to measure late-stage clinical development.

- **Marketed therapies**: We use the FDA's *Approved Drug Products with Therapeutic Equivalence Evaluations*, commonly known as the Orange Book, to identify patents associated with marketed drug products.[5] We link these products to DrugBank to recover their molecular targets, allowing marketed therapies to be represented in the same gene-disease space as patents and clinical trials. As a robustness check, we also construct an alternative measure in which genetic targets are identified directly from the text of the associated Orange Book patents.

[3] https://rest.kegg.jp
[4] https://clinicaltrials.gov/
[5] https://www.fda.gov/drugs/drug-approvals-and-databases/orange-book-data-files

- **Research investment in GWAS**: To estimate public funding associated with GWAS research, we query NIH RePORTER using the grants acknowledged by the GWAS papers in our sample.[6] We use these data to recover project-level funding records, link grants to publications, and estimate the extent of NIH investment in GWAS studies.

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.

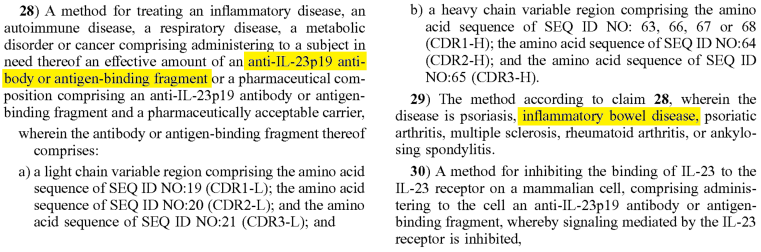
For this reason, we adopt an entity-based measurement strategy. Rather than tracing the impact of science through citation links between documents, we recover the substantive content of inventions from the entities discussed in patent text. This approach enables us to measure whether a scientific discovery affects downstream inventive activity in specific parts of the underlying knowledge space. Entities mentioned in a patent's text provide a 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

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

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



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 S1: Claims 28, 29, and 30 of patent application US 2016/0333091 A1. 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. [13], although it concerns the IL23R–IBD association first reported in that study. Instead, it cites later papers that themselves cite Duerr et al., consistent with the patent being part of the downstream research trajectory opened by the original finding.

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 S1 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 concerns the biological mechanism involving IL23R and inflammatory bowel disease (IBD). That association was first reported in a 2006 GWAS by Duerr and colleagues [13]. Nonetheless, the patent does not directly cite Duerr et al. [13]. 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,

[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.

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.

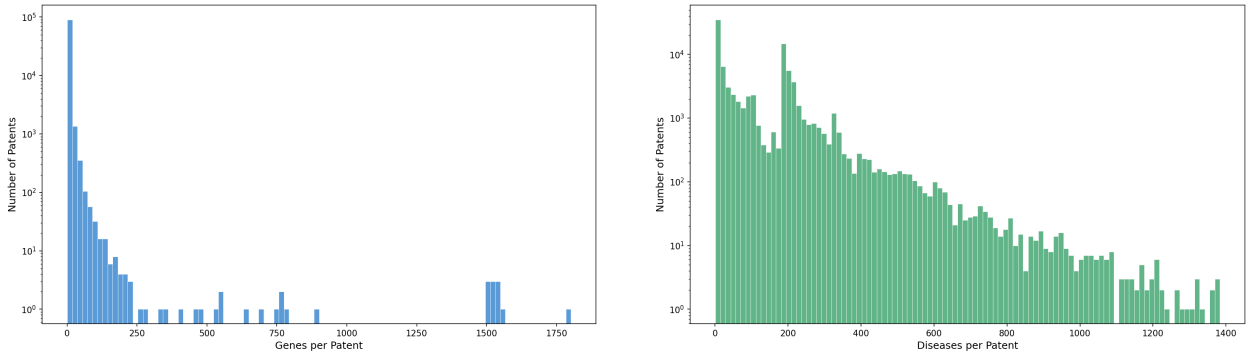
We clean the GWAS data so that scientific discoveries can be merged with the patent panel. While genes are reported by Gene ID, the GWAS Catalog reports traits using the EFO disease ontology. We therefore translate GWAS traits into MeSH using the official EFO-MeSH crosswalk; for traits lacking a direct 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.[9] For each gene-disease pair discovered by GWAS, we retain the earliest publication year in which it is reported, and we restrict attention to associations first reported between 2005 and 2020. The resulting file identifies the set of gene-disease associations first reported in a GWAS, along with the year of its publication. 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 map our additional downstream outcomes to the same gene-disease-year structure. Clinical trials are assigned to gene-disease pairs using the genetic target of the intervention and its disease indication, and are dated by their reported start year. For marketed therapies, we identify patents listed in the FDA Orange Book. We link each protected product to DrugBank to recover its pharmacological target. In our preferred measure, the gene and disease indications are taken from DrugBank. As an alternative measure, we identify the gene and disease directly from the patent text, which adds more noise but gives us consistent results. This allows patents, clinical trials, and marketed therapies to be analyzed using a common gene-disease-year unit.

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. For descriptive purposes, these distance categories are mutually exclusive: a pair that qualifies at multiple distances is assigned to the shortest distance from a GWAS discovery. The regression analysis then uses cumulative biological neighborhoods. For each gene-disease pair and radius r, we identify the earliest GWAS linking the same disease to a gene no more than r pathway steps away. The publication year of that GWAS determines when the pair enters the corresponding treatment measure. The resulting treatment definitions are nested, progressively expanding from the focal pair to genes within one, two, and three pathway steps.

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 affects inventive activity among biologically related

[9] 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.



(a) Genes mentioned per U.S. patent

(b) Diseases mentioned per U.S. patent

Figure S2: 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.

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.

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. Similarly, we merge the clinical-trial and marketed-therapy outcomes onto the same panel. 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. 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 gene-disease-year observations spanning these pairs over 2000–2020.

The distribution of biological breadth across patents is skewed (Figure S2): the median patent mentions 2 genetic targets but 51 diseases, and some patents map to more than 1,000 diseases. A significant portion of this breadth arises on the disease side, when a generic term such as “cancer” is normalized to every specific MeSH category beneath it. Because these broad disease lists can cause a single patent to contribute to very large numbers of

Table S1: Summary of the panel by biological distance from a GWAS discovery (patents mentioning more than 100 gene–disease pairs dropped)

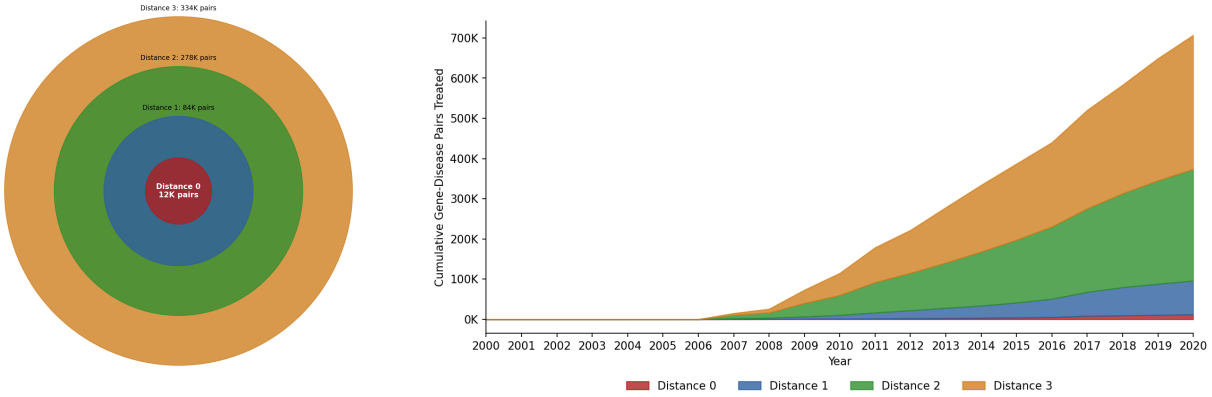
KEGG Distance from GWAS	Gene-Disease-Year Observations	Share of Full Panel	Unique Gene-Disease Pairs	Patents per Gene-Disease-Year	Total Patents per Gene-Disease Pair
0	247,485	0.17%	11,785	0.072	1.51
1	1,764,021	1.19%	84,001	0.042	0.87
2	5,830,755	3.94%	277,655	0.020	0.43
3	7,007,658	4.73%	333,698	0.014	0.29
Full Dataset	**148,124,655**	**100%**	**7,053,555**	**0.007**	**0.15**

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 1–3 denote pairs whose gene lies that many pathway steps from a GWAS gene for the same disease (non-cumulative rings). Patent counts exclude patents that mention more than 100 distinct gene-disease pairs. Total patents per pair sums over 2000–2020.

gene-disease pairs, our baseline outcome excludes patents mapped to more than 100 distinct pairs. This restriction retains 45,669 of the 92,611 patents. Among the retained patents, the median patent mentions 2 genes and 6 diseases, and the sample covers 10,160 genes, 2,396 diseases, and 455,022 gene-disease pairs with at least one patent. Importantly, this restriction changes only the patent count, not the set of gene-disease pairs in the balanced panel or their exposure to GWAS discoveries. As we show below, all results are robust to using the entire sample or different cuts.

