Technological Hype and Innovation: Evidence from the Human Genome

Manuel Hermosilla* and Jorge Lemus†

April 30, 2016

PROPOSAL PREPARED FOR THE NBER PROJECT:
"Economic Dimensions of Personalized and Precision Medicine"

Abstract

The completion of Human Genome Project in 2003 enabled the systematic study of gene-disease associations. Exploiting a rich dataset, and focusing on the first generation of studies on gene-disease associations, we evaluate the extent, pace, and outcomes of the translation of new scientific knowledge into drug discovery efforts. Our preliminary results support a "hype" hypothesis. Overly optimistic expectations about the potential of genetic-based methods for the identification of novel drug targets fueled initial fast-paced translation. The extent of translation gradually slowed down and ultimately vanished as lackluster clinical trial performance began to be observed. Some of the steps required to advance this project include: (i) to refine the causality framework by developing a finer linkage between studies and discovery attempts and instruments for market potential; (ii) to generate an external metric for technological hype from media articles and leveraging tools of sentiment analysis; (iii) to strengthen our analysis of clinical trial performance; and (iv) to assemble and incorporate into the analysis a data set of patent applications in the US.

*The Johns Hopkins Carey Business School. Email: mh@jhu.edu
†University of Illinois, Urbana-Champaign. Email: jalemus@illinois.edu

"If you read them now, the claims made for genomics in the 1990s sound a bit like predictions made in the 1950s for flying cars and anti-gravity device." Jack Scannell, 2014.[1]

1 Introduction

The beginning of the 21st century brought a big promise for the future of medicine. It was announced on the 26th of June of 2000, when President Bill Clinton welcomed Francis Collins and Craig Venter into the East Room of the White House to present the completion of first draft of the Human Genome Project (HGP). Francis Collins had lead a decade-long publicly-funded effort to sequence all the genes in human DNA, coordinating the effort of scientists from the US, the UK, Japan, Germany, France, and China. Backed by powerful new gene-sequencing technology, Craig Venter initiated in 1999 a parallel private initiative. These two teams eventually turned into cooperation completing the HGP ahead of time and under budget. The resulting compendium—a characterization of the most basic elements responsible for variation in human trait expression—has been considered a major achievement for science, compared by some to the moon-landing or Copernicus' heliocentric theory. Recognizing the magnitude of the contribution, Venter stated that day 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."[2] After 13 years of sequencing and a $3 billion investment, the final draft was unveiled in 2003.

Many scientists immediately predicted a revolution in the prevention, diagnosis, and treatment of diseases. For example, Randy Scott of Incyte Genomics stated that "in 10 years, we will understand the molecular basis for most human diseases." Lander et al. (2001) concluded that "The scientific work will have profound long-term consequences for medicine, leading to the elucidation of the underlying molecular mechanisms of disease and thereby facilitating the design in many cases of rational diagnostics and therapeutics targeted at those mechanisms." The genome map would allow scientists to link genetic mutations to the manifestation of diseases, and once these were identified, targeted therapies could be designed to replace genes, manipulate their expression, counteract protein imbalances, among others.

But in 2016, there is still not a full understanding of the biology of most human diseases—particularly for the most burdensome ones—and therapies providing fully effective cures are few and far between. The leading causes of death remain the similar to those in the 2000's and new therapies have not substantially increased life expectancy.[3] Voices of skepticism have gradually started to rise, with some people claiming that the project has not delivered as promised and that the industry should move past genetically targeted approaches.

In this project we develop a systematic evaluation of the translation of basic genetic

[1] "Why Are So Few Blockbuster Drugs Invented Today?"
[2] http://transcripts.cnn.com/TRANSCRIPTS/0006/26/bn.01.html
[3] "Where Are All The Miracle Drugs?"

research into the discovery of new therapies for human disease. We focus on a first generation of gene-disease research, the genome-wide association studies (GWAS), the first of which was published in 2005, shortly after the completion of the HGP. Our empirical strategy exploits rich disease-level temporal variation in the stocks and flows of new such studies to explain patterns of drug discovery. Our results suggest that, during the 2006-2010 period, the industry was quickly adopted new genetic insights, significantly increasing discovery efforts for therapies targeting the diseases whose genetic roots began to be clarified. In particular, our estimates suggest that during this period, each new study was associated with between 0.2 and 0.45 new therapies entering the discovery stage. However, clinical trial results did not support the alleged therapeutic superiority of genetically-based targets, gradually eroding the level of enthusiasm, and with it, the pace of translation. Indeed, after 2010, our results show that the new arrival of genetic insights did not translate into additional drug discovery attempts.

Overall, our estimates suggest an industry "hype". That is, the pharmaceutical industry reacted promptly to the initial belief about the superior therapeutic potential of genetically targeted medicines. However, the payoff of these investments was not as high as expected and over time the industry reacted slowly to novel academic publications.

Abundant anecdotal evidence supports an industry reaction driven by technological "hype" fueled by the media and the scientific enthusiasm that followed the completion of the project. For instance, Derek Lowe, former Bayer employee, claimed: "The whole industry went crazy with it. Bayer committed half a billion dollars into human genome research, and they got nothing for it. Nothing at all" .[4] The director of National Human Genome Research Institute, Eric Green, addressed this sentiment directly in 2011: "I feel bad if our enthusiasm and euphoria over completing the genome project was misinterpreted to mean that we would have cures 10 years later."[5] Others haven gone further, overtly faulting the media for inflating expectations beyond justification: "Journalists wrote about how we were going to have drugs for all these diseases in the next decade. Somebody was smoking something. This was just nuts."[6] Further evidence of this "hype" is discussed in Alpert and Chen (2012) and in Evans et al. (2011).

In hindsight, the failure to meet the initial expectations seems to have been caused from under-estimating the complexity of the problem. We now know that genetic variants are rarely mapped one-to-one into the most burdensome diseases (e.g., cancer, Alzheimer's, diabetes). Thus, even when targets are identified (i.e., a genetic variation has a statistically significant association with a disease), it is hard to develop a therapy that safely and effectively "hits" the target.[7] Furthermore, the genome has revealed new layers of biological complexity, delaying advances in applied science as basic science gaps are being

[4] "Why Are So Few Blockbuster Drugs Invented Today?"

[5] The Future of the Human Genome.

[6] http://blogs.nature.com/boston/2011/02/23/nnb-report-ten-years-later-harvard-assesses-the-genome-map

[7] This problem does not exist for "Mendelian" diseases (e.g., Huntington's Disease,Cystic Fibrosis), characterized by a one-to-one mapping between gene mutations and diseases expression, so reliable genetic tests and powerful predictive tools are available in those cases. Regrettably, most high-burden (and profitable) diseases are caused by the mutation and interaction of more than one single gene (as well as the influence of environmental factors).

filled. This point was eloquently made by the prominent UK researcher Roger Highfield, with an analogy between the genome and a book: if the HGP 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."[8]

The slow progress of translating science into successful drugs was anticipated by the more rigorous analysts and it does not negate a scientific revolution. For example, in 2000, Francis Collins emphasized that progress would primarily unfold towards the end of a four-decade period.[9] Even in the relatively short span of time since the completion of the HGP, genetics have enabled the new field of personalized medicine, and delivered advances in basic biology that offer strong support to Venter's prognosis of a better understanding of "our place in the biological continuum."[10] Effective genetically-based therapies have reached the market, including treatments for Chronic Myelogenous Leukemia and early-onset breast cancer diagnostic tests.

Accordingly, and due to the relatively short span of time since the completion of the HGP, we view our study as a "Progress Report." Our results do not inform an evaluation of the long-term impacts HDP on the quality and abundance of human therapeutics. However, they do suggest that that the pharmaceutical industry may energetically reacted to discontinuous advancements in the study of gene-disease associations. Such progress may indeed be currently unfolding, as a second-generation of gene-disease association studies ("Next Generation Studies") start to gain traction at the hand of declining computing power costs (Nordhaus et al., 2007).

Our results shed light on how misleading expectation could have an enormous welfare consequence. The initial hype could have distorted the direction of innovation, having firms gradually moving away from a "trial and error" to a "target-based" method to discover new drugs. Our findings can be rationalized by the observation that competitive pressures give incentives to firms to invest too soon, relative to the efficient timing. It may have been more efficient to allow for the basic science to develop further before investing in the novel approach.

The rest of this proposal is organized as follows. In the next section, we review the relevant literature. In Section 3, we provide background information on gene-disease studies and describe the publications database. In Section 4, we describe our database of pharmaceutical trials. In Section 5, we show our preliminary results and in Section 6 we interpret them to shed light on the discussion of the genome-bubble. Finally, in Section 7, we summarize and discuss our preliminary results and we offer further steps for our research.

[8] "Life just got a lot more complicated."

[9] "Reading the Book of Life: The Doctor's World; Genomic Chief Has High Hopes, and Great Fears, for Genetic Testing."

