NBER WORKING PAPER SERIES

THERAPEUTIC TRANSLATION OF GENOMIC SCIENCE: OPPORTUNITIES AND LIMITATIONS OF GWAS

Manuel I. Hermosilla
Jorge A. Lemus

Working Paper 23989
http://www.nber.org/papers/w23989

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2017, Revised March 2018

Previously circulated as "Therapeutic Translation in the Wake of the Genome." We thank Ernst Berndt, Dana Goldman, and John Rowe, and the participants of the NBER conference on Economic Dimensions of Personalized and Precision Medicine. Errors and omissions are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Therapeutic Translation of Genomic Science: Opportunities and limitations of GWAS
Manuel I. Hermosilla and Jorge A. Lemus
NBER Working Paper No. 23989
November 2017, Revised March 2018
JEL No. I1,L65,O30

ABSTRACT

Many scientists predicted a swift revolution in human therapeutics after the completion of the Human Genome Project ("HGP"). This revolution, however, has been slow to materialize in spite of the scientific advances. We investigate the role of biological complexity in slowing down this revolution. Our test relies on a disease-specific measure of biological complexity, constructed by drawing on insights from Network Medicine (Barabási et al., 2011). According to our measure, more complex diseases are associated with a larger number of genetic mutations—higher centrality in the Human Disease Network (Goh et al., 2007). With this measure in hand, we estimate the rate of translation of new science into early-stage drug innovation by focusing on a leading type of genetic epidemiological knowledge (Genome-Wide Association Studies), and employing standard methods for the measurement of R&D productivity. For less complex diseases, we find a strong and positive association between cumulative knowledge and the amount of innovation. This association weakens as complexity increases, becoming statistically insignificant at the extreme. Our results suggest that biological complexity is, in part, responsible for the slower-than-expected unfolding of the therapeutical revolution set in motion by the HGP.

Manuel I. Hermosilla
Carey Business School
Johns Hopkins University
100 International Dr
Baltimore, MD 21202
mh@jhu.edu

Jorge A. Lemus
University of Illinois, Urbana-Champaign
1407 W Gregory Drive
Urbana, IL 61874
jalemus@illinois.edu

1 Introduction

> If the Human Genome Project gave us a book, scientists are now learning how to read it (...) and biologists are beginning to face up to the uncomfortable truth that they have only been looking at the nouns (...) now we are reading the spaces in between—verbs, adverbs, adjectives, pronouns and the rest, and they are complicated indeed.
>
> Roger Highfield (Former Editor, *New Scientist*)[1]

The 2003 completion of the Human Genome Project ("HGP") blew a whirlwind of hope for the future of biology and medicine. Presenting the project's first draft, Craig Venter stated that "the basic knowledge that we're providing the world will have a profound impact on the human condition and the treatments for disease and our view on our place in the biological continuum." At the same venue, President Clinton remarked that the HGP would "revolutionize the diagnosis, prevention and treatment of most, if not all, human diseases."[2] The Human Genome was envisioned as a discovery platform, which would greatly facilitate the understanding of disease biology and in turn illuminate, sharpen, and speed up the drug discovery/design process (Daiger, 2005; Lander et al., 2001). Many scientists and analysts agreed with Venter and Clinton and predicted a swift revolution in human therapeutics—some even suggesting that it would materialize within the decade. For example, Randy Scott of Incyte Genomics claimed that "in 10 years, we will understand the molecular basis for most human diseases" (Palmer, September 30, 2013). However, despite large scientific advances, this revolution in human therapeutics remains manifestly unfulfilled almost 15 years after the HGP's completion (Lander, 2011; Mardis, 2011): disease mortality continues to be largely driven by the same causes of two decades ago, and the molecular basis for most important diseases has not yet been fully elucidated (Wade, June 12, 2010; Palmer, September 30, 2013).

One reason that may justify the slower-than-expected progress is the large amount of biological complexity that has been progressively revealed by the so-called "genetic revolution" (Wade, June 12, 2010; Hayden, 2010). For example, only recently we learned that common mutations may explain a relatively small percent of predicted genetic variance (Manolio et al., 2009), that non protein-coding mutations may regulate protein-coding genes (Li et al., 2016), and that genetic mutations rarely map one-to-one into diseases

[1] "Life just got a lot more complicated." *The Daily Telegraph*, June 19, 2007.

[2] `http://transcripts.cnn.com/TRANSCRIPTS/0006/26/bn.01.html`.

(Bauer-Mehren et al., 2011).[3] The recent "omnigenic" hypothesis (Boyle et al., 2017) complicates matters further by suggesting that seemingly unrelated "peripheral" genes (located outside core pathways, and which cannot be easily categorized based on known biology) may drive disease through cellular networks. Hayden (2010) illustrates this general idea by writing "(the HGP) opened the door to a vast labyrinth of new questions," and "the complexity of biology has seemed to grow by orders of magnitude."

In this article, we investigate the extent to which biological complexity has mediated the translation of genetic epidemiological science into early stage drug innovation, during the ten years that followed the HGP's completion. Our analysis focuses on the knowledge created by the leading type of genetic epidemiological research, the Genome-Wide Association Studies ("GWAS"), which takes a prominent role in drug discovery (Manolio, 2013). These studies search for genetic mutations underlying the manifestation of diseases, validating associations only if stringent statistical standards are met. Exploiting variation at targeted-disease level, we investigate the impact of GWAS knowledge accumulation on the number of therapies that enter the drug development process. We explore both the opportunities created by GWAS and its limitations.

To characterize the variability in biological complexity across diseases, we draw on insights from the emerging field of *Network Medicine* (Barabási et al., 2011), and rely on implementations of the *Human Disease Network* ("HDN"; Goh et al., 2007). In the HDN, diseases correspond to nodes, and disease-causing genes (or variants thereof) correspond to edges. We exploit the idea that therapeutic translation may be more complex for diseases that are caused by a larger set of genes, or for those that are connected with a larger number of diseases in the HDN. Intuitively, a disease caused by a large number of genetic mutations could manifest itself through many biological pathways, making it less likely to find a "silver bullet" treatment. A similar reasoning applies with respect to HDN connectivity. Therapeutic innovation targeting highly-connected diseases runs the risk of interfering with related biological processes, and with it, the possibility of causing adverse side effects or other biological imbalances. Thus, to prove its safety, such therapy may have to overcome greater challenges than one targeting a scantly-connected disease. At a broader level, it can be argued that translational complexity increases with these factors because they require developers to consider a larger set of biological factors in

[3] Boyle et al. (2017) illustrate the dramatic paradigm shift referencing the case of Autism: whereas the prediction of 15 or more responsible mutations of Risch et al. (1999) was perceived as "strikingly high at the time" (Boyle et al., 2017), based on recent research (Weiner et al., 2017) this number "seems quaintly low now" (Boyle et al., 2017).

the process of discovering, designing, and testing new compounds. Indeed, to this point Bauer-Mehren et al. (2011) write: “At the end of the day, how a disease is caused and thus how it can be treated can only be studied on the basis of the entire body of knowledge including all genes that are associated with the disease and their interactions through biological pathways.”

We construct these complexity metrics using the largest publicly-available repository of human gene- and variant-disease associations, DisGeNET, which contains results from tens of thousands of academic publications. Translation rates are estimated from a large sample of pharmaceutical pipelines, which cover over thirteen hundred targeted diseases, spanning nineteen therapeutic areas. Our results suggest that complexity plays an important role in moderating therapeutic translation. In particular, for less complex diseases, we find a strong and positive association between cumulative knowledge and the amount of new therapies entering the discovery process each year. This association weakens as complexity increases and becomes statistically insignificant for highly complex diseases. We perform several checks to verify that our results are not driven by the influence of unobservable variables.

At a conceptual level, our research is related to Fleming and Sorenson (2001, 2004), who also address the interplay of science and complexity in the context of technological innovation. In the framework laid out by Fleming and Sorenson (2004), science is useful because it helps to navigate the complexity that arises from the (recombinant) search process over a “technology landscape.” Instead, we view complexity as the defining trait of the “landscape’s topography,” and as a barrier to the practical applicability of scientific knowledge. By highlighting and empirically documenting the role of complexity, we also contribute to the literature that studies empirically the extent to which academic science translates into productivity growth or innovation (Rosenberg, 1974; Sveikauskas, 1981; Jaffe, 1989; Adams, 1990; Mansfield, 1995; Stephan, 1996; Cohen et al., 2002; Ahmadpoor and Jones, 2017), and specifically to the literature with a focus on the pharmaceutical industry (Henderson and Cockburn, 1994; Gambardella, 1995; Ward and Dranove, 1995; Zucker and Darby, 1996; Zucker et al., 1998; Cockburn and Henderson, 1998; Toole, 2012; Azoulay et al., 2015).[4] Our contribution relative to the latter set of articles is to measure and ascertain the role of biological complexity on innovative productivity, which

[4] Among these, Toole (2012) is the closest to the research herein. Relative to the work of Toole—which addresses the relationship between basic research funding and drug approvals, aggregating these at the therapeutic area level—our research enables a more translucent analysis, by focusing on the relationship between two variables that are more directly related (scientific publication and early-stage development), and by exploiting variation defined at a thinner aggregation level (targeted diseases).

has so far been treated as unobserved heterogeneity.

By many accounts, genetics have already shaped a new era of drug innovation. For example, genetics are now routinely used to identify drug targets and populations at higher risk of developing adverse events (Pollack, June 14, 2010). Nevertheless, failure to meet the more optimistic expectations from the early 2000s has lead to impatience and criticism. Some observers have questioned the worthiness of the endeavor (Evans et al., 2011), and others even suggested that a "genomic bubble" may have "temporarily bogged down the drug industry with information overload" (Pollack, June 14, 2010). Our findings suggest that biological complexity may be partly to blame, which is one of the limitation of GWAS. On the other hand, for less complex diseases, GWAS presents an opportunity to develop new therapies. By explicitly accounting for biological complexity, we provide a novel assessment of the progress made so far, while suggesting that polarized assessments can be reconciled.

The rest of the article is organized as follows. Section 2 describes and contextualizes GWAS science. Section 3 describes data sources and processing. Section 4 lays out the empirical strategy. Results are presented in Section 5 and conclusions in Section 6.

2 Genome-Wide Association Studies

A Genome-Wide Association study compares the DNA of a population that carries a certain trait (e.g., weight, aggressive personality, diabetes, acne, etc.) against that of a control population without it, under the assumption that individuals with a trait present similar genetic variations. Although the effectiveness of GWAS has been criticized, their scientific impact is widely recognized. For example, Visscher et al. (2012) states: "(..) the GWAS experimental design in human populations has led to new discoveries about genes and pathways involved in common diseases and other complex traits, has provided a wealth of new biological insights, has led to discoveries with direct clinical utility, and has facilitated basic research in human genetics and genomics."

