# How do patents affect follow-on innovation?
# Evidence from the human genome[*]

Bhaven Sampat
Columbia and NBER

Heidi L. Williams
MIT and NBER

February 12, 2015

## Abstract

In markets where innovation is cumulative, in the sense that discoveries are inputs into follow-on discoveries, optimal patent policy in part depends on how patents on existing technologies affect follow-on innovation. In this paper, we investigate whether patents on one specific technology — human genes — have affected follow-on scientific research and product development. Using administrative data on successful and unsuccessful patent applications submitted to the US Patent and Trademark Office, we link the exact gene sequences claimed in each patent application with data measuring gene-related scientific research (publications) and commercial investments (clinical development). We first document evidence of selection into patenting: in the full sample of human genes, patented genes appear more valuable than non-patented genes prior to being patented based on measures of both scientific and commercial value, thus motivating two quasi-experimental approaches. First, we present a simple comparison of follow-on innovation across genes claimed in successful versus unsuccessful patent applications. Second, we use the "leniency" of the assigned patent examiner as an instrumental variable for whether the patent application was granted a patent. Both approaches suggest that on average gene patents have had no effect on follow-on innovation. Our confidence intervals on these estimates are meaningfully precise: we are able to reject the effect sizes estimated in past papers investigating the effect of non-patent forms of intellectual property on follow-on innovation.

[*] Daron Acemoglu, Josh Angrist, David Autor, Stefan Bechtold, Nick Bloom, Tim Bresnahan, Joe Doyle, Dan Fetter, Amy Finkelstein, Alberto Galasso, Joshua Gans, Aaron Kesselheim, Pat Kline, Amanda Kowalski, Mark Lemley, Petra Moser, Ben Olken, Ariel Pakes, Jim Poterba, Arti Rai, Mark Schankerman, Scott Stern, Mike Whinston, and seminar participants at Chicago Booth, Clemson, Dartmouth, Duke, the Federal Reserve Board, Harvard, MIT, Northwestern Kellogg, the NBER (Productivity and Public Economics), Stanford, and UC-Santa Barbara provided very helpful comments. We are very grateful to Ernie Berndt for help with accessing the Pharmaprojects data; to Osmat Jefferson, the CAMBIA Lens initiative, Lee Fleming, and Guan-Cheng Li for sharing USPTO-related data; and to Lizi Chen, Cirrus Foroughi, Yunzhi Gao, Grant Graziani, Kelly Peterson, Lauren Russell, Sophie Sun, Nicholas Tilipman, and Hanwen Xu for excellent research assistance. Research reported in this publication was supported by the National Institute on Aging and the NIH Common Fund, Office of the NIH Director, through Grant U01-AG046708 to the National Bureau of Economic Research (NBER); the content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NBER. This work/research was also funded by the Ewing Marion Kauffman Foundation; the contents of this publication are solely the responsibility of the Grantee. Financial support from NIA Grant Number T32-AG000186 to the NBER, NSF Grant Number 1151497, and the NBER Innovation Policy and the Economy program is also gratefully acknowledged. Contact: bns3@columbia.edu, heidiw@mit.edu.

# 1 Introduction

Competitive markets may under-incentivize innovation due to the public good nature of new ideas. Intellectual property rights, such as patents, aim to address this under-investment problem by allowing inventors to capture a higher share of the social returns to their research investments. By awarding inventors a temporary right to exclude others from marketing their invention, patents aim to allow inventors to earn quasi-rents — temporarily — as a way to re-coup their research and development costs, thus providing dynamic incentives for investments in new technologies. Dating back at least to analyses such as Nordhaus (1969), optimal patent policy design has traditionally been framed as a trade-off between this benefit of providing incentives for the development of new technologies and the cost of deadweight loss from higher prices during the life of the patent.

Nordhaus-style models of optimal patent policy design have traditionally modeled innovations as isolated discoveries. However, in practice many or most innovations are cumulative, in the sense that any given discovery is also an input into later follow-on discoveries. When innovation is cumulative, optimal patent policy design also depends on how patents on existing technologies affect follow-on innovation. One example that has been prominent in recent policy debates is a patented human gene sequence. Sequenced genetic data is a research input into subsequent follow-on discoveries: by analyzing sequenced genetic data scientists may discover links between genetic variations and diseases, and such knowledge can be applied to commercialize medical technologies such as pharmaceutical treatments and diagnostic tests. If patents on discoveries such as human genes affect follow-on innovation, that effect is a key additional input into optimal patent policy design. In this paper, we investigate whether patents on human genes have affected follow-on scientific research and product development. Our broad goal is to inform whether the Nordhaus-style trade-off between ex ante incentives and deadweight loss is sufficient for optimal patent policy design, or whether — at least in this context — the effects of patents on follow-on innovation need to be considered.

A well-developed theoretical literature has documented ambiguous predictions for how patents will affect follow-on innovation. If a gene patent holder can develop all possible follow-on inventions herself, then all socially desirable follow-on inventions will be developed. However, as stressed by Scotchmer (2004), the patent holder may not know all potential opportunities for follow-on innovation. If ideas are scarce in this sense, and if gene patents are sufficiently broad that follow-on inventions require a license, then follow-on inventions will require cross-firm licensing agreements. If ex ante licensing agreements can be signed prior to follow-on inventors sinking their R&D investments, all socially desirable follow-on inventions will still be developed. However, if there are impediments to ex ante licensing — such as asymmetric information — then follow-on innovation will be lower than socially desirable due to the problem of how to divide the profit across firms (Scotchmer, 1991; Green and Scotchmer, 1995). This profit division problem could be exacerbated if transaction costs hinder cross-firm licensing agreements

(Heller and Eisenberg, 1998; Anand and Khanna, 2000; Bessen, 2004).[1] On the other hand, work dating back at least to Kitch (1977) has argued patents may facilitate investment and technology transfers across firms (Arora, 1995; Arora, Fosfuri and Gambardella, 2001; Kieff, 2001; Gans, Hsu and Stern, 2002, 2008), which may increase incentives for follow-on research and commercialization. Taken together, this motivates the need for empirical evidence to inform the question of how patents affect follow-on innovation.

While many have raised concerns that patents are hindering follow-on innovation in a variety of markets — ranging from software to biomedical research tools to business methods — little empirical evidence exists to either support or refute that view. Two empirical studies have documented evidence that non-patent forms of intellectual property which restrict access to research materials can substantially reduce follow-on academic research (Murray et al., 2008) and commercial development (Williams, 2013). However, patents as commonly used disclose discoveries and retain open access to materials, raising the question of whether these estimates for non-patent forms of intellectual property generalize to patents. This question of generalizability is important from a policy perspective. If non-patent restrictions discourage follow-on research, efforts to encourage innovation should focus on negotiations that enable researchers to have more open access to research materials, as in the DuPont case analyzed in Murray et al. (2008). On the other hand, if patents discourage follow-on research, reforms to the patent system may be appropriate.

The assumption that patents hinder follow-on innovation has been influential in shaping policy decisions. For example, Kesselheim, Cook-Deegan, Winickoff and Mello (2013) argue that the assumption that patents hinder follow-on innovation has informed a series of recent US Supreme Court decisions that moved to restrict the set of discoveries eligible for patent protection.[2] Our empirical estimates speak in particular to a recent set of high-profile legal rulings on the case *Association for Molecular Pathology v. Myriad Genetics*.[3] The firm Myriad Genetics was granted patents on human genes correlated with risks of breast and ovarian cancer, and in June 2013 the US Supreme Court unanimously ruled to invalidate a subset of Myriad's gene patent claims. Importantly for this paper, although gene patents were controversial for a variety of reasons — including concerns about the ethics of "owning" human genes — in the end the Court decision argued that DNA patents *"would 'tie up'...such tools and...inhibit future innovation."* As argued by a 2006 National Academies report and by Caulfield et al. (2013),

[1] Consistent with this view is an oft-cited anecdote that pharmaceutical firm Bristol Myers abandoned research on more than fifty cancer-related proteins due to conflicts with gene patent holders (Pollack, 2001).

[2] Specifically, Kesselheim and co-authors highlight three decisions since 2010 in which the US Supreme Court has expanded the set of discoveries excluded from patentability. First, in *Bilski v. Kappos* the Court invalidated patent claims on an investment strategy, announcing it supported a "high enough bar" on patenting abstract ideas that it would not "put a chill on creative endeavor and dynamic change." Second, in *Mayo v. Prometheus*, the Court invalidated patent claims on methods of using genetic variation to guide pharmaceutical dosing, expressing concern that "patent law not inhibit further discovery by improperly tying up the future of laws of nature." Finally, in *AMP v. Myriad* (as we will discuss more), the Court invalidated patent claims on genomic DNA, concluding that "[g]roundbreaking, innovative, or even brilliant" discoveries of natural phenomena are not patentable given that such patents "would 'tie up' the use of such tools and thereby inhibit future innovation premised upon them." Since the publication of Kesselheim, Cook-Deegan, Winickoff and Mello (2013), in *Alice Corp v. CLS Bank* the Court invalidated patent claims on software based on similar arguments.

[3] Appendix C provides some additional background on this case.

prior to this paper there was little empirical evidence available to either support or refute the assertion that gene patents were impeding follow-on innovation.[4] Gene patents are arguably also a case study of economic interest given the potential value of gene-related research for human health, and also offer a particularly clean empirical setting in which to test the relevant economic theory: research in this field largely starts from sequenced genes, so a quasi-experiment that randomly assigns some genes to be patented and others to not be patented provides a good approximation to the thought experiment of interest (as also argued by Garber and Romer (1996)).

The lack of empirical evidence on how patents affect follow-on innovation reflects three challenges. First, linking technologies embodied in patents with data identifying those technologies in "real-world" markets is difficult. Second, measuring follow-on innovation is also difficult. Third, a selection bias challenge arises because we may expect inventors to be more likely to file for and obtain patents on technologies that are more valuable. This type of selection raises the concern that any measured differences in follow-on innovation across patented and non-patented technologies may in part reflect the selection of which technologies are patented.

The contribution of this paper is to construct new data and develop two new quasi-experimental approaches to investigate how patents on one type of discovery — human genes — have affected follow-on scientific research and product development. Our data construction builds on administrative records from the US Patent and Trademark Office (USPTO) on the census of patent applications, both successful and unsuccessful.[5] We take advantage of the fact that patent applications claiming human genes as intellectual property must disclose the exact DNA sequences claimed in the text of the patent. By applying bioinformatics methods (Jensen and Murray, 2005), these DNA sequences can be annotated with gene identifiers, and these gene identifiers can in turn be linked to standard scientific and medical datasets providing measures of follow-on scientific research and product development related to each gene. Specifically, we measure the scientific publications related to each gene as an indicator of scientific research investments, and measure the use of genes in pharmaceutical clinical trials and diagnostic tests as indicators of commercial research investments. The key advantage of this DNA sequence-based match is that it enables the construction of a unique dataset which precisely links the technologies embodied in patent applications (namely, DNA sequences) with "real-world" measures of the same technologies.

These data allow us to document new facts relating patents, scientific research, and commercial investments

[4] Despite not directly investigating the effects of gene patents, the Williams (2013) evidence — documenting that a non-patent form of database protection on the human genome hindered follow-on innovation — was cited in several briefs filed in the *AMP v. Myriad* case, such as `http://www.americanbar.org/content/dam/aba/publications/supreme_court_preview/briefs-v2/12-398_pet.authcheckdam.pdf` and `http://www.yaleisp.org/wp-content/uploads/2012/01/31253-BRCA-AMICUS-BRIEF-FINAL.pdf`; see Marshall (2013) for a recent discussion.

[5] Importantly, this data set holds the disclosure of inventions constant: both accepted and rejected inventions are disclosed in the published patent applications that comprise our sample. This implies that our analysis corresponds to a partial equilibrium analysis (holding disclosure constant), whereas in general equilibrium firms may choose to protect their inventions with e.g. trade secrecy or other strategies rather than with patents if patent protection is not available. For example, in response to the *Mayo v. Prometheus* ruling discussed above, the law firm Goodwin Proctor issued a "client alert" arguing that the ruling called for "a renewed look at trade secret law for protection of diagnostics innovations as an alternative to patents"; see `http://www.goodwinprocter.com/Publications/Newsletters/Client-Alert/2012/0323_Lessons-from-Mayo-v-Prometheus.aspx`. We discuss these issues more in Section 6.

on the human genome. We first document evidence of selection into patenting: in the full sample of human genes, patented genes appear more valuable than non-patented genes based on measures of scientific and commercial value, prior to receiving patents.

Motivated by this evidence of selection, we develop two quasi-experimental methods for estimating how patents have affected follow-on innovation. First, we present a simple comparison of follow-on innovation across genes claimed in accepted versus rejected patent applications.[6] In our context, this method is reasonable if — conditional on being included in a patent application — whether a gene is granted a patent is as good as random.[7] Empirically, genes claimed in accepted and rejected patent applications appear similar on characteristics fixed at the time of application, providing some evidence for the validity of this empirical approach.

Second, we develop a novel instrumental variable for which patent applications are granted patents: namely, the "leniency" of the assigned patent examiner. Patent examiners are charged with a uniform mandate: grant patents to patent-eligible, novel, non-obvious, and useful inventions. However, prior research has documented that in practice this mandate leaves patent examiners a fair amount of discretion (Cockburn, Kortum and Stern, 2003; Lemley and Sampat, 2010, 2012). We leverage the interaction of this across-examiner heterogeneity with the quasi-random assignment of patent applications to examiners as a source of variation in which patent applications are granted patents. Past qualitative evidence — and new empirical evidence we document — supports the assertion that the assignment of patent applications to examiners is plausibly random conditional on some covariates (such as application year and technology type), suggesting that the leniency of the patent examiner to which a patent application is assigned can provide a valid instrument for whether the patent application is granted a patent.

In contrast with what one would infer from a naive comparison of follow-on innovation on patented and non-patented genes, both of our quasi-experimental approaches suggest that gene patents have not had quantitatively important effects on either follow-on scientific research or follow-on commercial investments. The estimates from our first quasi-experimental approach — comparing follow-on innovation across genes claimed in successful versus unsuccessful patent applications — document economically small and meaningfully precise estimates. Across a variety of specifications, our 95% confidence intervals tend to reject declines or increases in follow-on scientific research and commercial investment on the order of more than 5-15%. Given that the previous literature (Murray et al., 2008; Williams, 2013) documented evidence that non-patent forms of intellectual property reduced follow-on innovation on the order of 30%, our estimates suggest meaningfully smaller effects in this context.[8]

[6] Strictly speaking, patent applications are never formally rejected by the USPTO, only abandoned by applicants (Lemley and Sampat, 2008). For brevity, we colloquially refer to such applications as "rejected" throughout the text.

[7] As we will discuss in more detail, this approach does *not* require that patents are as good as randomly assigned to patent applications; rather, the relevant assumption is at the gene level, for genes included in accepted and rejected applications.

[8] Our estimates are consistent with the recent contemporaneous contribution of Galasso and Schankerman (forthcoming), who document no evidence that patent invalidations increase patent citations in the technology field that includes most of our gene patents. Specifically, the most common patent class in our sample is 435 (Chemistry: molecular biology and microbiology), which is included in the NBER patent technology field "drugs and medical"; Galasso and Schankerman do not find evidence that patent invalidations increase follow-on

Our first quasi-experimental approach has the advantage of generating meaningfully precise estimates, but has the cost of relying on a less transparent source of variation. Our second quasi-experimental approach — using the leniency of the assigned patent examiner as an instrument for which patent applications are granted patents — has the advantage of relying on a clean and transparent source of variation, but generates less precise estimates (although even these estimates can usually reject the 30% declines documented in the prior literature). Taken together, the fact that these two approaches generate similar conclusions gives us confidence in focusing our interpretation on the more precise estimates generated by the first approach.

Taken together, these results suggest two conclusions. First, for the case of human genes, the traditional patent trade-off of ex ante incentives versus deadweight loss may be sufficient to analyze optimal patent policy design, as the effects of patents on follow-on innovation appear to be quantitatively small. Second, the available evidence suggests there may be important heterogeneity in how and when intellectual property policies affect follow-on innovation.[9] A variety of evidence suggests that non-patent policies governing access to materials appear to have important effects on follow-on innovation (Murray et al., 2008; Furman and Stern, 2011; Williams, 2013), and the evidence from Galasso and Schankerman (forthcoming) suggests that patents may deter follow-on innovation in some specific contexts. This growing body of empirical evidence suggests policy-relevant distinctions that should inform future theoretical and empirical research on cumulative innovation.

Methodologically, our two quasi-experimental approaches build on similar applications in labor economics and public finance. Our comparison of accepted and rejected patent applications builds on Bound (1989) and von Wachter, Song and Manchester (2011)'s investigations of the disincentive effects of disability insurance on labor supply, and Aizer, Eli, Ferrie and Lleras-Muney (2013)'s investigation of the effects of cash transfers on mortality. Likewise, our approach of using examiner "leniency" as a source of variation in which patent applications are granted patents builds on past work investigating the effects of incarceration length using variation across judges (Kling, 2006), the effects of foster care using variation across foster care case workers (Doyle, 2007, 2008), and the disincentive effects of disability insurance on labor supply using variation across disability insurance examiners (Maestas, Mullen and Strand, 2013). Especially given the relatively small number of patent law changes in countries like the US in recent years, these two new sources of quasi-experimental variation may provide valuable opportunities to investigate the effects of patents in a variety of other applications.

Section 2 describes our data. Section 3 documents evidence of selection in which genes are patented, motivating our two quasi-experimental approaches. Section 4 presents estimates from our first quasi-experimental ap-

innovation in that subcategory. Unfortunately, the variation used by Galasso and Schankerman (forthcoming) — namely, the quasi-random assignment of court cases to judges at the US Court of Appeals for the Federal Circuit, as an instrument for patent invalidations — is not directly applicable in our context because very few human gene patents have ever been asserted in court (Holman, 2007, 2008).

[9] We here do not focus on the estimates from Murray and Stern (2007), who documented evidence that patent grants hindered subsequent citations to matched scientific papers (a measure of follow-on innovation), because subsequent work by the authors (Fehder, Murray and Stern, 2014) documented that this result was specific to the first few years after the founding of the journal they analyzed.

proach, comparing follow-on innovation across genes claimed in successful and unsuccessful patent applications. Section 5 presents estimates from our second quasi-experimental approach, using the leniency of the assigned patent examiner as an instrumental variable for whether the patent application was granted a patent. Section 6 interprets our empirical results in light of the growing literature investigating the impact of intellectual property rights in markets where innovation is cumulative, and Section 7 concludes.

## 2 Data

This section describes our data construction.[10] To fix ideas, we start by describing an example patent application — USPTO patent application 08/483,554 — claiming intellectual property over the BRCA1 gene, and describe our measures of follow-on innovation in the context of that example (Section 2.1). We then describe in more detail how we construct our sample of USPTO patent applications claiming intellectual property over human genes (Section 2.2), and our gene-level measures of follow-on scientific research and product development (Section 2.3).

### 2.1 Example: USPTO patent application 08/483,554

On 7 June 1995, Myriad Genetics filed USPTO patent application number 08/483,554: *17Q-Linked Breast and Ovarian Cancer Susceptibility Gene*. This application was subsequently granted a patent on 5 May 1998 (US patent number 5,747,282), and would later become a focus of the US Supreme Court case *Association for Molecular Pathology v. Myriad Genetics*.

This patent claimed intellectual property over an isolated sequence of nucleotide bases (adenine, cytosine, guanine, and thymine), the sequence of which is listed explicitly in the patent (*SEQ ID NO:1: ATG GAT TTA...*). By comparing this sequence of nucleotide bases to the census of human gene sequences, we can uncover that this sequence corresponds to the BRCA1 gene.

The text of the Myriad patent describes how variation in the precise sequence of nucleotide bases in the BRCA1 gene can induce variation in an individuals' risk of developing breast and ovarian cancers. For example, women with certain abnormal types of BRCA1 or BRCA2 genes have a 40 to 80 percent lifetime risk of developing breast cancer, relative to about 12 percent in the general population.[11]

Such links between genetic variation and diseases are referred to as genotype-phenotype links. In the case of the BRCA1 gene, scientific papers have investigated links between BRCA1 and breast cancer (141 publications), ovarian cancer (96 publications), and pancreatic cancer (2 publications).[12] Once scientific knowledge of a given genotype-phenotype link has been documented, this knowledge can be applied to develop medical technologies

[10] Appendix A describes our data construction in more detail.

[11] These statistics are drawn from GeneReviews, published by the US National Institutes of Health (NIH).

[12] These data are drawn from the NIH Online Mendelian Inheritance in Man (OMIM) database, described below.

such as pharmaceutical treatments and gene-based diagnostic tests. In the case of the BRCA1 gene, more than 200 pharmaceutical clinical trials have been conducted that focus on mutations in the BRCA1 gene, and a BRCA1 genetic test is marketed by the firm Myriad Genetics.[13] The goal of our data construction is to trace these measures of follow-on scientific research and product development for each human gene, and to link these data to records of which human genes have been included in USPTO patent applications and granted patents.