[10] For example, the genome has lead to the surprising discovery that the number of genes is not related in an obvious way to the complexity of the organism. Humans have only about 21,000 genes, while a fruit fly –a much simpler organism– has around 17,000. A mouse has around 23,000.

2 Related Literature

An idea that dates back at least to Nelson (1959) and Arrow (1962) is that the creation of knowledge will be under-provided by markets due to its public-good nature. Even more, profit maximizing firms will under-provide the creation of or knowledge that does not have an immediate application (basic science). Bryan and Lemus (2016) explain that the public good nature of knowledge in combination with the difficulty of making new discoveries will not only change the amount of investment, but also the direction of research. Even more, Furman et al. (2012) shows that different patterns of funding can have significant impact on the choice of research of scientists.

Knowledge needs to be created, but also preserved and diffused. Institutions play a major role in all of these aspects. For instance, the human genome project would have not been possible without the financial support and oversee of several institutions. Furman and Stern (2011) present an empirical setting to estimate the role of institutions in the creation, maintenance, certification and accumulation of knowledge. Pisano (2006) discusses whether private companies, specifically biotech companies, can profit from the creation of basic knowledge.

Our project closely relates to the literature that explores how private and publicly-funded research is captured by firms in the industry. Cockburn and Henderson (1998), focusing on the biotech industry, highlights the role played by the network of scientific connections of a firm on the absorption of basic research. We focus on the events that followed the completion of the Human Genome Project, specifically on genome-wide association studies. Our project does not focus on patenting—for example, Jensen and Murray (2005) studies the gene patenting and Hawkins (2011) studies the impact of patents on the development of genetic tests—but rather on how academic publications is translated into the developing of new therapies. In a future version, we will link GWAS patents and papers as in Murray (2002).

We focus on GWAS since it is arguably the most innovative method to identify drug targets. Drews (2000) examines the history of drug discovery and explains that drugs used to be developed using chemistry, but over time biology and genetics became the main tools to guide the discovery of new drugs. In fact, the completion of the human genome project was followed by a significant economic impact, which is analyzed in Tripp and Grueber (2011). New techniques have allowed scientists to identified mutations for more than 2,900 protein-coding genes in humans (Chong et al., 2015; Brinkman et al., 2006). Visscher et al. (2012) evaluates the impact of GWAS from a broader perspective and concludes that it has had a positive impact on new discoveries about genes and new biological insights. Sanseau et al. (2012) discusses the use of GWAS for drug repositioning, and Nelson et al. (2015) studies the use of GWAS as evidence to obtain FDA approval for new therapies. However, despite all the progress and new technology, Scannell et al. (2012) have pointed out that pharmaceutical R&D efficiency—measured as the number of new drugs brought to market by the global biotechnology and pharmaceutical industries per billion US dollars of R&D spending—has steadily declined. Agarwal et al. (2013) explores the competition in the development of novel drug targets, finding that about half of them are pursued by just one company.

Finally, in our database we distinguish whether research is privately funded, since conflict of interest in reporting has been an issue in this industry. An academic institution offers different incentives to a scientists compared to an industry job. Scientists self-select into one or the other given their individual preferences, which can be heterogeneous in their taste for having control and freedom of research (Aghion et al., 2008) or in their ability and their preference for non-pecuniary returns (Agarwal and Ohyama, 2013). In terms of publications, Fugh-Berman (2013) points out the pervasive involvement of the pharmaceutical industry in basic research and emphasizes the consequences of under-reporting of negative results and misrepresenting reported results. Similarly, Bennett et al. (2010) studies the association between pharmaceutical industry funding and basic science research for a specific therapy (erythropoiesis-stimulating agents). The authors find that when the pharmaceutical industry supports an study, researchers are more less likely to find potentially harmful effects of the therapy.

3 Identifying the Genetic Roots of Human Disease

3.1 Scientific Approaches

There are three different approaches to study the relation between genes and diseases: Linkage analysis (LA), Genome-wide association studies (GWAS), and next-generation sequencing (NGS).[11]

The LA and GWAS identify genomic *regions* in the chromosomes associated with diseases, although in some cases it has been possible to pin-point the specific genetic variant causing the disease. NGS allows for higher resolution, identifying specific mutations.

Linkage analysis is a technique that has been successfully applied, even before the HGP, 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 to identify other Mendelian disorders. However, LA studies rely on related individuals and they do not provide "high resolution" (meaning, they identify broad regions of variations) and they also have limited statistical power. In particular, linkage analysis is limited to study complex disorders where not too many related individual carry the disease.

The completion of the HGP allowed scientists to use an alternative approach to linkage analysis. A genome-wide association study (GWAS) compares DNA regions of a population that carries certain trait (i.e. weight, aggressive personality, diabetes, acne, etc.) with the genome of a control population without the trait. GWAS operate under the assumption that the same disease are driven by the same (possible multiple) genetic variations. Although the effectiveness of GWAS has been criticized, Visscher et al. (2012) highlights its importance: "...the GWAS experimental design in human populations has led to new discoveries about genes and pathways involved in common diseases and other

[11] Londin et al. (2013) for a more detailed explanation.

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." GWAS has served as a starting point to understand the biology behind diseases.[12]

Next-generation sequencing is the most recent method to identify genomic variants and consists on sequencing the entire genome (or specifically targeted regions). Some recent findings have shown that NGS are superior to LA, but NGS struggle to determine the causes of complex diseases given the large cost of sequencing the complete genome for large samples.[13]

Our analysis only consider the GWAS database, since it contains many studies that have significantly impacted the way scientists think about the mechanisms behind certain diseases.

3.2 GWAS database

We collected all the GWAS publications from the open-source GWAS Catalog,[14] made available by Burdett et al. (2016). The first successful GWAS was published in 2005. Figure 1 shows the number of GWAS publications over time. Each paper in the GWAS database will study typically only one trait.[15] For each one of the trait studied in a paper, there are several genes associated. Mendelian diseases (or single-gene diseases) are characterized by a single gene. However, GWAS publications show that each trait is generally associated to multiple genes, although less than 5% of these studies identify more than 30 genes per trait. Figure 2 presents a histogram of the number of genetic variations found by GWAS papers for each trait.

Figure 2: Number of genes identified in each trait studied in the GWAS papers.



[12] http://blog.goldenhelix.com/grudy/have-we-wasted-7-years-and-100-million-dollars-on-gwas-studies/

[13] NGS has been prohibitively costly in the past, but recent advances in technology are allowing for large studies to relying on NGS. Some recent advances include the combination of GWAS and NGS.

[14] https://www.ebi.ac.uk/gwas/

[15] About 83% of the papers study a single trait; 9% of the papers study 2 traits and only 8% of the papers study 3 or more traits

Figure 1: Number of papers on GWAS by month.

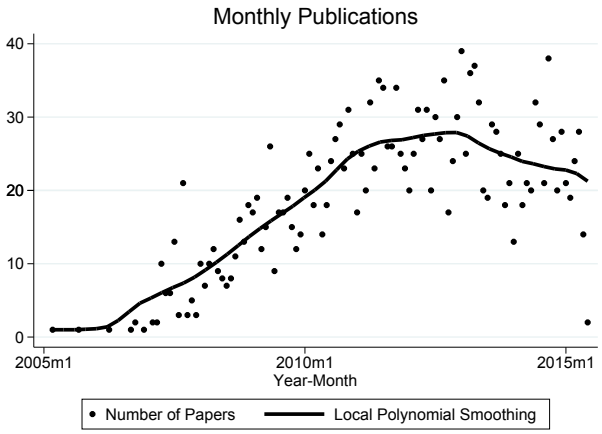
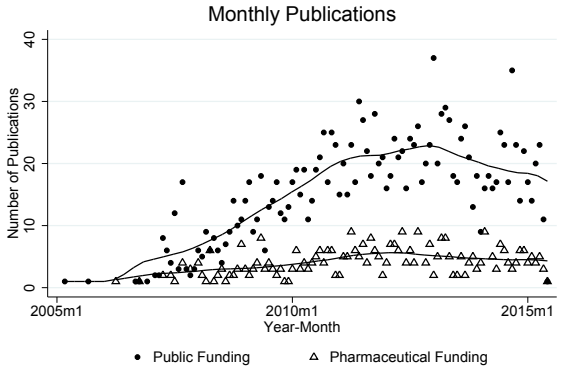