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. The biological neighborhood surrounding these focal associations is substantially larger. As Table S1 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 S3 visualizes this structure. Panel (a) shows that direct GWAS discoveries are greatly outnumbered by nearby exposed 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.

Across the full panel, the average gene-disease pair receives 0.007 patents per year and 0.15 patents cumulatively over the 2000–2020 period (Table S1). These averages are higher for gene-disease pairs linked, directly or indirectly, to GWAS discoveries. Direct GWAS pairs receive 0.072 patents per year on average, while pairs at distances 1, 2, and 3 receive 0.042, 0.020, and 0.014 patents per year, respectively. In cumulative terms, the corresponding totals are 1.51, 0.87, 0.43, and 0.29 patents per pair. These differences are descriptive and combine activity before and after GWAS discovery. Nonetheless, they illustrate both the concentration of inventive activity near GWAS discoveries and the much larger number of biologically related opportunities over which potential spillovers can operate.



(a) Unique treated pairs by distance

(b) Cumulative treated pairs over time by distance

Figure S3: 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.

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 proved extremely important for downstream therapeutic innovations. We focus on the 2006 GWAS by Duerr and colleagues [13], which first reported an association between IL23R (interleukin-23 receptor, NCBI Gene ID: 149233) and IBD.

This example is informative 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 network involving cytokines, receptors, and transcription factors involved in mucosal inflammation (Figure S4). 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. Figure S4 shows the underlying biological pathway, while Figure 1 in the main text overlays inventive activity on this local neighborhood.

Between 2006 and 2020, 85 unique U.S. patents explicitly cite the GWAS by Duerr and colleagues. Over the same period, our BioBERT-based measure identifies 386 U.S. patents targeting IBD through IL23R or genes in its surrounding Th17 neighborhood. Most of this inventive activity involves genes other than IL23R itself. This case study illustrates the logic of our measurement strategy, since most downstream innovation can occur across biologically related targets without retaining an observable citation to the scientific study that first identified the focal association.

6) Empirical design and baseline estimates of corporate patents generated by GWAS studies (2005–2020)

We estimate the effect of GWAS discoveries using the gene-disease-year panel described above. For each biological radius r, we estimate:

$$Y_{gdt} = \beta_r PostGWAS^r_{gdt} + \alpha_{gd} + \lambda_{gt} + \mu_{dt} + \varepsilon_{gdt}, \tag{S1}$$

where Y_{gdt} is the number of U.S. corporate patents involving gene g and disease d in year t. $PostGWAS^r_{gdt}$ equals one beginning in the first year in which the pair is exposed to a GWAS discovery within a biological radius r. We estimate separate specifications for the focal gene-disease pair and for cumulative neighborhoods extending one, two, and three pathway steps from a GWAS gene associated with the same disease. The coefficient β_r measures the average change in patenting among pairs exposed within radius r compared to pairs not exposed to any discovery. All specifications include gene-disease fixed effects, which absorb persistent differences in inventive activity across pairs; gene-year fixed effects, which absorb trends affecting the same gene across diseases; and disease-year fixed effects, which absorb trends affecting the same disease across genes. Identification comes from the pair-specific change in patenting when a gene-disease pair first becomes exposed to a GWAS discovery, relative to contemporaneous patenting for pairs not yet exposed. Standard errors are clustered two-way by gene and disease.

Table S2 reports the baseline OLS estimates for U.S. corporate patenting. At the focal gene-disease pair, a GWAS discovery increases patenting by 0.038 patents per year, equivalent to 69% of the pre-GWAS mean among treated pairs. The response remains substantial as the biological neighborhood expands. For pairs within one pathway step of a GWAS gene, the

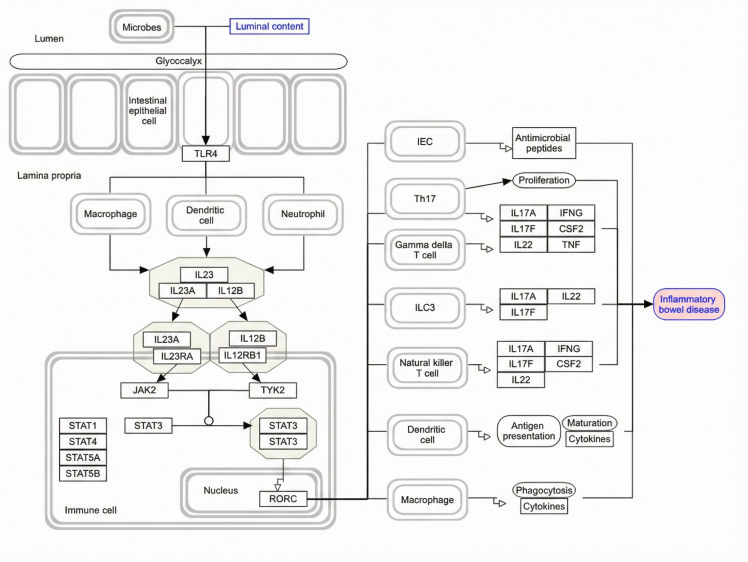


Figure S4: Biological neighborhood surrounding IL23R in the Th17 cell differentiation pathway. The figure shows the local pathway structure used to define biological proximity in the IBD case study. IL23R (reported here with the alternative name *IL23RA*) is the focal gene first linked to IBD by Duerr et al. [13]. Figure from WikiPathways.

Table S2: OLS Regression results: Baseline estimates of the effect of GWAS discoveries on U.S. corporate patenting (patents mentioning more than 100 gene-disease pairs dropped).

	Number of U.S. patents targeting a gene-disease pair			
	(1)	(2)	(3)	(4)
GWAS (Focal gene)	0.0383***			
	(0.0127)			
GWAS (Distance 1)		0.0127***		
		(0.0030)		
GWAS (Distance 2)			0.0037***	
			(0.0009)	
GWAS (Distance 3)				0.0028***
				(0.0006)
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
Pre-GWAS Mean	0.0555	0.0373	0.0221	0.0168

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered two-way by gene and disease in parentheses. Each column reports estimates from a separate OLS panel regression of the yearly number of U.S. corporate patents involving gene i and disease j on exposure to a GWAS discovery. Column (1) defines exposure as the first GWAS reporting the focal gene–disease association. Columns (2)–(4) use cumulative biological neighborhoods and define exposure as the first GWAS linking the same disease to a gene within one, two, or three KEGG pathway steps of gene i, respectively. All specifications include gene–disease, gene–year, and disease–year fixed effects. "Pre-GWAS mean" is the mean of the outcome among pairs eventually exposed in the corresponding specification, measured in years before their first exposure. The outcome excludes patents mentioning more than 100 distinct gene–disease pairs.

estimated increase is 0.013 yearly patents per pair, or 34% of the pre-GWAS mean. Within two and three pathway steps, the estimates fall to 0.0037 and 0.0028 patents per year, equivalent to approximately 17% of pre-GWAS patenting at each radius. 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 that rely solely on the original publication would miss much of this downstream impact.

Figure 2 in the main text reports the corresponding event-study estimates. The figures are obtained from OLS models using the same fixed effects as in equation S1, but with the post-discovery indicator replaced by indicators for years relative to first exposure, with the year before exposure as the omitted category. The event studies show no evidence of differential patenting trends before exposure and a sustained increase afterward across the focal pair and broader biological neighborhoods. The absence of differential pre-trends supports the identifying assumption underlying the research design.

7) Robustness to alternative patent measures

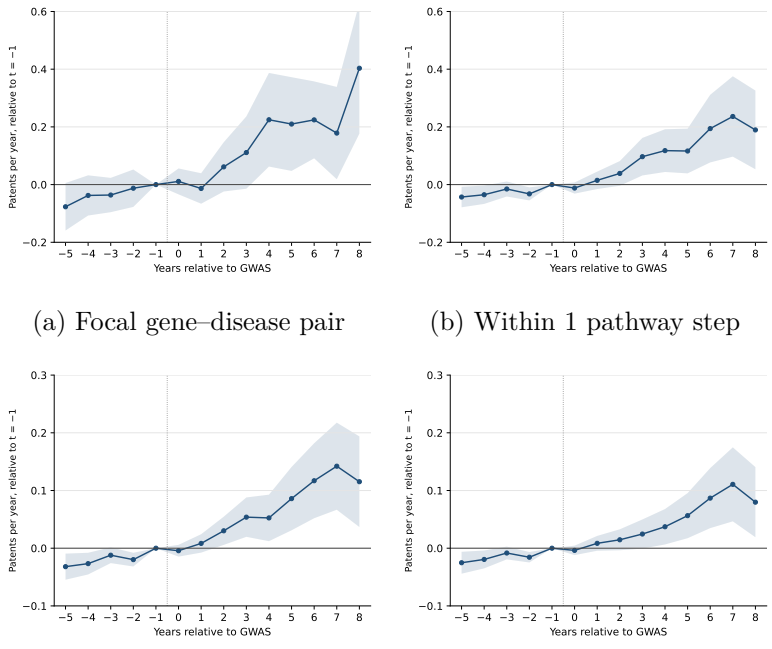
Our baseline patent outcome excludes patents that mention more than 100 gene-disease pairs because broad disease terminology can map a single patent to many more pairs. In

this section, we examine the robustness of our findings to this restriction by using three alternative samples.