## 2.2 Constructing USPTO patent application sample

Our quasi-experimental approaches require constructing data on the census of published USPTO patent applications that claim intellectual property over human genes — both successful (granted patents) and unsuccessful (not granted patents).[14] Traditionally, unsuccessful USPTO patent applications were not published. However, as part of the American Inventors Protection Act of 1999, the vast majority of USPTO patent applications filed on or after 29 November 2000 are published in the public record — regardless of whether they are granted patents — eighteen months after the filing date.[15]

From the census of USPTO patent applications, we identify the subset of applications claiming intellectual property over human genes. To do this, we follow the methodology proposed by Jensen and Murray (2005), which can be briefly summarized as follows.[16] Since the early 1990s DNA sequences have been listed in USPTO patent applications in a standard format, labeled with the text "SEQ ID NO" (sequence identification number). This standardized format allows for DNA sequences to be cleanly extracted from the full text of USPTO published patent applications. Once extracted, standard bioinformatics methods can be used to compare these sequences against the census of human gene DNA sequences in order to annotate each sequence with standard gene identifiers that can be linked to outside databases.[17]

We apply this Jensen and Murray (2005) methodology to construct two samples. First, we construct a "full

[13] These data are drawn from the Citeline Pharmaprojects database and the NIH GeneTests.org database, described below.

[14] To give a sense of scale, in 2010 the USPTO was staffed by over 6,000 patent examiners, and received over 490,000 applications; see GAO (2010) and `http://www.uspto.gov/web/offices/ac/ido/oeip/taf/appl_yr.htm`.

[15] For more details, see `http://www.uspto.gov/web/offices/pac/mpep/s1120.html` and the discussion in Lemley and Sampat (2010). Most applications not published eighteen months after filing are instead published sixty months after filing. Some US patent applications opt out of publication: Graham and Hegde (2013) document that around 8% of US applications opted for pre-grant secrecy of patent applications. For the NBER patent technology field "drugs and medical," which includes the most common patent class in our sample, the share was 3.5%.

[16] While we focus on the Jensen and Murray (2005) definition of gene patents, there are many different types of DNA-related patent claims — e.g., protein-encoding sequences, expressed sequence tags (ESTs), single-nucleotide polymorphisms (SNPs), sequence-based claims, and method claims that pertain to specific genes or sequences; see Scherer (2002) and Holman (2012) for more discussion. As one point of comparison, nearly all of the patent applications in our sample are included in the *DNA Patent Database* (`http://dnapatents.georgetown.edu/SearchAlgorithm-Delphion-20030512.htm`), which is constructed using a completely different methodology; we are grateful to Robert Cook-Deegan and Mark Hakkarinen for sharing the patent application numbers included in the *DNA Patent Database*, which enabled this comparison. Our empirical strategies could of course be applied to any empirically implementable definition of gene patents, but we are not aware of data on any other alternative definitions.

[17] Following Jensen and Murray (2005), we focus attention on the subset of DNA sequences that are explicitly listed in the claims of the patent applications (excluding DNA sequences that are referenced in the text of the patent application — and hence listed in the SEQ ID NO format — but not explicitly claimed as intellectual property).

sample" that aims to include the census of published USPTO patent applications that claim any DNA sequences — not just human gene sequences — in their patent claims (Bacon et al., 2006). Second, we construct a "human gene sample" that includes only the subset of published USPTO patent applications that claim human genes (as opposed to non-human genes, such as mouse genes; Lee et al. (2007)). This human gene sample is the focus of our analysis, given the focus in this paper on investigating how patents on human genes have affected follow-on innovation. However, this sample includes far fewer patent applications than the full sample: we have around 1,500 applications in the human gene patent sample, relative to around 30,000 applications in the full sample. This matters for our second quasi-experimental design, because it is much more difficult to reliably estimate how patent grant propensities vary across examiners in the human gene sample alone. Reflecting this challenge, we use the full sample to estimate the examiner patent grant propensities, and apply those estimates to the examiners in our human gene sample.

Table 1 presents patent application-level summary statistics on our full sample and our human gene sample. Our full sample includes 30,243 USPTO patent applications with DNA sequences listed in their claims, while our human gene sample includes 1,582 USPTO patent applications with human genes listed in their claims. Our data primarily includes US patent applications filed from 2001-2010, given that the requirement for unsuccessful patent applications to be published came into force for patent applications filed on or after 29 November 2000. A small share (<0.01%) of our sample has an application year prior to 2000; we retain these observations in our sample in the absence of a clear reason to exclude them. The human gene sample ends with application year 2005, marking the last year of data used as an input into the data construction done by Lee et al. (2007). In both samples, about 30 percent of the patent applications had been granted patents by 2010, although this raw statistic does not account for censoring.[18]

While our quasi-experimental approaches can only be implemented over the time period when unsuccessful patent applications were published (that is, applications filed after 29 November 2000), for the purposes of our descriptive analyses we also apply the Jensen and Murray (2005) methodology to identify granted USPTO patents that claim human genes but were filed prior to 29 November 2000.[19]

[18] This grant rate is also a simple grant rate that does not account for the fact that US patent applications that are rejected can spawn closely related "new" applications (continuations or divisionals). Carley, Hegde and Marco (forthcoming) use internal USPTO data to calculate simple and "family" (including patents granted to continuations or divisionals) grant rates in the universe of new (not related to any previously filed applications) utility patent applications filed at the USPTO from 1996-2005. In this sample, 55.8% of applications were granted patents directly (without the use of continuations), and including patents granted to children increases the allowance rate to 71.2%. For the NBER patent technology field "drugs and medical," which includes the most common patent class in our sample, the progenitor allowance rate is 42.8% and the family allowance rate is 60.7%.

[19] These data are also drawn from Lee et al. (2007), and include USPTO patents granted through 2005.

## 2.3 Measuring follow-on innovation

We collect data on three measures of gene-level follow-on innovation: scientific publications as a measure of scientific research effort; and two measures of product commercialization: gene-related pharmaceutical research, and gene-based diagnostic tests.

We collect data on the scientific publications related to each gene from the Online Mendelian Inheritance in Man (OMIM) database, which catalogs scientific papers that have documented evidence for links between genetic variation and phenotypes. Using this data, we construct a count of the number of scientific papers published related to each gene — across all phenotypes — in each year.

Because of the long time lags between basic drug discovery and the marketing of new drugs, new approvals of drugs that take advantage of sequenced genetic data are just barely starting to enter the market (Wade, 2010). Given these time lags, rather than using drug approvals as a measure of pharmaceutical research, we instead focus on an intermediate measure of drug discovery — namely, drug compounds currently under development, as disclosed in clinical trials. Specifically, we measure gene-related pharmaceutical clinical trials using the Citeline Pharmaprojects database, a privately-managed competitive intelligence database that tracks drug compounds in clinical trials and — critically for our project — assigns gene identifiers to compounds related to genetic variation on specific genes. Using this data, we construct a count of the number of clinical trials related to each gene in each year.

Finally, we collect data on gene-based diagnostic tests from the GeneTests.org database, a genetic testing registry. Some gene-based tests provide individuals with information about disease risk, such as the BRCA tests related to risks of breast and ovarian cancers; other gene-based tests identify individuals as relatively more or less appropriate for a given medical treatment, such as a test which predicts heterogeneity in the side effects of the widely-prescribed blood thinner warfarin. Using the GeneTests.org data, we construct an indicator for whether a gene is included in any gene-based diagnostic test as of 2012. Unfortunately, this data is only available in a cross-section (not a panel).

A priori, the impact of patents on follow-on scientific research could differ from the impact of patents on product development. For example, many have argued that most patented inventions are made available to academic researchers on sufficiently favorable licensing terms that academics are able to continue their research (USPTO, 2001), so that even if transaction costs hinder licensing agreements for commercial applications we could expect to see no impact of patents on measures of follow-on academic research such as scientific publications. Alternatively, patents may change researchers' incentives of whether to disclose the results of their research through academic publications (Moon, 2011), in which case observed differences in scientific publications could be explained by differences in disclosure rather than by differences in the true amount of underlying scientific research. By contrast, we would not expect the product development outcomes we measure to be affected by such disclosure preferences given that the measures we observe are revealed in the natural course of firms commercializing and selling their technologies.

## 3 Comparison of patented and non-patented human genes

Our interest in this paper is in comparing follow-on innovation across patented and non-patented inventions. In order to investigate this question, we first examine the selection process in order to better understand which inventions are patented. On one hand, we may expect inventors to be more likely to both file for and be granted patents on technologies that are more valuable: inventors may be more willing to pay the time and monetary cost of filing patent applications for inventions that are more valuable, and patent applications claiming intellectual property over more valuable inventions may also be more likely to "clear the bar" and be granted patents. On the other hand, the USPTO grants patents based on criteria — patent-eligibility, novelty, non-obviousness, and usefulness — that may not closely correspond with measures of scientific and commercial value (Merges, 1988). Hence, it is an empirical question whether inventions with higher levels of scientific and commercial value are more likely to be patented.

There is a remarkable absence of empirical evidence on this question, largely due to a measurement challenge: in most markets, it is very difficult to measure the census of inventions, and to link those inventions to patent records in order to identify which inventions are and are not patented. The only other paper we are aware of which has undertaken such an exercise is Moser (2012), who constructs a dataset of innovations exhibited at world's fairs between 1851 and 1915 and documents that "high quality" (award-winning) exhibits were more likely to be patented, relative to exhibits not receiving awards. Given this dearth of previous estimates, documenting evidence on the selection of inventions into patenting is itself of interest. In addition, if we do observe evidence of selection into patenting, that would imply that any measured differences in follow-on innovation across patented and non-patented genes may in part reflect the selection of which genes are included in patent applications and granted patents (as opposed to an effect of patents on follow-on innovation).

For this exercise, we start with the full sample of human genes (N=26,440). As measured in our data, approximately 29 percent of human genes have sequences that were explicitly claimed in granted US patents (N=7,717).[20] Figure 1 documents trends in follow-on innovation by year separately for genes that ever receive a patent (triangle-denoted solid blue series), and for genes that never receive a patent (circle-denoted dashed red series). For scientific publications (Figure 1(a)), we plot the average log number of scientific publications by year in each year from 1970

[20] As a point of comparison, Jensen and Murray (2005) document that as of 2005, approximately 20 percent of human genes had sequences that were explicitly claimed as granted patents. Because our sample includes patent applications that were granted patents after 2005, we would expect our estimate to be mechanically larger.

to 2012.[21] For clinical trials (Figure 1(b)), we plot the average log number of clinical trials by year in each year from 1995 to 2011.[22]

Because by construction our human gene patents are measured starting in the mid-1990s (when the SEQ ID NO notation was introduced), the cleanest test of selection into patenting is a comparison of pre-1990 follow-on innovation across patented and non-patented genes. While the clinical trials data only starts later (in 1995), the scientific publications data is available prior to 1990 so can be used for this comparison. Looking at the data series in Figure 1(a) from 1970 to 1990 provides clear evidence of positive selection: genes that will later receive patents were more scientifically valuable — based on this publications measure — prior to being patented. Moreover, even within the 1970 to 1990 period this positive selection appears not just in the levels of follow-on innovation, but also in the trends: genes that will later receive patents appear to have divergent trends in scientific publications relative to genes that will never receive patents, even before any patents are granted.

These patterns — a level difference, and a divergence in trends — also appear in the clinical trials data presented in Figure 1(b), although it is not possible to cleanly separate selection and treatment because that data series starts in 1995. Likewise, if we tabulate the probability that genes are used in diagnostic tests for patented and non-patented genes, 13.45% of patented genes are used in diagnostic tests, compared to 6.39% of non-patented genes.

While we only observe our diagnostic test outcome in a cross-section, for our clinical trial outcomes we can replicate a version of this analysis which focuses on the set of gene patents whose applications were filed in or after 2001, in which case 1995 to 2000 can serve as a "pre-period" that isolates a selection effect. Appendix Figure D.1 Panels (e) and (f) document that we see very similar patterns of selection on levels as well as trends using this alternative sample of gene patents.

Taken together, these patterns suggest strong evidence of positive selection: genes that will later receive patents appear more scientifically and commercially valuable prior to being granted patents, relative to genes that will never receive patents. This evidence is important for two reasons. First, this evidence suggests that our measures of value (pre-patent filing scientific publications and clinical trials) are correlated with patenting activity, which provides some evidence that these measures can provide a meaningful basis for assessing selection in our two quasi-experimental approaches. Second, this evidence implies that a simple comparison of follow-on innovation across patented and non-patented genes is unlikely to isolate an unbiased estimate of how gene patents affect follow-on

[21] We focus on the average log number of scientific publications by year because even within a calendar year, the number of publications per human gene is quite right-skewed. The pattern of selection that we document is unchanged if we instead plot the share of genes with at least one scientific publication by year (Appendix Figure D.1 Panel (a)), or the average number of scientific publications by year (Appendix Figure D.1 Panel (c)). Here and elsewhere, we add one to the outcome variables in order to include observations with no observed follow-on innovation.

[22] These years — 1995 to 2011 — are the only years for which the Pharmaprojects data are available. As with the scientific publications measure, we focus on the average log number of clinical trials by year because this variable is quite right-skewed. Again, the pattern of selection that we document is unchanged if we instead plot the share of genes with at least one clinical trial by year (Appendix Figure D.1 Panel (b)), or the average number of clinical trials by year (Appendix Figure D.1 Panel (d)).

innovation. While a naive comparison that did not account for selection would conclude based on Figure 1 that patents encourage follow-on innovation, as we will see in Sections 4 and 5 our two quasi-experimental approaches will suggest a different conclusion.

# 4 Comparison of accepted and rejected patent applications

Our first quasi-experimental source of variation investigates a simple idea, which is whether genes that were included in unsuccessful patent applications can serve as a valid comparison group for genes that were granted patents.

A comparison of follow-on innovation on genes included in accepted and rejected patent applications will be valid if — conditional on being included in a patent application — whether a gene is granted a patent is as good as random. A priori, it is not clear that this could offer a valid comparison. The USPTO is responsible for assessing whether patent applications should be granted patents based on five criteria: patent-eligibility (35 U.S.C. §101), novelty (35 U.S.C. §102), non-obviousness (35 U.S.C. §103), usefulness (35 U.S.C. §101), and that the text of the application satisfies the disclosure requirement (35 U.S.C. §112). Any given invention in our sample will claim intellectual property rights over at least one gene, and what is required for our first quasi-experimental approach to be valid is that conditional on being included in a patent application, whether a gene is granted a patent is as good as randomly assigned. Importantly, this strategy does not require assuming that patents are as good as randomly assigned across applications; rather, the relevant assumption is that genes that are included in patent applications that are granted patents are comparable to genes that are included in patent applications that are not granted patents. Empirically, we will document in this section that genes claimed in successful and unsuccessful patent applications appear similar on observable characteristics fixed at the time of application, providing some evidence for the validity of this empirical approach. While this comparison is quite simple, we will see that the resulting estimates are similar to the estimates from our second quasi-experimental source of variation (the examiner leniency variation, presented in Section 5).

## 4.1 Graphical analysis

For this exercise, we start with the sample of genes included in at least one patent application in our USPTO patent applications sample (N=15,530; 59% of the full sample of 26,440 human genes). Of this sample, 4,858 genes are claimed in a patent application that is subsequently granted a patent (31%), relative to 10,672 genes that are never observed to be subsequently granted a patent (69%).[23] Figure 2(a) illustrates the time pattern of when the patented

[23] Note that this 4,858 figure is mechanically lower than the 7,717 figure in Section 3, because we here focus only on patents granted on patent applications filed after 29 November 2000 (the date when unsuccessful applications began to be published).

group receives its (first) patent grant over time. Over half of these genes have received a patent by 2005, and (by construction) all have received a patent by 2010.

Figures 2(b) and 2(c) document trends in follow-on innovation by year. As in Figure 1, we plot the average log number of scientific publications by year in each year from 1970 to 2012 (Figure 2(b)), and the average log number of clinical trials by year in each year from 1995 to 2011 (Figure 2(c)).[24] The solid blue triangle-denoted line represents genes claimed in at least one granted patent, and the dashed red circle-denoted line represents genes claimed in at least one patent application but never in a granted patent.

As a point of comparison, the dashed green square-denoted line represents genes never claimed in a patent application (N=10,910; 41% of the full sample of 26,440 human genes). Comparing this group of genes to the other two groups of genes, we see clear evidence of selection into patent filing: genes included in successful and unsuccessful patent applications are much more valuable both scientifically (publications) and commercially (clinical trials) prior to the patent application filing compared to genes that are never claimed in a patent application. In terms of interpreting the results in Section 3, this suggests that the major source of selection in which genes are patented is selection in which genes are included in patent applications (as opposed to which genes are granted patents, conditional on being included in patent applications).

The key comparison of interest in this section is across the first two data series: genes included in successful patent applications, and genes included in unsuccessful patent applications. Strikingly, we see little evidence of selection in pre-2001 levels or trends of our two follow-on innovation measures once we limit the sample to genes included in patent applications. For scientific publications (Figure 2(b)), the two groups follow each other quite closely from 1970 to 1990, and diverge slightly in trends from 1990 to 2000 — with genes that will subsequently be included in unsuccessful patent applications having slightly more scientific publications.[25] For clinical trials (Figure 2(c)), the two groups follow each other quite closely in both levels and trends over all available pre-2001 years of data. Taken at face value, the similarity of these two groups in pre-2001 outcomes provides some evidence for the validity of this empirical approach. A priori, one might have expected genes that were more scientifically or commercially valuable to have been more likely to receive patents. However, conditional on being included in a patent application, this appears not to be the case.

Looking at the post-2001 time period, we see that although these two groups of genes diverge (by construction) in whether they are claimed in granted patents (Figure 2(a)), we do not see any evidence of a divergence in follow-on innovation outcomes. That is, these figures suggest that gene patents have not had a quantitatively important

[24] Appendix Figure D.2 documents analogous figures if we instead plot the share of genes with at least one scientific publication or clinical trial by year (Appendix Figure D.2 Panels (a) and (b)) or the average number of scientific publications or clinical trials by year (Appendix Figure D.2 Panels (c) and (d)).

[25] Note that this slight divergence is not apparent in the robustness check documented in Appendix Figure D.2 Panel (c), where we plot the average number of scientific publications by year.

effect on either follow-on scientific research or on follow-on commercialization.

## 4.2 Regression analysis

We quantify the magnitudes of these differences in a regression framework in Table 2. Because our scientific publication and clinical trial outcomes are quite skewed, a proportional model or binary outcome (measuring "any follow-on innovation") is more appropriate than modeling the outcome in levels. Our baseline specifications focus on the log of follow-on innovation and (separately) an indicator for any follow-on innovation; Appendix Table D.1 presents robustness of these results to alternative specifications.

Given the absence of strong visual evidence for a difference in follow-on innovation across patented and non-patented genes, our focus here is on what magnitudes of effect sizes can be ruled out by our confidence intervals. Across these specifications, our 95% confidence intervals tend to reject declines or increases in follow-on innovation on the order of more than 5-15%. For brevity, we focus on interpreting the log coefficients. For our measures of follow-on scientific research (publications; Panel A of Table 2) and commercialization (clinical trials; Panel B of Table 2), the 95% confidence intervals can reject declines or increases of more than 2%. For our measure of diagnostic test availability (only measured as a binary indicator; Panel C of Table 2), we estimate that genes receiving patents had a statistically insignificant 0.9 percentage point decrease in the likelihood of being included in a diagnostic test as of 2012 relative to genes included in patent applications but not granted patents. Our 95% confidence interval can reject declines of greater than 2 percentage points and reject increases of more than 0.2 percentage points. Relative to a mean of 12%, this confidence interval suggests that we can reject declines in this outcome of greater than 17%.[26]

To summarize these findings, we also present an average standardized treatment effect: $\Sigma_{j \in J} \frac{1}{J} \frac{\beta_j}{\sigma_j}$, where $\sigma_j$ is the standard deviation of the outcome $y_j$ in the control group and $\beta_j$ is the coefficient estimates for outcome $j$. We estimate this via a pooled ordinary-least-squares regression stacking data for each outcome variable in a single regression equation. The 95% confidence intervals for these estimates, presented in Panel D of Table 2, can reject declines or increases in follow-on innovation of greater than 0.03 standard deviations.

The main caveat with these estimates is that they rely on a relatively non-transparent source of variation, since it is not clear why some gene patent applications are granted patents and others are not. The next section explores the underlying sources of this variation more directly.

[26] As a point of comparison, only 3% of genes never included in a patent application are included in a diagnostic test as of 2012.