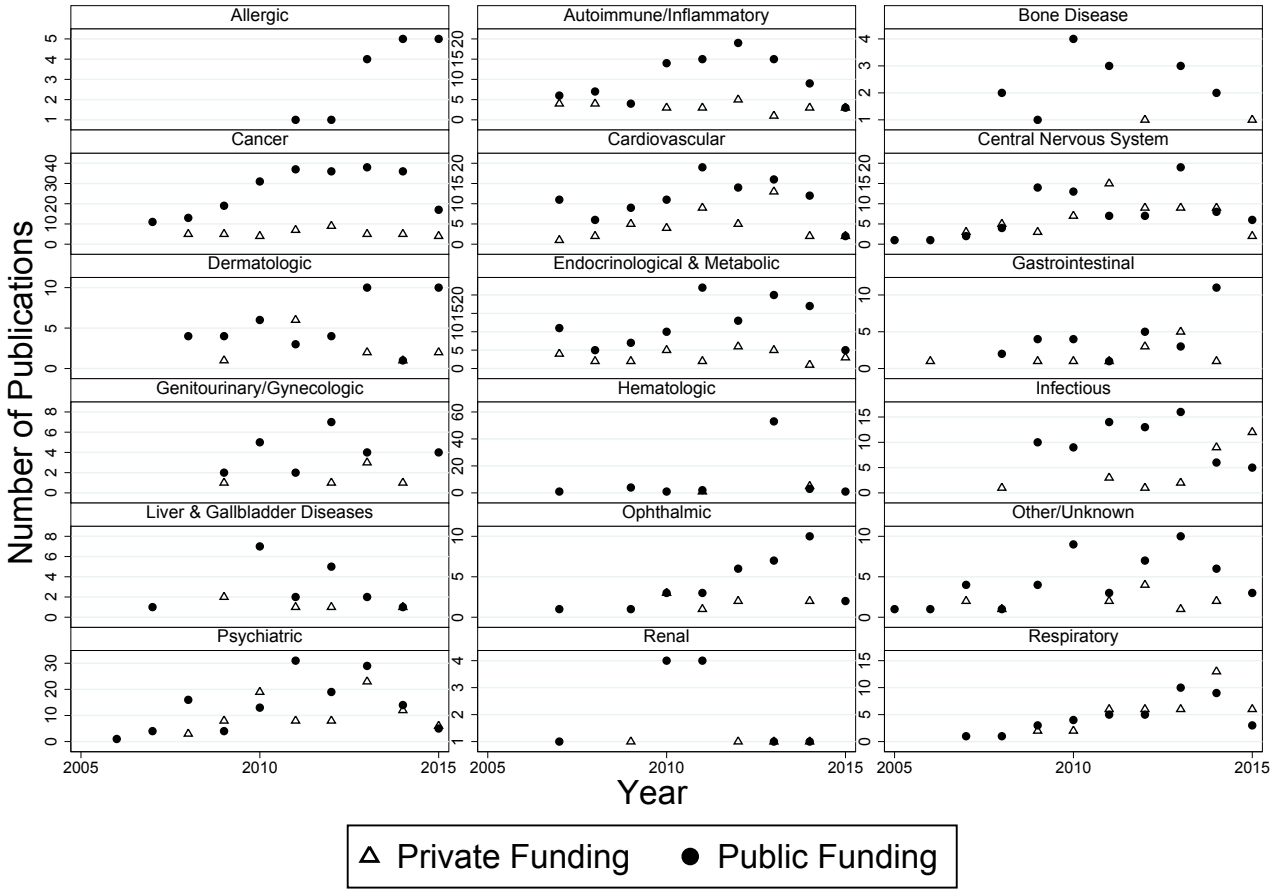
Figure 3: GWAS papers in which Pharmaceutical companies are sources of funding or conflict of interest.



To study the involvement of pharmaceutical companies in the development of basic science, we performed a textual analysis of the papers in GWAS. We manually processed each paper in the database to determine the sources of funding. In some journals, the sources of funding are explicitly reported. However, in the large majority of the papers, the funding was found in the acknowledgment section. Apart from looking for funding, we collected information (when reported) about potential conflicts of interest of some of the authors. We matched a set of 604 large pharmaceutical companies with the GWAS database to determine whether these companies funded research or were mentioned as a potential conflict of interest. We find that 80.53% of the papers do not declare private involvement, while the remaining 19.47% explicitly declare one of the pharmaceutical companies as a source of funding or as a conflict of interest for one or more authors. Figure 3 shows the number of publications over time by source of funding.[16]

[16]We refer as 'private funding' when a paper was either explicitly funded by a pharmaceutical company or when we found a declaration of conflict of interest.

Figure 4: Publication rates by funding source



Apart from the volume of publications funded by pharmaceutical companies, we explore whether studies funded by pharmaceutical companies are concentrated in some specific therapeutic areas. Figure 4 shows a decomposition of monthly publications by the sources of funding by therapeutic area. It is interesting to note the lack of private involvement in cancer and autoimmune/inflammatory studies. This observation is in-line with the recent finding in Budish et al. (2015).

4 Pharmaceutical Pipelines Data

4.1 Source

We obtained data on pharmaceutical pipelines from Thomson Reuters' Life Sciences Cortellis. This company collects information on the pipelines of a large number of biotechnology and pharmaceutical from printed/online media, academic conferences and publications, etc. The full data sample includes development histories of over 90,000 therapies

(i.e., compound/targeted-disease combinations) that have entered the discovery process from the mid 1970's around the world. The development of therapies is tracked since they are first discussed in an academic conference or publication, or reported by a clinical trial submitted to www.clinicaltrials.gov, or featured by the media or regulatory updates, or announced in the sponsoring firm's website, or when they are coded as the firm as entering the "discovery" stage.

We restrict the sample to new therapies developed in the United States entering the discovery stage during or after the year 2000 (about 55% of the full sample). As we explain in the following sections, we restrict the sample even further to avoid mismatched errors and we end up with a set of 16,518 therapies entering the discovery stage between 2000 and 2014.

4.2 The Drug Development Process

Drug development may take over a decade from the initial formulation until commercialization, and the process entails a set of sequential stages (Pisano, 2006).[17] Drug development begins at the discovery stage. If the results at this stage are promising, the process continues with clinical trials on human subjects (phases I to III). At any of these stages, testing results may suggest lack of efficacy or safety, leading to the termination of the development. Although we observe the progress of a therapy throughout stages (as well as the stage at which the process is terminated/discontinued), many dates are missing in our dataset. We only keep therapies for which observe the date of entry into the discovery stage.

The discovery stage is itself comprised of several sub-stages and may take many years to complete. The first step in discovery is to identify of "druggable targets," i.e. chemical or molecular entities that will elicit a biological response in a controllable way (Hughes et al., 2011).

The druggable targets identified by the firms create candidate molecules that are stored in vast libraries and kept under zealous secrecy.[18] Firms will draw from this pool of candidate molecules and continue with the following steps: target validation, high throughput screening, validation, hit identification. Once all these steps are completed, the drug candidate is ready to be tested on human subjects in clinical trials (Hughes et al., 2011). The information from Cortellis does not allow us to identify the specific discovery sub-stage therapies are at when they first appear in the data.

[17] A detailed description of the process is beyond the scope of this study, but available from this reference.

[18] For example, Merck is known for maintaining a library containing 4 million such entities.

4.3 Linking Cortellis Therapies to GWA Studies

As of March 2016, the set of studies in GWAS related to a total of 1,123 genetic-based human traits. We kept the set of these corresponding to human diseases, which left us with 666 traits.[19] The matching procedure described below resulted in 934 gene-disease matches, with 744 GWAS traits matched to 313 targeted diseases by Cortellis' therapies (out of the 1,912 in the full Cortellis sample), but we dropped 10 of them because their associated GWAS studies were published in 2015. Thus, our analysis sample contains 303 different targeted diseases. Given the total of 16,518 therapies in the sample, each condition is associated to an average of 54.5 therapies across years. However, as we show in the next section, there is significant heterogeneity across therapeutic areas.

Linking these to the conditions targeted by the new therapies in the Cortellis dataset presented two main challenges. The first and obvious one 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 one stems from imperfect matches. In particular, many of the traits studied by GWAS studies can be thought to inform the understanding of the biology of diseases targeted by Cortellis' therapies in an indirect but strongly related way. For example, GWAS' studies for the "longevity" trait could inform the design of compounds targeting "aging." Similarly, the GWAS trait "5-htt brain serotonin transporter levels" is thought to underlie a variety of neuropsychiatric disorders, hence could inform the design of compounds targeting "post-natal depression."

To identify these matches in the data we assembled a team of experts. A first expert (Ph.D. in Epidemiology) helped us to create a list of match morphologies. We then asked two independent coders (M.D. residents) to identify as many matches as possible from the data, classifying each of these according to the morphology. The first expert then curated these independent codings, dropping unjustified matches, including missing ones, and resolving conflicts. Descriptive statistics of the resulting data are shown in table 1. The total number of unique entries for traits and conditions is smaller than the sum of unique entries by match type due to overlap across categories.

4.4 Rates of Discovery

Figure 5 presents the distribution of therapies by therapeutic area. Cancer is the by far the most prominent of all, with about 5,500 therapies. Infectious, autoimmune/inflammatory, endocrinological & metabolic, central nervous system, cardiovascular, and psychiatric constitute a fringe of followers, each with between 1,100 and 1,800 therapies in total.[20]

[19] Traits that are not associated human disease include "economic and political preferences," "educational attainment," "freckles," "hand grip strength," "reading or mathematical ability," "red/white wine liking," "self-employment," among others. To identify human diseases, We use Merriam-Webster's definition: "an illness that affects a person, animal, or plant: a condition that prevents the body or mind from working normally," keeping the human disease meaning only.