First, we remove the restriction entirely and re-estimate the event-study specifications using all 92,611 patents in our data. Figure S5 shows that the dynamic pattern is similar to the baseline results. Across biological radii, there is little evidence of differential patenting trends before a GWAS discovery, followed by a sustained increase afterward. The response is largest at the focal gene-disease pair and becomes progressively smaller as the biological neighborhood expands. The corresponding difference-in-differences estimates are 0.176 additional patents per year at the focal pair and 0.084, 0.039, and 0.017 within one, two, and three pathway steps, respectively. The broader estimates are less precise when all patents are retained, consistent with the substantial noise introduced by patents mapped to very large numbers of gene-disease pairs.

Second, we retain all patents but count them fractionally according to the number of gene-disease pairs they mention. Specifically, a patent mapped to k pairs contributes $1/k$ to each pair, so that each patent receives a total weight of one. Table S3 again shows evidence of spatial spillovers. The estimated effect is 0.0024 at the focal pair and declines to 0.0007, 0.0002, and 0.0002 as the biological radius expands from one to three pathway steps. All four estimates are statistically significant. Thus, our main finding does not depend on assigning the full weight of broadly written patents to every gene-disease pair they mention.

Finally, Table S4 relaxes the baseline restriction, excluding only patents that mention



(a) Focal gene–disease pair (b) Within 1 pathway step

(c) Within 2 pathway steps (d) Within 3 pathway steps

Figure S5: Event studies by pathway distance, all patents (no cutoff on the number of gene–disease pairs per patent). Each panel reports coefficients on relative-year indicators from an OLS regression with gene–disease, gene–year and disease–year fixed effects, relative to the year before the first GWAS at that distance; 90% confidence bands from standard errors clustered two-way by gene and disease. Each panel has its own vertical axis.

Table S3: OLS regression results, all patents with fractional counting

	Number of U.S. patents targeting a gene-disease pair			
	(1)	(2)	(3)	(4)
GWAS (Focal gene)	0.0024**			
	(0.0010)			
GWAS (Distance 1)		0.0007***		
		(0.0002)		
GWAS (Distance 2)			0.0002***	
			(0.0001)	
GWAS (Distance 3)				0.0002***
				(0.0000)
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
Pre-GWAS Mean	0.0046	0.0026	0.0016	0.0013

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered two-way by gene and disease in parentheses. Each column reports estimates from a separate OLS panel regression of the yearly number of U.S. corporate patents involving gene i and disease j on exposure to a GWAS discovery. Column (1) defines exposure as the first GWAS reporting the focal gene–disease association. Columns (2)–(4) use cumulative biological neighborhoods and define exposure as the first GWAS linking the same disease to a gene within one, two, or three KEGG pathway steps of gene i, respectively. All specifications include gene–disease, gene–year, and disease–year fixed effects. "Pre-GWAS mean" is the mean of the outcome among pairs eventually exposed in the corresponding specification, measured in years before their first exposure. Patents are counted fractionally, so a patent mentioning k gene–disease pairs contributes $1/k$ to the outcome for each pair.

more than 1,000 gene-disease pairs. The estimated effect rises to 0.118 additional patents per year at the focal pair and declines to 0.054, 0.017, and 0.010 at one, two, and three pathway steps, respectively. The coefficients are positive and statistically significant, but become noisier with increasing distance from the focal discovery.

8) Other outcomes

We next ask whether the spillovers observed in patenting persist as projects move further through the development process. Using ClinicalTrials.gov, we map 69,449 trials to gene-disease pairs based on the intervention's genetic target and disease indication. The estimates in Table S5 show that they do. A GWAS discovery increases clinical activity by 0.0074 trials per year at the focal pair. The effect remains similar in relative terms as the biological neighborhood expands, with increases of 0.0064, 0.0036, and 0.0024 trials within one, two, and three pathway steps, respectively. These spillover estimates are statistically significant and correspond to roughly 17–20% of pre-GWAS clinical activity among treated pairs.

Interestingly, the spillovers also extend to later stages of clinical development. Table S6 separates trials by their highest registered phase. Although the estimates become less precise once trials are divided by phase, the effects remain positive across biological neighborhoods and are visible even among late-stage trials. For Phase 3/4 trials, our estimates correspond to increases of roughly 27%, 24%, and 17% of pre-GWAS clinical activity.

Table S4: Robustness check: Estimates excluding patents mentioning more than 1,000 gene-disease pairs

	Number of U.S. patents targeting a gene-disease pair			
	(1)	(2)	(3)	(4)
GWAS (Focal gene)	0.1180***			
	(0.0408)			
GWAS (Distance 1)		0.0541***		
		(0.0151)		
GWAS (Distance 2)			0.0165**	
			(0.0078)	
GWAS (Distance 3)				0.0100
				(0.0062)
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
Pre-GWAS Mean	0.2447	0.2271	0.1642	0.1430

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered two-way by gene and disease in parentheses. Each column reports estimates from a separate OLS panel regression of the yearly number of U.S. corporate patents involving gene i and disease j on exposure to a GWAS discovery. Column (1) defines exposure as the first GWAS reporting the focal gene–disease association. Columns (2)–(4) use cumulative biological neighborhoods and define exposure as the first GWAS linking the same disease to a gene within one, two, or three KEGG pathway steps of gene i, respectively. All specifications include gene–disease, gene–year, and disease–year fixed effects. "Pre-GWAS mean" is the mean of the outcome among pairs eventually exposed in the corresponding specification, measured in years before their first exposure. The outcome excludes patents mentioning more than 1,000 distinct gene–disease pairs.

We finally examine therapies that reach the market. Table S7 reports two measures based on patents listed in the FDA Orange Book of approved therapeutic products. Our preferred measure assigns each patent to the pharmacological targets of the protected drug using DrugBank (columns 1–4). GWAS discoveries increase this outcome by 0.0022 at the focal pair and by 0.0020, 0.0011, and 0.0008 within one, two, and three pathway steps, respectively. Relative to pre-GWAS activity, these estimates correspond to increases of 72%, 53%, 49%, and 43%, and are statistically significant throughout the biological neighborhood. As an alternative, columns 5–8 identify the gene directly from Orange Book patent text. The focal estimate is smaller and imprecisely estimated, but the spillover coefficients remain positive and statistically significant. The consistency across the two mappings shows that the spatial pattern is not driven by how we assign genetic targets to marketed drugs and remains visible at the final stage of therapeutic development.

9) Robustness to alternative representations of biological space

Our baseline analysis measures biological distance using the KEGG network as of 2025. A potential concern is that this representation incorporates biological knowledge accumulated after the GWAS discoveries in our sample. If some of the relationships recorded in KEGG were characterized partly in response to those discoveries, the network could mechanically align with the downstream response we seek to measure. In this section, we repeat the

Table S5: Alternative outcome: new clinical trials (ClinicalTrials.gov / AACT data)

	Outcome per gene-disease pair and year			
	(1)	(2)	(3)	(4)
GWAS (Focal gene)	0.0074*			
	(0.0044)			
GWAS (Distance 1)		0.0064**		
		(0.0030)		
GWAS (Distance 2)			0.0036***	
			(0.0013)	
GWAS (Distance 3)				0.0024**
				(0.0010)
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
Pre-GWAS Mean	0.0382	0.0320	0.0186	0.0139

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered two-way by gene and disease in parentheses. Each column reports estimates from a separate OLS panel regression of the yearly number of new clinical trials involving gene i and disease j on exposure to a GWAS discovery. Column (1) defines exposure as the first GWAS reporting the focal gene–disease association. Columns (2)–(4) use cumulative biological neighborhoods and define exposure as the first GWAS linking the same disease to a gene within one, two, or three KEGG pathway steps of gene i, respectively. All specifications include gene–disease, gene–year, and disease–year fixed effects. "Pre-GWAS mean" is the mean of the outcome among pairs eventually exposed in the corresponding specification, measured in years before their first exposure.