# 5 Analyzing examiner-level variation in patent grant propensities

Our second source of quasi-experimental variation constructs an instrumental variable for whether a patent application is granted a patent. Our key idea is to build on previous research which has established that although patent examiners are charged with a uniform mandate, in practice examiners have a fair amount of discretion, and this discretion appears to translate into substantial variation in the decisions different examiners make on otherwise similar patent applications (Cockburn, Kortum and Stern, 2003; Lichtman, 2004; Lemley and Sampat, 2010, 2012).[27]

Of course, the existence of variation in leniency across examiners on its own is not sufficient for examiner-level leniency to be a valid instrumental variable for patent grants: the exclusion restriction requires that the instrument only affect the outcome (follow-on innovation) through the likelihood that a gene is patented. As we will describe, the institutional context suggests that the assignment of patent applications to USPTO patent examiners should be effectively random conditional on some covariates (such as application year and technology type); consistent with this assumption, we will document that genes assigned to 'lenient' and 'strict' examiners look similar on observable characteristics fixed at the time of patent application.[28]

To motivate our empirical specification, Section 5.1 provides some qualitative background on the key institutional features underlying our empirical strategy,[29] after which we present our empirical estimates.

## 5.1 Assignment of patent applications to patent examiners

A central USPTO office assigns application numbers to incoming applications, as well as patent class and subclass codes detailing the type of technology embodied in the application.[30] These class and subclass numbers determine which of ~300 so-called Art Units — specialized groups of examiners — will review the application.[31] Within an Art Unit, a supervisory patent examiner assigns the application to a patent examiner for review. While the patent application process up to this point is quite structured, from this point forward substantial discretion is left in the

[27] One of the individuals interviewed by Cockburn, Kortum and Stern (2003) described this variation informally by saying, "there may be as many patent offices as there are patent examiners." Similarly, a trade publication written by a former USPTO patent examiner and current patent agent (Wolinsky, 2002) described this variation by saying: *"The successful prosecution of a patent application at the USPTO requires not only a novel invention and adequate prosecution skills, but a bit of luck...If you knew the allowance rate of your examiner, you could probably estimate your odds of getting your patent application allowed."*

[28] While conditional random assignment of applications to examiners assuages many potential concerns about this exclusion restriction, some additional issues remain. In particular, while we focus on variation in patent grant propensity, examiner heterogeneity may also manifest itself in other ways, such as the breadth of patent grants (in terms of the number or strength of allowed claims) and time lags in grant decisions (Cockburn, Kortum and Stern, 2003).

[29] The discussion in this section draws heavily on Cockburn, Kortum and Stern (2003), Lemley and Sampat (2012), Frakes and Wasserman (2014), and GAO (2005). See Appendix B for more detail on the USPTO patent examination process.

[30] There are currently over 450 patent classes, and more than 150,000 subclasses; see `http://www.uspto.gov/patents/resources/classification/overview.pdf`.

[31] See `http://www.uspto.gov/patents/resources/classification/art/index.jsp`. For the current version of the class/subclass-to-Art Unit concordance, see `http://www.uspto.gov/patents/resources/classification/caau.pdf`. The main Art Units in our sample are from the 1600 group (Biotechnology and Organic Chemistry).

hands of individual examiners, who are responsible for determining which — if any — claims in the application are patentable.

Because no “standard” method for the within-Art Unit assignment of applications to examiners is uniformly applied in all Art Units, Lemley and Sampat (2012) conducted written interviews with roughly two dozen current and former USPTO examiners to inquire about the assignment process. While the results of these interviews suggested that there is not a single “standard” assignment procedure that is uniformly applied in all Art Units, these interviews revealed no evidence of deliberate selection or assignment of applications to examiners on the basis of characteristics of applications other than those observed in standard USPTO datasets. For example, in some Art Units supervisors reported assigning applications to examiners based on the last digit of the application number; because application numbers are assigned sequentially in the central USPTO office, this assignment system — while not purposefully random — would be functionally equivalent to random assignment for the purposes of this study. In other Art Units, supervisors reported placing the applications on master dockets based on patent classes and subclasses, with examiners specializing in those classes (or subclasses) being automatically assigned the oldest application from the relevant pool when requesting a new application. Our key conclusion from this institutional context is that effective conditional random assignment of applications to examiners — within Art Unit and application year — is plausible. Consistent with this assumption, we will document that genes assigned to ‘lenient’ and ‘strict’ examiners look similar on observable characteristics fixed at the time of patent application.

From a practical perspective, it is worth noting that informational barriers limit the extent to which we would expect patent applications to be systematically sorted across examiners in a way that would be problematic for our empirical specifications. Because of the limited attention given to patents prior to their assignment to a specific examiner, and the judgment required to determine the patentability of a given invention, it seems plausible that informational barriers would impose real constraints on sorting (this argument has been made in more detail by Merges (2001)).[32] In particular, the “patentability” of applications is difficult to assess *ex ante*, and there is no evidence that supervisory patent examiners attempt to do so before assigning applications to particular examiners.

## 5.2 Measure of examiner leniency

The institutional context outlined above motivates — conditional on Art Unit-by-application year fixed effects — the following candidate instrumental variable for application $i$ filed in year $t$ assigned to examiner $j$ in Art Unit $a$:

$$Z_{itja} = \left( \frac{n_{\text{grant}_j} - \mathbf{1}(\text{grant}_i = 1)}{n_{\text{application}_j} - 1} \right)$$

[32] Merges (2001) also argues that although sorting of patent applications to examiners may be efficient, an additional barrier to such sorting is the strong “all patents are created equal” tradition at the USPTO, which cuts strongly against any mechanism for separating and sorting patents.

where $n_{\text{grant}_j}$ is the number of patents granted by examiner $j$, $n_{\text{application}_j}$ is the number of patent applications reviewed by examiner $j$, and $\text{grant}_i$ is an indicator equal to 1 if patent application $i$ was granted a patent. This expression measures the patent grant rate of examiner $j$, to which application $i$ was assigned, for all gene patent application cases except application $i$ itself. Motivated by the institutional context, we include Art Unit-by-application year fixed effects in our regressions, so that this instrument measures an examiner's patent grant propensity relative to other examiners reviewing applications in that Art Unit in that application year.[33]

For our baseline specification, we calculate this instrument based on the patent grant status of each gene patent application as of 2010. One concern is that this measure is censored for patent applications that are still in the examination process. To investigate the relevance of censoring, Appendix Figure D.5 plots — for our first five cohorts of gene patent applications, filed in 2001-2005 — the distribution of time between the application filing and the patent grant date, for the set of granted patents in our full sample. The average time to patent grant is around four years, and the cumulative share flattens out around six years. Given that we observe patent grant outcomes through 2010, this suggests that focusing on the human gene sample of patent applications filed in 2005 or earlier should be relatively free of censoring concerns.

A separate concern is that if we only observed a small number of applications per examiner, our estimate of the variation in leniency across examiners would be overstated. To address this concern, we limit the sample to examiners that saw at least ten gene patent applications (Heckman, 1981; Greene, 2001). Together with two additional restrictions — Art Unit-years with at least two examiners, and Art Unit-years reviewing at least 10 applications — this restriction drops ~3% of our sample.

Appendix Figure D.4(a) presents a histogram of the residualized instrument $Z_{itja}$ in the full sample. The standard deviation of the instrument is 0.10, with a range from -0.56 to 0.64. Appendix Figure D.4(b) presents histograms of the residualized instrument $Z_{itja}$ in the human gene sample, as estimated in the full sample and (separately) the human gene sample. When estimated on the full sample, the standard deviation of the instrument is 0.09, with a range from -0.33 to 0.35; when estimated on the human gene sample, the standard deviation and range are similar.

### 5.3 Graphical analysis

We first present a simple graphical analysis that provides an analog to Figure 2. While our main analysis in this section will exploit the more continuous sources of variation available in our data, these graphs provide a more transparent way of displaying our main results graphically.

Figure 3 is constructed as follows. We start with the human gene sample of patent applications (N=1,582).

[33] In our descriptive statistics and graphical analyses, we residualize the instrument by these Art Unit-by-application year fixed effects in order for the descriptive statistics and figures to reflect the same variation as our regression analyses.

Because patent applications can claim intellectual property over more than one human gene, and one human gene can be claimed in more than one patent application, we expand this data to patent application-human gene level observations. We classify each observation as being assigned to an above (triangle-denoted solid blue series) or below (circle-denoted dashed red series) median examiner based on the value of the examiner instrument for that application, as estimated in the full sample. To approximate the natural experiment described in Section 5.1, we condition on art unit-by-application year fixed effects. We then collapse this data into two groups, by above-median and below-median examiner leniency.

Figure 3(a) illustrates the time pattern of when each group receives patent grants over time. Consistent with the first stage estimates we document in Section 5.4, patent applications assigned to more lenient examiners are more likely to be granted patents. Consistent with the selection analysis we present in Section 5.5, comparing the pre-2001 levels and trends of both scientific publications (Figure 3(b)) and clinical trials (Figure 3(c)) suggests little evidence of selection: genes assigned to more lenient examiners appear to have similar levels of follow-on innovation to genes assigned to less lenient examiners prior to the application filing date.[34] Finally, consistent with our estimates in Section 5.6, although these two groups of genes diverge in whether they are claimed by patents, we do not see any evidence of a divergence in follow-on innovation outcomes after these patents start to be granted (post-2001). That is, just as we saw in our comparison of successful and unsuccessful patent applications in Section 4, these figures suggest that gene patents have not had a quantitatively important effect on either follow-on scientific research or on follow-on commercialization.

## 5.4 First stage estimates

Figure 4(a) presents a binned scatterplot relating our instrument (residualized by Art Unit-by-application year fixed effects) to the probability of patent grant in our full sample of patent applications (focus for now on the dark red plot; we will discuss the lighter yellow overlaid plot below). Visually, there is a strong relationship. Figure 4(b) plots this relationship in our human gene sample of patent applications, binning application-gene level data. The slopes of the estimated linear fit lines in the two graphs are quite similar.

To quantify the magnitude of these upward slopes, we estimate the following equation for a patent application $i$ examined by patent examiner $j$ filed in year $t$ assigned to Art Unit $a$:

$$\mathbf{1}\left(\text{patent grant}\right)_{itja} = \alpha + \beta \cdot Z_{itja} + \Sigma_{at}\mathbf{1}\left(\text{art unit}\right)_{ta} + \varepsilon_{itja}$$

where the outcome variable $\mathbf{1}\left(\text{patent grant}\right)_{itja}$ is an indicator variable equal to one if patent application $i$ was

[34] Appendix Figure D.3 documents analogous figures if we instead plot the share of genes with at least one scientific publication or clinical trial by year (Appendix Figure D.3 Panels (a) and (b)) or the average number of scientific publications or clinical trials by year (Appendix Figure D.3 Panels (c) and (d)).

granted a patent, $Z_{itja}$ is our leave-one-out-mean (LOOM) gene patent grant rate instrument (as defined in Section 5.2), and $\Sigma_{at}\mathbf{1}\left(\text{art unit}\right)_{ta}$ are a set of Art Unit-by-application year fixed effects. We estimate heteroskedasticity robust standard errors, clustered at the examiner level.

Column (1) of Panel A of Table 3 presents the estimate which corresponds to Figure 4(a). The point estimate implies that a 10 percentage point increase in an examiner's average patent grant rate is associated with a 6.8 percentage point increase in the likelihood that a patent application is granted a patent. The F-statistic is on the order of 400, well above the rule of thumb for weak instruments (Stock, Wright and Yogo, 2002). As a robustness check, in Appendix Table D.2 we replace our Art Unit-by-year fixed effects with Art Unit-by-year-by-class-by-subclass fixed effects, on the subsample for which these finer fixed effects can be estimated. The point estimates from this more stringent specification are very similar to and not statistically distinguishable from our baseline point estimates in Table 3, suggesting that at least in our context, variation in the measured leniency of different examiners is unlikely to be generated by the systematic sorting of patent applications that are in more versus less 'patentable' technology classes or subclasses.

In our sample of human gene patent applications — estimated on a dataset of unique application-level observations — we obtain a first stage coefficient of 0.865 (Column (1) of Panel B of Table 3) — slightly larger in magnitude although less precise than the corresponding estimate in Panel A. If we expand our human gene patent application sample to patent application-gene level observations — a change which gives more weight to patent applications which claim intellectual property over more genes — we obtain a first stage coefficient of 0.563 (Column (1) of Panel C of Table 3). As expected given the smaller sample of applications in the human gene samples, the estimates in Panel B and Panel C are less precise than the estimates in Panel A. The key first stage coefficient that will provide the basis for our empirical estimate is in Column (1) of Panel C; while statistically significant at the 1% level, the F-statistic is on the order of 8, just below the conventional cut-off of 10 for weak instruments.

As discussed in Section 5.2 our patent grant outcome is measured as of 2010 and is censored for patent applications that are still in the process of being examined, but this censoring should be less of a concern for our five early cohorts of gene patent applications (2001-2005). The point estimates on this sample are very similar to our baseline point estimates (results not shown), suggesting that censoring appears to not substantively affect the magnitude of the estimated first stage coefficient. Note that this is likely because the Art Unit-by-application year fixed effects largely account for differences in the probability of patent grant that are mechanically related to time since application. Given this similarity, we retain all cohorts of gene patent applications to retain a larger sample size.

## 5.5 Investigating selection

In order for the examiner leniency measure constructed in Section 5.2 to be a valid instrumental variable for the likelihood that a given gene patent application is granted a patent, it must satisfy the exclusion restriction: the instrument can only affect follow-on innovation outcomes through the likelihood that a gene is patented. The institutional details described in Section 5.1 suggest that the assignment of applications to examiners is plausibly random conditional on Art Unit-by-application year fixed effects, lending some a priori credibility to the exclusion restriction. In this section, we aim to empirically assess whether this assumption is reasonable by investigating whether patent applications assigned to 'lenient' and 'strict' examiners look similar on observable characteristics fixed at the time of patent application.

Ideally, we would empirically assess selection using variables that are correlated with the 'patentability' of the application at the time of filing, so that we could test whether applications that appear more patentable tend to be assigned to more lenient examiners. As discussed by Lemley and Sampat (2012), it is difficult to identify variables that measure the 'patent-worthiness' of an invention.[35] We here focus on two imperfect predictors of patentability: first, measures of the value of patent applications that are fixed at the time of application; and second, measures of the value of genes that are claimed in the patent application.

### 5.5.1 Selection tests based on the value of patent applications

A variety of metrics have been proposed as measures of the value of granted patents: forward citations (Trajtenberg, 1990), patent renewal behavior (Pakes, 1986; Schankerman and Pakes, 1986; Bessen, 2008), patent ownership reassignments (Serrano, 2010), patent litigation (Harhoff, Scherer and Vopel, 2003), and excess stock return values (Kogan, Papanikolaou, Seru and Stoffman, 2013). For our purposes, not all of these measures are appropriate: we need measures of patent value that are defined for patent applications (not just for granted patents), and also want a measure that is fixed at the time of application (and hence unaffected by subsequent grant decisions). For these reasons, we focus on two value measures which fit these criteria: patent family size and claims count.

Generally stated, a patent "family" is defined as a set of patent applications filed with different patenting authorities (e.g. US, Europe, Japan) that refer to the same invention. The key idea is that if there is a per-country cost of filing for a patent, firms will be more likely to file a patent application in multiple countries if they perceive the patent to have higher private value. Past work starting with Putnam (1996) has documented evidence that patent family size is correlated with other measures of patent value. We define patent family size as the number of unique countries in which the patent application was filed (as measured in the INPADOC patent family).

[35] In their paper, they show that two observable characteristics fixed at the time of application — the number of pages in the application and the patent family size — are not correlated with a measure of examiner experience (years of employment at the USPTO). That evidence provides some support for our exclusion restriction.

We use claims count as an alternative value measure that is fixed at the time of patent application, as proposed by Lanjouw and Schankerman (2001). The key idea here is that patents list "claims" over specific pieces of intellectual property, and that patents with more claims may be more valuable.

For our purposes, there are two key empirical questions we want to investigate using these measures. First, do patent family size and/or claims count predict patent grant? While clearly imperfect metrics of patentability, we document in Appendix Table D.3 that these variables are predictive of patent grants. Second, is the predicted probability of patent grant — predicted as a function of family size and claims count — correlated with our examiner leniency instrument? We document in Column (2) of Table 3 that if we regress the predicted probability of patent grant (predicted as a function of family size and claims count) on our examiner leniency instrument (residualized by art unit-by-application year fixed effects), we estimate economically small (and — in two of three panels — 'wrong signed') statistically insignificant coefficients. This relationship for Panel A is displayed non-parametrically in the lighter yellow plot in Figure 4(a): consistent with the regression estimate, there is no visual relationship between the predicted probability of patent grant and our instrument.

### 5.5.2 Selection tests based on the value of claimed genes

We also investigate an alternative set of covariates: namely, characteristics of the genes that are included in the patent applications. Specifically, we use our measure of the level of scientific research activity on each gene in the years *prior* to the patent application (1970 to 1999) as a measure of the state of scientific knowledge about that gene.[36] Because one patent application often covers more than one human gene, this selection exercise relies on patent application-by-gene level data (since the covariates are defined at the gene level).

Do these gene-level covariates predict the 'patentability' of a given patent application? If we regress an indicator variable for whether a patent application is granted a patent on these covariates in our sample of human gene patent applications, we find that these covariates are predictive of patent grants: the p-value from an F-test of the scientific research variables as predictors of patent grant is less than 0.001.[37] Hence, although these covariates are clearly imperfect proxies for patentability, they do appear to be useful for the purpose of this exercise. If we regress the predicted probability of patent grant (predicted as a function of our gene-level scientific research variables) on our examiner leniency instrument (residualized by art unit-by-application year fixed effects), we estimate an economically small and statistically insignificant coefficient of 0.0009 (s.e. 0.0053).[38] This relationship is displayed

[36] We also compile a host of other gene-level observable variables that measure a gene's location on the genome — such as the chromosome on which a gene is located. Whereas the scientific research variables are defined for all observations in our sample, the genome location covariates are often missing; hence, we analyze these two sets of covariates separately so as to retain as many observations as possible, and focus on the scientific research variables.

[37] This p-value is also less than 0.001 if we add the location covariates, and estimate this regression on the sub-sample of data with non-missing data on the location covariates.

[38] This estimate is not shown in Table D.3, given that that table focuses on investigating selection based on patent application-level covariates.

non-parametrically in the lighter yellow plot in Figure 4(b): consistent with the regression estimate, there is no visual relationship between the predicted probability of patent grant and our instrument.

Taken together, the analysis in this section provides indirect support of our exclusion restriction in the following sense: we find no evidence that applications which appear more likely to be patented based on measures that are fixed at the time of filing are differentially assigned to more lenient examiners. Hence, the variation in grant rates across examiners appears to reflect differences in the decisions made on ex ante similar applications.

### 5.6 Reduced form and instrumental variables estimates

Table 4 documents our reduced form estimates, relating our examiner leniency instrument to follow-on innovation outcomes, and Table 5 documents the analogous instrumental variable estimates. Appendix Tables D.4 and D.5 present robustness of these results to alternative specifications.

For brevity, we focus on interpreting the log instrumental variable estimates. For our measures of follow-on scientific research (publications; Panel A of Table 5) and commercialization (clinical trials; Panel B of Table 5), the 95% confidence intervals can reject declines or increases of more than 10%. For our measure of diagnostic test availability (Panel C of Table 5), the 95% confidence interval can reject declines of greater than 7 percentage points and reject increases of more than 2.6 percentage points. Relative to a mean of 9%, this confidence interval suggests that we can reject declines in this outcome of greater than 75% and reject increases of greater than 28%. The 95% confidence intervals on our standardized treatment effect estimates (Panel D of Table 5) can reject declines or increases in follow-on innovation of greater than 0.2 standard deviations.

The estimates from our first quasi-experimental approach — comparing follow-on innovation across genes claimed in successful versus unsuccessful patent applications — are much more precise than these estimates, but have the disadvantage of relying on a less transparent source of variation. However, our confidence in interpreting those estimates as causal is strengthened by the fact that the examiner leniency instrument — which appears to generate a clean ’experiment’ — generates similar results, albeit which are much less precise. From an economic perspective, the estimates from our first approach can cleanly reject the effect sizes documented in the prior literature of how non-patent forms of intellectual property affect follow-on innovation (Murray et al., 2008; Williams, 2013), whereas the estimates from our second approach can only sometimes reject these effect sizes.