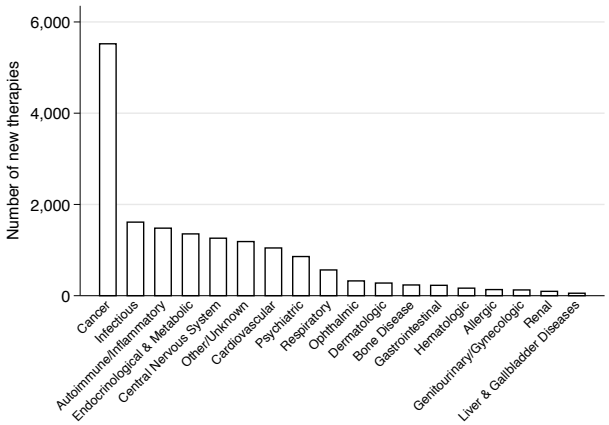
[20] The distribution is roughly consistent with industry aggregates of pipeline composition reported by Giovanetti and Jaggi (2013).

Table 1: Descriptive statistics of the matching procedure

Type	Number of unique entries			Examples	
	Matches	GWAS traits	Cortellis cond.	GWAS trait	Cortellis condition
1. Perfect /Synonym	178	170	165	autism	autism
				ileal carcinoids	carcinoid tumor
				arterial stiffness	arteriosclerosis
2. Dual	101	54	63	obesity and osteoporosis	obesity
				obesity and osteoporosis	osteoporosis
				ulcerative colitis or crohn's disease	ulcerative colitis
				ulcerative colitis or crohn's disease	chron's disease
3. Epidemiological	147	127	106	prion diseases	creutzfeldt jakob disease
				enteric fever	salmonella typhi infection
				hdl cholesterol	Hypercholesterolemia
4. Treatment	128	77	49	hepatitis b vaccine response	hepatitis b virus infection
				response to antidepressants	depression
				response to clopidogrel therapy	blood clotting disorder
5. Subclass	258	245	96	age-related hearing impairment	hearing loss
				alzheimer's disease (late onset)	alzheimer's disease
				allergic rhinitis in non-asthmatics	allergic rhinitis
6. Biomarker/Phenotype/Trait	17	16	14	hippocampal atrophy	alzheimer's disease
				antibody level in response to infection	infectious disease
				kidney function decline traits	Renal disease
				osteoporosis-related phenotypes	osteoporosis
				periodontal disease-related phenotypes	periodontal disease
Total unique entries	829	678	313		

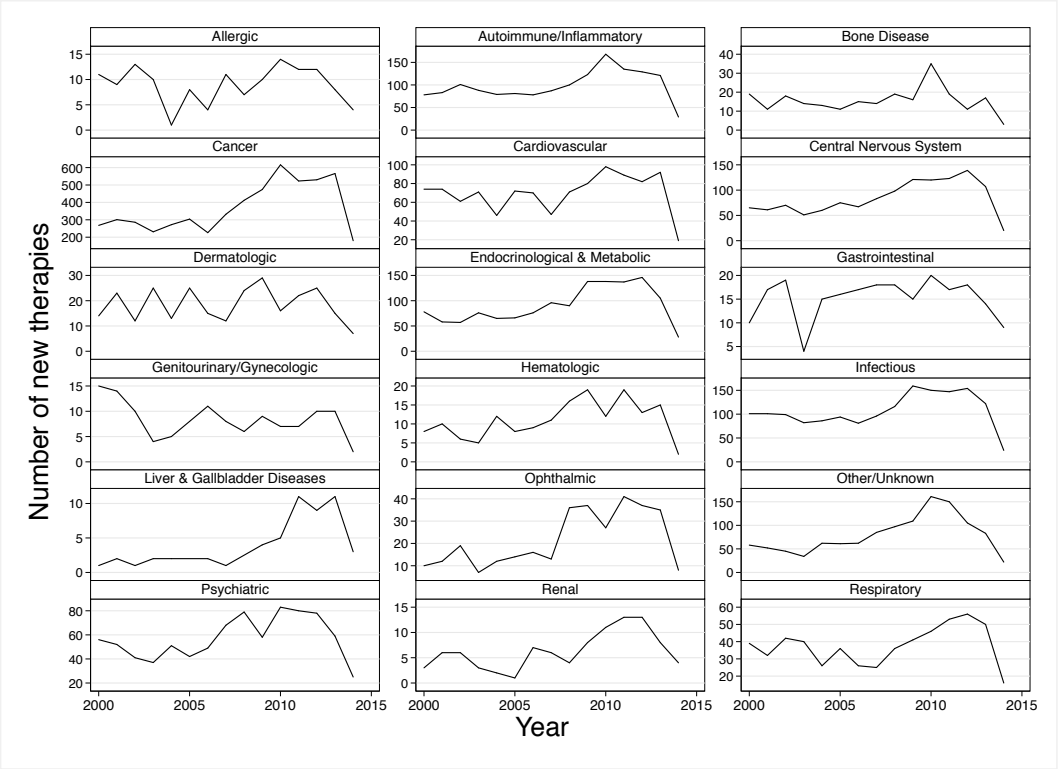
Figure 6 shows the discovery patterns across years, by therapeutic area. Many therapeutic areas exhibit a marked increase in the rate of discovery towards the second half of the sample period, consistent with the boost in expenditure for prescription drug for diseases prevalent among the elderly, mainly fueled by the enactment of Medicare Part D (Blume-Kohout and Sood, 2013; Dranove et al., 2014). All categories exhibit a decline in the number of therapies entering discovery in 2014, which can be justified on reporting lags.[21]

Figure 5: Distribution of new therapies in the sample across therapeutic areas



[21] Our dataset was downloaded from the Cortellis website during the first months of 2015.

Figure 6: Distribution of new therapies in the sample across years



4.5 Clinical Trial Performance (Survival)

Figure 7 describes the share of therapies entering the discovery stage each year. In our sample, between 20% and 25% of the therapies that entered discovery in the 2000-2007 period are reach phase I. Due to both attrition and right truncation, reaching phases II are III ocurr less frequently, although the percentage of therapies is relatively flat over time. Starting with the 2008 cohort of therapies entering discovery, the percentages observed reaching each clinical trial stage steadily decrease.

Somewhat surprisingly, some of the therapies first observed in discovery after 2012 are observed entering clinical trials in the sample. Although we have not yet thoroughly investigated this feature of the data, we believe that it may in part stem from re-purposing compounds or compounds "taken off the shelf." These are compounds whose safety and basic efficacy has been previously tested, so they can skip preliminary stages of discovery.

To assess developmental performance we again consider the probability that a therapy will reach each successive clinical trial stage, but drop all those observations that represent ongoing stages. We are thus left with a dataset that allows us to directly plot the probability of survival. Figure 8 presents the results by therapeutic area. An average of 26% of compounds that enter the discovery survive until phase I; 16% until phase II, and

5% until phase III.[22]

Figure 7: Clinical trial phase reach by discovery cohort

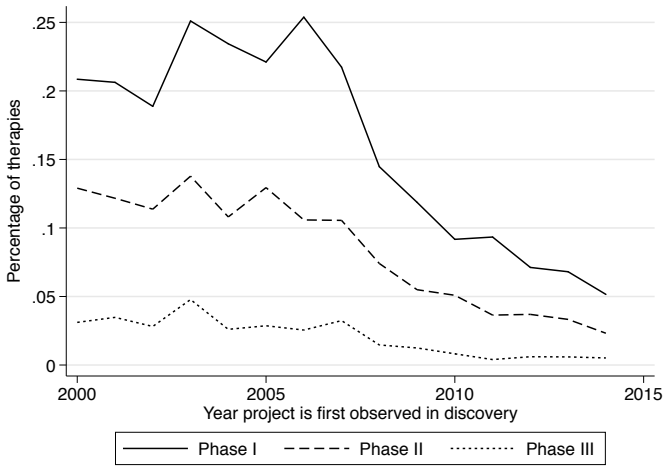
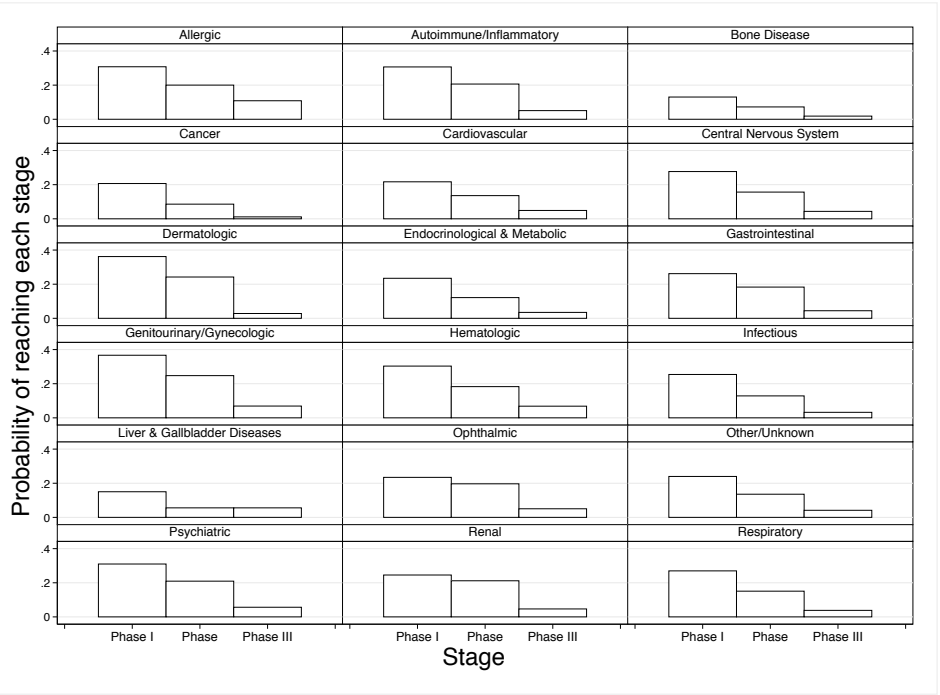