Table S6: New clinical trials by phase (ClinicalTrials.gov data)

	Outcome per gene-disease pair and year											
	New Phase 1 trials				New Phase 2 trials				New Phase 3/4 trials			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
GWAS (Focal gene)	0.0012				0.0037*				0.0017			
	(0.0011)				(0.0020)				(0.0023)			
GWAS (Distance 1)		0.0014				0.0027*				0.0020***		
		(0.0009)				(0.0016)				(0.0008)		
GWAS (Distance 2)			0.0009**				0.0015**				0.0010***	
			(0.0004)				(0.0008)				(0.0004)	
GWAS (Distance 3)				0.0006**				0.0012**				0.0005**
				(0.0003)				(0.0006)				(0.0002)
Gene-Disease FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Gene-Year FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Disease-Year FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Observations	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888
Pre-GWAS Mean	0.0074	0.0061	0.0037	0.0029	0.0157	0.0168	0.0096	0.0072	0.0123	0.0073	0.0042	0.0030

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered two-way by gene and disease in parentheses. Each column reports estimates from a separate OLS panel regression of the yearly number of new clinical trials involving gene i and disease j on exposure to a GWAS discovery. Trials are ClinicalTrials.gov studies mapped to the gene–disease pair (AACT); the outcome does not use patent data. Within each panel, the first column defines exposure as the first GWAS reporting the focal gene–disease association, and the next three use cumulative biological neighborhoods, defining exposure as the first GWAS linking the same disease to a gene within one, two, or three KEGG pathway steps of gene i, respectively. All specifications include gene–disease, gene–year, and disease–year fixed effects. "Pre-GWAS mean" is the mean of the outcome among pairs eventually exposed in the corresponding specification, measured in years before their first exposure.

Table S7: FDA Orange Book patents, with the drug target gene taken from the drug's DrugBank targets (columns 1–4) or from the patent text (columns 5–8)

	Outcome per gene-disease pair and year							
	Orange Book patents, drug-target gene				FDA Orange Book patents			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
GWAS (Focal gene)	0.00215**				0.000226			
	(0.00103)				(0.000186)			
GWAS (Distance 1)		0.00198***				0.000355**		
		(0.00044)				(0.000170)		
GWAS (Distance 2)			0.00114***				0.000172***	
			(0.00023)				(0.000054)	
GWAS (Distance 3)				0.00076***				0.000128***
				(0.00017)				(0.000041)
Gene-Disease FE	YES	YES	YES	YES	YES	YES	YES	YES
Gene-Year FE	YES	YES	YES	YES	YES	YES	YES	YES
Disease-Year FE	YES	YES	YES	YES	YES	YES	YES	YES
Observations	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888
Pre-GWAS Mean	0.00298	0.00371	0.00230	0.00177	0.000133	0.000281	0.000175	0.000130

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered two-way by gene and disease in parentheses. Each column reports estimates from a separate OLS panel regression of the yearly number of FDA Orange Book patents mapped to gene i and disease j on exposure to a GWAS discovery. Columns (1) and (5) define exposure as the first GWAS reporting the focal gene–disease association. Columns (2)–(4) and (6)–(8) use cumulative biological neighborhoods and define exposure as the first GWAS linking the same disease to a gene within one, two, or three KEGG pathway steps of gene i, respectively. All specifications include gene–disease, gene–year, and disease–year fixed effects. "Pre-GWAS mean" is the mean of the outcome among pairs eventually exposed in the corresponding specification, measured in years before their first exposure. Columns (1)–(4) take the gene from the DrugBank targets of the drug protected by the Orange Book patent. Columns (5)–(8) identify the gene directly from the patent text. The outcome excludes patents mentioning more than 100 distinct gene–disease pairs.

analysis using biological networks based on knowledge available at the beginning of the GWAS era.

We first use the 2005 release of STRING [38]. This network is useful for two reasons. First, it predates the diffusion of GWAS and therefore measures biological proximity using relationships known before the discoveries in our sample. Second, STRING is constructed from protein–protein associations rather than curated pathway membership, providing a different representation of how genetic targets are functionally connected. We consider a broad network using a combined-score threshold of 400 and, as a more conservative specification, a high-confidence network using a threshold of 700. The results are reported in Table S8. Under the broader definition, the estimated increase in patenting is 0.0063 within one step, 0.0011 within two steps, and 0.0010 within three steps. Raising the confidence threshold yields corresponding estimates of 0.0097, 0.0023, and 0.0013. All coefficients are positive and statistically significant at the 1% level. Although the magnitude varies with network density, both specifications recover consistent findings.

As a separate robustness check, we construct a network from gene relationships extracted from 827 pathway diagrams published before 2006 [20]. This representation is substantially sparser than either KEGG or STRING and relies on a different source of biological knowledge. The estimated effect is 0.0307 within one step, 0.0129 within two steps, and 0.0074 within three steps, corresponding to 49%, 34%, and 26% of pre-GWAS patenting among

Table S8: Alternative gene networks: STRING 2005 at combined score ≥400 (columns 1–4) and ≥700 (columns 5–8)

	Number of U.S. patents targeting a gene-disease pair							
	STRING 2005, score ≥400				STRING 2005, score ≥700			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
GWAS (Focal gene)	0.0383***				0.0383***			
	(0.0127)				(0.0127)			
GWAS (Distance 1)		0.0063***				0.0097***		
		(0.0011)				(0.0018)		
GWAS (Distance 2)			0.0011***				0.0023***	
			(0.0003)				(0.0004)	
GWAS (Distance 3)				0.0010***				0.0013***
				(0.0003)				(0.0003)
Gene-Disease FE	YES	YES	YES	YES	YES	YES	YES	YES
Gene-Year FE	YES	YES	YES	YES	YES	YES	YES	YES
Disease-Year FE	YES	YES	YES	YES	YES	YES	YES	YES
Observations	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888
Pre-GWAS Mean	0.0555	0.0239	0.0115	0.0096	0.0555	0.0306	0.0148	0.0108

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered two-way by gene and disease in parentheses. Each column reports estimates from a separate OLS panel regression of the yearly number of U.S. corporate patents involving gene i and disease j on exposure to a GWAS discovery. Columns (1) and (5) define exposure as the first GWAS reporting the focal gene–disease association. Columns (2)–(4) and (6)–(8) use cumulative biological neighborhoods and define exposure as the first GWAS linking the same disease to a gene within one, two, or three STRING pathway steps of gene i, respectively. All specifications include gene–disease, gene–year, and disease–year fixed effects. "Pre-GWAS mean" is the mean of the outcome among pairs eventually exposed in the corresponding specification, measured in years before their first exposure. STRING columns replace the KEGG pathway network with the STRING v6.3 (2005) protein-association network at the stated combined-score threshold. The outcome excludes patents mentioning more than 100 distinct gene–disease pairs.

treated pairs. In conclusion, these tests show how the same spatial gradient appears across representations of the biological space that differ in source, coverage, and density.

Figure S6 reports the corresponding event-study estimates. Across KEGG, STRING, and the pathway-figure network, pre-exposure coefficients are small and the patenting response emerges only after the GWAS discovery. The same temporal pattern across networks that differ substantially in their sources and densities makes it unlikely that our findings are an artifact of the specific representation of biological space chosen.

10) Entry and expansion of research effort after GWAS discoveries

The increase in patenting following GWAS discoveries that we document could come from firms already working on these targets or from new firms entering the opportunities revealed by these discoveries. Table S10 distinguishes these two margins by measuring both the total number of assignees patenting on each gene-disease pair and the number of assignees that had never previously patented on that pair. We find evidence that GWAS discoveries increase both margins. New assignees increase by 0.089 at the focal pair, more than doubling relative to the pre-GWAS mean. The response attenuates with biological distance but remains positive and statistically significant through three pathway steps. The total number of assignees follows a nearly identical spatial gradient. Thus, the patenting response partly reflects firms beginning to invest in R&D targeting gene-disease relationships where they

Table S9: Alternative gene networks: open-literature pathway figures through 2006

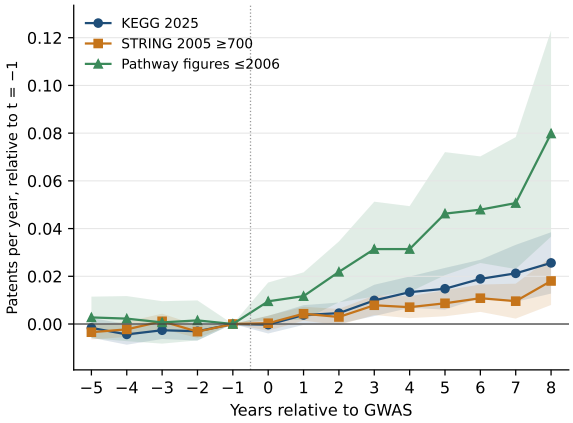
	Number of U.S. patents targeting a gene-disease pair			
	(1)	(2)	(3)	(4)
GWAS (Focal gene)	0.0383***			
	(0.0127)			
GWAS (Distance 1)		0.0307***		
		(0.0072)		
GWAS (Distance 2)			0.0129***	
			(0.0025)	
GWAS (Distance 3)				0.0074***
				(0.0015)
Gene-Disease FE	YES	YES	YES	YES
Gene-Year FE	YES	YES	YES	YES
Disease-Year FE	YES	YES	YES	YES
Observations	148,119,888	148,119,888	148,119,888	148,119,888
Pre-GWAS Mean	0.0555	0.0629	0.0379	0.0280

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered two-way by gene and disease in parentheses. Each column reports estimates from a separate OLS panel regression of the yearly number of U.S. corporate patents involving gene i and disease j on exposure to a GWAS discovery. Column (1) defines exposure as the first GWAS reporting the focal gene–disease association. Columns (2)–(4) use cumulative biological neighborhoods and define exposure as the first GWAS linking the same disease to a gene within one, two, or three pathway steps of gene i, respectively. All specifications include gene–disease, gene–year, and disease–year fixed effects. "Pre-GWAS mean" is the mean of the outcome among pairs eventually exposed in the corresponding specification, measured in years before their first exposure. Pathway-figure columns replace the KEGG pathway network with gene–gene edges read from 827 pathway diagrams in PMC articles published through 2006; the focal-gene column does not depend on the network. The outcome excludes patents mentioning more than 100 distinct gene–disease pairs.