Our empirical analysis relies on data available from the open-source GWAS Catalog (Burdett et al., 2016). The earliest study in this catalog was published in 2005. GWAS publications have since significantly impacted the way scientists think about the biological mechanisms behind certain diseases.[5] An important reason to focus on GWAS publications is that these studies are "hypothesis-free." This means that GWAS do not rely on assumptions to design experiments, but instead, they are a statistical analysis looking for high correlation between regions of DNA and diseases. In fact, mutations are validated under a strict threshold of statistical significance ($p < 5 \times 10^{-8}$) and the results must be replicated in an independent sample before they are incorporated into the GWAS Catalog (McCarthy et al., 2008). These strict requirements make GWAS publications a source of credible results, which are recognized by most scientists.

Apart from GWAS, there are alternative methods to study the link between genes and diseases (Londin et al., 2013). Linkage analysis (LA) is a technique used to identify genetic variants for Mendelian disorders—i.e. mutations caused by a single gene. Following the success of Kerem et al. (1989) in identifying the gene responsible for cystic fibrosis, LA studies have proved useful in identifying other Mendelian disorders. However, LA studies rely on related individuals: they do not provide "high resolution" (meaning, they identify broad regions of variations) and they have limited statistical power. Next-generation sequencing (NGS), the most recent method, has the advantage of identifying mutations at a high resolution (i.e. it allows researchers to identify specific gene mutations and variants). The main drawback of NGS has been the high cost of sequencing the complete genome for large samples (Koboldt et al., 2013). Recent advances in computer speed and storage capacity have lowered the costs, enabling large NGS studies.

3 Data and Variables

3.1 Therapies

We obtained pharmaceutical pipeline data from Thomson Reuters Cortellis, a subscription service that offers pipeline information for a large number of biotechnology and pharmaceutical firms. The full data sample includes development histories of over 90,000 therapies (i.e., compound/targeted-disease combinations) entering the development process around the world, since the mid 1970s. A "new therapy" in our data corresponds

[5] For instance, Cao and Moult (2014) explores the the use of GWAS in identifying drug targets. Visscher et al. (2017) and Zheng et al. (2009) review the remarkable discoveries that have been facilitated by GWAS publications.

to a novel indication entering the earliest stage in the process, the "discovery" stage. At this stage, therapies are optimized: they are evaluated analytically and in animal models to assess further development.[6]

We restrict the sample to new therapies that were first observed entering the discovery stage between 2003 and 2012, and to the set of diseases for which at least one new therapy was observed during this period. This selected sample covers 1,306 different diseases and includes 26,120 new therapies, distributed across 19 therapeutic areas. Cortellis also identifies each therapy's molecular target, which allows us to differentiate between protein- and gene-targeted therapies. The former are designed to interact with proteins along the cell-signaling cascade. The latter, to regulate or modify the expression of protein-encoding genes. Figure 1 describes the distribution of new therapies across therapeutic areas and emphasizes the imbalance of innovation efforts accross diseases. It also shows that the number of protein-targeted therapies in our sample exceeds that of gene-targeted therapies by about one order of magnitude.

The dependent variable in our econometric analysis, denoted by N_{dkt}, corresponds to the total number of new therapies for disease d that enter the discovery stage during year t employing target $k \in \{p \text{ (protein)}, g \text{ (gene)}\}$. Figure 2 describes the temporal evolution of the total number of new protein- and gene-targeted therapies observed in the data. These series display a similar pattern—they are roughly stable through 2006, but increasing between 2007-2012.

3.2 Knowledge Stocks

In March of 2016, we downloaded genetic association studies from GWAS Central."[7] The data include 2,044 studies, covering 1,362 traits. We restricted our attention only to traits that correspond to human diseases.[8] These studies associate a human disease with genetic variants (single-nucleotide polymorphisms): a *variant-disease association* ("VDA"). For

[6] Cortellis identifies these by collecting information of new therapies discussed in academic conferences or scientific publications, reported by a clinical trial submitted to www.clinicaltrials.gov or other websites, featured by the media or regulatory updates, or announced in the sponsoring firm's website.

[7] `http://www.gwascentral.org/`.

[8] Traits that are not associated with human disease include, for example, "economic and political preferences," "educational attainment," "freckles," "hand grip strength," among others. We retain only those pertaining human diseases as defined by Merriam-Webster's: "an illness that affects a person, animal, or plant: a condition that prevents the body or mind from working normally."

our empirical analysis, we assume that each VDA associating a disease to a set of variants adds one unit to the cumulative stock of knowledge for the disease, at the time when the corresponding GWAS is published.

Matching VDA data to the Cortellis pipeline presented two main challenges. The first regards differences in spelling and use of synonyms (e.g., peanut allergy and peanut hypersensitivity, Wilms' tumor and Nephroblastoma, etc). The second, and more challenging, stemmed from differences between the disease ontologies used by GWAS Central and Cortellis. For example, we noticed that the GWAS trait "longevity" could inform the design of Cortellis therapies targeting "aging." Similarly, the GWAS trait "5-htt brain serotonin transporter levels"—which is thought to underlie a variety of neuropsychiatric disorders—could inform the design of Cortellis therapies compounds targeting "post-natal depression." To systemically bridge these two ontologies we assembled a team of experts. Two independent coders (M.D. residents) were asked to identify as many matches as possible from the data. A third expert (Ph.D. in Epidemiology) then curated these lists and resolved conflicts. As a result, 17% of the diseases targeted by the therapies in the Cortellis sample were matched to at least one GWAS VDA. Figure 3 shows the number of GWAS VDAs and the number of different diseases recorded for each therapeutic area.

We construct the variable VDAFLOW_{dt} as the total number of VDAs published for a disease d in year t. Following the approach of Adams (1990) and Toole (2012), we define a knowledge-stock variable VDASTOCK as:

$$\text{VDASTOCK}_{dt} = \log\Big(1 + \sum_{t'=2003,..,t} (1-\delta)^{t-t'}\text{VDAFLOW}_{dt}\Big),$$

where $\delta \in [0,1]$ corresponds to an "obsolescence rate" that accounts for the depreciation of the knowledge embedded in GWAS publications over time. The log-transformation incorporates the idea that knowledge accumulation may be subject to marginally decreasing impacts on innovation. GWAS began to be published in 2005, so VDASTOCK equals zero for all diseases in 2003 and 2004. The dashed line in Figure 2 corresponds to VDASTOCK (assuming $\delta = 0$). The figure shows a high correlation between VDASTOCK and the number of new therapies.

3.3 Human Disease Network and Biological Complexity

We construct proxies for biological complexity from an implementation of the Human Disease Network ("HDN")—an undirected network in which diseases are connected to each other through common gene mutations (Goh et al., 2007). To build the most complete (up-to-date) representation of this network, we retrieved data from DisGeNET,[9] an aggregator that is widely considered the largest publicly-available repository of scientific results linking human diseases to their genetic underpinnings (Piñero et al., 2017). DisGeNET aggregates VDAs (like GWAS) and the coarser gene-disease associations ("GDAs"), from an array of specialized sources that focus on specific diseases or scientific approaches. At the time of data download, DisGeNET included 561,119 GDAs and 135,588 VDAs, covering over 20,000 diseases.

Table 1 presents descriptive statistics for the different data sources from where DisGeNET aggregates associations. These sources can be grouped into three categories shown in panels A, B, and C, respectively.[10] Curated sources (Panel A) include the GWAS Catalog, CTD Human, CLINVAR, HPO, ORPHANET, PSYGENET, and UNIPROT. Although all of these rely on findings submitted by individual scientific groups, they differ in terms of their focus and curation process. For example, CTD Human (Comparative Toxicogenomics Database) focuses on promoting the understanding of the effects of chemicals on human health, while ORPHANET focuses on rare diseases. In terms of the extent of curation, some data sources may select entries based on statistical significance (GWAS) and possibly reinterpret results for "accurate and comprehensive representation of biological knowledge" (UNIPROT), whereas others accept all submitted GDAs (insofar as supporting evidence is provided) and abide by the interpretations provided by the submitting group (CLINVAR). Panel B describes sources of results predicted from genomic analysis on laboratory mice and rats,[11] while Panel C describes sources which compile GDAs and VDAs by text mining the scientific literature.

The HDN can be implemented by considering either the set of available GDAs or VDAs. In particular, a network can be constructed based on the premise that any two diseases that are associated with the same gene or variant thereof should be connected in the respective network, whereas any two diseases which do not share associations,

[9] `http://www.disgenet.org`. We retrieved DisGeNET version 4.0 data on 6/12/17.

[10] This Table is reproduced with permission from the DisGeNET website. Minor formatting changes have been introduced for clarity.

[11] MGD and RGD respectively stand for corresponds to Mouse and Rate Genome Database.

should appear as disconnected. For example, our data shows that Parkinson's disease and Waldenström's disease are both associated to the EPO gene. These diseases will thus be connected in a network implemented with GDA data. We also observe that Parkinson's disease and Myopia are both associated to the HGF, KRAS, and PTEN genes. For simplicity, our implementations will assume that the "strength" or "validity" of the connections between Parkinson's and Waldenström's diseases, and Parkinson's disease and Myopia are the same. On the other hand, although Anemia is associated to several genes, none of these is also associated with Parkinson's disease. Thus, a network implementation based on GDA data would portray them as disconnected diseases.

We construct independent HDN versions using both types of association data. We label the network implementation based on GDAs as "GHDN" and that based on VDAs, as "VHDN." Differences between these arise not only because they rely on non-overlapping sets of scientific results, but more importantly, because VHDN imposes a more stringent requirement to establish connectedness between diseases.[12] As a result, VHDN presents a much more sparse structure, with lower overall levels of connectedness. Indeed, the number of connections in the VHDN is only about 2% that of the GHDN. Furthermore, whereas about 18% of diseases are "isolated"(disconnected from all other diseases) in the latter, 42% are isolated in the former.

Following the insights of previous research (e.g., Wachi et al., 2005; Jonsson and Bates, 2006; Bauer-Mehren et al., 2011; Silverman and Loscalzo, 2012), we use two network statistics to proxy for the biological complexity of each disease d. In particular, we define: (i) d's total number of genetic associations (NASSOC), and (ii) d's degree of centrality (CENTRALITY). For the GHDN, NASSOC corresponds to the total number of genes associated with d; for the VHDN, to the total number of associated variants. CENTRALITY corresponds to the total number of diseases $d' \neq d$ to which d is directly linked through networks' respective connectors.

Table 2 presents the distribution of NASSOC and CENTRALITY under the GHDN implementation (Panel a) or the VHDN implementation (Panel B). In both cases, there is a wide dispersion and a long right tail, i.e., there is a small number of diseases characterized by high biological complexity. Consistent with the higher sparsity of the VHDN network, the centrality measure is much lower than the centrality under the GHDN im-

[12] For two diseases to be connected in GHDN they ought to be associated to *some* mutation of the same gene. For them to be connected in VHDN, they need to be associated to *the same* mutation of the same gene.

plementation, but in both cases centrality measures are highly correlated at the disease level.[13] Interestingly, for both implementations, values of CENTRALITY are generally larger than those of NASSOC. This occurs due to the existence of clusters of highly interconnected diseases, where one gene or variant enables the connection of one disease with many others. Figure 4 presents averages by therapeutic area. Among others, patterns in this figure suggest that variants of Cancer rank high in both number of associations and network centrality.