Figure 8: Survival



[22]These estimates are roughly consistent with those reported by imply DiMasi et al. (2003). In particular, our estimates imply that, out of all therapies reaching phase I, 62% reach phase II and 31% reach phase III. The figures reported by DiMasi et al. (2003) are 71% and 31%, respectively.

5 Preliminary Results

5.1 Intensity of Discovery

We present results suggesting a swift translation of basic science knowledge into the discovery of new therapies. The dependent variable in our analysis is the reported number of new therapies for disease i entering the discovery stage in year t. In our model, two independent variables measure the *stock* and the *recent* knowledge about the biology of disease i at year t. In particular, to measure the stock of knowledge on disease i at year t, we define the variable $\text{PUBSTOCK}_{i,t-3}$ as the number of studies on disease i published up to year $t-3$. To measure the amount of knowledge recently created, we define the variable $\text{NEWPUBS}_{i,t-2:t-1}$ as the total number of studies on disease i published during years $t-1$ and $t-2$.

We estimate a negative binomial[23] model with the following specification:

$$\begin{aligned} \mathbb{E}[n_{it}] &= f(\alpha \cdot \text{NEWPUBS}_{i,t-2:t-1} + \beta \cdot \text{PUBSTOCK}_{i,t-3} \\ &\quad + \gamma \cdot \text{REVENUES}_{i,t-1} + \phi \cdot \text{MISSINGREVENUES}_{i,t-1} \\ &\quad + \delta_t + \eta_{a(i)}), \end{aligned} \tag{1}$$

where δ_t is a time fixed effect and $\eta_{a(i)}$ is a therapeutic area fixed effect. To control for variation in market potential across targeted conditions and time, we include the variable $\text{REVENUES}_{i,t-1}$. We compute this variable from patterns of prescription drug expenditures, reported by the Medical Expenditure Panel Survey[24] at the 3-digit ICD9 level. Because these expenditures are missing for many of the ICD9 categories of the targeted diseases in our sample, we also include the indicator $\text{MISSINGREVENUES}_{i,t-1}$, which equals one if revenue data is not available for a particular ICD9/year cell (in which case REVENUES is set to zero), and zero otherwise. We cluster standard errors by therapeutic area.

Table 2 presents the results of several variants of this specification. Estimates of the first column suggest that new scientific knowledge quickly permeates into drug discovery. Within two years of publication, each new study linking disease i with its genetic roots is associated with about 0.18 new therapies for that disease entering the discovery stage. The stock of "old" knowledge is not significantly related to the number of new therapies in discovery at a given year. A possible interpretation for the short life-cycle of new scientific knowledge would be that the knowledge embedded on each publication is specific, in that it produces a narrow set of therapeutic opportunities which are quickly exploited by the industry.

This basic result remains robust across the specifications of Columns 2-5, all of which

[23] The poisson specification was rejected by over-dispersion tests in all cases.

[24] http://meps.ahrq.gov/mepsweb/

include market potential controls. Throughout the models, the variable that measures revenues has a significantly positive coefficient, suggesting that one additional billion dollars of total prescription drug's expenditures on year $t-1$ are associated with between 0.06 and 0.09 new therapies in discovery on year t, which is systematically smaller than the effect of NEWPUBS.[25] We rationalize this results by the high attrition rates in the industry (DiMasi et al., 2003), which suggest that marginal reductions in the probability of developmental failure may be more valuable in expected terms than marginal increases in marginal potential. The negative coefficient of MISSINGREVENUES is expected to the extent that missing revenues in the Medical Expenditure Panel Survey data is likely to reflect low overall drug consumption expenditure.[26]

Results of Column 3 are obtained from a zero-inflated specification, since about 50% of the dependent variable's observations are zero. In Column 4, we exclude all cancer therapies and in Column 5, by excluding cancer therapies and restricting the sample period to 2008-2014. The coefficient on NEWPUBS remains significant and increases when cancer is included. Interestingly, PUBSTOCK acquires statistical significance, suggesting that in therapeutic areas other than cancer, developers may not be so quick to rip all low hanging fruits associated to new scientific knowledge as they are in cancer. These findings can be interpreted in light of the relatively larger potential payoffs for cancer therapies that reach the market, which may have lead to a more crowded discovery space, as well as the development of a richer pre-genome scientific knowledge base. Both of these effects suggest that the scientific findings of new genome-wide association publications have a lower marginal value in this therapeutic area.

We have so far been reluctant to attach causal interpretation to the effects of NEWPUBS. Indeed, such interpretation would not be warranted if scientific publications were better conceptualized as an output (as opposed to input) of the drug discovery process. We investigate this possibility in the next section.

5.2 Research Funding and Knowledge Flows

We begin by evaluating differences between the effects or publicly and privately funded studies. We do so by distinguishing NEWPUBS and PUBSTOCK according to funding source. Table 3 presents the results of a model including all these as explanatory variables.

The flow of both types of publications is associated with non-negligible, statistically significant positive effects on the rate of discovery. The significance of the effect of publicly funded publications is reassuring, as the direction of this type of research tends to be more sensitive to public health concerns than to market conditions (Ward and Dranove, 1995).

[25] All drug expenditure's figures are CPI-adjusted to year 2000.

[26] We acknowledge that prescription drug expenditure is not a fully adequate to proxy for the market potential of many of the therapies in our data, since these may be designed for administration in in-patient contexts. In these cases, expenditures would not be picked up by our data. We are currently evaluating alternative data sources and methodologies. For now, we are somewhat reassured of our results noting that the effects are robust to the exclusion of cancer therapies, where in-patient drug administration is most prominent.

Table 2: The number of new therapies in discovery as a function of scientific publications

	(1)	(2)	(3)	(4)	(5)
NEWPUBS	0.177***	0.151***	0.149***	0.177***	0.190***
	(0.032)	(0.032)	(0.032)	(0.036)	(0.035)
PUBSTOCK	0.073	0.068	0.062	0.083*	0.129***
	(0.054)	(0.050)	(0.045)	(0.050)	(0.043)
REVENUES		0.092***	0.091***	0.060**	0.066**
		(0.034)	(0.031)	(0.029)	(0.033)
MISSINGREVENUES		-0.803***	-0.797***	-0.568**	-0.423*
		(0.254)	(0.263)	(0.253)	(0.255)
Sample	All	All	All	Cancer excluded	Cancer excluded 2008 and after
Zero-inflation	No	No	Yes	No	No
N	4,545	4,545	4,545	2,121	1,799

All models include therapeutic area and year fixed effects. Robust standard errors (clustered by therapeutic area) are presented in parentheses. Legend $^{*}p$ <0.10, $^{**}p$ <0.05, $^{***}p$ <0.01

In addition, because our specifications of Table 2 include market potential controls, this finding suggests that at least part of the overall association between publication flows has a causal interpretation. In the appendix we show that this result remains robust when we weight the importance of each paper according to its citations the reported p-values of each association.

Not surprisingly, privately funded publications are associated with a stronger effect than publicly funded ones. One interpretation of this result is that firms are more likely to fund research that relates to the types of targets of compounds stored in their vast libraries. In these cases, the scientific validation of a target reduces technological uncertainty and leads firms to quickly "jump the gun," that is, quickly initiate optimization and preclinical testing for the compounds associated to that target. Even if privately funded publications represented an output (as opposed to input) of new drug discovery, it would be difficult to rationalize this association without granting the production of new scientific knowledge an instrumental role in the discovery of new drugs.

We explore this reverse causation hypothesis by reverting the specification of equation (1). In particular, we regress the number of new publications for disease i in year t based on the flows and stocks of new therapies entering the discovery stage, defined analogously to NEWPUBS and PUBSTOCK. We label these variables NEWCOMPS and COMPSTOCK, respectively. Results are presented in Table 4. The dependent variable in Columns 1 and 2 is the number of publicly funded publications, whereas that in Columns 3 and 4, the number of privately funded publications. Because publications are observed starting in 2005, we restrict the sample period to 2005-2014.