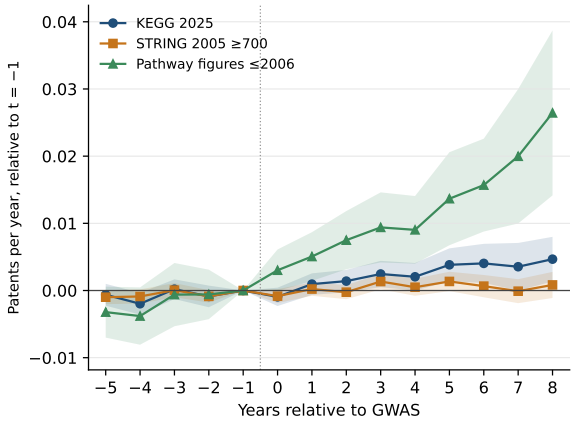
had not previously worked on.

We see the same pattern in scientific research. Table S11 uses PubTator3 to count papers linking a gene to a disease and to identify last authors publishing on that pair for the first time. GWAS discoveries increase publications by 1.1 papers per year at the focal association, with smaller but still significant effects across broader biological neighborhoods. The number of new last authors rises by 0.65 at the focal pair, roughly 70% relative to the pre-GWAS mean, and remains nearly 40% higher within three pathway steps. Thus, GWAS discoveries draw additional researchers toward both the reported association and nearby biological targets. This pattern is consistent with our interpretation that new scientific information changes the direction of subsequent research effort.

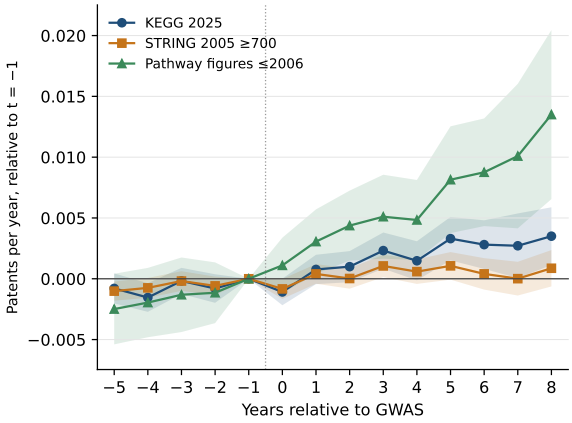
However, our estimates also leave open a separate possibility. Increased activity around genes identified in a GWAS could partly reflect the reallocation of a fixed amount of research effort away from other targets for the same disease. To examine this possibility, Table S12 collapses the data to the disease-year level and relates total activity to the cumulative number of gene associations reported by GWAS for that disease. If the responses documented above primarily reflected within-disease reallocation, greater GWAS exposure need not be accompanied by an increase in total R&D efforts at the disease level. Instead, cumulative GWAS discoveries are positively associated with total activity across all stages, from scientific publications and new researchers to new assignees and clinical trials. The coefficients are



(a) Within 1 pathway step



(b) Within 2 pathway steps



(c) Within 3 pathway steps

Figure S6: Cumulative spillover event studies for different gene networks: KEGG 2025, STRING 2005, and open-literature pathway figures before 2006. Patents mentioning more than 100 pairs are dropped. 90% confidence bands. Note that each panel has its own vertical axis.

positive and statistically significant for all six outcomes. These disease-level estimates are descriptive and do not use the same gene-year variation as our main design. Nonetheless, the fact that total activity rises across scientific, inventive, and clinical outcomes is inconsistent with a pure reallocation of effort across genetic targets within a disease.

11) Timing of GWAS discoveries and pre-existing inventive activity

Our empirical design allows the timing of GWAS research to differ systematically across diseases, but requires that the timing of a particular gene-disease association not be driven by pair-specific changes in inventive activity. However, GWAS are typically organized around a disease and then scan genetic variation broadly across the genome. Disease-year fixed effects absorb changes in research and inventive activity that affect the disease as a whole, while identification comes from the timing of *when* particular genes become associated with that

Table S10: New and all assignees on the pair's corporate patents after GWAS discoveries

	Outcome per gene-disease pair and year							
	New assignees, patents ≤100 pairs				Assignees, patents ≤100 pairs			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
GWAS (Focal gene)	0.0890**				0.1167**			
	(0.0384)				(0.0517)			
GWAS (Distance 1)		0.0383***				0.0470***		
		(0.0107)				(0.0138)		
GWAS (Distance 2)			0.0140***				0.0169***	
			(0.0029)				(0.0037)	
GWAS (Distance 3)				0.0099***				0.0122***
				(0.0020)				(0.0025)
Gene-Disease FE	YES	YES	YES	YES	YES	YES	YES	YES
Gene-Year FE	YES	YES	YES	YES	YES	YES	YES	YES
Disease-Year FE	YES	YES	YES	YES	YES	YES	YES	YES
Observations	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888
Pre-GWAS Mean	0.0742	0.0566	0.0329	0.0248	0.0897	0.0675	0.0387	0.0290

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered two-way by gene and disease in parentheses. Each column reports estimates from a separate OLS panel regression of the effects of GWAS on the yearly number of distinct assignees (harmonized firm names from Google Patents) for U.S. corporate patents involving gene i and disease j. Columns (1)–(4) count only assignees with no earlier patent on the pair, i.e., entry into the pair; columns (5)–(8) count all assignees on the pair's patents that year. Columns (1) and (5) define exposure as the first GWAS reporting the focal gene–disease association. Columns (2)–(4) and (6)–(8) use cumulative biological neighborhoods and define exposure as the first GWAS linking the same disease to a gene within one, two, or three KEGG pathway steps of gene i, respectively. All specifications include gene–disease, gene–year, and disease–year fixed effects. "Pre-GWAS mean" is the mean of the outcome among pairs eventually exposed in the corresponding specification, measured in years before their first exposure. Patents mentioning more than 100 distinct gene–disease pairs are dropped before counting names, as in the paper's main sample.

disease.

Figure S7 examines this distinction using patenting between 2000 and 2004, before the first GWAS discoveries in our sample. The Figure shows substantial sorting across diseases: diseases studied earlier by GWAS had considerably more pre-existing patenting. This pattern is precisely the type of disease-level variation absorbed by our disease-year fixed effects and thus not confounding our estimates. Differences across genes are much smaller.

Figure S8 shows the gene-disease pairs eventually discovered by GWAS. In the raw data, pairs with greater prior patenting tend to be discovered somewhat earlier. However, when we compare pairs within the same disease, this relationship becomes considerably weaker and not statistically significant. This evidence complements the event studies in Figure 2, which show little evidence that patenting at treated pairs was already increasing before the GWAS discovery. Together, these patterns indicate that the strongest selection in GWAS timing occurs across diseases rather than across gene-disease pairs within a disease, which is the variation used by our design.