## 6 Discussion

Two prior empirical studies documented evidence that non-patent forms of intellectual property that restrict access to research materials can substantially reduce follow-on innovation. Murray et al. (2008) analyze a package of patents and other licensing restrictions that the private firm Dupont held on genetically engineered mice, and

Williams (2013) analyzes a non-patent form of database protection that the private firm Celera held on their version of the sequenced genome. Importantly, both Dupont and Celera's forms of intellectual property restricted "openness": in order to gain access to Dupont's genetically engineered mice, or to Celera's sequenced genome, researchers needed to navigate legal contracts and restrictions. This feature stands in sharp contrast with how patents are commonly used in practice. For example, in our context, patented human genes were freely available to researchers with no restrictions — indeed, data on the entirety of the sequenced human genome was available throughout our time period in public open-access databases.

Both theoretical models (Aghion et al., 2008) and other sources of empirical evidence (Furman and Stern, 2011) suggest that limiting access to research materials may discourage basic research as well as the creation of new research lines. However, patents - unlike the non-patent forms of intellectual property analyzed in these past papers - disclose discoveries and generally retain open access to materials (Walsh et al., 2003a,b). Academics report in surveys that they very rarely investigate whether research materials that they already have access to are covered by patents.[39] One interpretation of this is that "tolerated infringement" is a common feature of the use of patented materials by academic researchers (Walsh, Arora and Cohen, 2003b; Walsh, Cho and Cohen, 2005). Empirically, many private firms seem to openly tolerate infringement by academic researchers.[40] Models such as Murray et al. (2008) and Bessen and Maskin (2009) provide rationales for why this behavior may be optimal for firms: if ideas are scarce in the sense that the initial patent holder does not know all potential opportunities for follow-on innovation, then allowing academics open access to research materials may expand the set of possible commercial applications. Of course, once basic research produces new commercial leads, licensing agreements for follow-on inventions must be negotiated. Hence, in order to rationalize our empirical estimates on product market outcomes it must also be the case that licensing markets operate relatively efficiently. Consistent with this is the fact that there has been very little gene patent-related litigation (Holman, 2008).

Against this background, perhaps the most natural interpretation of the available empirical evidence is the following. Patents per se neither restrict access to materials for basic researchers, nor generate quantitatively important transaction costs in licensing markets, and hence do not deter either follow-on scientific research nor follow-on commercialization. In contrast, non-patent policies that restrict basic researchers' access to materials can reduce follow-on scientific research, thus reducing the number of research lines that can be pursued commercially (and hence also reducing follow-on commercialization). Consistent with this interpretation, theoretical models such as Murray et al. (2008) predict that private firms should prefer patents to other forms of intellectual property that restrict access to materials, and at least in the case of Celera this seems to be an accurate description of the

[39] The Walsh, Cho and Cohen (2005) survey suggested that only about 5% of academic bench scientists reported regularly checking for patents on research inputs.

[40] For example, in the famous case of Human Genome Sciences' CCR5 patent, then-CEO William Haseltine publicly stated: "We would not block anyone in the academic world from using this for research purposes."

historical record: Celera attempted (but largely failed) to obtain patents on its sequenced genetic data, and in the absence of patent protection used their non-patent form of intellectual property as the next best available alternative means of capturing returns to their investment in sequencing the human genome.

Of course, this interpretation is at best suggestive, and given the available data it is not possible to rule out alternative explanations. A first alternative interpretation is that while gene patents have widely been perceived to be sufficiently broad that follow-on inventions require gene patent licenses, it could be that in practice this is not the case. Two examples may be helpful in providing more texture on this interpretation. First, in a now-classic paper Heller and Eisenberg (1998) documented that more than 100 issued US patents included the term 'adrenergic receptor' in the claim language, and pointed to this as an example of the complex biomedical patent landscape. However, in a comment on the Heller-Eisenberg paper published in the same issue of *Science*, Seide and MacLeod (1998) argued that based on a cursory patent clearance review, at most only a small number of licenses might be required.[41] Second, we can return to the example in the introduction where pharmaceutical firm Bristol Myers reported abandoning research on more than fifty cancer-related proteins due to conflicts with gene patent holders (Pollack, 2001). This example was actually reported as part of a licensing agreement Bristol Myers brokered with another firm (Athersys) for a method enabling the use of protein targets without infringing gene patents. Empirically, this alternative interpretation is difficult to test because very few gene patents have been litigated in court, and patent breadth is determined not only by the text of the claims approved by the patent office but also by courts' interpretations of these claims. But importantly, from a theoretical perspective narrow patents should not deter follow-on innovation, so if if this view is correct then our analysis would be most useful in highlighting that documentation of biomedical patenting frequency should not necessarily be interpreted as evidence of a problem: analyses such as the one in this paper are needed in order to investigate whether these patents are sufficiently broad to actually be affecting real economic behavior. A second alternative interpretation - argued by Lemley (2008) - is that both researchers and private firms simply ignore patents to a large degree, a view that is also consistent with our data. In order to test among these alternative theories, we would need to measure licensing agreements, data which is almost always kept confidential.

## 7 Conclusion

The contribution of this paper is to investigate whether patents on human genes have affected follow-on scientific research and product development. Using administrative data on successful and unsuccessful patent applications submitted to the US Patent and Trademark Office (USPTO), we link the exact gene sequences claimed in each

[41] A similar point has been made in the context of gene patents by Holman (2012): although a large number of human genes are claimed by US patents, he argues that a reading of the claims suggests that few licenses would be required for common types of follow-on innovation.

patent application with data measuring gene-related scientific research (publications) and commercial investments (clinical development). Building on these data, we develop two new sources of quasi-experimental variation: first, a simple comparison of follow-on innovation across genes claimed in successful versus unsuccessful patent applications; and second, use of the "leniency" of the assigned patent examiner as an instrumental variable for whether the patent application was granted a patent. Both approaches suggest that — on average — gene patents have not had quantitatively important effects on follow-on innovation.

This empirical evidence speaks against two existing views. First, there has been widespread concern that patents on human genes may hinder follow-on innovation. For example, in the recent *Association for Molecular Pathology v. Myriad Genetics* case, the US Supreme Court invalidated patent claims on genomic DNA, arguing that such patents "would 'tie up' the use of such tools and thereby inhibit future innovation premised upon them." Our empirical estimates do not provide support for patents hindering follow-on innovation in the context of human genes. Second, dating back at least to the academic work of Kitch (1977), many have argued that patents on basic discoveries play an important role in facilitating subsequent investment and commercialization. Our empirical estimates do not provide support for patents spurring follow-on innovation in the context of human genes.

Taken together with the prior literature, our evidence suggests two conclusions. First, for the case of human genes, the traditional patent trade-off of ex ante incentives versus deadweight loss may be sufficient to analyze optimal patent policy design, because the effects of patents on follow-on innovation appear to be quantitatively small. Second, the available evidence suggests there may be important heterogeneity in how and when intellectual property rights affect follow-on innovation. A variety of evidence suggests that non-patent policies governing access to materials appear to have important effects on follow-on innovation (Murray et al., 2008; Furman and Stern, 2011; Williams, 2013), and the evidence from Galasso and Schankerman (forthcoming) suggests that patents may deter follow-on innovation in some specific contexts. This growing body of empirical evidence suggests policy-relevant distinctions that should inform future theoretical and empirical research on cumulative innovation.

Of course, we are hesitant to generalize these conclusions to other markets given that in this paper we analyze the effects of patents on one specific type of technology (human genes). However, a broader contribution of this paper is to develop two new sources of quasi-experimental variation that may provide valuable opportunities to investigate the effects of patents in a variety of other applications. An analogy to the literature in labor economics investigating the disincentive effects of disability insurance on labor supply may be helpful. An influential early literature related labor supply outcomes to observational variation in disability insurance (Parsons, 1980), and subsequent papers provided refined approaches by drawing comparisons between accepted and rejected applicants (Bound, 1989; von Wachter et al., 2011) and using variation in the leniency of disability insurance examiners as an instrument for disability insurance receipt (Maestas, Mullen and Strand, 2013). Our paper provides an example of how each of these approaches can be applied to investigate the effects of patents, using administrative data on successful and unsuccessful patent applications and — in the case of the examiner leniency approach — information on the patent examiner assigned to review the patent application. Especially given the relatively small number of patent law changes in countries like the US in recent years, these two new sources of quasi-experimental variation may provide valuable opportunities to investigate the effects of patents in a variety of other applications.

## References

**Aghion, Philippe, Mathias Dewatripont, and Jeremy Stein**, “Academic freedom, private-sector focus, and the process of innovation,” *RAND Journal of Economics*, 2008, *39* (3), 617–635.

**Aizer, Anna, Shari Eli, Joseph Ferrie, and Adriana Lleras-Muney**, “The effects of childhood means-tested cash transfers on mortality,” 2013. unpublished mimeo, available `http://www.econ.ucla.edu/alleras/research/papers/childhood%20transfers%20v10-september%202013.pdf`.

**Anand, Bharat and Tarun Khanna**, “The structure of licensing contracts,” *Journal of Industrial Economics*, 2000, *48* (1), 103–135.

**Arora, Ashish**, “Licensing tacit knowledge: Intellectual property rights and the market for know-how,” *Economics of Innovation and New Technology*, 1995, *4* (1), 41–60.

**_ , Andrea Fosfuri, and Alfonso Gambardella**, *Markets for Technology: The Economics of Innovation and Corporate Strategy*, MIT Press, 2001.

**Bacon, Neil, Doug Ashton, Richard Jefferson, and Marie Connett**, “Biological sequences named and claimed in US patents and patent applications: CAMBIA Patent Lens OS4 Initiative,” 2006. `http://www.patentlens.net` (last accessed 2 January 2012).

**Bessen, James**, “Holdup and licensing of cumulative innovations with private information,” *Economics Letters*, 2004, *82* (3), 321–326.

_ , “The value of U.S. patents by owner and patent characteristics,” *Research Policy*, 2008, *37* (5), 932–945.

_ **and Eric Maskin**, “Sequential innovation, patents, and imitation,” *RAND Journal of Economics*, 2009, *40* (4), 611–635.

**Bound, John**, “The health and earnings of rejected disability insurance applicants,” *American Economic Review*, 1989, *79* (3), 482–503.

**Burk, Dan L.**, “Are human genes patentable,” *International Review of Intellectual Property and Competition Law*, 2013, *44* (7), 747–749.

**Carley, Michael, Deepak Hegde, and Alan Marco**, “What is the probability of receiving a US patent?,” *Yale Journal of Law & Technology*, forthcoming.

**Caulfield, Timothy, Subhashini Chandrasekharan, Yann Joly, and Robert Cook-Deegan**, “Harm, hype and evidence: ELSI research and policy guidance,” *Genome Medicine*, 2013, *5* (21).

**Cockburn, Iain, Samuel Kortum, and Scott Stern**, “Are all patent examiners equal? Examiners, patent characteristics, and litigation outcomes,” in Wesley Cohen and Stephen Merrill, eds., *Patents in the Knowledge-Based Economy*, National Academies Press, 2003.

**Doyle, Joseph J.**, “Child protection and child outcomes: Measuring the effects of foster care,” *American Economic Review*, 2007, *97* (5), 1583–1610.

_ , “Child protection and adult crime: Using investigator assignment to estimate causal effects of foster care,” *Journal of Political Economy*, 2008, *116* (4), 746–770.

**Fehder, Daniel, Fiona Murray, and Scott Stern**, “Intellectual property rights and the evolution of scientific journals as knowledge platforms,” *International Journal of Industrial Organization*, 2014, *36*, 83–94.

**Frakes, Michael and Melissa Wasserman**, "Is the time allocated to review patent applications inducing examiners to grant invalid patents? Evidence from micro-level application data," 2014. NBER working paper.

**Furman, Jeffrey and Scott Stern**, "Climbing atop the shoulders of giants: The impact of institutions on cumulative research," *American Economic Review*, 2011, *101* (5), 1933–1963.

**Galasso, Alberto and Mark Schankerman**, "Patents and cumulative innovation: Causal evidence from the courts," *Quarterly Journal of Economics*, forthcoming.

**Gans, Joshua, David Hsu, and Scott Stern**, "When does start-up innovation spur the gale of creative destruction?," *RAND Journal of Economics*, 2002, *33* (4), 571–586.

**_ , _ , and _** , "The impact of uncertain intellectual property rights on the market for ideas: Evidence from patent grant delays," *Management Science*, 2008, *54* (5), 982–997.

**Garber, Alan and Paul Romer**, "Evaluating the federal role in financing health-related research," *Proceedings of the National Academy of Sciences*, 1996, *93* (23), 12717–12724.

**Golden, John M. and William M. Sage**, "Are Human Genes Patentable? The Supreme Court Says Yes and No," *Health Affairs*, 2013, *32* (8), 1343–1345.

**Graham, Stuart and Deepak Hegde**, "Do inventors value secrecy in patenting? Evidence from the American Inventor's Protection Act of 1999," 2013. unpublished USPTO mimeo.

**Green, Jerry and Suzanne Scotchmer**, "On the division of profit in sequential innovation," *RAND Journal of Economics*, 1995, *26* (1), 20–33.

**Greene, William**, "Estimating econometric models with fixed effects," 2001. unpublished mimeo.

**Harhoff, Dietmar, F.M. Scherer, and Katrin Vopel**, "Citations, family size, opposition, and the value of patent rights," *Research Policy*, 2003, *32* (8), 1343–1363.

**Harrison, Charlotte**, "Isolated DNA patent ban creates muddy waters for biomarkers and natural products," *Nature Reviews Drug Discovery*, 2013, *12*, 570.

**Heckman, James**, "The incidental parameter problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process," in Charles Manski and Daniel McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, 1981.

**Heller, Michael and Rebecca Eisenberg**, "Can patents deter innovation? The anticommons in biomedical research," *Science*, 1998, *280* (5364), 698–701.

**Holman, Christopher**, "The impact of human gene patents on innovation and access: A survey of human gene patent litigation," *University of Missouri-Kansas City Law Review*, 2007, *76*, 295–361.

_ , "Trends in human gene patent litigation," *Science*, 2008, *322* (5899), 198–199.

_ , "Debunking the myth that whole-genome sequencing infringes thousands of gene patents," *Nature Biotechnology*, 2012, *30* (3), 240–244.

**Jensen, Kyle and Fiona Murray**, "Intellectual property landscape of the human genome," *Science*, 2005, *310* (5746), 239–240.

**Kesselheim, Aaron, Robert Cook-Deegan, David Winickoff, and Michelle Mello**, "Gene patenting - The Supreme Court finally speaks," *New England Journal of Medicine*, 2013, *369* (9), 869–875.

**Kieff, Scott**, “Property rights and property rules for commercializing inventions,” *Minnesota Law Review*, 2001, *85*, 697–754.

**Kitch, Edmund**, “The nature and function of the patent system,” *Journal of Law and Economics*, 1977, *20* (2), 265–290.

**Kling, Jeffrey**, “Incarceration length, employment, and earnings,” *American Economic Review*, 2006, *96* (3), 863–876.

**Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman**, “Technological innovation, resource allocation, and growth,” 2013. MIT mimeo.

**Lanjouw, Jean and Mark Schankerman**, “Characteristics of patent litigation: A window on competition,” *RAND Journal of Economics*, 2001, *32* (1), 129–151.

**Lee, Byungwook, Taehyung Kim, Seon-Kyu Kim, Kwang H. Lee, and Doheon Lee**, “Patome: A database server for biological sequence annotation and analysis in issued patents and published patent applications,” *Nucleic Acids Research*, 2007, *35* (Database issue), D47–D50.

**Lemley, Mark**, “Rational ignorance at the patent office,” *Northwestern University Law Review*, 2001, *95* (4), 1495–1532.

_ , “Ignoring Patents,” *Michigan State Law Review*, 2008, *19* (1), 19–34.

_ **and Bhaven Sampat**, “Is the patent office a rubber stamp?,” *Emory Law Journal*, 2008, *58*, 181–203.

_ **and** _ , “Examining patent examination,” *Stanford Technology Law Review*, 2010, *2*.

_ **and** _ , “Examiner characteristics and patent office outcomes,” *Review of Economics and Statistics*, 2012, *94*, 817–827.

**Lichtman, Doug**, “Rethinking prosecution history estoppel,” *University of Chicago Law Review*, 2004, *71* (1), 151–182.

**Maestas, Nicole, Kathleen Mullen, and Alexander Strand**, “Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt,” *American Economic Review*, 2013, *103* (5), 1797–1829.

**Marshall, Eliot**, “Lock up the genome, lock down research,” *Science*, 2013, *342*, 72–73.

**Merges, Robert**, “Commercial success and patent standards: Economic perspectives on innovation,” *California Law Review*, 1988, *76*, 803–876.

_ , “As many as six impossible patents before breakfast: Property rights for business concepts and patent system reform,” *Berkeley Technology Law Journal*, 2001, *14*, 577–615.

**Moon, Seongwuk**, “How does the management of research impact the disclosure of knowledge? Evidence from scientific publications and patenting behavior,” *Economics of Innovation and New Technology*, 2011, *20* (1), 1–32.

**Moser, Petra**, “Innovation without patents: Evidence from World’s Fairs,” *Journal of Law and Economics*, 2012, *55* (1), 43–74.

**Murray, Fiona and Scott Stern**, “Do formal intellectual property rights hinder the free flow of scientific knowledge? An empirical test of the anti-commons hypothesis,” *Journal of Economic Behavior and Organization*, 2007, *63* (4), 648–687.

_ , **Philippe Aghion, Mathias Dewatripont, Julian Kolev, and Scott Stern**, "Of mice and academics: Examining the effect of openness on innovation," 2008. unpublished MIT mimeo.

**National Academy of Sciences**, *Reaping the Benefits of Genomic and Proteomic Research: Intellectual Property Rights, Innovation, and Public Health*, National Academies Press, 2006.

**Nordhaus, William**, *Invention, Growth, and Welfare: A Theoretical Treatment of Technological Change*, MIT Press, 1969.

**Pakes, Ariel**, "Patents as options: Some estimates of the value of holding European patent stocks," *Econometrica*, 1986, *54* (4), 755–784.

**Parsons, Donald**, "The decline in male labor force participation," *Journal of Political Economy*, 1980, *88* (1), 117–134.

**Pollack, Andrew**, "Bristol-Myers and Athersys make deal on gene patents," *New York Times*, 2001, *8 January.*

**Putnam, Jonathan**, "The value of international patent protection," 1996. Yale PhD dissertation.

**Rai, Arti and Robert Cook-Deegan**, "Moving beyond 'isolated' gene patents," *Science*, 2013, *341*, 137–138.

**Sampat, Bhaven**, "USPTO patent application data, Version 2," 2011. `http://hdl.handle.net/1902.1/16402UNF:5:HFkba86yNYacbyMItlKPVQ==BHAVENSAMPAT` (last accessed 2 January 2012).

**Schankerman, Mark and Ariel Pakes**, "Estimates of the value of patent rights in European countries during the post-1950 period," *Economic Journal*, 1986, *96* (384), 1052–1076.

**Scherer, Frederic Michael**, "The economics of human gene patents," *Academic Medicine*, 2002, *77* (12), 1348–1367.

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

_ , *Innovation and Incentives*, MIT Press, 2004.

**Seide, Rochelle and Janet MacLeod**, "Comment on Heller and Eisenberg," 1998. ScienceOnline: `http://www.sciencemag.org/feature/data/980465/seide.dtl`.

**Serrano, Carlos**, "The dynamics of the transfer and renewal of patents," *RAND Journal of Economics*, 2010, *41* (4), 686–708.

**Stock, James, Jonathan Wright, and Motohiro Yogo**, "A survey of weak instruments and weak identification in generalized method of moments," *Journal of Business and Economic Statistics*, 2002, *20* (4), 518–529.

**Trajtenberg, Manuel**, *Economic Analysis of Product Innovation: The Case of CT Scanners*, Harvard University Press, 1990.

**United States Supreme Court**, "Association for Molecular Pathology et al. v. Myriad Genetics Inc. et al.," 2013. 12-398.

**US General Accounting Office (GAO)**, "Intellectual property: USPTO has made progress in hiring examiners, but challenges to retention remain," 2005.

_ , "U.S. Patent and Trademark Office: Performance management processes," 2010.

**USPTO**, "Utility examination guidelines," *Federal Register*, 2001, *66* (4), 1092–1099.

**von Wachter, Till, Jae Song, and Joyce Manchester**, “Trends in employment and earnings of allowed and rejected applicants to the Social Security Disability Insurance program,” *American Economic Review*, 2011, *101* (7), 3308–3329.

**Wade, Nicholas**, “A decade later, genetic map yields few new cures,” *New York Times*, 2010, *12 June.*

**Walsh, John, Ashish Arora, and Wesley Cohen**, “Research tool patenting and licensing and biomedical innovation,” in Wesley Cohen and Stephen Merrill, eds., *Patents in the Knowledge-Based Economy*, National Academy Press, 2003.