4 Empirical Strategy

To motivate our empirical strategy, Figure 5 shows the patterns of innovation and accumulation of GWAS publications for cardiovascular diseases. Panel A in Figure 5 shows the cumulative number of published GWAS VDAs available each year for each of the 98 cardiovascular diseases.[14] Panel B and Panel C show the number of gene-targeted and protein-targeted therapies that enter the discovery stage. A visual inspection of these patterns suggests a rough correlation between the accumulation of published VDAs and the amount of innovation for each disease. Our empirical analysis distills this relationship by controlling for observed and unobserved conditioning factors employing count-data models.

Given that the data exhibits over-dispersion, we estimate Negative Binomial specifications. Furthermore, Figure 5 shows that there is a large number of observations associated with $N_{dkt} = 0$. Although these occurrences primarily manifest for gene-targeted therapies, they are not rare among protein-targeted therapies. To account for this feature of the data, we use a zero-inflated specification of the Negative Binomial model, which allow us to separately capture the determinants of $N_{dkt} = 0$ and $N_{dkt} > 0$ outcomes.[15]

Another distinctive pattern of Figure 5 is that the number of new therapies observed each year is unevenly distributed, and temporally persistent, across diseases. That is, some diseases exhibit systematically larger values of N_{dkt}. Moreover, this heterogeneity

[13] The correlation for CENTRALITY is 0.73 ($p < 0.01$). The correlation for NASSOC is 0.63 ($p < 0.01$).

[14] Within the sample period, GWAS VDAs became available for about 30% of diseases in this area.

[15] In the full sample, about 78% N_{dkt} observations equal 0 (61% for gene-targeted therapies, 95% for protein-targeted ones). The zero-inflated specification is supported by the Vuong test. The inflation model is specified to include a constant and an indicator for gene-targeted therapies.

is observed for both gene- and protein-targeted therapies. To account for this form of "*dk*-specific," time-invariant unobserved heterogeneity we employ a "pre-sample mean estimator" approach of Blundell et al. (2002), where average pre-sample values of the dependent variable are used to proxy for unobserved heterogeneity. In our context, this amounts to including the average of the logged dependent variable in a pre-sample period as an independent variable in our regressions, while constraining its coefficient to one. We compute this pre-sample mean using data from the 1990-2001 period.

For our econometric analysis, we estimate several versions of the following specification:

$$N_{dkt} \quad = \quad f(\Theta X_{dkt} + \lambda_t + \eta_{a(d)} + \hat{\mu}_{dk}), \tag{1}$$

where f corresponds to the zero-inflated negative-binomial functional form, λ_t is a year fixed effect, $\eta_{a(d)}$ is a therapeutic area fixed effect, and $\hat{\mu}_{dk}$ corresponds to the disease/target-specific pre-sample level, given by:

$$\hat{\mu}_{dk} = \log(1 + \bar{N}_{dk}), \text{ with } \bar{N}_{dk} = \frac{1}{12} \sum_{t=1990,..,2001} N_{dkt}$$

In equation (1), Θ is a vector of coefficients for the variables contained in X. Along with the first lag of VDASTOCK, X includes an indicator that identifies gene-targeted therapies (GENETARGET), the disease-specific network statistics (NASSOC, CENTRALITY) that proxy for translational complexity, and their interactions with VDASTOCK's first lag.

To account for economic and public health "pulling forces," we also include in X the first lags of MEPSPATS and MEPSEXPND, which respectively proxy for the epidemiological pervasiveness and market size associated with each disease, and are constructed using data from the Medical Expenditure Panel Survey (MEPS). MEPSPATS corresponds to the log total number of patients (in millions) in the US that report suffering from condition d during year t; MEPSEXPND, corresponds to the log total amount spent on prescription drugs for the condition during the same year (measured in billions of dollars, CPI-adjusted to year 2000).[16]

[16] MEPS (`https://meps.ahrq.gov/mepsweb/`) is a large and representative sample of health care usage and insurance in the US. MEPSPATS is constructed using data from MEPS' yearly "Medical Conditions Files," which report the incidence of diseases on individuals at the 3-digit ICD9 level. Thus, all diseases associated with a single 3-digit ICD9 code are awarded the same value for MEPSPATS. MEPSEXPND is constructed with data from the yearly "Prescribed Medicines Files" using the same procedure. In both cases, individual variables are aggregated at the year level using individual representativeness weights.

The parameters of interest in equation (1) are those corresponding to VDASTOCK's lag and those corresponding to the interaction between VDASTOCK's lag and the network statistics for diseases. In particular, a statistically significant and positive coefficient for VDASTOCK's lag would indicate that larger stocks of GWAS science increase the rate of therapeutic innovation. The coefficient's specific value would then illustrate the extent of this translational effect. The parameter for its interactions with network statistics would identify the extent to which this effect is moderated by each disease's network environment.

In the analysis, we avoid imposing assumptions regarding the relative adequacy of GHDN and VHDN as a means to characterize biological complexity. Our approach is to first show that the main promoted effects hold when each of these are considered independently, and then, that they continue to hold when the joint variation of GHDN and VHDN is summarized by an ordering of diseases, which we derive through a flexible, data-driven clustering method.

Lastly, by the structure of DisGeNET data, the computation of network statistics from the GHDN or VHDN do not hinge on GWAS science. This is suggested by Figure 6, which compares the NASSOC and CENTRALITY values computed with and without accounting for GWAS results in the construction of VHDN (in red, the 45 degree line).[17] Although for some diseases GWAS results account for a non-negligible share of observed associations, they do not significantly distort the overall ordering. Together, these observations suggest that GWAS science does not overtly condition our measurement of biological complexity. In subsection 5.2 we analyze the potential inferential confounds introduced by this and other issues, finding no evidence to suggest that they drive our main results.

[17] To facilitate the comparison, values are normalized by each variables' largest values when all DisGeNET results are considered.

5 Results

5.1 Translational Complexity

Table 3 presents the estimates of different versions of equation (1).[18] Column 1 presents the simplest specification, which does not include our measure of biological complexity, nor the pre-sample proxy for unobserved heterogeneity. In Column 1, the coefficient for GENETARGET is negative, large, and strongly significant, which is consistent with the systematically smaller number of gene-targeted therapies. The estimated coefficient for VDASTOCK implies that a 1 percent increase in the stock of GWAS knowledge is associated with a 1.4 percent increase in the number of new therapies entering the discovery stage. The estimated coefficient for the interaction between VDASTOCK and GENETARGET suggest that scientific knowledge stocks have a larger impact on the innovation of protein-targeted rather than gene-targeted therapies. Furthermore, consistent with the results of Toole (2012), the coefficient estimates for MEPSPATS and MEPSEXPND are positive, indicating a disease's epidemiological pervasiveness and market size both increase the rate of therapeutical innovation. Although both of these are estimated precisely by this specification, MEPSPATS loses its statistical significance in the more comprehensive specifications.

In Column 2 we control for disease-target type-specific unobserved heterogeneity through the coefficient-constrained inclusion of logged pre-sample means. Although most coefficients retain their sign and statistical significance, their magnitude becomes smaller, suggesting that this type of unobserved heterogeneity plays a relevant role in innovation rates. This is particularly noteworthy for the coefficient of VDASTOCK, which now is about half the estimate of Column 1, implying that a 1 percent larger knowledge stock can be linked to only a 0.7 percent increase in new therapies entering the drug development process.

The specifications of Columns 3 and 4 incorporate our measure of biological complexity, which is computed with the GHDN network implementation. Both CENTRALITY and NASSOC are measured in hundreds and these two measures are (by construction)

[18] These and subsequent results are obtained by setting the "obsolescence rate" $\delta = 0.05$. This value was determined by comparing information-based criteria of specifications estimated on a grid for plausible δ values.

highly correlated. To avoid multicollinearity issues we considered these variables separately in Columns 3 and 4. The specification of Column 3 considers diseases' network centralities. The positive and significant coefficient for CENTRALITY points to a particular dimension of unobserved heterogeneity, whereby more central diseases are more frequently the focus of therapeutical innovation. The positive coefficient for its interaction with GENETARGET suggests that the effect is more pronounced for gene-targeted therapies.

This baseline effect of CENTRALITY could be rationalized by a combination of supply- and demand-side factors. The latter could ensue if MEPSEXPND underestimates the "true" market potential for more central diseases. Noting that Cancer variants tend to have higher network centralities (see Figure 4), such underestimation is a real possibility in this context. This is because MEPSEXPND is computed from prescription drug data, which may omit much of the expenditure on drugs used for the treatment of Cancer (typically administered via injections and thus, possibly, not available through prescription on a systematic basis).[19] At a more fundamental level, the underestimation of market potential could be grounded in the possibility that new Cancer therapies provide a particularly significant improvement to the standard, compared to other therapeutic areas, and so unlock value that is unaccounted for by historical spending patterns. Table 4 presents results obtained by reproducing the above analysis, but on a sample that omits diseases in the Cancer area. Because results remain largely unchanged, they attenuate the concerns stemming from these potential confounds. Supply-side factors justifying the positive coefficient of CENTRALITY could be rooted in potentially larger knowledge spillovers or scrap values for therapies targeting more central diseases. Under this view, the return to investment of these therapies may, in part, be driven by the broader usefulness of applied knowledge generated in the process, or by the ability to re-purpose therapies for use in the treatment of different, bu related, conditions.

In Column 3, the coefficient associated with the interaction of VDASTOCK and CENTRALITY is negative and statistically significant. This suggests that the impacts of larger knowledge stocks on innovation rates are smaller for more central diseases. To the extent that CENTRALITY is accepted as a proxy for biological complexity, this coefficient shows evidence that new genetic epidemiological science has a smaller innovative impact

[19] Dranove et al. (2014) provide some facts that suggest that measures for market potential for cancer indications that are based on prescription drug expenditures may not be completely inadequate. For example, many of the top-selling biotechnology drugs are covered by Medicare prescription drug insurance. Some of this coverage may operate through the practice of "brown bagging," by which patients purchase drugs in retail pharmacies and then have them immediately administered in an outpatient setting.

among the more complex diseases. The same conclusion can be drawn from the estimates of Column 4, which account for biological complexity through NASSOC. Columns 5 and 6 reproduce the analysis of Columns 3 and 4, but use complexity metrics computed with the VHDN implementation. These estimates offer further support for the moderating role of complexity in translation.[20]

Although these results broadly support our main insight—that biological complexity mediates the extent of translation of new GWAS knowledge into therapeutical innovation—they also entail the possibility that larger knowledge stocks may deter innovation. Concretely, the coefficient estimates of specifications 3 through 5 all imply that, evaluated at a large enough percentile of CENTRALITY or NASSOC, an increase in VDASTOCK could deter innovation.