12) Aggregate importance of spillovers

The regressions in this paper are estimated at the gene-disease-year level. To assess the aggregate importance of spillovers, we combine the estimated response within each biological radius with the number of gene-disease pairs that have become exposed by the end of

Table S11: Publications and new last authors on the pair after GWAS discoveries

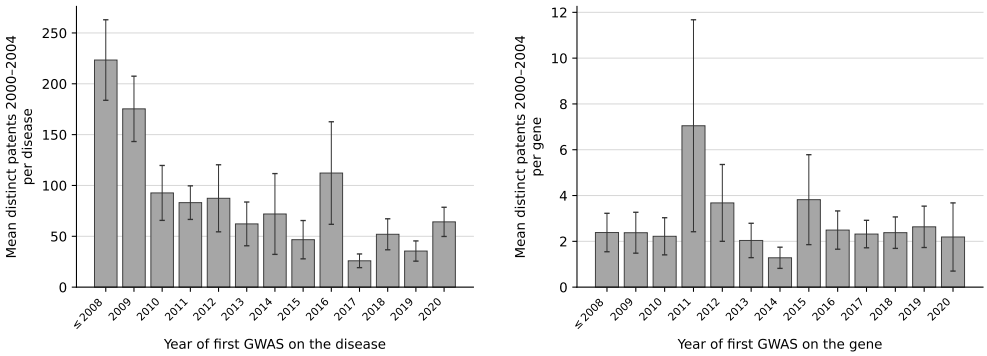
	Outcome per gene-disease pair and year							
	Papers				New last authors			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
GWAS (Focal gene)	1.0997***				0.6512***			
	(0.3983)				(0.1892)			
GWAS (Distance 1)		0.4537***				0.2951***		
		(0.1180)				(0.0642)		
GWAS (Distance 2)			0.1364***				0.0964***	
			(0.0311)				(0.0191)	
GWAS (Distance 3)				0.0785***				0.0557***
				(0.0187)				(0.0117)
Gene-Disease FE	YES	YES	YES	YES	YES	YES	YES	YES
Gene-Year FE	YES	YES	YES	YES	YES	YES	YES	YES
Disease-Year FE	YES	YES	YES	YES	YES	YES	YES	YES
Observations	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888	148,119,888
Pre-GWAS Mean	1.2816	0.5793	0.2590	0.1762	0.9314	0.4545	0.2111	0.1456

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered two-way by gene and disease in parentheses. Each column reports estimates from a separate OLS panel regression of a yearly measure of scientific activity on gene i and disease j on exposure to a GWAS discovery. Columns (1)–(4) count distinct PubMed articles in which PubTator3 extracts a relation between the gene and the disease. Columns (5)–(8) count distinct last-author names with no earlier publications on the pair. Columns (1) and (5) define exposure as the first GWAS reporting the focal gene–disease association. Columns (2)–(4) and (6)–(8) use cumulative biological neighborhoods and define exposure as the first GWAS linking the same disease to a gene within one, two, or three KEGG pathway steps of gene i, respectively. All specifications include gene–disease, gene–year, and disease–year fixed effects. "Pre-GWAS mean" is the mean of the outcome among pairs eventually exposed in the corresponding specification, measured in years before their first exposure.

Table S12: Total activity per disease and GWAS discoveries (disease × year panel)

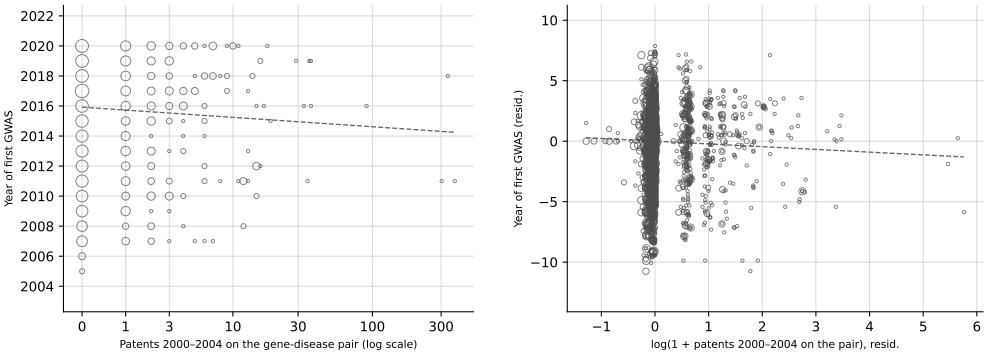
	log(1 + outcome)					
	Papers	New last authors	Patents	New assignees	Clinical trials	Orange Book patents
	(1)	(2)	(3)	(4)	(5)	(6)
log(1 + cumulative GWAS hits)	0.073***	0.067***	0.052***	0.140***	0.077***	0.397***
	(0.008)	(0.007)	(0.008)	(0.009)	(0.017)	(0.021)
Disease FE	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES
Observations	51,135	51,135	51,135	51,135	51,135	51,135
Pre-GWAS Mean	379.254	314.247	23.936	61.606	34.198	8.806

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered by disease in parentheses. One observation per disease × year, 2000–2020 (2,435 diseases, of which 531 receive at least one GWAS discoveries by 2020). Each cell reports the coefficient from a separate OLS regression $\log(1+Y_{d,t}) = \alpha+\beta X_{d,t}+\gamma_d+\delta_t+\epsilon_{d,t}$, where $Y_{d,t}$ is the disease's total yearly activity summed over all its genes. $X_{d,t}$ is log(1+cumulative GWAS hits), where cumulative GWAS hits is the number of disease d's genes with a GWAS association reported by year t inclusive. Papers are PubMed articles linking a gene to the disease; new last authors are last authors with no earlier paper on the corresponding gene–disease pair; patents are U.S. pharmaceutical patents, counted once per disease; new assignees are assignees with no earlier patent on the corresponding pair; clinical trials are distinct ClinicalTrials.gov studies mapped to the pair; Orange Book patents are FDA Orange Book-listed patents. Pre-GWAS mean is the mean of $Y_{d,t}$ across diseases that are ever hit, in the year before their first hit. These are descriptive within-disease comparisons.



(a) By disease (b) By gene

Figure S7: Timing of the first GWAS and pre-2005 patenting by disease and gene. Panel (a) reports the mean number of distinct patents over 2000–2004 per disease, grouped by the year of the first GWAS on that disease. Panel (b) reports the corresponding measure by gene, grouped by the year of the first GWAS on that gene. Each patent is counted once per disease or gene it mentions. GWAS years 2005–2008 are pooled, and whiskers report 90% confidence intervals of the mean.



(a) Treated pairs, raw (b) Treated pairs, residualized on disease

Figure S8: Timing of the first GWAS and pre-2005 patenting among treated gene–disease pairs. Each panel contains one observation for each of the 11,785 gene–disease pairs first reported by a GWAS and plots the year of first discovery against patenting on the pair over 2000–2004. Panel (a) reports the raw relationship. Panel (b) residualizes both discovery year and prior patenting on disease fixed effects. Patenting is expressed as $\log(1 + \text{count})$. The dashed line reports the OLS regression.

our sample. For each cumulative radius around a GWAS discovery, we multiply the corresponding coefficient by the number of gene-disease pairs exposed within that radius. This calculation measures the annual response associated with the full set of opportunities exposed by GWAS discoveries by 2020. Table S13 reports these calculations for patents, clinical trials, and marketed therapies.

For patents, the 11,785 gene-disease associations directly identified by GWAS correspond to 451.3 additional pair-level patent appearances per year when evaluated over the set of focal associations exposed by 2020. Extending the calculation to all pairs within three pathway steps yields 1,989.9 annual pair-level appearances. Thus, 77% of the estimated response

Table S13: Back-of-the-Envelope Calculation of GWAS Impact Across Outcomes

Cumulative Level	OLS Coefficient β	Gene-Disease Pairs N	Effect $\beta \times N$
Panel A. U.S. corporate patents			
0-distance only	0.0382926	11,785	451.3
0+1-distance	0.0126656	95,786	1,213.2
0+1+2-distance	0.0037340	373,441	1,394.4
0+1+2+3-distance (All)	0.0028140	707,139	1,989.9
Panel B. Clinical trials			
0-distance only	0.0073631	11,785	86.8
0+1-distance	0.0063905	95,786	612.1
0+1+2-distance	0.0035735	373,441	1,334.5
0+1+2+3-distance (All)	0.0024169	707,139	1,709.0
Panel C. Orange Book patents, DrugBank target gene			
0-distance only	0.0021509	11,785	25.3
0+1-distance	0.0019800	95,786	189.7
0+1+2-distance	0.0011371	373,441	424.6
0+1+2+3-distance (All)	0.0007559	707,139	534.5
Panel D. Orange Book patents, gene from patent text			
0-distance only	0.0002264	11,785	2.7
0+1-distance	0.0003546	95,786	34.0
0+1+2-distance	0.0001716	373,441	64.1
0+1+2+3-distance (All)	0.0001281	707,139	90.6

Notes: For each outcome and cumulative biological radius, the table multiplies the corresponding OLS coefficient β by the number of gene-disease pairs N exposed within that radius. Panel A uses the baseline U.S. corporate patent outcome, which excludes patents mentioning more than 100 gene-disease pairs. Panel B uses new ClinicalTrials.gov trials mapped to gene-disease pairs and dated by trial start year. Panel C uses patents listed in the FDA Orange Book and assigns the gene using the DrugBank targets of the protected drug. Panel D uses the same Orange Book patents but identifies the gene directly from patent text. The resulting quantities are annual pair-level responses and therefore need not correspond one-for-one to distinct patent documents, clinical trials, or marketed products. Within three pathway steps, 77% of the patenting response, 95% of the clinical-trial response, 95% of the Orange Book response using DrugBank targets occur beyond the focal gene-disease associations.

across the opportunities exposed by 2020 occurs beyond the gene-disease relationships directly identified by GWAS.