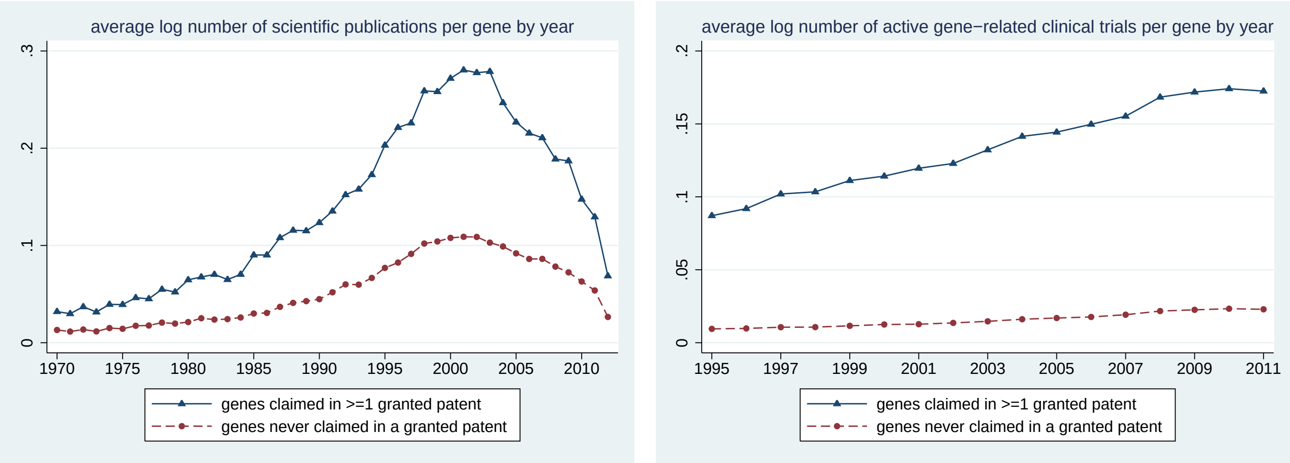
**_ , _ , and _** , “Working through the patent problem,” *Science*, 2003, *299* (5609), 1021.

**_ , Charlene Cho, and Wesley Cohen**, “View from the bench: Patents and material transfers,” *Science*, 2005, *309* (5743), 2002–2003.

**Williams, Heidi**, “Intellectual property rights and innovation: Evidence from the human genome,” *Journal of Political Economy*, 2013, *121* (1), 1–27.

**Wolinsky, Scott**, “An inside look at the patent examination process,” *The Metropolitan Corporate Counsel*, 2002, *10* (9), 18.

Figure 1: **Follow-on innovation on patented and non-patented human genes**

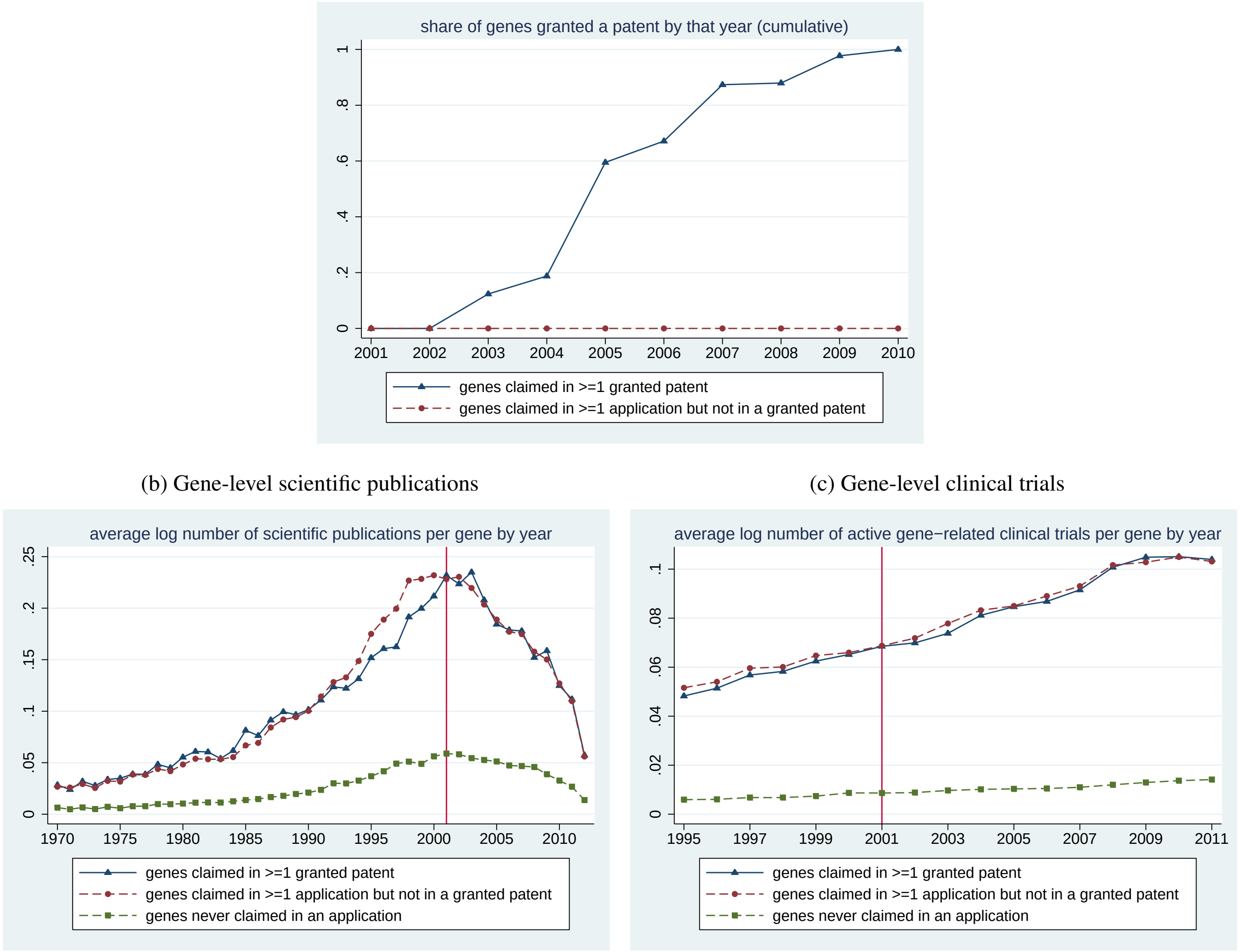


(a) Gene-level scientific publications

(b) Gene-level clinical trials

*Notes*: This figure plots trends in follow-on innovation by year separately for genes that ever receive a patent, and for genes that never receive a patent. The figure is constructed from gene-level data. Panel (a) uses gene-level scientific publications as a measure of follow-on innovation, and plots the average log number of scientific publications by year in each year from 1970 to 2012. Panel (b) uses gene-level clinical trials as a measure of follow-on innovation, and plots the average log number of clinical trials by year in each year from 1995 to 2011. We add one to both outcome variables in order to include observations with no observed follow-on innovation.

Figure 2: **Patents and follow-on innovation on human genes claimed in accepted/rejected patent applications**



(a) Share of genes receiving a patent grant

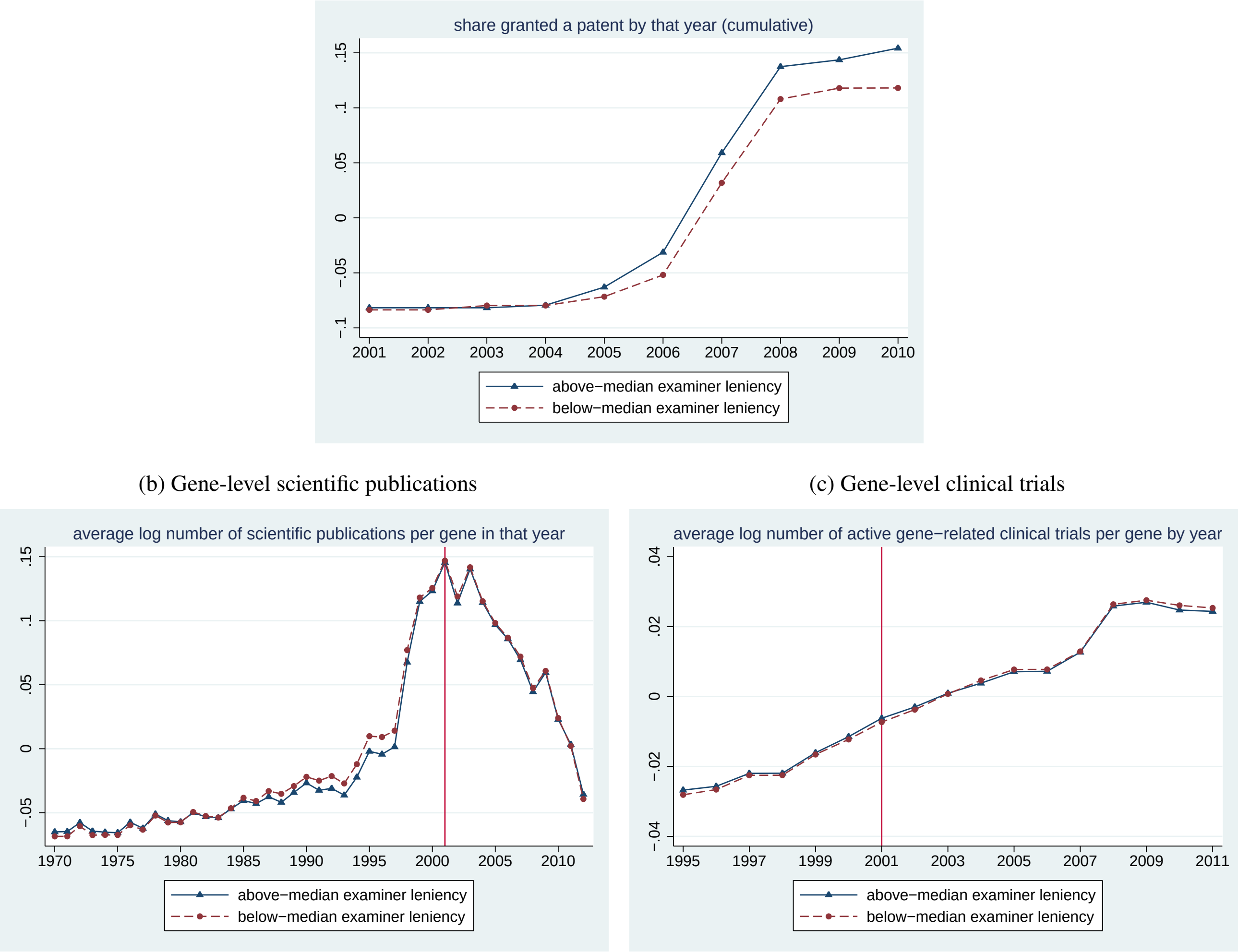
(b) Gene-level scientific publications

(c) Gene-level clinical trials

*Notes*: This figure plots trends in patenting and follow-on innovation by year separately for three groups of genes: genes claimed in at least one granted patent; genes claimed in at least one patent application but never in a granted patent; and (in Panels (b) and (c)) genes never claimed in a patent application. The figure is constructed from gene-level data. Panel (a) documents the share of genes receiving a patent grant by year; by construction, this is zero for the circle-denoted red dashed line in all years, is zero for the triangle-denoted blue line in 2001, and reaches one for the triangle-denoted blue line in 2010; the intermediate years simply illustrate the time path of patent grants between 2001 and 2010 for the triangle-denoted blue line. Panel (b) uses gene-level scientific publications as a measure of follow-on innovation, and plots the average log number of scientific publications by year in each year from 1970 to 2012. Panel (c) uses gene-level clinical trials as a measure of follow-on innovation, and plots the average log number of clinical trials by year in each year from 1995 to 2011. The vertical line in the calendar year 2001 in Panels (b) and (c) denotes that because this figure focuses on patents that were filed in or after 2001, all years prior to 2001 can be considered a pre-period and used to estimate the selection of genes into patenting based on pre-filing measures of scientific research (publications) and commercialization (clinical trials). In Panels (b) and (c), we add one to both outcome variables in order to include observations with no observed follow-on innovation.

Figure 3: **Patents and follow-on innovation on human genes assigned to above/below leniency examiners**

(a) Share receiving a patent grant

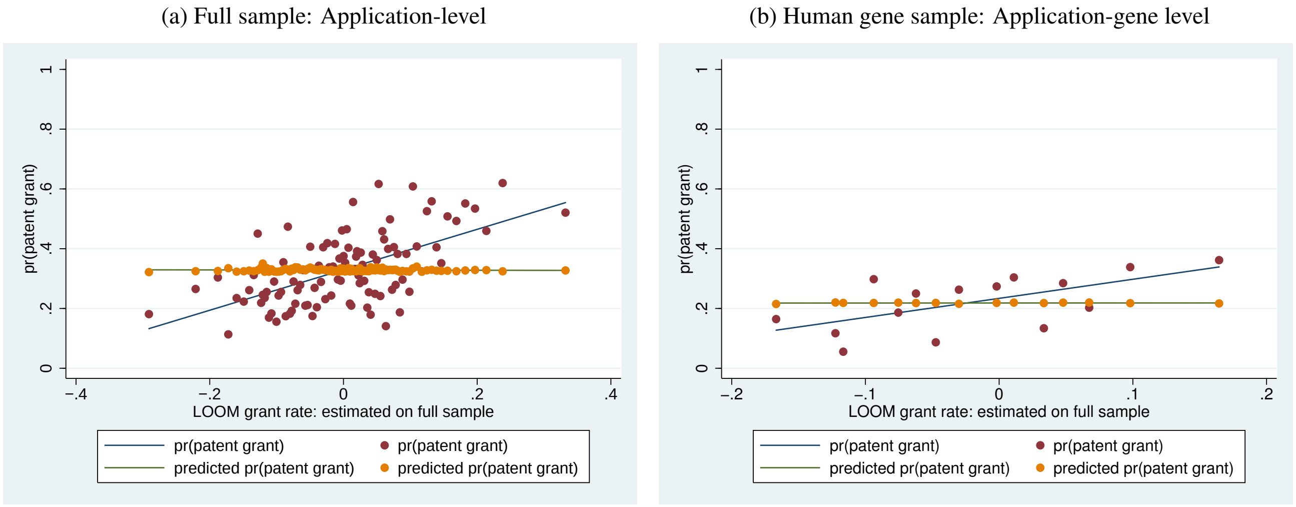


(b) Gene-level scientific publications

(c) Gene-level clinical trials

*Notes*: This figure plots trends in patenting and follow-on innovation by year separately for two groups of genes: genes claimed in patent applications that were assigned to above-median leniency patent examiners, and genes claimed in patent applications that were assigned to below-leniency patent examiners. The figure is constructed from patent application-gene level data. Panel (a) documents the share of genes receiving a patent grant by year. Panel (b) uses gene-level scientific publications as a measure of follow-on innovation, and plots the average log number of scientific publications by year in each year from 1970 to 2012. Panel (c) uses gene-level clinical trials as a measure of follow-on innovation, and plots the average log number of clinical trials by year in each year from 1995 to 2011. The vertical line in the calendar year 2001 in Panels (b) and (c) denotes that because this figure focuses on patents that were filed in or after 2001, all years prior to 2001 can be considered a pre-period and used to estimate the selection of genes into patenting based on pre-filing measures of scientific research (publications) and commercialization (clinical trials). In Panels (b) and (c), we add one to both outcome variables in order to include observations with no observed follow-on innovation.

Figure 4: **Probability of patent grant, by examiner leniency instrument** ($Z_{itja}$)



(a) Full sample: Application-level

(b) Human gene sample: Application-gene level

*Notes*: This figure plots the first stage relationship between the probability of patent grant and our examiner leniency instrument, and plots estimates from two of our selection tests. Panel (a) is constructed from patent application-level data for our full sample of patent applications; Panel (b) is constructed from patent application-gene level data for our human gene sample of patent applications. Panel (a) presents a binned scatter plot relating our leave-one-out-mean (LOOM) grant rate instrument, residualized by Art Unit-by-application year fixed effects, to two variables: (1) the patent grant rate; and (2) the predicted patent grant rate, where we predict patent grant as a function of our two measures of patent value fixed at the time of application (patent family size and claims count). Panel (b) presents a binned scatter plot relating our LOOM grant rate instrument, residualized by Art Unit-by-application year fixed effects and estimated on the full sample, to two variables: (1) the patent grant rate; and (2) the predicted patent grant rate, where we predict patent grant as a function of our gene-level scientific research variables (scientific publications in each year from 1970 to 1999). The second series is similar if we also include gene location covariates as predictors, and estimate the prediction on the sub-sample of data with non-missing data on the location covariates.

Table 1: **Patent application-level summary statistics**

| | mean | standard deviation | minimum | maximum | number of observations |
|---|---|---|---|---|---|
| **Panel A: Full sample** | | | | | |
| application year | 2004 | 2.6122 | 1994 | 2010 | 30,243 |
| 0/1, patent granted as of 2010 | 0.3282 | 0.4696 | 0 | 1 | 30,243 |
| **Panel B: Human gene sample** | | | | | |
| application year | 2002 | 0.7156 | 1998 | 2005 | 1,582 |
| 0/1, patent granted as of 2010 | 0.2541 | 0.4355 | 0 | 1 | 1,582 |

*Notes*: This table shows summary statistics for our patent application-level data in each of our two samples: Panel A for the full sample of patent applications, and Panel B for the human gene sample of patent applications.

Table 2: **Patents and follow-on innovation on human genes claimed in accepted/rejected patent applications: Regression estimates**

| outcome: | log of follow-on innovation in 2011/12 | any follow-on innovation in 2011/12 | |
|---|---|---|---|
| | (1) | (2) | |
| **Panel A: Scientific publications** | | | |
| patent granted | 0.0019 | -0.0013 | |
| | (0.0060) | (0.0054) | |
| mean of dependent variable | 0.2238 | 0.1094 | |
| number of observations | 15,530 | 15,530 | |
| **Panel B: Clinical trials** | | | |
| patent granted | 0.0008 | -0.0012 | |
| | (0.0080) | (0.0043) | |
| mean of dependent variable | 0.5424 | 0.0657 | |
| number of observations | 15,530 | 15,530 | |
| **Panel C: Diagnostic test** | | | |
| patent granted | - | -0.0091 | * |
| | | (0.0055) | |
| mean of dependent variable | - | 0.1199 | |
| number of observations | - | 15,530 | |
| **Panel D: Standardized treatment effect** | | | |
| patent granted | 0.0037 | -0.0124 | |
| | (0.0123) | (0.0099) | |
| number of observations | 31,060 | 46,590 | |

*Notes*: This table estimates differences in follow-on innovation on genes claimed in at least one granted patent relative to genes claimed in at least one patent application but never in a granted patent. The sample for these regressions is constructed from gene-level data, and includes genes claimed in at least one patent application in our USPTO human gene patent application sample (N=15,530). Each coefficient is from a separate regression. The standardized treatment effect estimates in Panel D are from a stacked regression, as described in the text (hence the difference in the number of observations). In Column (1), we add one to all outcome variables to include observations with no observed follow-on innovation. Estimates are from ordinary-least-squares models. Heteroskedasticity robust standard errors. *: p<0.10; **: p<0.05; ***: p<0.01.

Table 3: **First stage and falsification test estimates: Patent grants by examiner leniency instrument**

| outcome: | 0/1, =1 if patent granted | | predicted pr(patent grant) |
|---|---|---|---|
| | (1) | | (2) |
| **Panel A: Full application-level sample** (mean = 0.3295) | | | |
| LOOM grant rate | 0.6817 | *** | -0.0031 |
| | (0.0331) | | (0.0037) |
| number of observations | 29,197 | | 29,139 |
| **Panel B: Human gene application-level sample** (mean = 0.2554) | | | |
| LOOM grant rate | 0.8655 | *** | -0.0137 |
| | (0.1850) | | (0.0334) |
| number of observations | 1,566 | | 1,562 |
| **Panel C: Human gene application-gene level sample** (mean = 0.2187) | | | |
| LOOM grant rate | 0.5630 | *** | 0.0263 |
| | (0.1997) | | (0.0422) |
| number of observations | 301,642 | | 301,483 |

*Notes*: This table presents first stage estimates and falsification test estimates, relating the probability of patent grant and the predicted probability of patent grant to our examiner leniency instrument. The sample for these regressions is constructed from patent application-level data in Panels A and B, and from patent application-gene level data in Panel C. Panel A documents estimates for our full sample of application-level data. Panel B documents estimates for our human gene sample of application-level data. Panel C documents estimates for our human gene sample of application-gene level data. The outcome variable in Column (1) is an indicator variable for whether the application has been granted a patent as of 2010. The outcome variable in Column (2) is the predicted probability of patent grant, predicted as a function of two proxy variables for patent value that are fixed at the time of application: family size and claims count. The slightly smaller sample sizes in Column (2) (relative to Column (1)) reflect a small amount of missing data on the patent value measures. All specifications condition on Art Unit-by-application year fixed effects. Heteroskedasticity robust standard errors are clustered at the examiner level. *: p<0.10; **: p<0.05; ***: p<0.01.