Table 3: Publicly and privately funded scientific publications and the intensity of discovery

NEWPUBS (Public funding)	0.129***
	(0.034)
PUBSTOCK (Public funding)	0.083
	(0.073)
NEWPUBS (Private funding)	0.212***
	(0.048)
PUBSTOCK (Private funding)	0.021
	(0.079)
REVENUES	0.092***
	(0.033)
MISSINGREVENUES	-0.798***
	(0.256)
N	4,545

Includes therapeutic area and year fixed effects. Robust standard errors (clustered by therapeutic area) are presented in parentheses. Legend $^{*}p$ <0.10, $^{**}p$ <0.05, $^{***}p$ <0.01

Table 4: The number of scientific publications as a function of new therapy discovery

	(1)	(2)	(3)	(4)
NEWCOMPS	-0.016**	0.006	0.000	0.007
	(0.007)	(0.013)	(0.012)	(0.015)
COMPSTOCK	0.008**		0.002	
	(0.004)		(0.005)	
REVENUES	0.065***	0.069***	0.090***	0.093***
	(0.015)	(0.019)	(0.009)	(0.008)
MISSINGREVENUES	-0.486***	-0.525***	-0.672**	-0.690**
	(0.173)	(0.156)	(0.324)	(0.327)
Publication type	Publicly funded	Publicly funded	Privately funded	Privately funded
N	3,030	3,030	3,030	3,030

All models are estimated on the 2005-2014 sample, and include therapeutic area and year fixed effects. Robust standard errors (clustered by therapeutic area) are presented in parentheses. Legend $^{*}p$ <0.10, $^{**}p$ <0.05, $^{***}p$ <0.01

5.3 Survival

To evaluate the flow of new scientific knowledge we introduce a second time index, τ, which points to the year a stage is resolved. This indexing is important because, in

addition to depending on the scientific evaluation of testing results, survival may depend on the economic outlook at the time results are observed. Thus, we use this index to specify market controls the $\text{REVENUES}_{\tau-1}$ and $\text{MISSINGREVENUES}_{\tau-1}$.

As before, the stock and flow of scientific publications is measured at the time the therapy is observed entering discovery. A therapy's year of introduction to discovery (its cohort) is denoted as $t(i)$, so knowledge flows are proxied by $\text{NEWPUBS}_{i,t(i)-2:t(i)-1}$ and $\text{PUBSTOCK}_{i,t(i)-3}$. Therefore, this specification aims to measure the impact of new scientific knowledge on the design of new therapy cohorts.

We specify a series of dichotomic dependent variables $y_{i,s,\tau}$ representing various types of transitions. All of these variables equal one if a the transition is successful (i.e., therapy i is observed entering stage s), and zero otherwise, and explain their variation with a probit model. In every case, we drop all those observations representing development stages that are still in progress, so our estimates are clean of right truncation bias. The specification is given by:

$$\begin{aligned}\mathbb{E}[y_{i,s,\tau}] \;=\; & \text{Probit}(\alpha \cdot \text{NEWPUBS}_{i,t(i)-2:t(i)-1} + \beta \cdot \text{PUBSTOCK}_{i,t(i)-3} \\ & +\gamma \cdot \text{REVENUES}_{i,\tau-1} + \phi \cdot \text{MISSINGREVENUES}_{i,\tau-1} \\ & +\delta_{t(i)} + \eta_{a(i)})\end{aligned} \tag{2}$$

As before, the model includes therapeutic area fixed effects and clusters standard errors at the same level. Notice also that time fixed effects are attached to therapy cohorts (discovery year), so that our survival specifications will account for the fact that drugs entering the discovery on a later period are intrinsically less likely to be observed entering more advanced stages.

Table 5 presents the results for various types of transitions. In all cases, except for Column 1, the greater availability of scientific knowledge at the time a therapy is discovered does not seem to increase survival rates. Although the coefficient of PUBFLOW in Column 1 is statistically significant at the 10% confidence level, its implied marginal effect is relatively small. It suggests that each additional publication increases the probability of reaching clinical trials by about 0.007. Evaluated at the average value of NEWPUBS (2005 and after), the estimate implies that new scientific knowledge for a given disease increases the number of therapies reaching clinical trials by about 12%.

These results suggest that the effect of scientific discoveries embedded in GWA studies improved firms' expectations regarding therapeutic performance, reducing uncertainty and fostering the advancement of project into clinical trials. Those expectations, however, were not met by clinical trials' outcomes. In the next section we provide additional context to evaluate this hypothesis.

It important is to highlight that these results are preliminary and should be interpreted cautiously. Figure 7 above showed that the percentages of therapies entering discovery after 2005 is relatively small. In addition, statistics at the bottom of table 5 suggest

Table 5: Survival

	(1) Discovery to Phase I	(2) Discovery to Phase II	(3) Discovery to Phase III	(4) Phase I to Phase II	(5) Phase II to Phase III
PUBFLOW	0.024*	-0.001	0.003	0.006	-0.014
	(0.013)	(0.021)	(0.021)	(0.023)	(0.032)
PUBSTOCK	0.016	-0.008	-0.015	-0.042	0.004
	(0.014)	(0.018)	(0.028)	(0.031)	(0.087)
REVENUES	0.015***	-0.015	0.026**	-0.015	0.025
	(0.004)	(0.011)	(0.013)	(0.013)	(0.016)
MISSINGREVENUES	1.147***	-2.385***	-1.972***	0.277**	0.606***
	(0.087)	(0.046)	(0.088)	(0.122)	(0.153)
Fraction of obs. with PUBFLOW>0	0.19	0.18	0.17	0.16	0.10
Fraction of obs. with PUBSTOCK>0	0.10	0.09	0.08	0.08	0.04
N	9,230	8,731	8,391	1,726	801

Probit models using as dependent variable an indicator for successful transition. All models include therapeutic area and (discovery) year fixed effects. Robust standard errors (clustered by therapeutic area) are presented in parentheses. Legend $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$

that the extent of variation among the metrics for scientific knowledge flows is, at best, moderate, potentially translating into low-powered inference. We are currently investigating alternative regression designs and avenues for additional data collection to help us ameliorate this problem.

6 The "Hype" after the Human Genome Project

6.1 Background

There is vast evidence documenting the expectation that the completion of the human genome survey would unleash a scientific and therapeutic revolution. For example, the Nobel Laureate James D. Watson (discoverer of the DNA structure) stated in 1990 that "The ability to sequence DNA quickly and cheaply will also provide the technological basis for a new era in drug development."

Although emphasizing that some results may take several decades to be observed, the lead of the HGP, Francis Collins, predicted in 2000 great advances in the prevention, diagnosis and treatment of human disease.[27] Lander et al. (2001) concluded that "The scientific work will have profound long-term consequences for medicine, leading to the elucidation of the underlying molecular mechanisms of disease and thereby facilitating the design

[27] http://www.nytimes.com/2000/06/27/science/reading-book-life-doctor-s-world-genomic-chief-has-high-hopes-great-fears-for.html

in many cases of rational diagnostics and therapeutics targeted at those mechanisms." These expectations quickly diffused across the medical community, leading most medical schools to include genetic courses in their curricula (Marshall, 2011).

Today, there are much lower expectations for the development of new drugs and for a big revolution in medicine. Lander et al. (2001) overtly faults the media for creating unwarranted "hype" or, inflating expectations about the extent and speed of the therapeutic progress enabled by the unveiling of the genome: "Journalists wrote about how we were going to have drugs for all these diseases in the next decade. Somebody was smoking something. This was just nuts."[28]

Alpert and Chen (2012) and Evans et al. (2011) directly evaluate the extent to which these expectations have materialized. The studies concur on the idea that expectations were unrealistic, and a major impact on drug development has not yet been materialized. The authors point out that the field of genetics today is a much more basic than an applied discipline, and that solid although sluggish progress is being made.

The main difficulty translating genetics into new drugs stems from the fact that genetic variants are rarely mapped one-to-one into the most burdensome diseases (e.g., cancer, Alzheimer's, diabetes). Even when targets are identified—that is, a genetic variation has a statistically significant association with a disease—it is hard to develop a compound that safely and effectively "hits" the target. This problem does not exist when there is a one-to-one mapping between gene mutations and diseases,[29] so reliable genetic tests and powerful predictive tools are available in those cases. Regrettably, most high-burden (and profitable) diseases can be caused by the mutation and interaction of more than one single gene (as well as the influence of environmental factors.)