The same pattern is even more pronounced further downstream. For clinical trials, the estimated response rises from 86.8 at the focal associations to 1,709.0 within three pathway steps, implying that 95% occurs on biologically related targets. Using DrugBank targets to map Orange Book patents shows that the response increases from 25.3 at the focal pairs to 534.5 within three steps, with 95% occurring beyond the directly identified associations. The alternative mapping based on genes extracted from Orange Book patent text yields a corresponding share of 97%. Thus, the aggregate importance of spillovers persists from patenting through clinical development and marketed therapies.

13) Estimating investment in GWAS research

To place these effects in relation to the resources devoted to GWAS, 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 study composition changes substantially over time. Craig et al. [11] estimate that a 4,000-person GWAS costs approximately US$1.8 million, implying a cost of roughly US$450 per participant in

Table S14: 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.

2009. Similarly, Visscher et al. [43] 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. Mills and Rahal [31] further estimate that approximately 47.6 million individuals were genotyped globally between 2007 and 2018. Applying a cost of US$500 per genotyped individual would imply expenditures of roughly US$23.8 billion.

Two countervailing forces make it difficult to extrapolate these benchmarks over the full GWAS era. On the one hand, genotyping technology becomes cheaper over time, lowering the cost per participant. On the other hand, GWAS sample sizes increase sharply, raising the number of observations collected by each study. We therefore use 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 uses NIH grant funding as acknowledged in GWAS publications. Throughout this section, we use the full set of 4,491 GWAS reported in the GWAS Catalog, rather than only the 1,402 studies with novel findings that we use in our regression sample, because the objective is to characterize investment in the broader scientific effort from which these discoveries emerged.

Method 1: Genotyping costs

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 S14. Because 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 assign an exact platform price to each study. Specifically, we fit a linear time trend to the observed prices, shown in Figure S9, to capture the sharp decline in per-sample genotyping costs during the GWAS era.

We apply these year-specific unit costs to each GWAS in the Catalog. For study s published in year t, we compute direct genotyping expenditures as $\text{Cost}_s = N_s \times c_t$, where N_s is the reported sample size and c_t is the estimated per-sample genotyping cost in year t. This calculation therefore allows later studies to face lower per-person genotyping costs while also accounting for their substantially larger sample sizes.

Table S15 reports the resulting estimates under several filtering scenarios. In the full sample, 4,491 studies covering 190.8 million participant observations imply direct genotyping expenditures of US$10.76 billion. This figure likely overstates the cost of generating original

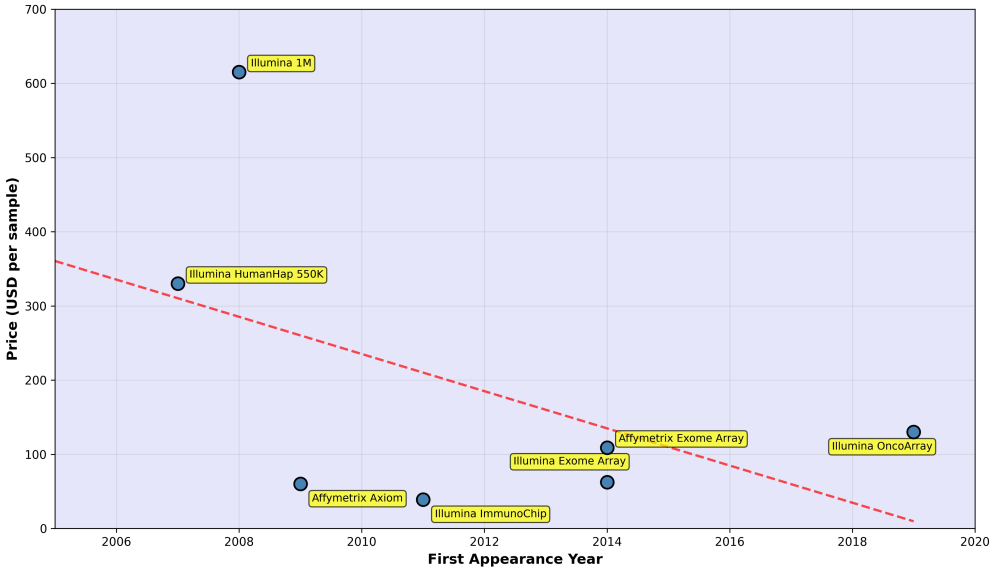


Figure S9: Estimated per-sample genotyping costs based on observed platform prices. The dashed line shows the fitted linear decline in the cost of genotyping a single individual over time.

Table S15: 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 are identified through keyword matching. Large studies are defined separately as studies with more than 100,000 participants and are excluded to reduce the influence of analyses based on shared biobank samples.

data because many GWAS reuse previously collected samples. We therefore construct more conservative estimates by separately excluding meta-analyses identified through keyword matching and studies with more than 100,000 participants, which are more likely to rely on shared biobank data. Applying both restrictions leaves 3,695 studies and 30.8 million participant observations and implies US$3.28 billion in direct genotyping expenditures. We use this quantity as our preferred benchmark for the direct genotyping costs of GWAS between 2005 and 2020.

Method 2: NIH funding

We compare the estimates above with a calculation based on NIH grants acknowledged by GWAS publications. To do so, we query the NIH RePORTER API for all grants acknowledged by the GWAS papers in our sample. Figure S10 shows that NIH support is present throughout the sample period. The main difficulty is attribution. NIH grants often support multiple projects and generate multiple publications, so the amount assigned to GWAS depends on how each grant is allocated across outputs.

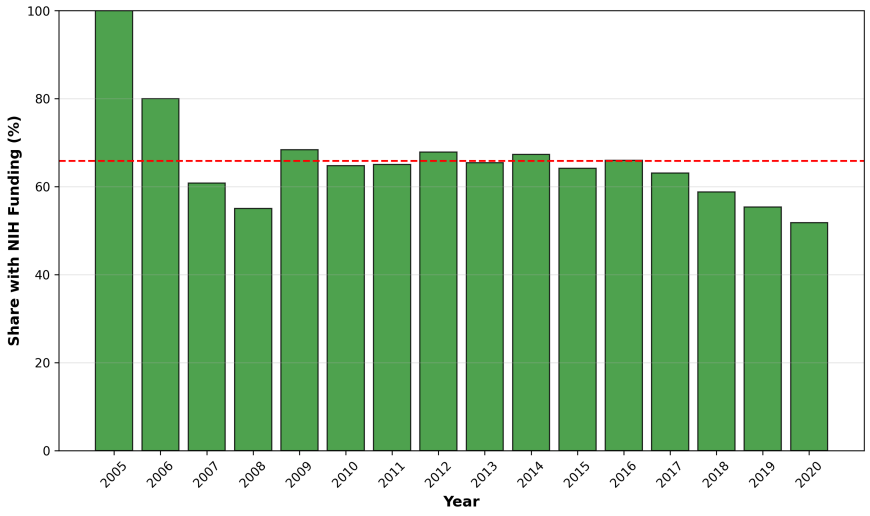


Figure S10: Share of GWAS papers acknowledging NIH funding, by publication year. The figure shows the proportion of studies acknowledging NIH funding across the full sample of 4,491 GWAS papers. The dashed line indicates the sample average.

Table S16: NIH Funding Estimates by Filtering Scenario (Full Sample)

Sample	Studies	Studies with 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: Two 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.

We report two attribution rules in Table S16. 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 across GWAS papers but likely overstates GWAS-specific funding because the same grants also support other lines of research. The second divides each grant by the total number of publications it supports, including non-GWAS outputs. Under our preferred study filters, the two procedures yield US$32.03 billion and US$1.01 billion, respectively. The latter is likely conservative because GWAS are substantially more capital-intensive than the average publication supported by the same grant.

We view these estimates as broad benchmarks for NIH support rather than alternative estimates of total GWAS costs. GWAS are also funded by other public agencies, foundations, and universities, while NIH grants also support activities beyond direct genotyping. For the research-productivity calculations below, we rely on the more transparent bottom-up estimate of direct genotyping expenditures.

14) Spillovers and measured research productivity

The aggregate patenting quantities in Section 12 are expressed in gene-disease pairs, which are the natural units of our empirical design, but do not map one-for-one to distinct patent documents because a single patent can mention multiple pairs. To obtain a back-of-the-envelope measure that can be compared with existing estimates of patent output per dollar of research spending, we convert the pair-level response into patent-equivalent units by dividing by the median number of gene-disease pairs per patent in our baseline sample, which is 12. Under this conversion, the focal associations generate approximately 37.6 patent-equivalent units per year. Extending the calculation to the neighborhood within three biological steps raises this quantity to approximately 165.8 per year, of which 128.2 arise on nearby targets rather than at the focal pair itself.