Table 4: **Patents and follow-on innovation on human genes by examiner leniency: Reduced form regression estimates**

| outcome: | log of follow-on innovation in 2011/12 | any follow-on innovation in 2011/12 |
|---|---|---|
| | (1) | (2) |
| **Panel A: Scientific publications** | | |
| LOOM grant rate | -0.0066 | -0.0035 |
| | (0.0141) | (0.0128) |
| mean of dependent variable | 0.1462 | 0.0894 |
| number of observations | 301,642 | 301,642 |
| **Panel B: Clinical trials** | | |
| LOOM grant rate | -0.0165 | -0.0107 |
| | (0.0217) | (0.0137) |
| mean of dependent variable | 0.2581 | 0.0505 |
| number of observations | 301,642 | 301,642 |
| **Panel C: Diagnostic test** | | |
| LOOM grant rate | - | -0.0125 |
| | | (0.0140) |
| mean of dependent variable | - | 0.0928 |
| number of observations | - | 301,642 |
| **Panel D: Standardized treatment effect** | | |
| LOOM grant rate | -0.0340 | -0.0337 |
| | (0.0503) | (0.0416) |
| number of observations | 603,284 | 904,926 |

*Notes*: This table presents reduced form estimates, relating follow-on innovation to our examiner leniency instrument. The sample for these regressions is constructed from application-gene-level data, and includes patent application-gene-level observations in our human gene sample (N=301,642). Each coefficient is from a separate regression. The standardized treatment effect estimates in Panel D are from a stacked regression, as described in the text (hence the difference in the number of observations). Estimates are from ordinary-least-squares models and condition on Art Unit-by-application year fixed effects. In Column (1), we add one to all outcome variables to include observations with no observed follow-on innovation. Heteroskedasticity robust standard errors are clustered at the examiner level. *: p<0.10; **: p<0.05; ***: p<0.01.

Table 5: **Patents and follow-on innovation on human genes by examiner leniency: Instrumental variable regression estimates**

| outcome: | log of follow-on innovation in 2011/12 | any follow-on innovation in 2011/12 |
|---|---|---|
| | (1) | (2) |
| **Panel A: Scientific publications** | | |
| patent granted (instrumented) | -0.0118 | -0.0062 |
| | (0.0261) | (0.0233) |
| mean of dependent variable | 0.1462 | 0.0894 |
| number of observations | 301,642 | 301,642 |
| **Panel B: Clinical trials** | | |
| patent granted (instrumented) | -0.0293 | -0.0189 |
| | (0.0384) | (0.0247) |
| mean of dependent variable | 0.2581 | 0.0505 |
| number of observations | 301,642 | 301,642 |
| **Panel C: Diagnostic test** | | |
| patent granted (instrumented) | - | -0.0222 |
| | | (0.0239) |
| mean of dependent variable | - | 0.0928 |
| number of observations | - | 301,642 |
| **Panel D: Standardized treatment effect** | | |
| patent granted (instrumented) | -0.0599 | -0.0599 |
| | (0.0917) | (0.0749) |
| number of observations | 603,284 | 904,926 |

*Notes*: This table presents instrumental variable estimates, relating follow-on innovation to whether a patent application was granted a patent, instrumented by our examiner leniency instrument. The sample for these regressions is constructed from application-gene-level data, and includes patent application-gene-level observations in our human gene sample (N=301,642). Each coefficient is from a separate regression. The standardized treatment effect estimates in Panel D are from a stacked regression, as described in the text (hence the difference in the number of observations). Estimates are from ordinary-least-squares models and condition on Art Unit-by-application year fixed effects. In Column (1), we add one to all outcome variables to include observations with no observed follow-on innovation. Heteroskedasticity robust standard errors are clustered at the examiner level. *: p<0.10; **: p<0.05; ***: p<0.01.

# A Appendix: Data construction

This appendix describes our data construction in more detail. A brief background of application, publication, and patent numbers is useful before describing our data from the United States Patent and Trademark Office (USPTO).

Application numbers: The USPTO assigns patent applications application numbers, which consist of a series code and a serial number.[42] The USPTO states that these application numbers are assigned by the Office of Patent Application Processing (OPAP) immediately after mail has been opened.[43] As suggested by this process, the USPTO and other sources note that application numbers are assigned chronologically.[44] While application serial numbers are six digits, the use and length of application series codes has changed over time: in recent years series codes are two digits, but previously these codes were one digit and historically series codes were not used.[45]

Publication numbers: Traditionally, unsuccessful patent applications were not published by the USPTO. However, as part of the American Inventors Protection Act of 1999, the vast majority of patent applications filed in the US on or after 29 November 2000 are published eighteen months after the filing date. There are two exceptions. First, applications granted or abandoned before eighteen months do not appear in this sample unless the applicant chooses to ask for early publication. Lemley and Sampat (2008) estimate that about 17 percent of patents are granted before eighteen months, of which about half (46 percent) are published pre-patent grant. Second, applications pending more than eighteen months can "opt out" of publication if they do not have corresponding foreign applications, or if they have corresponding foreign applications but also have priority dates pre-dating the effective date of the law requiring publication (Lemley and Sampat, 2008).[46] If the patent application is published, then the USPTO assigns the application a publication number of the form USYEARXXXXXXX: a 2-digit country code, always US; followed by a 4-digit year (denoting year of publication); followed by a 7-digit identifier.

Patent numbers: Applications that are granted patents are assigned patent numbers. The number of characters in the patent number varies by the type of patent.[47] Utility patent numbers are six or seven digits; reissue patents start with "RE" followed by six digits;[48] plant patents start with "PP" followed by six digits; design patents start with "D" followed by seven digits; additions of improvement patents start with "AI" followed by six digits;[49] X-series patents start with "X" followed by seven digits;[50] H documents start with "H" followed by seven digits;[51] and T documents start with "T" followed by seven digits.[52]

[42] For more details, see `http://www.uspto.gov/web/offices/ac/ido/oeip/taf/filingyr.htm`.

[43] See `http://www.uspto.gov/web/offices/pac/mpep/s503.html`: *"Application numbers consisting of a series code and a serial number are assigned by the Office of Patent Application Processing (OPAP) immediately after mail has been opened. If an application is filed using the Office's electronic filing system, EFS-Web provides an Acknowledgement Receipt that contains a time and date stamp, an application number and a confirmation number."*

[44] See `http://www.uspto.gov/web/offices/ac/ido/oeip/taf/filingyr.htm` (*"In general, patent application serial numbers are assigned chronologically to patent applications filed at the U.S. Patent and Trademark Office."*) and `http://www.thomsonfilehistories.com/docs/RESOURCES_Series_Codes.pdf` (*"US patent applications consist of a 2-digit series code and a 6-digit application serial that is assigned chronologically as they are received at the USPTO."*).

[45] Note that design applications, provisional applications, and reexamination (*ex parte* and *inter partes*) applications are assigned different series codes; reissue patent application numbers follow the utility and design application structures. See `http://www.uspto.gov/web/offices/pac/mpep/s503.html` for details on these series codes.

[46] For more details, see `http://www.uspto.gov/web/offices/pac/mpep/s1120.html` and the discussion in Lemley and Sampat (2010). Most applications not published eighteen months after filing are instead published sixty months after filing.

[47] For more details, see `http://www.uspto.gov/patents/process/file/efs/guidance/infopatnum.jsp`.

[48] For more details on reissue patents, see `http://www.uspto.gov/web/offices/pac/mpep/s1401.html`.

[49] Addition of improvement patents were issued between 1838 and 1861 and covered an inventor's improvement on his or her own patented device. For more details, see §901.04 of `http://www.uspto.gov/web/offices/pac/mpep/s901.html`.

[50] X-series patents were issued between 1790 and 1836. For more details, see §901.04 of `http://www.uspto.gov/web/offices/pac/mpep/s901.html`.

[51] H documents are part of the statutory invention registration series. For more details, see `http://www.uspto.gov/web/offices/ac/ido/oeip/taf/issudate.pdf`.

[52] T documents are part of the defensive publication series. For more details, see `http://www.uspto.gov/web/offices/ac/ido/oeip/taf/issudate.pdf`.

## Data on USPTO published patent applications

### USPTO Patent Document Pre-Grant Authority files

The USPTO makes available Patent Document Authority Files.[53] The Pre-Grant Authority files contain listings of all US published applications beginning 15 March 2001 ("Pre-Grant"). Our versions of these files were downloaded on 24 March 2014 and are up to date as of February 2014.

Our version of the Pre-Grant Authority File includes 3,681,468 observations. Nearly all are utility patent applications (kind code=A1; 3,672,723; 99.76% of observations). Of the remaining 8,745 observations, 3,228 (kind code=A9; 0.09%) are corrections of published utility applications; 1,207 (kind code=A2; 0.03%) are subsequent publications of utility patent applications; and 4,310 (kind code=P1; 0.12%) are plant patent applications. One observation is denoted as the USPTO missing a copy of the relevant document images, and 1,025 observations are denoted as applications for which pre-grant publications were withdrawn. The identifier in the Pre-Grant Authority File is the publication number.

### USPTO *Cassis* BIB published patent applications data

The USPTO's *Cassis* BIB Applications DVD (Sampat, 2011) includes US patent applications published between 15 March 2001 and 31 December 2010. The data include 2,625,950 observations.

There are two identifiers in the *Cassis* data: publication numbers and application numbers. Publication numbers uniquely identify observations, whereas application numbers do not: 2,744 application numbers appear more than once in the dataset. Of those 2,744 application numbers, most (2,699) appear twice; 40 appear three times; and 5 appear four times. This is not unexpected, as it is known that application numbers can be published under multiple publication numbers.[54] Merging to the USPTO Pre-Grant Authority File reveals that almost all of these cases of non-unique application numbers are explained by subsequent publications of utility applications (697 observations) or by corrections of published utility applications (2,094 observations). The three remaining duplicate application numbers (8957425, 9726661, and 9728294) can be confirmed as having two application numbers on e.g. Google Patents, although the reason is not clear. We retain these duplicate application numbers in our sample in the absence of a clear reason to exclude them.

If we check the *Cassis* data against the USPTO Pre-Grant Authority File, as expected all observations in the USPTO's *Cassis* BIB Applications DVD appear in the Pre-Grant Authority File. As expected given that the *Cassis* BIB Applications DVD includes observations only through 31 December 2010 whereas the Pre-Grant Authority File includes observations through February 2014, many patent applications published between 1 January 2011 and February 2014 appear in the Pre-Grant Authority File but not in the *Cassis* BIB Applications DVD. In addition, 2,241 observations published prior to 1 January 2011 appear in the Pre-Grant Authority File but not in the *Cassis* BIB Applications DVD. These observations are roughly evenly distributed across publication years from 2001-2010; it is not clear why these observations are missing, but we are unable to include them in our analysis.

### USPTO full-text published patent application files

Google currently hosts bulk downloads of US patent applications published between 15 March 2001 to present.[55]

### USPTO PAIR data

We also use data from the USPTO PAIR (Patent Application Information Retrieval) system. The PAIR data contains the key patent application-level variables needed for our analysis (*e.g.* examiner name). Google is currently crawling the universe of patent documents in the USPTO PAIR system and posts updated versions of the data as

[53] Available at: `http://www.uspto.gov/patents/process/search/authority/index.jsp`.

[54] See, for example, WIPO Standard ST.1: `http://www.wipo.int/export/sites/www/standards/en/pdf/03-01-01.pdf`.

[55] Available at `http://www.google.com/googlebooks/uspto-patents-applications-text.html`.

they become available.[56] Not all of our patent applications were available in this Google data; we collected some additional data from a USPTO posting on ReedTech and entered data on the remaining applications manually from the USPTO PAIR website.[57]

### Thomson Innovation published USPTO patent applications data

We use the Thomson Innovation data as an additional source of data on patent applications. Specifically, the Thomson Innovation database provides information on two measures of patent value: claims count (as proposed by Lanjouw and Schankerman (2001)) and patent "family size" (as developed in Jonathan Putnam's 1996 dissertation). Below, we briefly describe these two value measures in more detail.

Claims count. We use claims count as a first proxy for the ex ante value of a patent application that is fixed at the time of patent application, as proposed by Lanjouw and Schankerman (2001). The key idea here is that patents list "claims" over specific pieces of intellectual property, so that patents with more claims may be more valuable. Past work has documented mixed empirical evidence on whether that is a valid assumption based on correlations of claims counts with other value measures.

Patent family size. We use patent "family" size as a second proxy for the ex ante value of a patent application that is fixed at the time of patent application, as proposed by Putnam (1996). A patent "family" is a group of related patents covering the same invention. Conceptually, this includes two types of patents: first, within-country family members include continuations, continuations-in-part, and divisionals; and second, foreign family members include patent applications covering the same technology in other jurisdictions. We here briefly describe each group of patents to motivate our family size measure:

- *Within-country patent families.* Within a country, patent families may include continuations, continuations-in-part, and divisionals. Because our focus is on US patent applications, we focus here on describing within-country patent families only for the US. This description summarizes material in the USPTO's *Manual of Patent Examining Procedure*.[58] A "continuation" is a subsequent application covering an invention that has already been claimed in a prior application (the "parent" application). A "continuation-in-part" is an application filed which repeats some portion of the parent application but also adds in new material not previously disclosed. A divisional application arises when an applicant divides claims in a parent application into separate patent applications. Taken together, the use of continuations, continuations-in-part, and divisionals imply that more than one patent can issue from a single original patent application. Lemley and Sampat (2008) document that among utility patent applications filed in January 2001 and published by April 2006 (a sample of 9,960 applications), 2,016 "children" (continuations, continuations-in-part, or divisionals) had been filed by April 2006: around 30% were continuations, 20% were continuations-in-part, and 40% were divisionals (an additional 10% were of indeterminable types).

- *Foreign patent families.* Patent protection is jurisdiction-specific, in the sense that a patent grant in a particular jurisdiction provides the patent assignee with a right to exclude others from making, using, offering for sale, selling, or importing the invention into that jurisdiction during the life of the patent (subject to the payment of renewal fees). Hence, for any given patent application, applicants must choose how many jurisdictions to file patent applications in, and given that there is a per-jurisdiction cost of filing we would expect patents that are perceived by the applicant as more privately valuable to be filed in a larger number of jurisdictions.[59] The first patent application is referred to as the priority application, and the filing date of the first application is referred to as the priority date; while the priority application can be filed in any

[56] Available at `http://www.google.com/googlebooks/uspto-patents-pair.html`.

[57] The ReedTech data is available at `http://patents.reedtech.com/Public-PAIR.php`; the hand data entry was done based on data available in the USPTO PAIR system, available at `http://portal.uspto.gov/pair/PublicPair`.

[58] Available at `http://www.uspto.gov/web/offices/pac/mpep/s201.html`.

[59] Multi-national routes such as applications filed with the European Patent Office or Patent Cooperation Treaty applications are intermediate steps towards filings in specific jurisdictions.

jurisdiction, Putnam (1996) argues that the transaction costs involved with foreign filings (e.g. translation of the application) generally imply that domestic filing is cheaper than foreign filing, and that most priority applications are filed in the inventor's home country. Under the Paris Convention for the Protection of Industrial Property (signed in 1883), all additional filings beyond the priority application that wish to claim priority to the priority application must occur within one year of the priority date. Putnam (1996) argues that most foreign applications — if filed — are filed near the one-year anniversary of the home country filing.

- *Commonly used measures of patent family size.* The term patent family can be used to describe different constructs: a patent family can be defined to include only documents sharing exactly the same priority or combination of priorities, or as all documents having at least one common priority, or as all documents that can be directly or indirectly linked via a priority document.[60] There are three commonly-used measures of family size: Espacenet (produced by the European Patent Office), the Derwent World Patents Index (DWPI; produced by Thomson Reuters), and INPADOC (produced by the European Patent Office).[61] Researchers tend to rely on these measures because collecting data from individual non-USPTO patent authorities would be quite time-consuming.

  1. Espacenet uses a 'simple' patent family definition which defines a patent family as all documents with exactly the same priority or combination of priorities. This family is constructed based on data covering around 90 countries and patenting authorities.[62]

  2. DWPI uses a similar patent family definition which defines a patent family as all documents with exactly the same priority or combination of priorities, but also includes non-convention equivalents (e.g. applications filed beyond the 12 months defined by the Paris Convention). This family is constructed based on data covering around 50 patent authorities and defensive publications (international technology disclosures and research disclosures). Continuations and divisionals are not included in the DWPI family definition.

  3. INPADOC defines a patent family more broadly, defining an 'extended' patent family as all documents that can be directly or indirectly linked via a priority document even if they lack a common priority. This family is constructed based on the same data as the Espacenet measure.

- *Our measure of patent family size*. For the purpose of our study, we would like to use the general concept of family size to develop a proxy for patent value that is fixed at the time the patent application is filed. Given this objective, it is clear that we should exclude continuations, continuations-in-part, and divisionals from our family size measure: these applications arise — by construction — after the filing date of the original patent application. In addition — and potentially more concerning in our specific context — the propensity for applications to develop continuations, continuations-in-part, or divisionals may differ across examiners, and hence could be affected by the examiner. We define patent family size as the number of unique countries in which the patent application was filed (as measured in the INPADOC patent family).

[60] For details and examples, see `http://www.epo.org/searching/essentials/patent-families/definitions.html`.

[61] For details on details on Espacenet, see `http://www.epo.org/searching/essentials/patent-families/espacenet.html`; for details on DWPI, see `http://www.epo.org/searching/essentials/patent-families/thomson.html`; and for details on INPADOC, see `http://www.epo.org/searching/essentials/patent-families/inpadoc.html`.

[62] For a list, see `http://documents.epo.org/projects/babylon/eponet.nsf/0/2464E1CD907399E0C12572D50031B5DD/$File/global_patent_data_coverage_0711.pdf`.

## Data on USPTO granted patents

### USPTO Patent Document Patent Grant Authority Files

The USPTO makes available Patent Document Authority Files.[63] The Patent Grant Authority Files contain listings of all published or granted patent documents ("Patent Grant") found in USPTO Patent Image databases and not included in the Pre-Grant Authority Files. Our versions of these files were downloaded on 24 March 2014 and are up to date as of February 2014.

Our version of the Patent Grant Authority File includes 9,432,761 observations. The vast majority are utility patents (first character 0; 8,656,513; 91.77% of observations). Of the remaining 776,248 observations, 699,917 (first character D; 7.42%) are design patents; 44,742 (first two characters RE; 0.47%) are reissue patents; 24,257 (first two characters PP; 0.26%) are plant patents; 2,785 (first character X or first two characters RX; 0.03%) are X-series patents; 2,289 (first character H; 0.02%) are H documents; 1,968 (first character T; 0.02%) are T documents; and 290 (first two characters AI; <0.00%) are additions of improvement patents. Following the kind code in the US Patent and Trademark Office data files (listed as I2), we code re-issued X-series patents (which start with "RX" rather than "X") as X-series patents.[64] There are 32,572 observations (0.35%) missing USPTO kind codes; 154 observations (<0.00%) denoted as the USPTO missing a copy of the relevant document images; and 32,572 observations (0.35%) denoted as applications for which pre-grant publications were withdrawn. The identifier in the Patent Grant Authority File is the patent number.

### USPTO data on granted patents

In order to observe which US patent applications are granted patents, we use a USPTO administrative dataset covering patents issued from January 1, 1975 to December 31, 2010.[65] The data include 3,995,847 observations.

There are two identifiers in this data: application number and patent number. From the patent grant year variable in this data, we can infer whether and when a patent was granted as of December 31, 2010.

If we check these data against the USPTO Patent Grant Authority File, there are 12,770 reissue patents which appear in this USPTO grant data but not in the USPTO Patent Grant Authority File (none of the reissue patents in the USPTO grant data appear in the USPTO Patent Grant Authority File). There are 5,449,684 patents in the USPTO Patent Grant Authority File which do not appear in the USPTO grant data. Given that the USPTO Patent Grant Authority File includes all patent grants whereas the USPTO grant data includes only patents issued from 1975-2010, presumably these 5,449,684 patents were granted either prior to 1975 or after 2010, but as the USPTO Patent Grant Authority File does not include information on year of grant we have no direct way to verify that.

[63] Available at: `http://www.uspto.gov/patents/process/search/authority/index.jsp`.

[64] For more details on kind codes, see `http://www.uspto.gov/patents/process/search/authority/kindcode.jsp`.

[65] Available at `http://dvn.iq.harvard.edu/dvn/dv/boffindata/faces/study/StudyPage.xhtml;jsessionid=b8224569cfb3e22f8f305fcfdb51?globalId=hdl:1902.1/16412&studyListingIndex=0_b8224569cfb3e22f8f305fcfdb51`; under the "Data & Analysis" tab, select "saved original (stata binary)" from the drop-down menu.