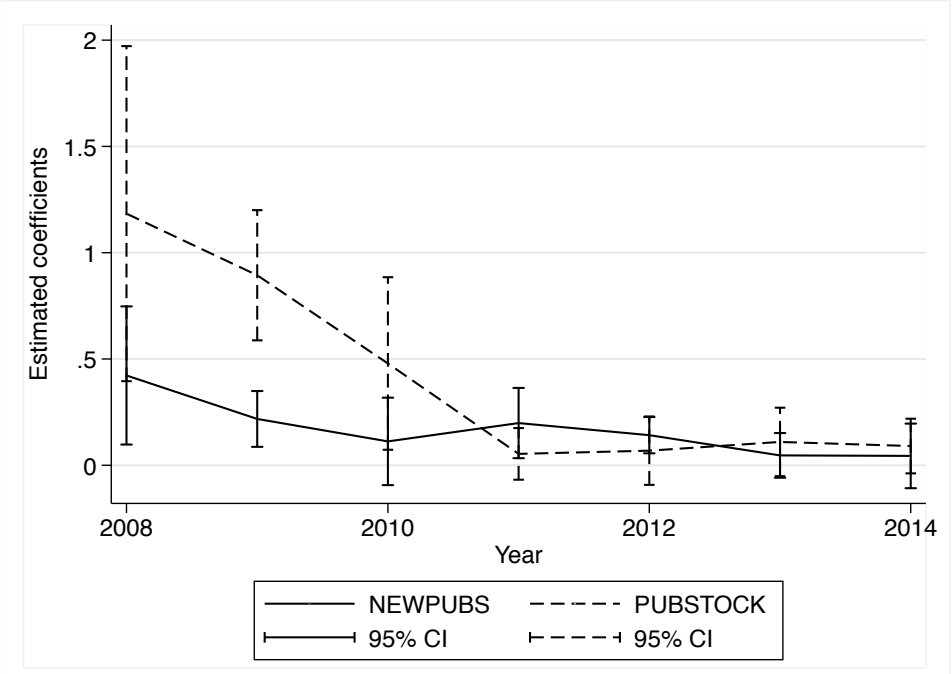
Despite of the this bleak assessment, there are success stories. For example, the targeted treatment Gleevec has proven itself effective in the fight of Chronic Myelogenous Leukemia. Several diagnostics have been designed to screen for specific mutations of the BRCA1 and BRCA2 genes, yielding a significant improvement in early-onset detection of breast cancer. Recently, gene therapy has been successfully applied to restore vision to blind patients. Furthermore, at a more basic level, the unveiling of the genome marked a before and after in the understanding of human biology, producing strong evidence to back Craig Venter's 2000 claim that the genome would help us decipher "our place in the biological continuum."[30]

[28] http://blogs.nature.com/boston/2011/02/23/nnb-report-ten-years-later-harvard-assesses-the-genome-map

[29] Diseases in this class, such as Huntington's disease an Cystic Fibrosis, are called "Mendelian."

[30] For example, the genome has lead to the surprising discovery that the number of genes is not related in an obvious way to the complexity of the organism. Humans have only about 21,000 genes, while a fruit fly –a much simpler organism– has around 17,000. A mouse has around 23,000.

Figure 9: Testing the "hype" hypothesis



6.2 Analysis

We investigate an empirical implication of the "hype" hypothesis that stems from expectations' gradual reversal to the mean. In particular, as the (lackluster) clinical trial performance of therapies whose design was informed by GWAS' studies start to be observed, firms begin to gradually adjust their priors down. This means that the marginal impact of new scientific knowledge characterizing a disease's biology progressively shrinks, leading to ever smaller impacts on the intensity of discovery.

To test this hypothesis we adjust our baseline model by allowing the parameters of NEWPUBS and PUBSTOCK to vary in time. In particular, we estimate the following negative-binomial specification:

$$\begin{aligned} \mathbb{E}[n_{it}] \quad = \quad & f(\alpha_t \cdot \text{NEWPUBS}_{i,t-2:t-1} + \beta_t \cdot \text{PUBSTOCK}_{i,t-3} \\ & + \gamma \cdot \text{REVENUES}_{i,t-1} + \phi \cdot \text{MISSINGREVENUES}_{i,t-1} \\ & + \delta_t + \eta_{a(i)}), \end{aligned} \tag{3}$$

The parameter set $\{\alpha_t, \beta_t\}_{t \geq 2008}$ captures the heterogeneous extents to which scientific knowledge flows impacted discovery rates across years. The "hype" hypothesis implies that these parameters will be decreasing in t.[31] To estimate the model we interact year

[31] Estimates are similar to those obtained when we also include $\{\alpha_t\}_{t=2006,2007}$, but the latter are estimated on few NEWPUBS>0 data points and translate into coefficients above 1, which we deem an

indicator variables with each publication metric. That is, we do not impose a functional form on the $\{\alpha_t, \beta_t\}_{t \geq 2008}$ sequences.

Figure 9 summarizes the results. Observed patterns support the "hype" hypothesis. In 2008, each new study increased was associated with about 0.45 additional therapies entering the discovery stage within two years of publication. One year later, in 2009, this effect was cut by about a third, then again cut in half in 2010. Although larger in 2011, it falls again in 2012 and losses statistical significance afterward.

A similar pattern is observed for the coefficients of PUBSTOCK. Although initial effects are larger, they are estimated imprecisely due to the relatively low number of publications before 2007. Effects loose statistical significance by 2011.

7 Discussion and Road Map

We have presented a preliminary analysis of the impact of the first generation of gene-disease association studies (i.e., genome-wide association studies) on drug discovery. We generate three main findings. First, scientific publications are associated to increased rates of discovery efforts. Although the causality of the effect is not demonstrated conclusively, various analysis and external evidence support it. Second, we find that scientific publications at the time a new therapy enters the discovery has a positive impact on the probability of reaching clinical trials, but does not affect the probability of clinical trial success. Lastly, we show that the effect of publications on the intensity of discovery efforts was initially (i.e., around 2008) high, but experience a monotonic and pronounced decrease in time, ultimately vanishing around 2011. After this year, the pace of discovery of new therapies for a given disease appears to be unaffected by the flow of new scientific knowledge linking that disease to its genetic roots.

We interpret these results as supportive of a "hype" hypothesis. We provide plenty of anecdotal evidence that also supports this view. For instance, speaking about genetically-identified targets, a former Bayer employee stated that "The whole industry went crazy with it. Bayer committed half a billion dollars into human genome research, and they got nothing for it. Nothing at all." Our results regarding the pace of translation across years map how the industry first "went crazy with it" and later became disenchanted, whereas our results regarding clinical trial performance illustrate how firms "got nothing for it."

There is still much work that remains to be done:

- We want to understand better the decline of GWAS publications over time (illustrated in Figure 1), and hopefully leverage this patterns for inference. One possible explanation is that as the "hype" was going away, there was less funding for GWAS research. An alternative story is that a new technique (NGS) has been substituting GWA studies, and therefore the perceived decline is a substitution effect rather

artifact of the low variation of the data in that period. We therefore drop them from our final specification.

than a decline in the overall number of publications. We have identified a dataset compiling NGS studies and are preparing it for analysis.

- We will generate a an external measure of technological hype along these lines: collect all media articles published in prominent media outlets (for example, The New York Times, Science, Business and Health sections), and look within them for references to genomic-based technologies, and classify them using sentiment analysis tools.

- We will attempt to directly identify the targets in GWAS publications and with those employed by new therapies. We have already systematized target information from Cortellis, but have not still been able to identify a gene-protein mapping required complete the linkage. This linkage would provide much more fidelity in the analysis, allowing us to specify diff-in-diff regressions in which the "treated group" entails therapies targeting the molecule produce by the gene whose connection with the specific disease has been identified by each GWAS study.

- We also envision assembling a dataset of patent filling activity. Doing this would allow us to explore a second layer of effects. In our current assessment of the project, this is not a strict priority and may be left for a follow-up project.

Finally, participating in this NBER initiative is an exceptional opportunity for young researchers like us. We would highly benefit from specialized and thorough feedback for this project, as well as for the follow-on ideas that we have in mind. We hope to contribute to this program and in general to the precision medicine literature. Please do not hesitate to contact us if you have any question.

References

Agarwal, Pankaj, Philippe Sanseau, and Lon R Cardon (2013) "Novelty in the target landscape of the pharmaceutical industry," *Nature Reviews Drug Discovery*, Vol. 12, pp. 575–576.

Agarwal, Rajshree and Atsushi Ohyama (2013) "Industry or academia, basic or applied? Career choices and earnings trajectories of scientists," *Management Science*, Vol. 59, pp. 950–970.

Aghion, Philippe, Mathias Dewatripont, and Jeremy C Stein (2008) "Academic freedom, private-sector focus, and the process of innovation," *The RAND Journal of Economics*, Vol. 39, pp. 617–635.

Alpert, Joseph S and Qin M Chen (2012) "Has the genomic revolution failed?" *Clinical cardiology*, Vol. 35, pp. 178–179.

Arrow, Kenneth (1962) "Economic welfare and the allocation of resources for invention," in *The rate and direction of inventive activity: Economic and social factors*: Princeton University Press, pp. 609–626.

Bennett, Charles L, Stephen Y Lai, Michael Henke, Sara E Barnato, James O Armitage, and Oliver Sartor (2010) "Association between pharmaceutical support and basic science research on erythropoiesis-stimulating agents," *Archives of internal medicine*, Vol. 170, pp. 1490–1498.