To further investigate the relationship between biological complexity and innovation, we cluster diseases into groups according to their measured network presence. In particular, we apply a *k-means* clustering algorithm on the list of all four available network statistics to group diseases into different subsamples. As a result, we obtain a partition of diseases without imposing assumptions regarding the relative importance of different network statistics or implementations.[21]

Table 5 characterizes the result of this clustering procedure. The first and largest group (subsample $\mathcal{S}_1$) includes 35% of the diseases in the sample. Diseases in $\mathcal{S}_1$ exhibit the lowest average values for all network metrics, i.e., they are the more peripheral or less connected than the remaining diseases in the sample. Thus, diseases in $\mathcal{S}_1$ are associated with lower amounts of biological complexity. Diseases in subsample $\mathcal{S}_i$ have lower average values for all network metrics compared to diseases in subsample $\mathcal{S}_{i+1}$. Also, subsample $\mathcal{S}_i$ contains more diseases than subsample $\mathcal{S}_{i+1}$. For example, subsample $\mathcal{S}_5$ includes 6% of the diseases in the sample, with average statistics exceeding those of $\mathcal{S}_1$ by at least two orders of magnitude. Thus, diseases in $\mathcal{S}_5$ correspond to the more central, more connected, and hence, the more complex diseases in the sample. Together, these statistics suggest that the clustering procedure yields a reasonable ordering of diseases into categories of distinct biological complexity.

We estimate a simplified version of specification (1) for each subsample. In particular,

[20] Differences in parameter values are largely driven by the different scaling of GHDN and VHDN metrics.

[21] We settled on five clusters because more clustering yields subsamples that are too small for estimation.

because there is relatively little variation in the network presence of diseases included within each subsample, we drop network metrics and their interactions from the set of dependent variables. The estimated coefficients for VDASTOCK in Table 6 indicate that the impact of new GWAS science on rates of therapeutical innovation is decreasing in complexity, which is consistent with the findings in Table 3. This suggest that, for the set of diseases associated to the lower levels of measured complexity (subsample $\mathcal{S}_1$), a 1 percent increase in VDASTOCK is associated with a 1.12 percent increase in the number of new therapies entering the discovery stage. The effect is generally decreasing with the average complexity of the subsamples. For diseases in $\mathcal{S}_5$, the effect is not significantly different from zero at usual statistical confidence levels. Thus, these results indicate that the negative coefficients associated with interactions of VDASTOCK and network metrics in Table 3 are primarily derived from variation at lower ranges of the considered network statistics, and cannot be taken to imply that larger stocks of GWAS knowledge could deter innovation.

A second aspect of interest in Table 6 corresponds to the sequence of coefficients for MEPSEXPND. In particular, these coefficients suggest that innovation for more peripheral (less connected) diseases is more responsive to market conditions when compared to innovation for more central (more connected) diseases. Based on the higher average centrality of Cancer diseases, we conjecture that this pattern may reflect a possible correlation between diseases' burden and network environment. If, like Cancer, more burdensome diseases are also more connected or central in the HDN, they may have also constituted the more frequent historical targets of the industry's innovation efforts. In this scenario, the less connected, more peripheral set of diseases would have fewer therapeutical alternatives. Guided by expected market profitability, pharmaceutical developers may have therefore seen this set of diseases as more lucrative for the application of novel genetic epidemiological science.

Table 7 displays the marginal effects of knowledge accumulation. These are computed by increasing in one the number of available GWAS publications for each disease, and then computing the implied percentage difference in the number of new therapies (averaged across diseases within each cluster). Measured both given the stocks of 2004 and 2012, these marginal effects largely coincide with the insights of Table 6.

Lastly, to provide a sense of the overall contribution of GWAS science to therapeutical innovation over the covered period, Figure 7 decomposes the bulk of new therapies observed in the sample, singling out the share of those which, according to the above model estimates, can be attributed to GWAS knowledge. Panel A focuses on gene-targeted therapies and Panel B, on protein-targeted ones. Because there was virtually no GWAS science published before 2007, their innovative impacts are perceivable only after 2007. Following the progressive accumulation of knowledge, a higher percentage of new therapies that can be linked to GWAS publications. By 2012, the contribution of GWAS is largest for gene-targeted therapies of lowest measured complexity ($\mathcal{S}_1$), at around 25%.

5.2 Do Unobservables Drive our Results?

Our inference may be confounded by two main factors: the conditional independence of GWAS knowledge accumulation and the computed measures of complexity. In this section we provide evidence suggesting that these concerns are unlikely to overturn our main conclusions.

We begin by addressing the conditional independence of the computed network statistics. Because the HDN implementations used above rely in part on DisGeNET research published during the sample period (including GWAS), one may worry about the existence of unobserved trends driving both the focus of this research, as well as that of the industry's innovative efforts.

To investigate this concern we replicate earlier results, but only construct NASSOC and CENTRALITY from DisGeNET research published no later than 2005. To carry out this analysis, we find the publication date of each article in the DisGeNET database. The 2005 threshold was selected in consideration of two factors. First, only one GWAS result in our sample was published before 2006. Second, selecting earlier thresholds significantly reduced the set of DisGeNET results available to implement GHDN and VHDN, yielding relatively little variation on NASSOC and CENTRALITY. Indeed, even with the 2005 threshold, the computed NASSOC and CENTRALITY variables present considerably less variation than in the original sample, the primary reason being that these data contain no associations for a much larger number of diseases. As a consequence, in this case, the resulting 5-cluster grouping yields an $\mathcal{S}_1$ subsample that includes 68% of diseases, whereas subsamples $\mathcal{S}_4$ and $\mathcal{S}_5$ include 7% and 3%, respectively. The higher degree of degeneracy of this partition prevents us from replicating the cluster-based analysis of Table 6, so we focus on the original specification (1) used by Table 3.

Estimation results are presented in Table 8. Although slightly smaller than in Table 6, the estimated coefficients of VDASTOCK have similar values and significance. Furthermore, because the smaller set of DisGeNET results used to implement the networks yields lower-valued NASSOC and CENTRALITY, their associated coefficients have generally larger values than in Table 6. Nevertheless, these retain their signs and statistical significance, suggesting that the considered trend does not drive the translational complexity effect.

Owing to the usefulness of GWAS knowledge for therapeutical innovation, a potential violation of their conditional independence is perhaps a bigger concern. In an extreme scenario, the documented positive impact of VDASTOCK on innovation could be entirely rooted in scientific or market trends that are unaccounted for by our analysis, but which prompt the conflux of higher innovation and GWAS publication rates among certain diseases. That is, in this extreme scenario, the identified translational rate could entirely represent the bias imposed by an omitted variable.

Before analyzing this issue, recall that our main insight—therapeutic translation rates are decreasing in biological complexity—relies on a comparison of translation rates across the support of measured biological complexity. We argue that this result is unlikely to be overturned by the presence of this type of trend, as the latter would be required to exhibit a rather specific structure. Namely, it should manifest more intensively among the less complex groups of diseases. The series of analysis performed in turn provide some support to this point.

We first implement a falsification test, based on the following rationale: if the large translational effects observed among low-complexity diseases are driven by the described omitted trend, we should continue to observe them when publication sequences are randomized within diseases with similar patterns of GWAS knowledge accumulation. This randomization would disband the empirically detectable causality of VDASTOCK on N, while roughly maintaining the structure of the alleged omitted trend.

We implement this test through tiered-resampling. Tier 1 includes the approximately 83% of diseases in the sample for which there are no GWAS VDAs in the sample. Tiers 2 and 3 partition the remaining set of diseases in groups of approximately equal size, based on the total number of available GWAS VDAs for each disease, and in such a way that Tier 3 diseases all have more GWAS VDAs than those in Tier 2. Publication sequences are then re-sampled (with replacement) within diseases of each tier, so maintaining the average number of published GWAS VDAs within each. (This average is always 0 for Tier 1). We generate 200 pseudo-samples following this procedure, reproducing the analysis for subsamples $\mathcal{S}_1$ and $\mathcal{S}_2$ (Columns 1 and 2 of Table 6) on each. Results indicate that the 1.14 and 0.95 estimates of Table 6 are largely improbable outcomes given the estimated parameter distributions: in both cases, they are larger than 99% of the obtained estimates. This analysis suggests that our main insight is not driven by the conflux of higher publication and innovation rates caused by an omitted trend.

We further note that, although the number of GWAS publications may be correlated with innovation series through an omitted trend, such a trend is likely to be a less important determinant of the informational content of published GWAS results. That is, although scientific and economic tendencies may prompt researchers to engage with specific research agendas at a certain times, they are less likely to determine the quality of these agendas' outcomes. Equivalently, these tendencies are less likely to determine the effective amount of usable knowledge that each GWAS publication adds to the knowledge base.

Based on this premise, and on extensive research (e.g., Garfield, 1972; Moed, 2006) suggesting that citation counts can be taken as a proxy for articles' contribution to existing knowledge base, we devise a test which exploits variation in GWAS articles' (forward) citations. Because a series of VDASTOCK constructed from the more cited articles would be less affected by the cited omitted trend, observing that our results continue to hold when this series is used would help to alleviate the concern at hand.

From the vast scientific and medical bibliography available from PubMed, we identified the set of articles that cite each GWAS publication in our data. Because articles published in earlier years have had a longer time to accumulate citations, we computed the number of citations observed within two years of publication. Considering the median number of citations obtained by articles contributing VDAs to each targeted disease, we next constructed two versions of VDASTOCK: one based on the articles that obtained a below-median citation count, and the other based on those which obtained an above-median citation count. Although these two versions of VDASTOCK are constructed based on an approximately similar number of articles for each disease, their values at a given point in time usually differ because high- and low-cited articles are not published at the same time. In the overall sample, there is no statistically significant difference between the average values of these two versions of VDASTOCK, suggesting that an article's impact is independent of its publication date.

Panel A of Table 9 presents the results obtained when we reproduce the specification of Table 6, but replace VDASTOCK with high- and low-cites articles (for exposition, other variables are excluded from the Table). Maintaining the basic result that complexity mediates translational rates, the set of estimated coefficients suggests that knowledge produced by the articles with more impact is associated with a generally larger effect on innovation. Thus, these results lend support to the idea that our main results are not driven by the influence of the described omitted trend.

We finish this section by considering a more specific form of omitted variable. Namely, the possibility that published GWAS VDAs are themselves the output of firms' decisions to innovate a new therapy. In particular, suppose that, in order to evaluate wether to introduce a candidate to the development process, pharmaceutical firms conduct the same type of analysis contained by GWAS publications. If this analysis demonstrates a genetic linkage for an specific disease, we may observe an increase in GWAS publications that precedes that for the introduction of new therapies into the development process. Such an effect could, by itself rationalize our results.