## Data on DNA-related USPTO published patent applications

### CAMBIA Patent Lens data

The CAMBIA Lens database provides a list of published USPTO patent applications associated with human and non-human protein and DNA sequences appearing in patent claims (Bacon et al., 2006).[66] This data construction was supported by the Ministry of Foreign Affairs of Norway through the International Rice Research Institute for CAMBIA's Patent Lens (the OS4 Initiative: Open Source, Open Science, Open Society, *Orzya sativa*).

Over the time period relevant for our analysis, US patent applications list DNA sequences in patent claims with a canonical 'sequence listing' label, e.g. SEQ ID NO:1 followed by the relevant DNA sequence. The CAMBIA Patent Lens data construction parses patent claims text for lists of SEQ ID NOs to determine which sequences are referenced in the claims. Importantly, CAMBIA makes available several versions of their data; following Jensen and Murray (2005), we focus on the dataset of nucleotide sequences (as opposed to amino acid sequences), and on the 'in-claims' subsample (as opposed to a larger dataset which includes DNA sequences listed in the text of the patent application, but not explicitly listed in the patent claims).

The CAMBIA Patent Lens data is updated over time; our version is current as of 8 February 2012. The level of observation is a patent-mRNA pair indexed by a publication/sequence number that combines the patent publication number and a mRNA sequence number. The patent publication numbers were extracted from the CAMBIA Patent Lens data using a Perl script.[67]

### Patome data

Patome annotates biological sequences in issued patents and published patent applications (Lee et al., 2007).[68] This data construction was supported by the Korean Ministry of Science and Technology (MOST).

Although the full Patome dataset contains issued patents and published patent applications from several jurisdictions — including Japan and Europe — in this paper we focus on the subsample of US published patent applications and granted patents. As in the CAMBIA Patent Lens data, the Patome data construction parses patent application texts for lists of SEQ ID NOs to determine which sequences are referenced in patent applications. Following the methodology pioneered by Jensen and Murray (2005), BLAST (Basic Local Alignment Search Tool) searches are used to compare listed sequences against a census of potential matches in order to identify regions of similarity. Using these BLAST searches, the DNA sequences are annotated with mRNA and gene identifiers (RefSeq and Entrez Gene numbers).

The Patome data includes some patent applications which do not correspond to the definition of human gene patents proposed by Jensen and Murray (2005); to follow the Jensen-Murray definition, we impose some additional sample restrictions. First, the Patome data includes sequences which appear in the text of patent applications but are not explicitly listed in patent claims; to follow the Jensen and Murray (2005) definition of gene patents, we exclude observations which do not appear in the patent claims.[69] Second, following Jensen and Murray (2005) we limit the sample to BLAST matches with an E-value of exactly zero; the goal of this conservative E-value is to prevent spurious matches. Finally, following Jensen and Murray (2005) we limit the sample to disclosed sequences that are at least 150 nucleotides in length; the motivation of this restriction is that this is the average length of one human exon and yet still small enough to capture EST sequences.

As in Jensen and Murray (2005), many patents claim multiple variants of the same gene (that is, multiple mRNA sequences corresponding to the same gene). Following their methodology, we focus on variation in patenting across human genes.

[66] Available at `http://www.patentlens.net/sequence/US_A/nt-inClaims.fsa.gz`.

[67] We are very grateful to Mohan Ramanujan for help extracting this data from FASTA format. This Perl script is available on request.

[68] Available at `http://verdi.kobic.re.kr/patome_int/`.

[69] Specifically, we merge the list of patent-mRNA numbers in the Patome data to the CAMBIA Patent Lens data, and drop observations that appear in Patome but not in the CAMBIA data; because our version of the CAMBIA data includes only patent-RNA observations listed in patent claims, this allows us to exclude Patome observations which are not explicitly listed in the claims of the patent applications.

## Data measuring innovation on the human genome

### Gene-level measures of scientific publications: OMIM data

We collect our measure of scientific research from the Online Mendelian Inheritance in Man (OMIM) database, a catalog of Mendelian traits and disorders. We use the full-text OMIM data and extract our variables of interest using a Python script.[70] One gene can be included in more than one OMIM record, and one OMIM record can involve more than one gene. We tally the total number of publications related to each gene in each year across all OMIM records related to that gene.

### Gene-level data on drug development: Pharmaprojects data

We collect data on gene-related drug development from the Pharmaprojects data.[71] According to the company Citeline, which compiles and sells the Pharmaprojects data, "There is continual two-way communication between Pharmaprojects staff and their contacts in the pharmaceutical and biotechnology industries; both to gather new data and importantly, to verify information obtained from other sources." Citeline employees gather drug information from company websites, reports, and press releases; annually every company covered in the database verifies information related to drugs in the development pipeline.

Pharmaprojects annotates a subset of the clinical trials in their data as related to specific Entrez Gene ID numbers. We construct a count of the number of clinical trials in which each gene is used as the basis for a pharmaceutical treatment compound in clinical trials in each year.

### Gene-level data on diagnostic tests: GeneTests.org data

We collect our measure of genes being used in diagnostic tests from the GeneTests.org database.[72] This data includes a laboratory directory that is a self-reported, voluntary listing of US and international laboratories offering in-house molecular genetic testing, specialized cytogenetic testing, and biochemical testing for inherited disorders. US-based laboratories listed in GeneTests.org must be certified under the Clinical Laboratory Improvement Amendment of 1988, which requires laboratories to meet quality control and proficiency testing standards; there are no such requirements for non-US-based laboratories.

We use the GeneTests.org data as of September 18, 2012, which lists OMIM numbers for which there is any genetic test available in the Genetests.org directory. As with the OMIM data above, one gene can appear in more than one OMIM record, and one OMIM record can involve more than one gene. We construct an indicator for whether a given gene is used in any diagnostic test as of 2012.

### Human genome-related covariates and crosswalks

We collect data on the two types of gene location covariates used by scientists: molecular location and cytogenetic location. Molecular location is recorded at the mRNA level; because a gene can link to more than one mRNA, we collapse this data to the gene level by taking an average of each variable (molecular location start, molecular location end, molecular location length, and orientation).[73] Cytogenetic location is recorded at the gene level, including map location variables (arm, region, band, and subband) and the chromosome.[74]

NCBI-generated crosswalks are used to link OMIM numbers indexing genotype-phenotype pairs to Entrez Gene ID numbers,[75] and to link mRNA-level RefSeq accession/version numbers to Entrez Gene ID numbers.[76]

[70] Available at `http://omim.org/downloads`.

[71] Pharmaprojects data is available for purchase through Citeline or the Pharmaprojects website: `http://www.pharmaprojects.com`.

[72] Available at `ftp://ftp.ncbi.nih.gov/pub/GeneTests/disease_OMIM.txt`.

[73] Available at `ftp://ftp.ncbi.nih.gov/refseq/release/release-catalog/release54.accession2geneid.gz`.

[74] Available at `ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/`.

[75] Available at `ftp://ftp.ncbi.nih.gov/gene/DATA/mim2gene_partial`.

[76] Available at `ftp://ftp.ncbi.nih.gov/refseq/release/release-catalog/release54.accession2geneid.gz`.

We also use an NCBI-generated crosswalk that links discontinued Entrez Gene ID numbers to current Entrez Gene ID numbers, which is useful for linking Pharmaprojects observations that list discontinued Entrez Gene ID numbers to our data.[77]

[77] Available at `ftp://ftp.ncbi.nih.gov/gene/DATA/gene_history.gz`.

# B Appendix: The USPTO patent examination process

In this appendix, we describe the USPTO patent examination process in more detail.[78]

## B.1 Overview of the USPTO patent examination process

The USPTO is responsible for determining whether inventions claimed in patent applications qualify for patentability. The uniform mandate for patentability is that inventions are patent-eligible (35 U.S.C. §101), novel (35 U.S.C. §102), non-obvious (35 U.S.C. §103), useful (35 U.S.C. §101), and the text of the application satisfies the disclosure requirement (35 U.S.C. §112).

Patent applications include a written description of the invention (the "specification"), declarations of what the application covers ("claims"), and a description of so-called prior art — ideas embodied in prior patents, prior publications, and other sources — that is known to the inventor and relevant to patentability. Once a patent application is received, as long as it satisfies a series of pre-examination formalities the Office of Patent Application Processing will assign it an application number, as well as a patent class and subclass.[79] These classifications, in turn, determine — based on a concordance between classes/subclasses and Art Units — the Art Unit to which the application is assigned, where Art Units are specialized groups of patent examiners that work on related subject matter.[80] Once assigned to an Art Unit, a supervisory patent examiner (SPE) then refines the patent classification if it is incorrect. In some cases, this means the application needs to be re-assigned to another Art Unit, though that is thought to be rare.

The SPE then assigns the application to a patent examiner for review (via a process described in more detail below). While the patent application process up to this point is quite structured, from this point forward substantial discretion is left in the hands of individual examiners. In particular, the examiner is responsible for determining which — if any — claims in the application are patentable in light of prior art disclosed by the applicant as well as prior art found through the examiner's own search. To give a sense of the time involved in the patent examination process, examiners have been estimated to spend an average of eighteen hours working on a given application (Lemley, 2001), although the process of evaluating a patent application takes several years.

As background for our empirical analysis, it is useful to outline the key steps in the examination process that occur from this point forward. The examiner sends a "first action" letter back to the applicant with an initial decision on the patent application. This initial decision can be an allowance, or (more commonly) a "non-final rejection." For non-final rejections, applicants have a fixed length of time (usually six months) during which to revise the application. After receiving the applicant's response, the examiner can then allow the application, negotiate minor changes, or send a second rejection. Counterintuitively, rejections in or after this second round are called "final rejections" but are in fact not final: there can be subsequent rounds of negotiation, and the only final resolution to a patent application that is not granted a patent is abandonment by the applicant (Lemley and Sampat, 2008).

## B.2 Within-Art Unit assignment of patent applications to patent examiners

The process outlined above clarifies that the assignment of patent examiners to applications is a function of at least two factors: first, the Art Unit to which the application is assigned; and second, the year the application is filed,

[78] The discussion in this appendix draws heavily on Cockburn, Kortum and Stern (2003), Lemley and Sampat (2012), and GAO (2005), among other sources referenced in the text.

[79] There are currently over 450 patent classes; the most common class in our sample is 435 (Chemistry: molecular biology and microbiology). There are currently more than 150,000 subclasses, which may be repeated among classes. For more details, see `http://www.uspto.gov/patents/resources/classification/overview.pdf`.

[80] There are over 300 Art Units; see `http://www.uspto.gov/patents/resources/classification/art/index.jsp`. For the current version of the class/subclass-to-Art Unit concordance, see `http://www.uspto.gov/patents/resources/classification/caau.pdf`. The main Art Units in our sample are from the 1600 group (Biotechnology and Organic Chemistry).

given that the group of examiners in an Art Unit will vary over time. In this section, we discuss how — within an Art Unit in a given application year — patent applications are assigned to patent examiners.

The USPTO does not publish rules regarding the assignment of applications within Art Units to particular examiners. Given this absence of formal written rules, Lemley and Sampat (2012) conducted written interviews with roughly two dozen current and former patent examiners and supervisory patent examiners to inquire about the assignment process. While the results of these interviews suggested that there is not a single "standard" assignment procedure that is uniformly applied in all Art Units, these interviews revealed no evidence of deliberate selection or assignment of applications to examiners on the basis of characteristics of applications other than those observed in standard USPTO datasets. In some Art Units, supervisors reported assigning applications to examiners based on the last digit of the application number; because application numbers are assigned sequentially in the central USPTO office, this assignment system — while not purposefully random — would be functionally equivalent to random assignment for the purposes of this study. In other Art Units, supervisors reported placing the applications on master dockets based on patent classes and subclasses, with examiners specializing in those classes (or subclasses) being automatically assigned the oldest application from the relevant pool when requesting a new application. Consistent with what we would expect given these types of assignment mechanisms, Lemley and Sampat present empirical evidence that observable characteristics of patent applications are uncorrelated with characteristics such as the experience of the examiner to which the application was assigned. Unfortunately, we do not have information on the specific set of assignment processes used by the Art Units most common in our sample over the relevant time period.[81] In the absence of such information, we rely on the interviews in Lemley and Sampat (2012) as a guide to designing our empirical specifications.

[81] We have tried, unsuccessfully, to track down individuals who were supervisory patent examiners in the Art Units most common in our sample over the relevant time period.

## C Appendix: Additional background on *AMP v. Myriad* case

This appendix provides some additional background information on the recent *AMP v. Myriad* case.

The private firm Myriad Genetics was granted patent rights on human genes correlated with risks of breast and ovarian cancer. In 2009, the American Civil Liberties Union (ACLU) and the Public Patent Foundation filed suit against Myriad, arguing that many of Myriad's patent claims were invalid on the basis that DNA should not be patentable. One technical detail that is critical to understanding the *AMP v. Myriad* case is that two types of nucleotide sequences were at issue: naturally occurring genomic DNA (gDNA), and complementary or cDNA, which is produced in a laboratory using gDNA as a template. After a series of lower court decisions, in June 2013 the US Supreme Court issued a unanimous ruling drawing a distinction between these two types of sequences: "A naturally occurring DNA segment is a product of nature and not patent eligible...but cDNA is patent eligible because it is not naturally occurring."[82]

The question of whether DNA is patent eligible may at first blush seem very far removed from the economics of gene patents. Yet in fact, the US Supreme Court decision was made in part on the basis of the research question examined in this paper: whether patents on human genes would impede follow-on innovation. A brief background on patent eligibility is helpful in clarifying this point. The patent eligibility criteria set out in the US Constitution (35 U.S.C. §101) has long been interpreted to exclude laws of nature, natural phenomena, and abstract ideas from patent eligibility. The *AMP v. Myriad* decision followed this precedent, arguing that "[g]roundbreaking, innovative, or even brilliant" discoveries of natural phenomena should be patent-ineligible, because patents "would 'tie up' the use of such tools and thereby inhibit future innovation premised upon them." As discussed by Rai and Cook-Deegan (2013), the Court decision essentially aimed to draw a line between patent-eligible and patent-ineligible discoveries based on the "delicate balance" between patents prospectively creating incentives for innovation and patent claims blocking follow-on innovation. In the end, the Court drew this line by ruling naturally occurring DNA patent-ineligible, and non-naturally occurring cDNA patent-eligible.

Numerous legal scholars have argued that the distinction between DNA and cDNA is "puzzling and contradictory" (Burk, 2013) given that "both isolated sequences and cDNA...have identical informational content for purposes of protein coding" (Golden and Sage, 2013); in interviews, patent attorneys expressed similar confusion (Harrison, 2013). A recent analysis of gene patent claims by Holman (2012) concluded that most human gene patents claimed cDNA, and would thus be unaffected by the Court ruling.

[82] The earlier decisions were a 2010 ruling by the US District Court for the Southern District of New York (see `http://www.pubpat.org/assets/files/brca/brcasjgranted.pdf`) and a 2011 ruling by the US Court of Appeals for the Federal Circuit (see `https://www.aclu.org/files/assets/10-1406.pdf`); a subsequent re-hearing of the case by the US Court of Appeals at the request of the US Supreme Court did not substantively change this decision.

# D Appendix: Additional results

Table D.1: **Patents and follow-on innovation on human genes claimed in accepted/rejected patent applications: Robustness**

| model: | OLS | OLS | OLS | OLS | OLS | OLS | OLS | OLS | QML-Poisson | OLS |
|---|---|---|---|---|---|---|---|---|---|---|
| outcome: | log(y+1) | log(y+0.01) | log(y+0.1) | log(y+10) | **1**(y>=1) | **1**(y>=3) | **1**(y>=9) | **1**(y>=12) | y | y |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| **Panel A: Scientific publications** | | | | | | | | | | |
| patent granted*2011 | 0.0019 | -0.0019 | 0.0009 | 0.0006 | -0.0013 | 0.0041 | -0.0003 | -0.0002 | 0.0359 | 0.0081 |
| | (0.0060) | (0.0278) | (0.0159) | (0.0012) | (0.0054) | (0.0028) | (0.0006) | (0.0004) | (0.0677) | (0.0154) |
| mean of dependent variable | | 0.2238 (mean in levels) | | | 0.1094 | 0.0252 | 0.0014 | 0.0006 | 0.2238 | 0.2238 |
| number of observations | 15,530 | 15,530 | 15,530 | 15,530 | 15,530 | 15,530 | 15,530 | 15,530 | 15,530 | 15,530 |
| patent granted*2012 | 0.0013 | -0.0023 | 0.0002 | 0.0005 | -0.0012 | 0.0022 | 0.0003 | 0.0001 | 0.0559 | 0.0065 |
| | (0.0044) | (0.0205) | (0.0117) | (0.0009) | (0.0040) | (0.0019) | (0.0005) | (0.0003) | (0.0977) | (0.0115) |
| mean of dependent variable | | 0.1151 (mean in levels) | | | 0.0566 | 0.0111 | 0.0006 | 0.0003 | 0.1151 | 0.1151 |
| number of observations | 15,530 | 15,530 | 15,530 | 15,530 | 15,530 | 15,530 | 15,530 | 15,530 | 15,530 | 15,530 |
| **Panel B: Clinical trials** | | | | | | | | | | |
| patent granted*2011 | 0.0008 | -0.0034 | -0.0007 | 0.0006 | -0.0012 | 0.0012 | -0.0005 | 0.0002 | 0.1088 | 0.0602 |
| | (0.0080) | (0.0255) | (0.0161) | (0.0028) | (0.0043) | (0.0031) | (0.0020) | (0.0017) | (0.1755) | (0.1001) |
| mean of dependent variable | | 0.5424 (mean in levels) | | | 0.0657 | 0.0337 | 0.0135 | 0.0102 | 0.5424 | 0.5424 |
| number of observations | 15,530 | 15,530 | 15,530 | 15,530 | 15,530 | 15,530 | 15,530 | 15,530 | 15,530 | 15,530 |

*Notes*: This table estimates differences in follow-on innovation on genes claimed in at least one granted patent relative to genes claimed in at least one patent application but never in a granted patent. The sample for these regressions is constructed from gene-level data, and includes genes claimed in at least one patent application in our USPTO human gene patent application sample (N=15,530). Each coefficient is from a separate regression: each column is a different specification, and each row is the relevant outcome variable measured in a given year. In all but column (9), estimates are from ordinary least squares models; in column (9) estimates are from a quasi-maximum likelihood Poisson model. Heteroskedasticity robust standard errors. *: p<0.10; **: p<0.05; ***: p<0.01.

Table D.2: **Robustness of first stage estimates: Patent grants by examiner leniency instrument**

| | | | | | | |
|---|---|---|---|---|---|---|
| outcome: | 0/1, =1 if patent granted | | 0/1, =1 if patent granted | | 0/1, =1 if patent granted | |
| fixed effects included: | Art Unit - by - Application Year | | Art Unit - by - Application Year | | Art Unit - by - Application Year - by - Class - by - Subclass | |
| | (1) | | (2) | | (3) | |
| **Panel A: Full application-level sample** (mean = 0.3295) | | | | | | |
| LOOM grant rate | 0.6817 | *** | 0.7083 | *** | 0.6753 | *** |
| | (0.0331) | | (0.0450) | | (0.0448) | |
| number of observations | 29,197 | | 15,253 | | 15,253 | |
| **Panel B: Human gene application-level sample** (mean = 0.2554) | | | | | | |
| LOOM grant rate | 0.8655 | *** | 0.6621 | *** | 0.6142 | ** |
| | (0.1850) | | (0.2431) | | (0.2484) | |
| number of observations | 1,566 | | 1,148 | | 1,148 | |
| **Panel C: Human gene application-gene level sample** (mean = 0.2187) | | | | | | |
| LOOM grant rate | 0.5630 | *** | 0.5620 | *** | 0.4721 | ** |
| | (0.1997) | | (0.2000) | | (0.2266) | |
| number of observations | 301,642 | | 301,523 | | 301,523 | |

*Notes*: This table presents robustness checks on our first stage estimates, relating the probability of patent grant to our examiner leniency instrument. The sample for these regressions is constructed from patent application-level data in Panels A and B, and from patent application-gene level data in Panel C. Panel A documents estimates for our full sample of application-level data. Panel B documents estimates for our human gene sample of application-level data. Panel C documents estimates for our human gene sample of application-gene level data. Column (1) documents estimates that condition on Art Unit-by-application year fixed effects. Column (3) replaces the Art Unit-by-application year fixed effects with Art Unit-by-application year-by-class-by-subclass fixed effects, estimated on the subsample of data meeting our sample restrictions. For ease of comparability, Column (2) documents estimates that condition on Art Unit-by-application year fixed effects but use the same sample as in Column (3). Heteroskedasticity robust standard errors are clustered at the examiner level. *: p<0.10; **: p<0.05; ***: p<0.01.