Blume-Kohout, Margaret E and Neeraj Sood (2013) "Market size and innovation: Effects of Medicare Part D on pharmaceutical research and development," *Journal of public economics*, Vol. 97, pp. 327–336.

Brinkman, Ryan R, Marie-Pierre Dubé, Guy A Rouleau, Andrew C Orr, and Mark E Samuels (2006) "Human monogenic disordersa source of novel drug targets," *Nature Reviews Genetics*, Vol. 7, pp. 249–260.

Bryan, Kevin A and Jorge Lemus (2016) "The Direction of Innovation."

Budish, Eric, Benjamin N. Roin, and Heidi Williams (2015) "Do Firms Underinvest in Long-Term Research? Evidence from Cancer Clinical Trials," *American Economic Review*, Vol. 105, pp. 2044–85, URL: http://www.aeaweb.org/articles/?doi=10.1257/aer.20131176, DOI: http://dx.doi.org/10.1257/aer.20131176.

Burdett, T, PN Hall, E Hastings, LA Hindorff, HA Junkins, AK Klemm, J MacArthur, TA Manolio, J Morales, H Parkinson, and Welter D (2016) "The NHGRI-EBI Catalog of published genome-wide association studies.."

Chen, Ruijun, Nihar R Desai, Joseph S Ross, Weiwei Zhang, Katherine H Chau, Brian Wayda, Karthik Murugiah, Daniel Y Lu, Amit Mittal, and Harlan M Krumholz (2016) "Publication and reporting of clinical trial results: cross sectional analysis across academic medical centers," *BMJ*, Vol. 352, DOI: http://dx.doi.org/10.1136/bmj.i637.

Chong, Jessica X, Kati J Buckingham, Shalini N Jhangiani, Corinne Boehm, Nara Sobreira, Joshua D Smith, Tanya M Harrell, Margaret J McMillin, Wojciech Wiszniewski, Tomasz Gambin et al. (2015) "The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities," *The American Journal of Human Genetics*, Vol. 97, pp. 199–215.

Cockburn, Iain M and Rebecca M Henderson (1998) "Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery," *The Journal of Industrial Economics*, Vol. 46, pp. 157–182.

DiMasi, Joseph A, Ronald W Hansen, and Henry G Grabowski (2003) "The price of innovation: new estimates of drug development costs," *Journal of health economics*, Vol. 22, pp. 151–185.

Dranove, David, Craig Garthwaite, and Manuel Hermosilla (2014) "Pharmaceutical Profits and the Social Value of Innovation,"Technical report, National Bureau of Economic Research.

Drews, Jürgen (2000) "Drug discovery: a historical perspective," *Science*, Vol. 287, pp. 1960–1964.

Evans, James P, Eric M Meslin, Theresa M Marteau, Timothy Caulfield et al. (2011) "Deflating the genomic bubble," *Science*, Vol. 331, pp. 861–862.

Fugh-Berman, Adriane (2013) "How basic scientists help the pharmaceutical industry market drugs," *PLoS Biol*, Vol. 11, p. e1001716.

Furman, Jeffrey L, Fiona Murray, and Scott Stern (2012) "Growing stem cells: the impact of federal funding policy on the US scientific frontier," *Journal of Policy Analysis and Management*, Vol. 31, pp. 661–705.

Furman, Jeffrey L. and Scott Stern (2011) "Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research," *American Economic Review*, Vol. 101, pp. 1933–63, URL: http://www.aeaweb.org/articles/?doi=10.1257/aer.101.5.1933, DOI: http://dx.doi.org/10.1257/aer.101.5.1933.

Giovanetti, Glenn and Gautam Jaggi (2013) "Beyond Borders: Biotechnology Industry Report," *Ernst & Young*.

Hawkins, Naomi (2011) "The impact of human gene patents on genetic testing in the United Kingdom," *Genetics in Medicine*, Vol. 13, pp. 320–324.

Hughes, JP, S Rees, SB Kalindjian, and KL Philpott (2011) "Principles of early drug discovery," *British journal of pharmacology*, Vol. 162, pp. 1239–1249.

Jensen, Kyle and Fiona Murray (2005) "Intellectual property landscape of the human genome," *Science(Washington)*, Vol. 310, pp. 239–240.

Kerem, Bat-sheva, Johanna M Rommens, Janet A Buchanan, Danuta Markiewicz, Tara K Cox, Aravinda Chakravarti, Manuel Buchwald, and Lap-Chee Tsui (1989) "Identification of the cystic fibrosis gene: genetic analysis," *Science*, Vol. 245, pp. 1073–1080.

Lander, Eric S, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh et al. (2001) "Initial sequencing and analysis of the human genome," *Nature*, Vol. 409, pp. 860–921.

Londin, Eric, Priyanka Yadav, Saul Surrey, Larry J Kricka, and Paolo Fortina (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.

Marshall, Eliot (2011) "Waiting for the Revolution," *Science*, Vol. 331, pp. 526–529, URL: http://science.sciencemag.org/content/331/6017/526, DOI: http://dx.doi.org/10.1126/science.331.6017.526.

Murray, Fiona (2002) "Innovation as co-evolution of scientific and technological networks: exploring tissue engineering," *Research policy*, Vol. 31, pp. 1389–1403.

Nelson, Matthew R, Hannah Tipney, Jeffery L Painter, Judong Shen, Paola Nicoletti, Yufeng Shen, Aris Floratos, Pak Chung Sham, Mulin Jun Li, Junwen Wang et al. (2015) "The support of human genetic evidence for approved drug indications," *Nature genetics*, Vol. 47, pp. 856–860.

Nelson, Richard R. (1959) "The Simple Economics of Basic Scientific Research," *Journal of Political Economy*, Vol. 67, pp. 297–306, URL: http://www.jstor.org/stable/1827448.

Nordhaus, William D et al. (2007) "Two centuries of productivity growth in computing," *Journal of Economic History*, Vol. 67, p. 128.

Pisano, Gary (2006) "Can science be a business?" *Harvard business review*, Vol. 10, pp. 1–12.

Pisano, Gary P (2006) *Science business: The promise, the reality, and the future of biotech*: Harvard Business Press.

Ross, Joseph S, Tony Tse, Deborah A Zarin, Hui Xu, Lei Zhou, and Harlan M Krumholz (2012) "Publication of NIH funded trials registered in ClinicalTrials. gov: cross sectional analysis," *Bmj*, Vol. 344, p. d7292.

Sanseau, Philippe, Pankaj Agarwal, Michael R Barnes, Tomi Pastinen, J Brent Richards, Lon R Cardon, and Vincent Mooser (2012) "Use of genome-wide association studies for drug repositioning," *Nature biotechnology*, Vol. 30, pp. 317–320.

Scannell, Jack W, Alex Blanckley, Helen Boldon, and Brian Warrington (2012) "Diagnosing the decline in pharmaceutical R&D efficiency," *Nature reviews Drug discovery*, Vol. 11, pp. 191–200.

Swinney, David C and Jason Anthony (2011) "How were new medicines discovered?" *Nature reviews Drug discovery*, Vol. 10, pp. 507–519.

Tripp, Simon and Martin Grueber (2011) "Economic impact of the human genome project," *Battelle Memorial Institute*.

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

Ward, Michael R and David Dranove (1995) "The vertical chain of research and development in the pharmaceutical industry," *Economic Inquiry*, Vol. 33, pp. 70–87.

Appendix

Apart from paper counts papers, we also look at their impacts. A novel way to measure of impact, which is specific to this industry, is to use the reported statistical significance of the findings. Each study in our database reports the p-values for the strongest SNP risk allele, which is a measure of how robust are the findings of the papers. It is reasonable to think that results with a higher degree of confidence provide more certainty regarding the mechanism that links genes and diseases, which could imply a positive relation between lower p-values (statistical significant results) and the number of new therapies. An additional way, is to weight papers according to their citations. Table 6 reproduces the results of Table 3 including these weighted measures. Qualitative results remain mostly unchanged, although they also suggest that there may be systematic differences in the types of science finance by private an public interest. We plan to address as the project unfolds.

Table 6: Weighted impacts of scientific publications on the intensity of discovery

	(1)	(2)
NEWPUBS (Public funding)	0.244***	0.174***
	(0.092)	(0.037)
PUBSTOCK (Public funding)	0.091	0.078
	(0.057)	(0.090)
NEWPUBS (Private funding)	0.590***	0.268***
	(0.188)	(0.059)
PUBSTOCK (Private funding)	0.265*	0.038
	(0.148)	(0.096)
REVENUES	0.088***	0.092***
	(0.033)	(0.034)
MISSINGREVENUES	-0.837***	-0.800***
	(0.247)	(0.258)
Weight	Per-year average citations	p-values
N	4,545	4,545

Both models include therapeutic area and year fixed effects. Robust standard errors (clustered by therapeutic area) are presented in parentheses. Legend $^{*}p$ <0.10, $^{**}p$ <0.05, $^{***}p$ <0.01