We analyze this issue based on the idea that this rationale is more likely to be reflected among GWAS publications funded by the industry, rather than based on those funded by public entities. If our main result primarily relied on VDASTOCK series constructed from the former, the validity of our main insight should be discounted. To implement the test we mined articles' acknowledgements and PubMed records in order to identify the set of GWAS publications in our data where industry funding is acknowledged. About 21% of the GWAS publications in the sample have these acknowledgements. Next, as before, we constructed two versions of VDASTOCK, one based on the articles that report this type of funding, and other based on those that do not, and estimated an analog specification. Panel B of the Table 9 presents the results. These, for the most part, suggest that the translational effect is attached to those articles that do not report industry funding. Therefore, these results do not offer support for the idea that our main result follows from the considered reverse causality hypothesis.

6 Conclusions

Notwithstanding a rich stream of research investigating the extent and mechanics by which basic science fuels and shapes pharmaceutical innovation,[22] the role of *biological complexity* has remained unexplored. We combine insights of *Network Medicine* with standard approaches for the measurement of R&D returns (Blundell et al., 2002; Toole, 2012) to take a step forward in this direction.

Our results posit that biological complexity is an important determinant of the rate of translation. This rate is large among diseases with lower measured complexity, decreasing as complexity rises, and indistinguishable from zero among diseases in the extreme of higher complexity. Particularly, in the current "genomic era," biological complexity stands out as a potentially important conditioning factor for the assessment of innovative productivity in the industry, and the allocation of funding by scientific agencies. It may also represent a useful construct to retrospectively assess the overall impacts of the Human Genome Project, as well as to fine tune expectations going forward.

As with much of the research oriented at measuring the returns of R&D, our analysis grapples with significant identification challenges (Hall et al., 2010). Here, these arise primarily because the *direction* of scientific research and therapeutical innovation are likely determined by common factors, which are not observable in the data, and cannot be fully controlled for empirically. We must therefore promote a cautious interpretation of the estimated coefficients. Nevertheless, a series of checks suggest that our main insight—that the translation rate is decreasing in biological complexity—is unlikely to be overturned by biases introduced in through these means. Empirical approaches that exploit exogenous variation stemming from the nuances of research funding rules (as in Azoulay et al., 2015) may be useful to further assert the validity of these results. This approach may also allow to address issues that we are forced to neglect here. For example, the speed of translation.

Two avenues for follow-up research stand out in our view. First, genetic epidemiological knowledge may be useful during clinical trial development, by guiding the identification of sub-populations at higher risk of developing adverse side effects.[23] These events point

[22] This literature is referenced in the introduction.

[23] Pollack (June 14, 2010) reports "Many drug companies now collect and analyze the DNA of patients in clinical trials, hoping to find genetic signatures that will allow drugs to be better tailored to specific patients." Pollack (June 14, 2010) rationalizes this trend with the case of the blockbuster antiplatelet drug "Plavix," for which a variation of the gene CYP-2C19 was found to render patients at higher risk

to a two-layered translational effect, one operating through the amount of innovation, the other through potentially higher rates of clinical trial success. It is not clear a-priori whether biological complexity will boost or temper the clinical trial success. Secondly, the presence and extent of scale and scope economies has been an important area of inquiry in the study of the pharmaceutical industry (e.g., Henderson and Cockburn, 1996; Cockburn and Henderson, 2001). However, most of this research utilizes pre-genomic datasets that are highly aggregated. By virtue of its rich and "exogenous" structure, the Human Disease Network permits the construction of "spillover weights" directly from the data, at the disease-pair level. Applied to contemporaneous data, this approach could enable a more translucent, fine-grained analysis of pharmaceutical scale and scope economies in the genomic era.

of heart attacks. The point is also illustrated by the 2004 market withdrawal of Merck's Cox-2 inhibitor "Vioxx" (Rofecoxib) due to adverse cardiovascular events. Years later, the research of Brune et al. (2008) and Ruff et al. (2011) found that these events were associated with patients exhibiting high levels of an aminoacid, which could be detected in advance through genetic diagnostics (Goldman et al., 2013).

References

Adams, J. D. (1990), 'Fundamental stocks of knowledge and productivity growth', *Journal of Political Economy* **98**(4), 673–702.

Ahmadpoor, M. and Jones, B. F. (2017), 'The dual frontier: Patented inventions and prior scientific advance', *Science* **357**(6351), 583–587.

Azoulay, P., Zivin, J. S. G., Li, D. and Sampat, B. N. (2015), Public r&d investments and private-sector patenting: evidence from nih funding rules, Technical report, National Bureau of Economic Research.

Barabási, A.-L., Gulbahce, N. and Loscalzo, J. (2011), 'Network medicine: a network-based approach to human disease', *Nature reviews. Genetics* **12**(1), 56.

Bauer-Mehren, A., Bundschus, M., Rautschka, M., Mayer, M. A., Sanz, F. and Furlong, L. I. (2011), 'Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases', *PloS one* **6**(6), e20284.

Blundell, R., Griffith, R. and Windmeijer, F. (2002), 'Individual effects and dynamics in count data models', *Journal of Econometrics* **108**(1), 113–131.

Boyle, E. A., Li, Y. I. and Pritchard, J. K. (2017), 'An expanded view of complex traits: From polygenic to omnigenic', *Cell* **169**(7), 1177–1186.

Brune, K., Katus, H. A., Moecks, J., Spanuth, E., Jaffe, A. S. and Giannitsis, E. (2008), 'N-terminal pro–b-type natriuretic peptide concentrations predict the risk of cardiovascular adverse events from antiinflammatory drugs: a pilot trial', *Clinical chemistry* **54**(7), 1149–1157.

Burdett, T., Hall, P., Hastings, E., Hindorff, L., Junkins, H., Klemm, A., MacArthur, J., Manolio, T., Morales, J., Parkinson, H. and D, W. (2016), 'The nhgri-ebi catalog of published genome-wide association studies.'.

Cao, C. and Moult, J. (2014), 'Gwas and drug targets', *BMC genomics* **15**(4), S5.

Cockburn, I. M. and Henderson, R. M. (1998), 'Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery', *The Journal of Industrial Economics* **46**(2), 157–182.

Cockburn, I. M. and Henderson, R. M. (2001), 'Scale and scope in drug development: unpacking the advantages of size in pharmaceutical research', *Journal of health economics* **20**(6), 1033–1057.

Cohen, W. M., Nelson, R. R. and Walsh, J. P. (2002), 'Links and impacts: the influence of public research on industrial r&d', *Management science* **48**(1), 1–23.

Daiger, S. P. (2005), 'Was the human genome project worth the effort?', *Science* **308**(5720), 362–364.

Dranove, D., Garthwaite, C. and Hermosilla, M. (2014), Pharmaceutical profits and the social value of innovation, Technical report, National Bureau of Economic Research.

Evans, J. P., Meslin, E. M., Marteau, T. M. and Caulfield, T. (2011), 'Deflating the genomic bubble', *Science* **331**(6019), 861–862.

Fleming, L. and Sorenson, O. (2001), 'Technology as a complex adaptive system: evidence from patent data', *Research policy* **30**(7), 1019–1039.

Fleming, L. and Sorenson, O. (2004), 'Science as a map in technological search', *Strategic Management Journal* **25**(8-9), 909–928.

Gambardella, A. (1995), *Science and innovation: The US pharmaceutical industry during the 1980s*, Cambridge University Press.

Garfield, E. (1972), 'Citation analysis as a tool in journal evaluation', *Science* **178**(4060), 471–479.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M. and Barabási, A.-L. (2007), 'The human disease network', *Proceedings of the National Academy of Sciences* **104**(21), 8685–8690.

Goldman, D. P., Gupta, C., Vasudeva, E., Trakas, K., Riley, R., Lakdawalla, D., Agus, D., Sood, N., Jena, A. B. and Philipson, T. J. (2013), The value of diagnostic testing in personalized medicine, *in* 'Forum for Health Economics and Policy', Vol. 16, pp. S87–S99.

Hall, B. H., Mairesse, J. and Mohnen, P. (2010), 'Measuring the returns to r&d', *Handbook of the Economics of Innovation* **2**, 1033–1082.

Hayden, E. (2010), 'Human genome at ten: life is complicated', *Nature News* **464**(7289), 664–667.

Henderson, R. and Cockburn, I. (1994), 'Measuring competence? exploring firm effects in pharmaceutical research', *Strategic management journal* **15**(S1), 63–84.

Henderson, R. and Cockburn, I. (1996), 'Scale, scope, and spillovers: the determinants of research productivity in drug discovery', *The Rand journal of economics* pp. 32–59.

Jaffe, A. B. (1989), 'Real effects of academic research', *The American economic review* pp. 957–970.

Jonsson, P. F. and Bates, P. A. (2006), 'Global topological features of cancer proteins in the human interactome', *Bioinformatics* **22**(18), 2291–2297.

Kerem, B.-s., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., Buchwald, M. and Tsui, L.-C. (1989), 'Identification of the cystic fibrosis gene: genetic analysis', *Science* **245**(4922), 1073–1080.

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. and Mardis, E. R. (2013), 'The next-generation sequencing revolution and its impact on genomics', *Cell* **155**(1), 27–38.

Lander, E. S. (2011), 'Initial impact of the sequencing of the human genome', *Nature* **470**(7333), 187.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. et al. (2001), 'Initial sequencing and analysis of the human genome', *Nature* **409**(6822), 860–921.

Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., Gilad, Y. and Pritchard, J. K. (2016), 'Rna splicing is a primary link between genetic variation and disease', *Science* **352**(6285), 600–604.

Londin, E., Yadav, P., Surrey, S., Kricka, L. J. and Fortina, P. (2013), 'Use of linkage analysis, genome-wide association studies, and next-generation sequencing in the identification of disease-causing mutations', *Pharmacogenomics: Methods and Protocols* pp. 127–146.

Manolio, T. A. (2013), 'Bringing genome-wide association findings into clinical use', *Nature reviews. Genetics* **14**(8), 549.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A. et al. (2009), 'Finding the missing heritability of complex diseases', *Nature* **461**(7265), 747–753.

Mansfield, E. (1995), 'Academic research underlying industrial innovations: sources, characteristics, and financing', *The review of Economics and Statistics* pp. 55–65.

Mardis, E. R. (2011), 'A decade's perspective on dna sequencing technology', *Nature* **470**(7333), 198.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. and Hirschhorn, J. N. (2008), 'Genome-wide association studies for complex traits: consensus, uncertainty and challenges', *Nature reviews. Genetics* **9**(5), 356.

Moed, H. F. (2006), *Citation analysis in research evaluation*, Vol. 9, Springer Science & Business Media.

Palmer, B. (September 30, 2013), 'Where are all the miracle drugs?', Slate. *Available at:* `http://www.slate.com/articles/health_and_science/human_genome/2013/09/human_genom`

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F. and Furlong, L. I. (2017), 'Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants', *Nucleic acids research* **45**(D1), D833–D839.