We then compare these quantities with the cost estimates from Section 13. Using our preferred estimate, the 3,695 GWAS in the filtered sample imply direct genotyping expenditures of approximately US$3.28 billion, or roughly US$205 million per year over 2005–2020. Combining this annual cost with the estimated patenting response implies approximately 8.1 patent-equivalent units per US$10 million when spillovers within three biological steps are included. Restricting attention to the focal gene-disease associations instead yields approximately 1.8 per US$10 million. For comparison, Azoulay et al. [5] estimate that US$10 million invested in NIH extramural grants yields approximately 2.7 additional corporate patents. Thus, counting only direct effects places GWAS below this benchmark, whereas including spillovers raises the estimate to roughly three times as large.

A separate comparison shows how much of the estimated impact would be missed by measures that require an observable link to the originating research. Across the full GWAS Catalog, only 534 studies, or 11.9%, are cited by at least one U.S. patent, yielding 1,781 citing patents between 2005 and 2020. Our estimates instead imply approximately 2,653 patent-equivalent units over the same period, once spillovers across biologically related targets are accounted for. The two quantities are not directly equivalent, but the comparison illustrates the measurement problem motivating our approach: much of the downstream response to basic research occurs outside the set of technologies that retain a direct documentary link to the original scientific study.

Note, however, that these calculations should be interpreted as benchmarks for research productivity rather than estimates of the full social return to GWAS.

References

[1] Pedro Aceves and James A Evans. Mobilizing conceptual spaces: How word embedding models can inform measurement and theory within organization science. *Organization Science*, 35(3):788–814, 2024.

[2] Ajay Agrawal, John McHale, and Alexander Oettl. Artificial intelligence and scientific discovery: A model of prioritized search. *Research Policy*, 53(5):104989, 2024.

[3] Mohammad Ahmadpoor and Benjamin F Jones. The dual frontier: Patented inventions and prior scientific advance. *Science*, 357(6351):583–587, 2017.

[4] Ufuk Akcigit, Douglas Hanley, and Nicolas Serrano-Velarde. Back to basics: Basic research spillovers, innovation policy, and growth. *The Review of Economic Studies*, 88 (1):1–43, 2021.

[5] Pierre Azoulay, Joshua S Graff Zivin, Danielle Li, and Bhaven N Sampat. Public r&d investments and private-sector patenting: evidence from nih funding rules. *The Review of Economic Studies*, 86(1):117–152, 2019.

[6] Elisabetta Bianchi and Lars Rogge. The il-23/il-17 pathway in human chronic inflammatory diseases—new insight from genetics and targeted therapies. *Genes & Immunity*, 20(5):415–425, 2019.

[7] Nicholas Bloom, Mark Schankerman, and John Van Reenen. Identifying technology spillovers and product market rivalry. *Econometrica*, 81(4):1347–1393, 2013.

[8] Ewen Callaway. New concerns raised over value of genome-wide disease studies. *Nature*, 546(7659):463–464, 2017.

[9] Maria Cerezo, Elliot Sollis, Yue Ji, Elizabeth Lewis, Ala Abid, Karatuğ Ozan Bircan, Peggy Hall, James Hayhurst, Sajo John, Abayomi Mosaku, et al. The nhgri-ebi gwas catalog: Standards for reusability, sustainability and diversity. *Nucleic Acids Research*, 53(D1):D998–D1005, 2025.

[10] Wesley M Cohen, Richard R Nelson, and John P Walsh. Links and impacts: the influence of public research on industrial r&d. *Management Science*, 48(1):1–23, 2002.

[11] Jamie E Craig, Alex W Hewitt, Amy E McMellon, Anjali K Henders, Lingjun Ma, Leanne Wallace, Shiwani Sharma, Kathryn P Burdon, Peter M Visscher, Grant W Montgomery, and Stuart MacGregor. Rapid inexpensive genome-wide association using pooled whole blood. *Genome Research*, 19(11):2075–2080, 2009.

[12] Ying Ding, Min Song, Jia Han, Qi Yu, Erjia Yan, Lili Lin, and Tamy Chambers. Entitymetrics: Measuring the impact of entities. *PloS one*, 8(8):e71416, 2013.

[13] Richard H Duerr, Kent D Taylor, Steven R Brant, John D Rioux, Mark S Silverberg, Mark J Daly, A Hillary Steinhart, Clara Abraham, Miguel Regueiro, Anne Griffiths, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*, 314(5804):1461–1463, 2006.

[14] Maya Durvasula, C Scott Hemphill, Lisa Larrimore Ouellette, Bhaven Sampat, and Heidi L Williams. The nber orange book dataset: A user's guide. *Research Policy*, 52 (7):104791, 2023.

[15] Lee Fleming. Recombinant uncertainty in technological search. *Management Science*, 47(1):117–132, 2001.

[16] Lee Fleming and Olav Sorenson. Science as a map in technological search. *Strategic Management Journal*, 25(8-9):909–928, 2004.

[17] Lee Fleming, Hillary Greene, G Li, Matt Marx, and Dennis Yao. Government-funded research increasingly fuels innovation. *Science*, 364(6446):1139–1141, 2019.

[18] Zvi Griliches. The search for R&D spillovers. *The Scandinavian Journal of Economics*, 94(Supplement):S29–S47, 1992.

[19] Bronwyn H Hall, Jacques Mairesse, and Pierre Mohnen. Measuring the returns to r&d. In *Handbook of the Economics of Innovation*, volume 2, pages 1033–1082. Elsevier, 2010.

[20] Kristina Hanspers, Anders Riutta, Martina Summer-Kutmon, and Alexander R Pico. Pathway information extracted from 25 years of pathway figures. *Genome Biology*, 21 (1):273, 2020.

[21] Adam B. Jaffe. Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits, and market value. *American Economic Review*, 76(5):984–1001, 1986.

[22] Adam B Jaffe, Manuel Trajtenberg, and Rebecca Henderson. Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics*, 108(3):577–598, 1993.

[23] Benjamin F Jones and Lawrence H Summers. A calculation of the social returns to innovation. In *Innovation and Public Policy*, pages 13–60. University of Chicago Press, 2022.

[24] Minoru Kanehisa, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, et al. Kegg for linking genomes to life and the environment. *Nucleic Acids Research*, 36(suppl_1):D480–D484, 2007.

[25] Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, et al. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic Acids Research*, 52(D1):D1265–D1275, 2024.

[26] Pat Langley, Herbert A. Simon, Gary L. Bradshaw, and Jan M. Zytkow. *Scientific Discovery: Computational Explorations of the Creative Processes*. MIT Press, Cambridge, MA, 1987.

[27] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[28] Danielle Li, Pierre Azoulay, and Bhaven N Sampat. The applied value of public investments in biomedical research. *Science*, 356(6333):78–81, 2017.

[29] Alan C Marco, Amanda Myers, Stuart JH Graham, Paul D'Agostino, and Kirsten Apple. The uspto patent assignment dataset: Descriptions and analysis. *USPTO Economic working paper 2015-2*, 2015.

[30] Matt Marx and Aaron Fuegi. Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal*, 41(9):1572–1594, 2020.

[31] Melinda C Mills and Charles Rahal. A scientometric review of genome-wide association studies. *Communications Biology*, 2(1):9, 2019.

[32] Kyle R Myers and Lauren Lanahan. Estimating spillovers from publicly funded r&d: Evidence from the us department of energy. *American Economic Review*, 112(7):2393–2423, 2022.

[33] Francis Narin, Kimberly S Hamilton, and Dominic Olivastro. The increasing linkage between us technology and public science. *Research Policy*, 26(3):317–330, 1997.

[34] Richard R Nelson. The simple economics of basic scientific research. *Journal of Political Economy*, 67(3):297–306, 1959.

[35] Paul M Romer. Endogenous technological change. *Journal of Political Economy*, 98(5, Part 2):S71–S102, 1990.

[36] Daniel Schliesmann and Matteo Tranchero. Learning about roads not taken. In *Academy of Management Proceedings*, 2026.

[37] Suzanne Scotchmer. Standing on the shoulders of giants: cumulative research and the patent law. *Journal of Economic Perspectives*, 5(1):29–41, 1991.

[38] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2023.

[39] Matteo Tranchero. Data-driven search and the birth of theory: Evidence from genome-wide association studies. *University of Pennsylvania*, 2026.

[40] Matteo Tranchero. Finding diamonds in the rough: Data-driven opportunities and pharmaceutical innovation. *University of Pennsylvania*, 2026.

[41] Matteo Tranchero, Charlie Guthmann, and Bikash Kumar Panda. Biological entities in us patents - data, figures and tables. *Zenodo*, 2026.

[42] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R Martin, Hilary C Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):1–21, 2021.

[43] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[44] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[45] Martin L Weitzman. Recombinant growth. *The quarterly journal of economics*, 113(2): 331–360, 1998.

[46] Deborah A Zarin, Tony Tse, Rebecca J Williams, Robert M Califf, and Nicholas C Ide. The clinicaltrials. gov results database—update and key issues. *New England Journal of Medicine*, 364(9):852–860, 2011.