Pollack, A. (June 14, 2010), 'Awaiting the genome payoff', The New York Times. *Available at:* `http://www.nytimes.com/2010/06/15/business/15genome.html?pagewanted=all&mcubz=`

Risch, N., Spiker, D., Lotspeich, L., Nouri, N., Hinds, D., Hallmayer, J., Kalaydjieva, L., McCague, P., Dimiceli, S., Pitts, T. et al. (1999), ‘A genomic screen of autism: evidence for a multilocus etiology’, *The American Journal of Human Genetics* **65**(2), 493–507.

Rosenberg, N. (1974), ‘Science, invention and economic growth’, *The Economic Journal* **84**(333), 90–108.

Ruff, C. T., Morrow, D. A., Jarolim, P., Ren, F., Contant, C. F., Kaur, A., Curtis, S. P., Laine, L., Cannon, C. P. and Brune, K. (2011), ‘Evaluation of nt-probnp and high sensitivity c-reactive protein for predicting cardiovascular risk in patients with arthritis taking longterm nonsteroidal antiinflammatory drugs’, *The Journal of rheumatology* **38**(6), 1071–1078.

Silverman, E. K. and Loscalzo, J. (2012), ‘Network medicine approaches to the genetics of complex diseases’, *Discovery medicine* **14**(75), 143.

Stephan, P. E. (1996), ‘The economics of science’, *Journal of Economic literature* **34**(3), 1199–1235.

Sveikauskas, L. (1981), ‘Technological inputs and multifactor productivity growth’, *The Review of Economics and Statistics* pp. 275–282.

Toole, A. A. (2012), ‘The impact of public basic research on industrial innovation: Evidence from the pharmaceutical industry’, *Research Policy* **41**(1), 1–12.

Visscher, P. M., Brown, M. A., McCarthy, M. I. and Yang, J. (2012), ‘Five years of gwas discovery’, *The American Journal of Human Genetics* **90**(1), 7–24.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. and Yang, J. (2017), ‘10 years of gwas discovery: biology, function, and translation’, *The American Journal of Human Genetics* **101**(1), 5–22.

Wachi, S., Yoneda, K. and Wu, R. (2005), ‘Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues’, *Bioinformatics* **21**(23), 4205–4208.

Wade, N. (June 12, 2010), ‘A decade later, genetic map yields few new cures’, The New York Times. *Available at:* `http://www.nytimes.com/2010/06/13/health/research/13genome.html?pagewanted=all&mc`

Ward, M. R. and Dranove, D. (1995), ‘The vertical chain of research and development in the pharmaceutical industry’, *Economic Inquiry* **33**(1), 70–87.

Weiner, D. J., Wigdor, E. M., Ripke, S., Walters, R. K., Kosmicki, J. A., Grove, J., Samocha, K. E., Goldstein, J. I., Okbay, A., Bybjerg-Grauholm, J. et al. (2017), ‘Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders’, *Nature genetics* .

Zheng, G., Marchini, J., Geller, N. L. et al. (2009), ‘Introduction to the special issue: Genome-wide association studies’, *Statistical Science* **24**(4), 387–387.

Zucker, L., Darby, M. and Brewer, M. (1998), 'Intellectual human capital and the birth of us biotechnology enterprises', *American Economic Review* **88**, 290–306.

Zucker, L. G. and Darby, M. R. (1996), 'Star scientists and institutional transformation: Patterns of invention and innovation in the formation of the biotechnology industry', *Proceedings of the National Academy of Sciences* **93**(23), 12709–12716.

Figures and Tables

Figure 1: Number of new therapies (by therapeutic area) observed entering the development process in 2003-2012.

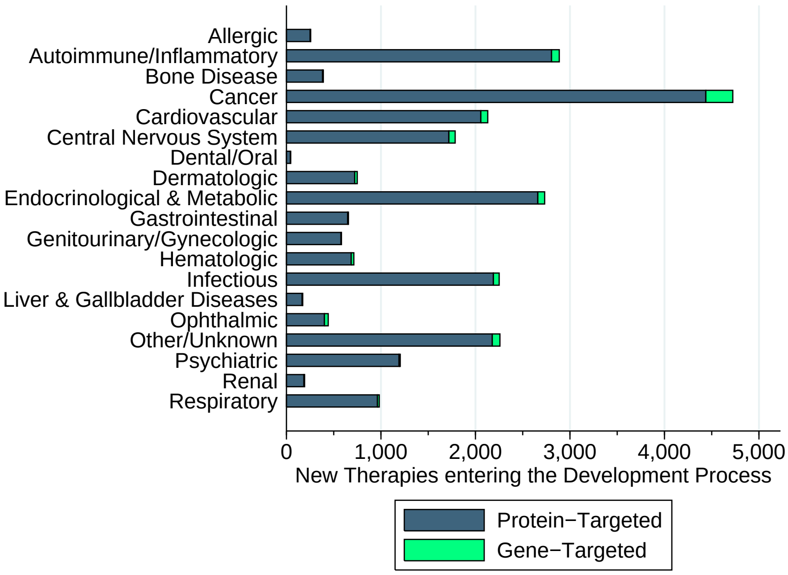


Figure 2: Temporal patterns of therapeutic innovation and GWAS VDA publication.

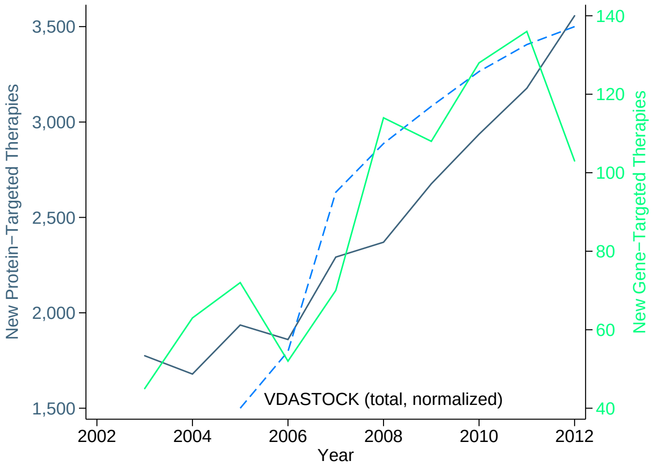


Figure 3: Number of Targeted Diseases and GWAS GDAs by Therapeutic Area.

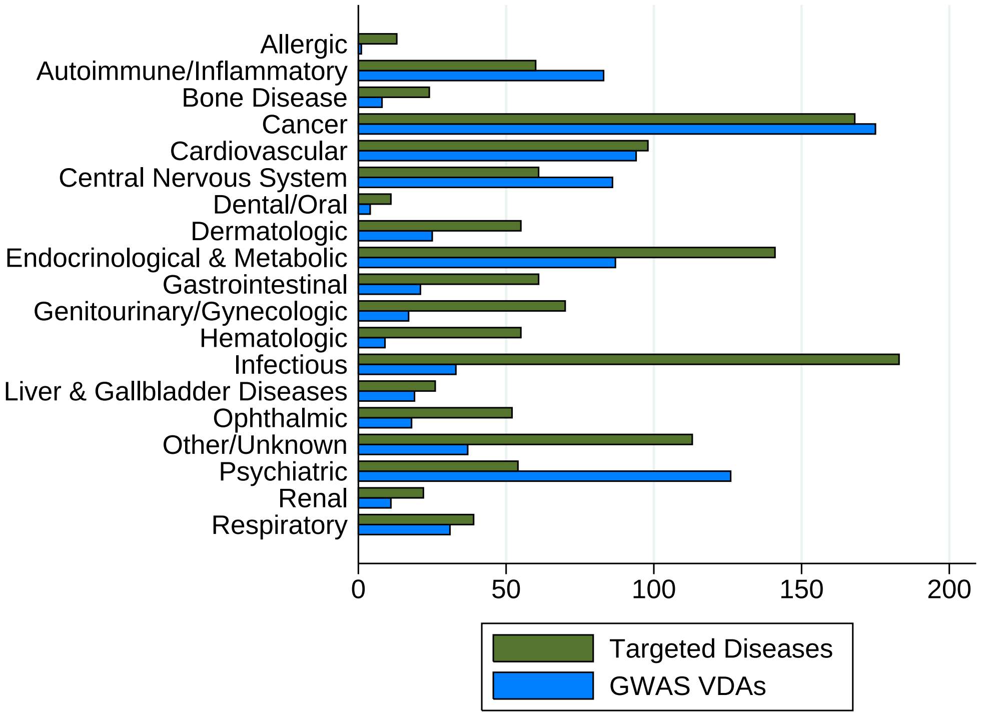
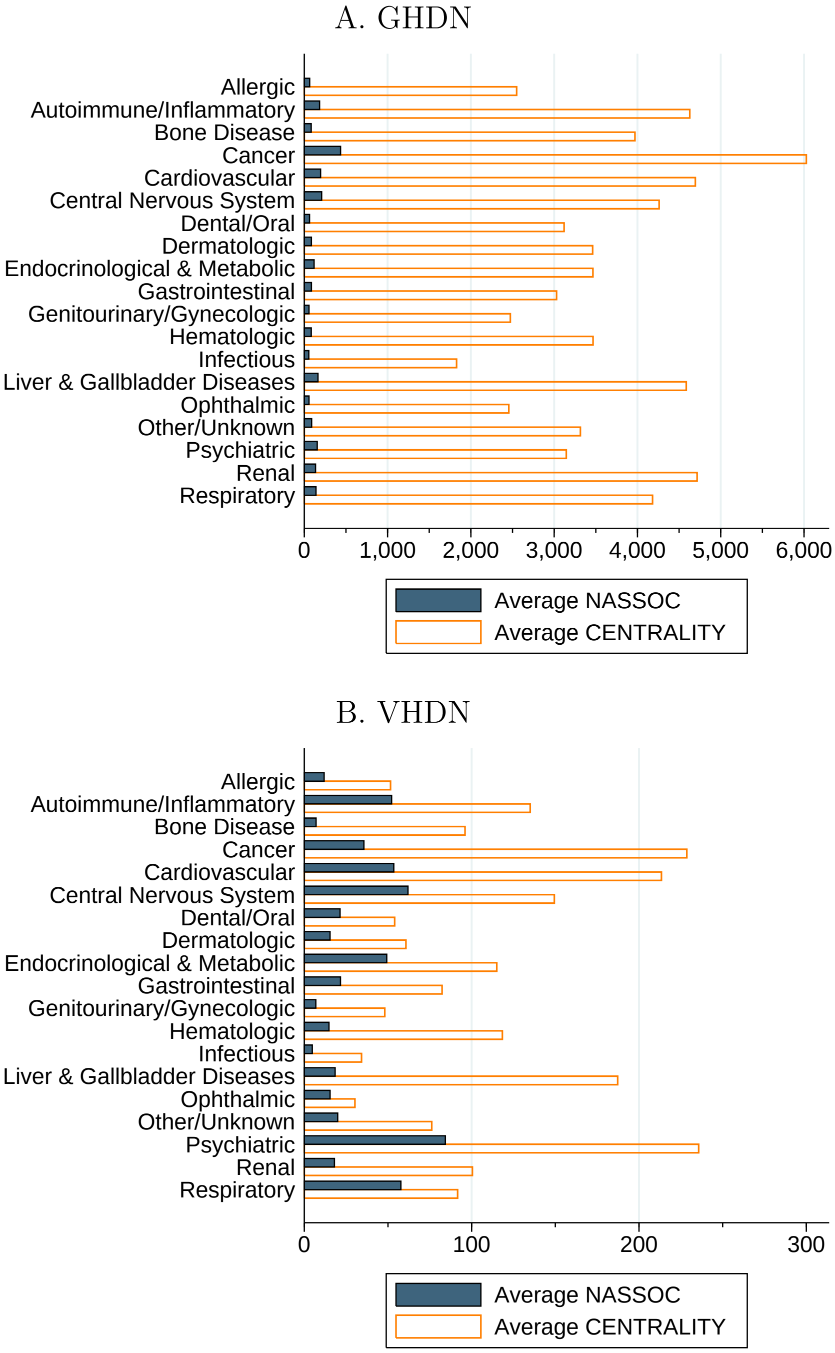


Table 1: DisGeNET GDA and VDA summary statistics.

Source	Gene-Disease Associations (GDAs)			Variant-Disease Associations (VDAs)		
	Genes	Diseases	Associations	Variants	Diseases	Associations
			A. Curated			
CTD human	7,787	4,929	25,975			
CLINVAR				45,546	5,639	54,888
GWASCAT				15,790	610	20,719
HPO	2,661	6,702	97,547			
ORPHANET	3,195	3,056	5,842			
PSYGENET	1,546	112	3,757			
UNIPROT	2,481	3,259	3,517	16,546	3,044	17,205
			B. Animal Models			
CTD mouse	63	107	168			
CTD rat	22	13	31			
MGD	1,464	1,323	1,994			
RGD	1,076	629	4,291			
			C. Literature			
GAD	8,173	2,689	56,821	5,145	410	6,242
LHGDN	5,941	1,799	31,468			
BEFREE	14,916	11,964	401,674	20,476	4,310	51,900
			Total			
	17,074	20,370	561,119	83,002	9,169	135,588

Reproduced with permission from the DisGeNet website. Retrieved 6/12/2017.

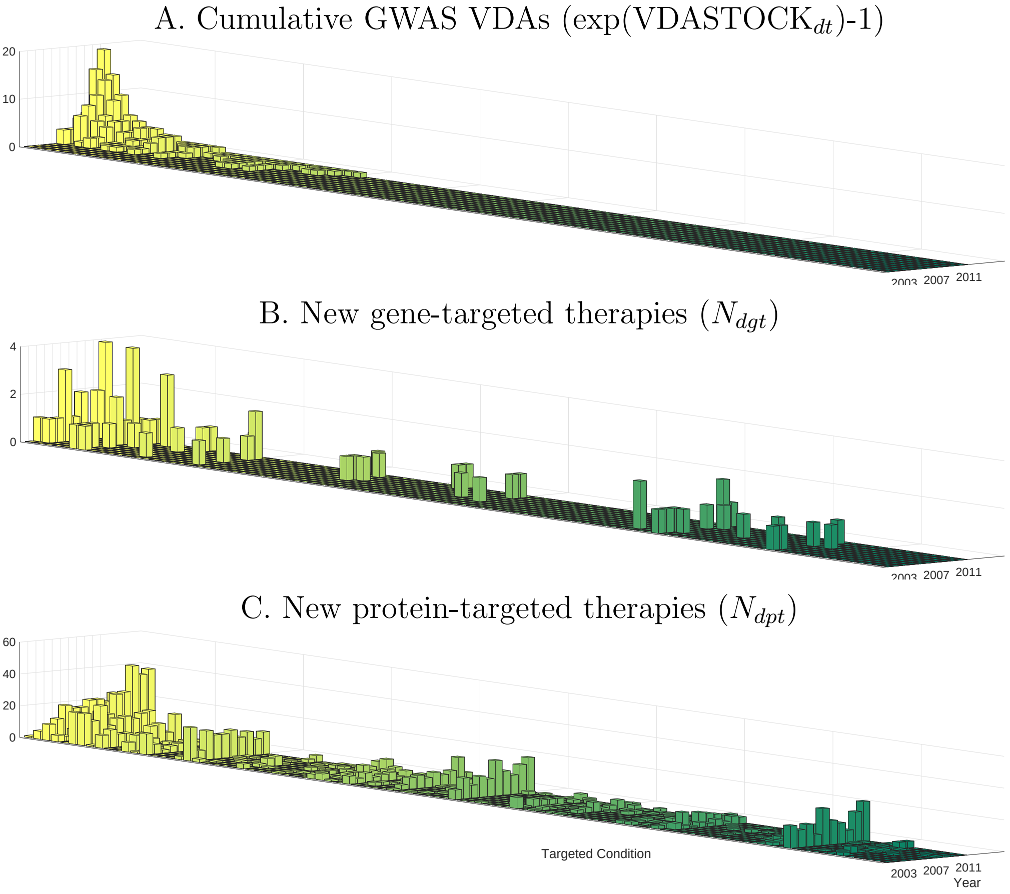
Figure 4: Average network statistics for diseases in the pipelines sample.



A. GHDN

B. VHDN

Figure 5: GWAS VDAs flows and innovation of cardiovascular therapies. Each disease is represented by a different color shade along the longer axis, whereas the temporal dimension unfolds along the depth of the graph. Diseases are arranged according to the total number of published VDAs in the sample period. Conditional on an equal number of published VDAs, diseases are ordered alphabetically. The same ordering of diseases is employed across all three panels.



A. Cumulative GWAS VDAs (exp(VDASTOCK$_{dt}$)-1)

B. New gene-targeted therapies (N_{dgt})

C. New protein-targeted therapies (N_{dpt})

Table 2: Distributions of Network Statistics.

	Percentile								
	1	5	10	25	50	75	90	95	99
	Panel A. GHDN								
NASSOC	0	0	0	3	32	131	390	750	1,774
CENTRALITY	0	0	0	720	2,982	5,740	8,540	10,420	12,827
	Panel B. VHDN								
NASSOC	0	0	0	0	1	15	70	145	495
CENTRALITY	0	0	0	0	4	145	412	617	1,004

Figure 6: Influence of GWAS Research on Computed Network Statistics

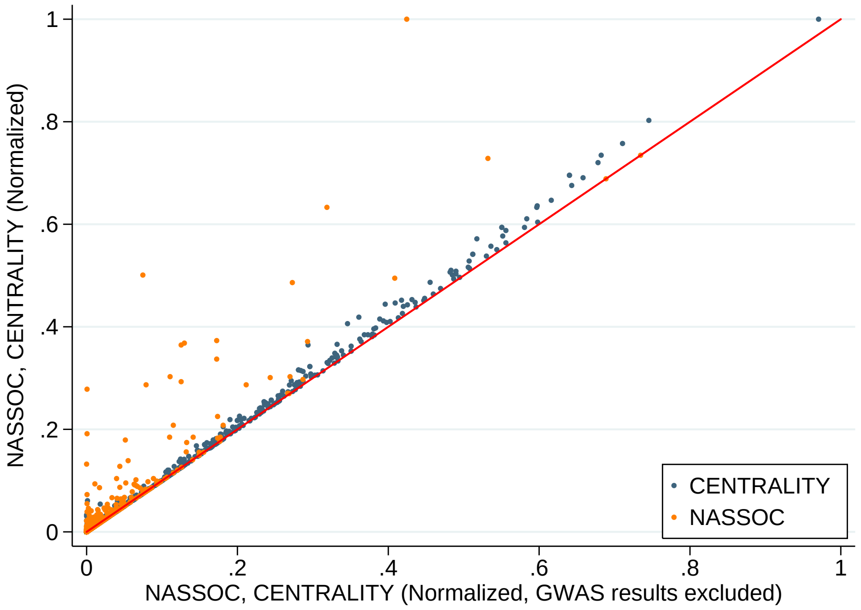
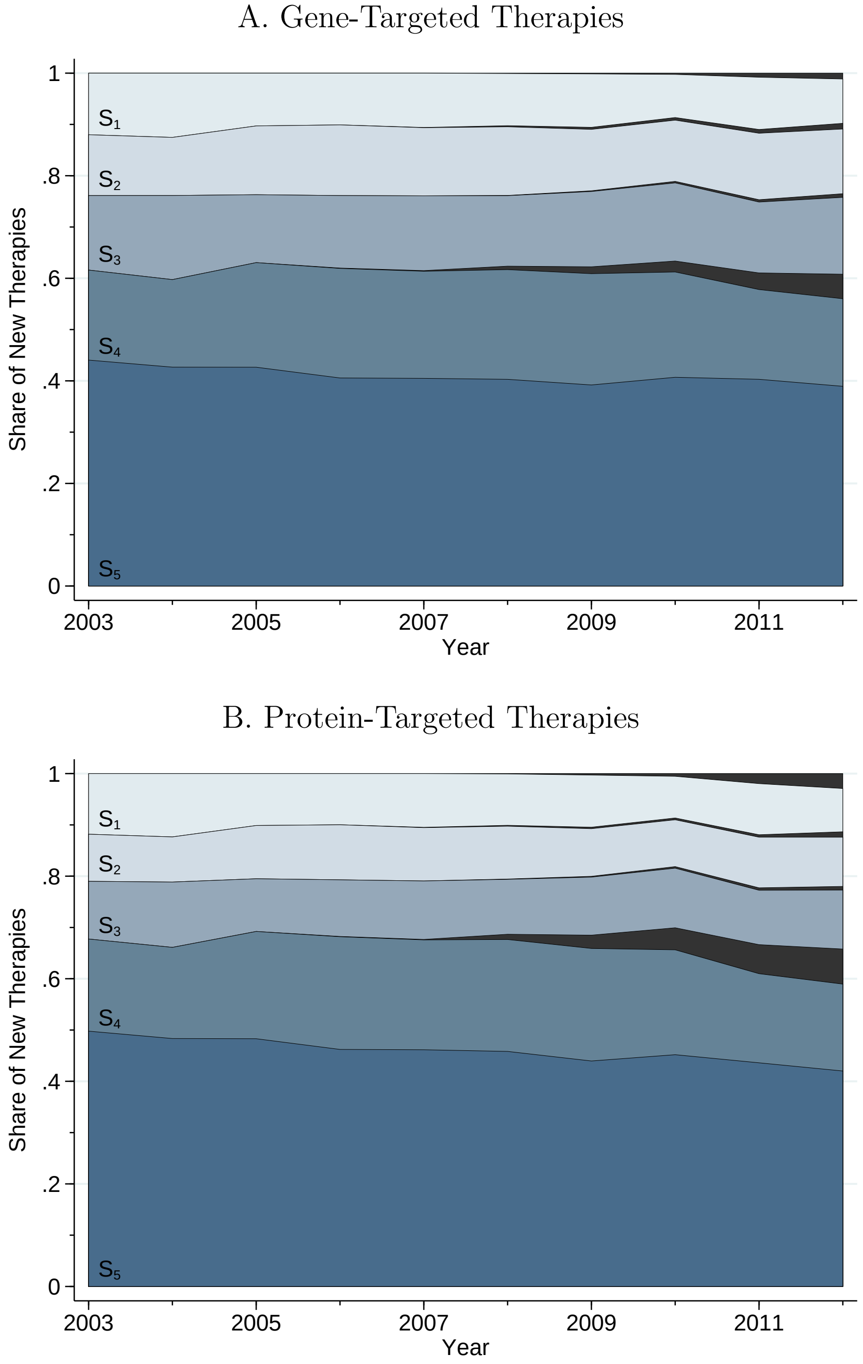


Table 3: Drivers of Therapeutic Translation.

	(1)	(2)	(3)	(4)	(5)	(6)
GENETARGET$_k$	-2.91***	-1.53***	-2.64***	-2.21***	-2.30***	-1.88***
	(0.23)	(0.25)	(0.42)	(0.21)	(0.39)	(0.27)
VDASTOCK$_{d,t-1}$	1.43***	0.71***	0.98***	0.85***	0.80***	0.71***
	(0.17)	(0.10)	(0.13)	(0.11)	(0.11)	(0.12)
GENETARGET$_k$×VDASTOCK$_{d,t-1}$	-0.39*	0.11	0.03	0.15	0.07	0.09
	(0.22)	(0.17)	(0.12)	(0.15)	(0.12)	(0.17)
CENTRALITY$_d$			0.01***		0.16***	
			(0.00)		(0.02)	
GENETARGET$_k$×CENTRALITY$_d$			0.01***		0.10***	
			(0.00)		(0.03)	
CENTRALITY$_d$×VDASTOCK$_{d,t-1}$			-0.01***		-0.09***	
			(0.00)		(0.01)	
NASSOC$_d$				0.10***		0.23***
				(0.02)		(0.05)
GENETARGET$_k$×NASSOC$_d$				0.04		0.10*
				(0.04)		(0.05)
NASSOC$_d$×VDASTOCK$_{d,t-1}$				-0.06***		-0.11***
				(0.01)		(0.02)
MEPSPATS$_{d,t-1}$	0.03***	0.01	0.00	0.00	0.00	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
MEPSEXPND$_{d,t-1}$	0.04***	0.02***	0.02***	0.02***	0.01***	0.02***
	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Network implementation	N/A	N/A	GHDN	GHDN	VHDN	VHDN
Pre-sample Estimator		✓	✓	✓	✓	✓
Observations	26,120	26,120	26,120	26,120	26,120	26,120

Note: Results from Negative Binomial, zero-inflated specifications. All specifications include fixed effects for therapeutic areas and years. Clustered standard errors are presented in parentheses. Legend: $^*p < 0.1,^{**} p < 0.05,^{***} p < 0.01$.

Figure 7: Innovation attributable to GWAS science. Black areas correspond to the share of new therapies associated with GWAS' VDAs.



A. Gene-Targeted Therapies

B. Protein-Targeted Therapies

Table 4: Drivers of Therapeutic Translation (Diseases in the Cancer area are excluded from the sample).

	(1)	(2)	(3)	(4)
GENETARGET$_k$	-2.28***	-1.99***	-1.88***	-1.66***
	(0.47)	(0.29)	(0.36)	(0.30)
VDASTOCK$_{d,t-1}$	0.95***	0.82***	0.86***	0.80***
	(0.15)	(0.12)	(0.10)	(0.10)
GENETARGET$_k$×VDASTOCK$_{d,t-1}$	-0.02	0.05	0.03	0.07
	(0.16)	(0.16)	(0.16)	(0.20)
CENTRALITY$_d$	0.01***		0.17***	
	(0.00)		(0.03)	
GENETARGET$_k$×CENTRALITY$_d$	0.01***		0.10**	
	(0.00)		(0.05)	
CENTRALITY$_d$×VDASTOCK$_{d,t-1}$	-0.01***		-0.10***	
	(0.00)		(0.01)	
NASSOC$_d$		0.15***		0.28***
		(0.02)		(0.07)
GENETARGET$_k$×NASSOC$_d$		0.08**		0.17*
		(0.04)		(0.09)
NASSOC$_d$×VDASTOCK$_{d,t-1}$		-0.07***		-0.15***
		(0.01)		(0.04)
MEPSPATS$_{d,t-1}$	0.01	0.01	0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
MEPSEXPND$_{d,t-1}$	0.01**	0.01*	0.01*	0.02***
	(0.01)	(0.01)	(0.01)	(0.01)
Network implementation	GHDN	GHDN	VHDN	VHDN
Pre-sample Estimator	✓	✓	✓	✓
Observations	22,760	22,760	22,760	22,760

Note: Results from Negative Binomial, zero-inflated specifications. All specifications include fixed effects for therapeutic areas and years. Clustered standard errors are presented in parentheses. Legend: $^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

Table 5: Complexity Clusters.

Subsample	Fraction of sample diseases	Average of GHDN CENTRALITY	GHDN NASSOC	VHDN CENTRALITY	VHDN NASSOC
$\mathcal{S}_1$	0.35	397.8	4.5	6.2	2.8
		(506.7)	(7.8)	(30.8)	(16.2)
$\mathcal{S}_2$	0.24	2,693.8	41.1	42.9	7.8
		(635.0)	(43.7)	(81.0)	(18.7)
$\mathcal{S}_3$	0.21	5,050.0	114.8	117.3	20.5
		(710.7)	(123.6)	(165.1)	(53.1)
$\mathcal{S}_4$	0.14	7,892.2	312.7	256.2	58.9
		(892.1)	(267.3)	(210.4)	(108.3)
$\mathcal{S}_5$	0.06	11,485.0	1,144.1	714.0	248.0
		(1262.0)	(708.2)	(285.3)	(309.9)

Note: Subsamples created through a k-means clustering procedure on all GHDN and VHDN network statistics. Within-subsample standard deviations are presented in parentheses.

Table 6: Therapeutical Translation Across Disease Clusters of Varying Complexity.

	Subsample $\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_4$	$\mathcal{S}_5$
VDASTOCK$_{d,t-1}$	1.12***	0.74***	0.32***	0.52***	0.06
	(0.17)	(0.17)	(0.09)	(0.15)	(0.08)
GENETARGET$_k$	-1.99***	-0.96	-0.78	-1.60***	-1.40***
	(0.76)	(0.58)	(0.53)	(0.30)	(0.08)
MEPSPATS$_{d,t-1}$	0.01	0.02*	-0.02	0.05***	-0.01
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
MEPSEXPND$_{d,t-1}$	0.03***	0.02*	0.01	0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Pre-sample Estimator	✓	✓	✓	✓	✓
Observations	9,140	6,220	5,420	3,680	1,660

Note: Results from Negative Binomial, zero-inflated specifications for the dependent variable N_{dkt}. All specifications include fixed effects for therapeutic areas and years. Clustered standard errors are presented in parentheses. Legend: $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$.

Table 7: Marginal Effects.

Year	Subsample				
	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_4$	$\mathcal{S}_5$
2004	1.17	0.65	0.27	0.43	0.04
2012	1.14	0.62	0.26	0.35	0.02

Note: Marginal effects are computed by increasing in one the number of available GWAS publications for each disease, and then computing the implied percentage difference in the number of new therapies (averaged across diseases within each cluster).

Table 8: Drivers of Therapeutic Translation.

	(1)	(2)	(3)	(4)
GENETARGET$_k$	-2.27***	-1.93***	-1.92***	-1.76***
	(0.40)	(0.27)	(0.34)	(0.26)
VDASTOCK$_{d,t-1}$	0.77***	0.71***	0.71***	0.70***
	(0.11)	(0.09)	(0.09)	(0.10)
GENETARGET$_k$×VDASTOCK$_{d,t-1}$	0.02	0.08	0.12	0.07
	(0.13)	(0.15)	(0.14)	(0.15)
CENTRALITY$_d$	0.09***		0.71***	
	(0.01)		(0.18)	
GENETARGET$_k$×CENTRALITY$_d$	0.06***		0.37*	
	(0.02)		(0.21)	
CENTRALITY$_d$×VDASTOCK$_{d,t-1}$	-0.04***		-0.36***	
	(0.01)		(0.06)	
NASSOC$_d$		1.06***		1.59***
		(0.25)		(0.31)
GENETARGET$_k$×NASSOC$_d$		0.43*		1.28***
		(0.26)		(0.49)
NASSOC$_d$×VDASTOCK$_{d,t-1}$		-0.51***		-0.76***
		(0.10)		(0.22)
MEPSPATS$_{d,t-1}$	0.00	0.00	0.00	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
MEPSEXPND$_{d,t-1}$	0.02***	0.02***	0.02***	0.02***
	(0.00)	(0.00)	(0.00)	(0.00)
Network implementation	GHDN	GHDN	VHDN	VHDN
Pre-sample Estimator	✓	✓	✓	✓
Observations	26,120	26,120	26,120	26,120

Note: Results from Negative Binomial, zero-inflated specifications. All specifications include fixed effects for therapeutic areas and years. Clustered standard errors are presented in parentheses. Legend: $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$. Network statistics computed with DisGeNET research published 2005 or earlier.

Table 9: Assessing the Influence of Unobservables.

GWAS articles used to construct VDASTOCK	(1)	(2)	(3)	(4)	(5)
			Subsample		
	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_4$	$\mathcal{S}_5$
			Panel A. Number of 2-Year Citations		
Below-median 2-year citations	0.80***	1.02**	0.16	0.12	0.04
	(0.28)	(0.48)	(0.18)	(0.22)	(0.08)
Above-median 2-year citations	1.13***	0.45***	0.36*	0.57**	0.04
	(0.22)	(0.11)	(0.21)	(0.28)	(0.06)
			Panel B. Funding Source		
No industry funding	1.11***	0.57***	0.47***	0.54***	-0.03
	(0.19)	(0.10)	(0.12)	(0.17)	(0.08)
Some industry funding	0.61	1.53**	-0.42	0.14	0.17
	(0.76)	(0.78)	(0.35)	(0.32)	(0.11)
Pre-sample Estimator	✓	✓	✓	✓	✓
Observations	9,140	6,220	5,420	3,680	1,660

Note: Results from Negative Binomial, zero-inflated specifications for the dependent variable N_{dkt}. All specifications include fixed effects for therapeutic areas and years, an indicator for gene-targeted therapies, and MEPS variables that proxy for epidemiological pervasiveness and market size. Clustered standard errors are presented in parentheses. Legend: $^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.