Table D.3: **Predicted probability of patent grant**

| outcome: | 0/1, =1 if patent granted | |
|---|---|---|
| | (1) | |
| **Panel A: Full application-level sample** (mean = 0.3295) | | |
| family size | 0.0027 | *** |
| | (0.0008) | |
| claims count | 0.0002 | ** |
| | (0.0001) | |
| number of observations | 30,183 | |
| **Panel B: Human gene application-level sample** (mean = 0.2554) | | |
| family size | 0.0019 | |
| | (0.0014) | |
| claims count | 0.0018 | *** |
| | (0.0005) | |
| number of observations | 1,578 | |
| **Panel C: Human gene application-gene level sample** (mean = 0.2187) | | |
| family size | 0.0046 | *** |
| | (0.0013) | |
| claims count | 0.0021 | *** |
| | (0.0006) | |
| number of observations | 303,914 | |

*Notes*: This table assesses whether patent application-level measures of value are predictive of patent grant. The sample is constructed from patent application-level data in Panels A and B, and from patent application-gene level data in Panel C. The reported coefficients are from a regression of an indicator for the patent application being granted a patent on two proxy variables for patent value that are fixed at the time of application: family size and claims count. Panel A documents estimates for our full sample of application-level data. Panel B documents estimates for our human gene sample of application-level data. Panel C documents estimates for our human gene sample of application-gene level data. The slightly larger sample sizes (relative to other tables) reflect that we do not impose our instrument-related sample restrictions. Heteroskedasticity robust standard errors are clustered at the examiner level. *: p<0.10; **: p<0.05; ***: p<0.01.

Table D.4: **Patents and follow-on innovation on human genes by examiner leniency: Robustness of reduced form regression estimates**

| model: | OLS | OLS | OLS | OLS | OLS | OLS | OLS | OLS | QML-Poisson | OLS |
|---|---|---|---|---|---|---|---|---|---|---|
| outcome: | log(y+1) | log(y+0.01) | log(y+0.1) | log(y+10) | **1**(y>=1) | **1**(y>=3) | **1**(y>=9) | **1**(y>=12) | y | y |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| **Panel A: Scientific publications** | | | | | | | | | | |
| LOOM grant rate*2011 | -0.0066 | -0.0227 | -0.0146 | -0.0014 | -0.0035 | -0.0038 | -0.0008 | -0.0001 | - | -0.0179 |
| | (0.0141) | (0.0666) | (0.0380) | (0.0026) | (0.0128) | (0.0044) | (0.0005) | (0.0001) | - | (0.0312) |
| mean of dependent variable | | 0.1462 (mean in levels) | | | 0.0894 | 0.0140 | 0.0004 | 0.0001 | 0.1462 | 0.1462 |
| number of observations | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 |
| LOOM grant rate*2012 | -0.0035 | -0.0232 | -0.0120 | -0.0006 | -0.0050 | 0.0008 | -0.0003 | -0.0002 ** | - | -0.0078 |
| | (0.0116) | (0.0643) | (0.0351) | (0.0019) | (0.0131) | (0.0025) | (0.0002) | (0.0001) | - | (0.0217) |
| mean of dependent variable | | 0.0720 (mean in levels) | | | 0.0464 | 0.0046 | 0.0002 | 0.0001 | 0.0720 | 0.0720 |
| number of observations | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 |
| **Panel B: Clinical trials** | | | | | | | | | | |
| LOOM grant rate*2011 | -0.0165 | -0.0615 | -0.0374 | -0.0045 | -0.0107 | -0.0044 | -0.0022 | -0.0018 | - | -0.0397 |
| | (0.0217) | (0.0802) | (0.0493) | (0.0057) | (0.0137) | (0.0082) | (0.0032) | (0.0036) | - | (0.1148) |
| mean of dependent variable | | 0.2581 (mean in levels) | | | 0.0504 | 0.0224 | 0.0086 | 0.0060 | 0.2581 | 0.2581 |
| number of observations | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 |

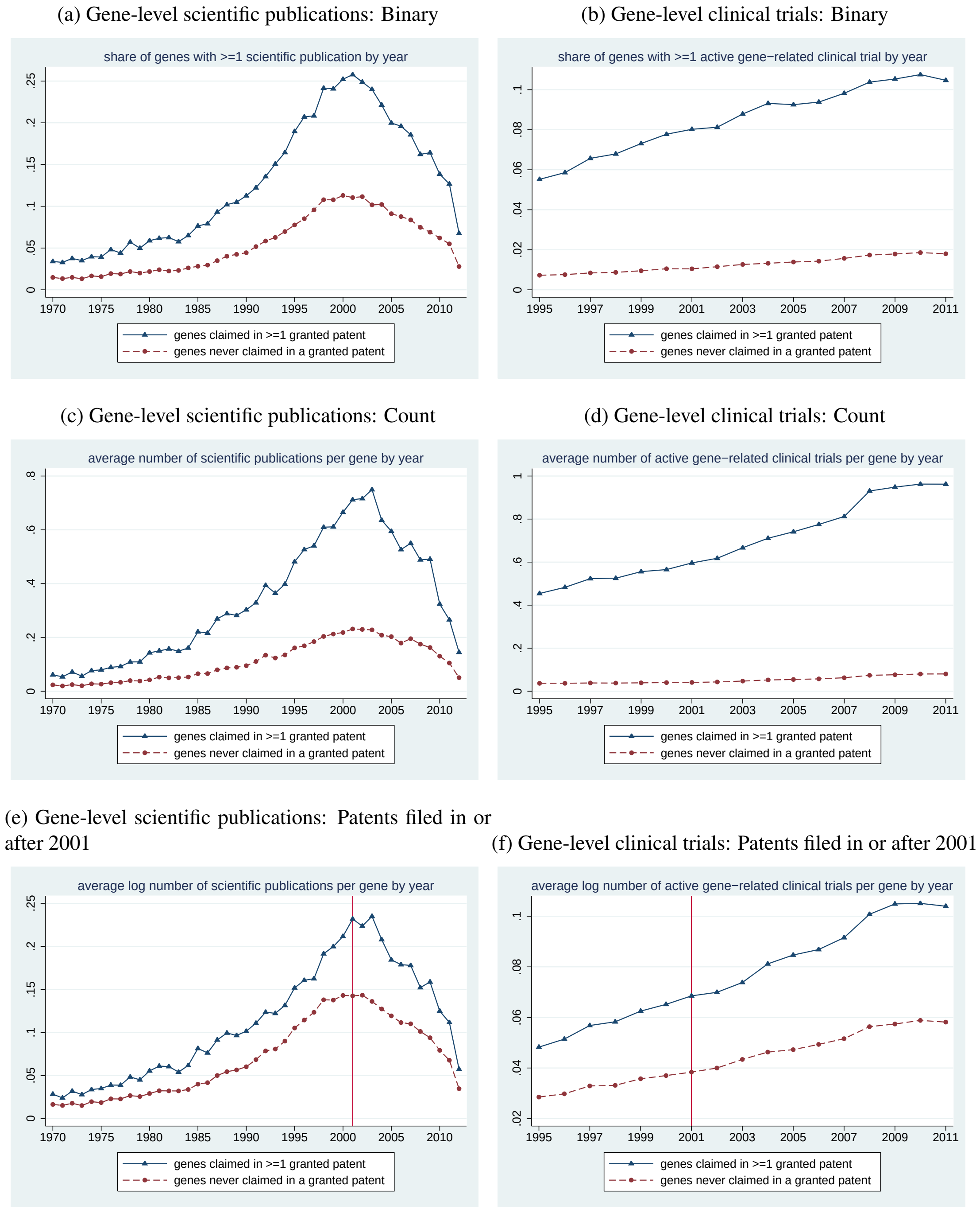
*Notes*: This table presents reduced form estimates, relating follow-on innovation to our examiner leniency instrument. The sample for these regressions is constructed from application-gene-level data, and includes patent application-gene-level observations in our human gene sample (N=301,642). Each coefficient is from a separate regression: each column is a different specification, and each row is the relevant outcome variable measured in a given year. In all but column (9), estimates are from ordinary least squares models; in column (9) estimates are to be added from a quasi-maximum likelihood Poisson model. Estimates condition on Art Unit-by-application year fixed effects. Heteroskedasticity robust standard errors are clustered at the examiner level. *: p<0.10; **: p<0.05; ***: p<0.01.

Table D.5: **Patents and follow-on innovation on human genes by examiner leniency: Robustness of instrumental variable regression estimates**

| model: | OLS | OLS | OLS | OLS | OLS | OLS | OLS | OLS | QML-Poisson | OLS |
|---|---|---|---|---|---|---|---|---|---|---|
| outcome: | log(y+1) | log(y+0.01) | log(y+0.1) | log(y+10) | **1**(y>=1) | **1**(y>=3) | **1**(y>=9) | **1**(y>=12) | y | y |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| **Panel A: Scientific publications** | | | | | | | | | | |
| patent granted (instrumented)*2011 | -0.0118 | -0.0403 | -0.0259 | -0.0026 | -0.0062 | -0.0068 | -0.0014 | -0.0003 | - | -0.0318 |
| | (0.0261) | (0.1213) | (0.0695) | (0.0049) | (0.0233) | (0.0084) | (0.0010) | (0.0003) | - | (0.0579) |
| mean of dependent variable | | 0.1462 (mean in levels) | | | 0.0894 | 0.0140 | 0.0004 | 0.0001 | 0.1462 | 0.1462 |
| number of observations | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 |
| patent granted (instrumented)*2012 | -0.0063 | -0.0412 | -0.0214 | -0.0010 | -0.0089 | 0.0014 | -0.0005 | -0.0004 | - | -0.0138 |
| | (0.0204) | (0.1129) | (0.0615) | (0.0034) | (0.0231) | (0.0046) | (0.0005) | (0.0003) | - | (0.0377) |
| mean of dependent variable | | 0.0720 (mean in levels) | | | 0.0464 | 0.0046 | 0.0002 | 0.0001 | 0.0720 | 0.0720 |
| number of observations | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 |
| **Panel B: Clinical trials** | | | | | | | | | | |
| patent granted (instrumented)*2011 | -0.0293 | -0.1092 | -0.0665 | -0.0080 | -0.0189 | -0.0079 | -0.0039 | -0.0033 | - | -0.0705 |
| | (0.0384) | (0.1439) | (0.0880) | (0.0098) | (0.0247) | (0.0143) | (0.0055) | (0.0062) | - | (0.1984) |
| mean of dependent variable | | 0.2581 (mean in levels) | | | 0.0504 | 0.0224 | 0.0086 | 0.0060 | 0.2581 | 0.2581 |
| number of observations | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 | 301,642 |

*Notes*: This table presents instrumental variable estimates, relating follow-on innovation to whether a patent application was granted a patent, instrumented by our examiner leniency instrument. The sample for these regressions is constructed from application-gene-level data, and includes patent application-gene-level observations in our human gene sample (N=301,642). Each coefficient is from a separate regression: each column is a different specification, and each row is the relevant outcome variable measured in a given year. In all but column (9), estimates are from ordinary least squares models; in column (9) estimates are to be added from a quasi-maximum likelihood Poisson model. Estimates condition on Art Unit-by-application year fixed effects. Heteroskedasticity robust standard errors are clustered at the examiner level. *: p<0.10; **: p<0.05; ***: p<0.01.

Figure D.1: **Follow-on innovation on patented and non-patented human genes: Robustness**



(a) Gene-level scientific publications: Binary

(b) Gene-level clinical trials: Binary

(c) Gene-level scientific publications: Count

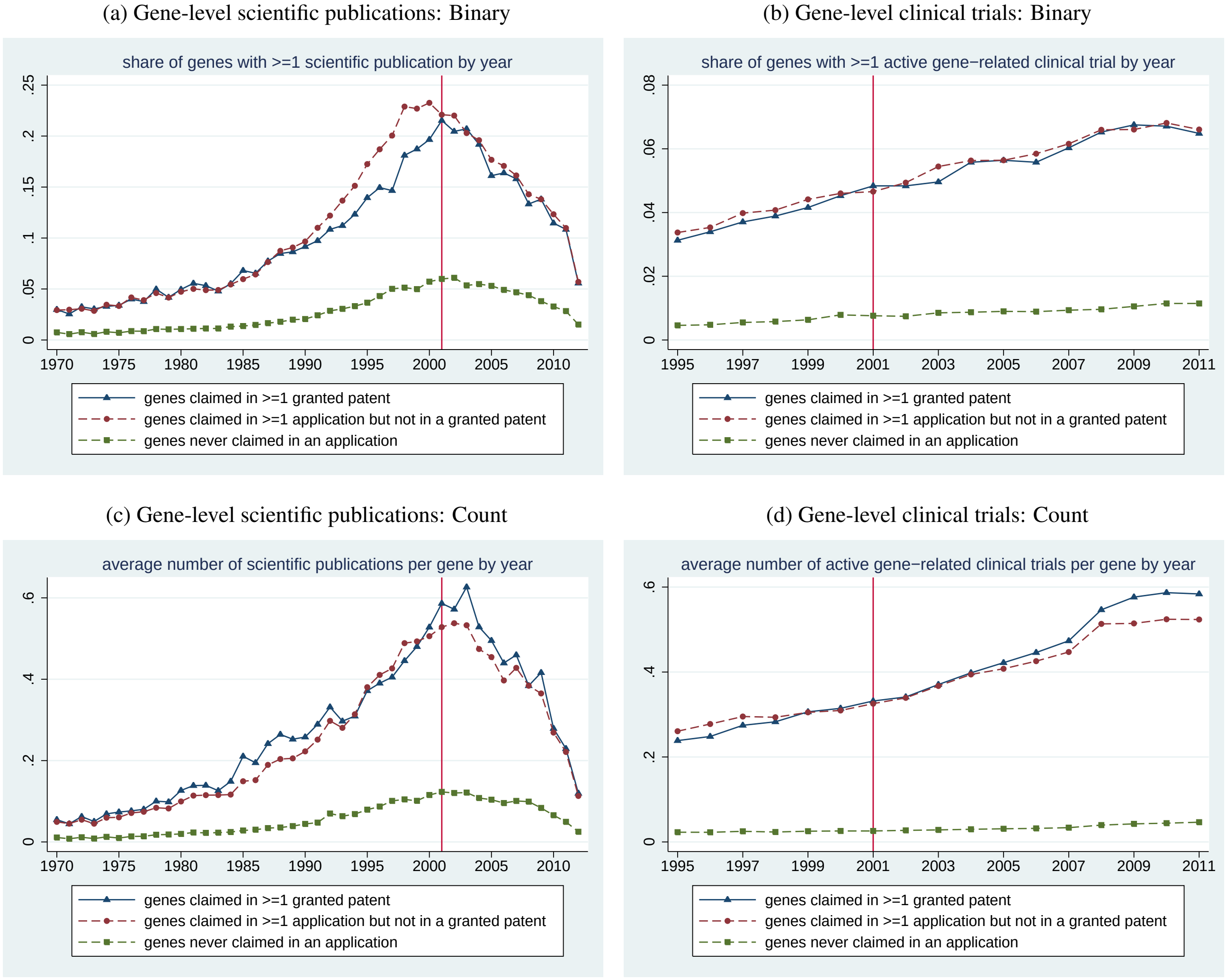
(d) Gene-level clinical trials: Count

(e) Gene-level scientific publications: Patents filed in or after 2001

(f) Gene-level clinical trials: Patents filed in or after 2001

*Notes*: This figure plots trends in follow-on innovation by year separately for genes that ever receive a patent, and for genes that never receive a patent. The figure is constructed from gene-level data. The right-hand side panels use gene-level scientific publications as a measure of follow-on innovation, and the left-hand-side panel uses gene-level clinical trials as a measure of follow-on innovation. The first row of figures plots the average of indicator variables for any follow-on innovation by year; the second row of figures plots the average number of each follow-on measure by year; and the third row of figures limits the sample of patents to patents filed in or after 2001. In the third row of figures, the vertical line in the calendar year 2001 denotes that because this figure focuses on patents that were filed in or after 2001, all years prior to 2001 can be considered a pre-period and used to estimate the selection of genes into patenting based on pre-filing measures of scientific research (publications) and commercialization (clinical trials). In Panels (e) and (f), we add one to both outcome variables in order to include observations with no observed follow-on innovation.

Figure D.2: **Patents and follow-on innovation on human genes claimed in accepted/rejected patent applications: Robustness**



(a) Gene-level scientific publications: Binary

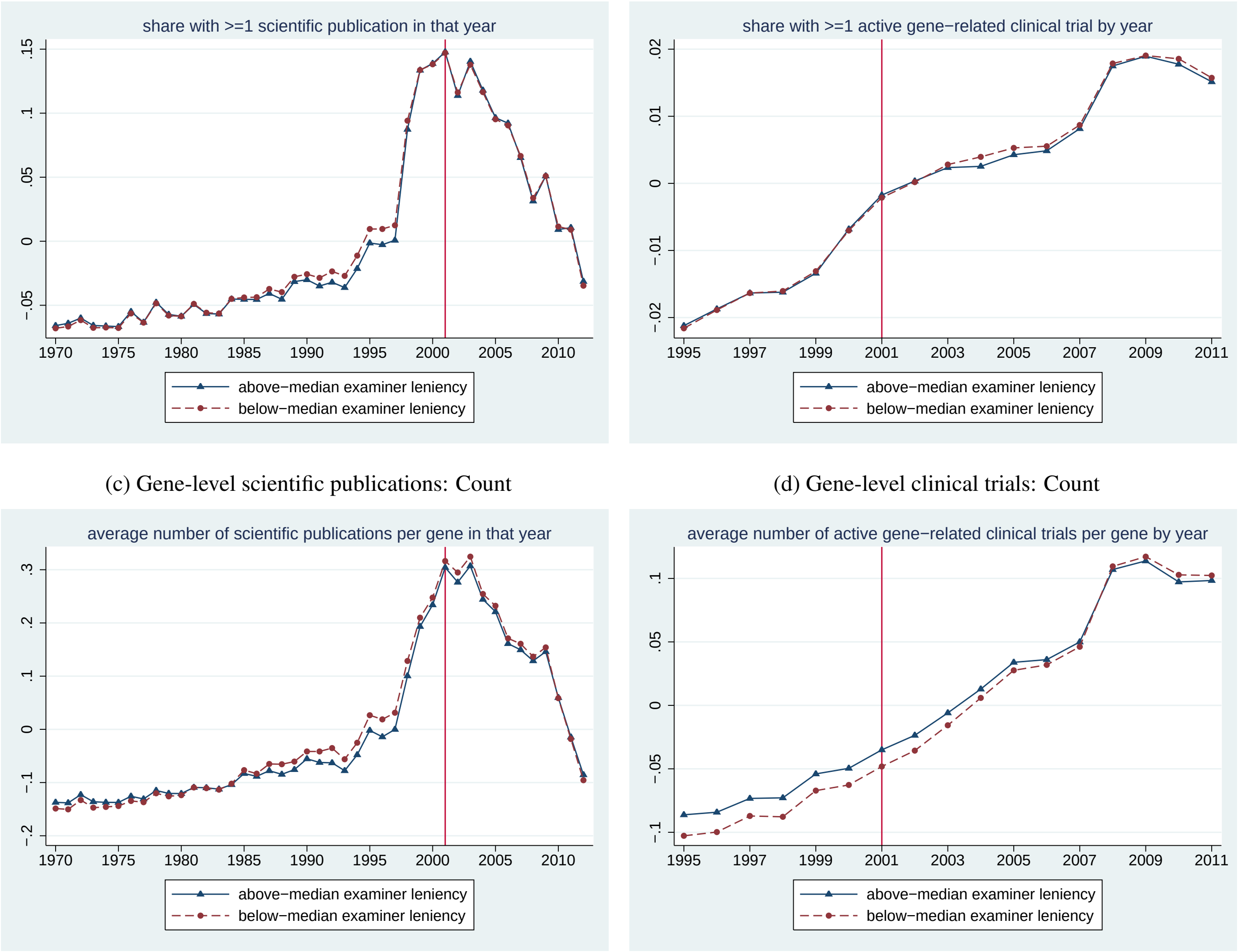
(b) Gene-level clinical trials: Binary

(c) Gene-level scientific publications: Count

(d) Gene-level clinical trials: Count

*Notes*: This figure plots trends in patenting and follow-on innovation by year separately for three groups of genes: genes claimed in at least one granted patent; genes claimed in at least one patent application but never in a granted patent; and genes never claimed in a patent application. The figure is constructed from gene-level data. The right-hand side panels use gene-level scientific publications as a measure of follow-on innovation, and the left-hand-side panel uses gene-level clinical trials as a measure of follow-on innovation. The first row of figures plots the average of indicator variables for any follow-on innovation by year, and the second row of figures plots the average number of each follow-on measure by year.

Figure D.3: **Patents and follow-on innovation on human genes assigned to above/below leniency examiners: Robustness**



(a) Gene-level scientific publications: Binary

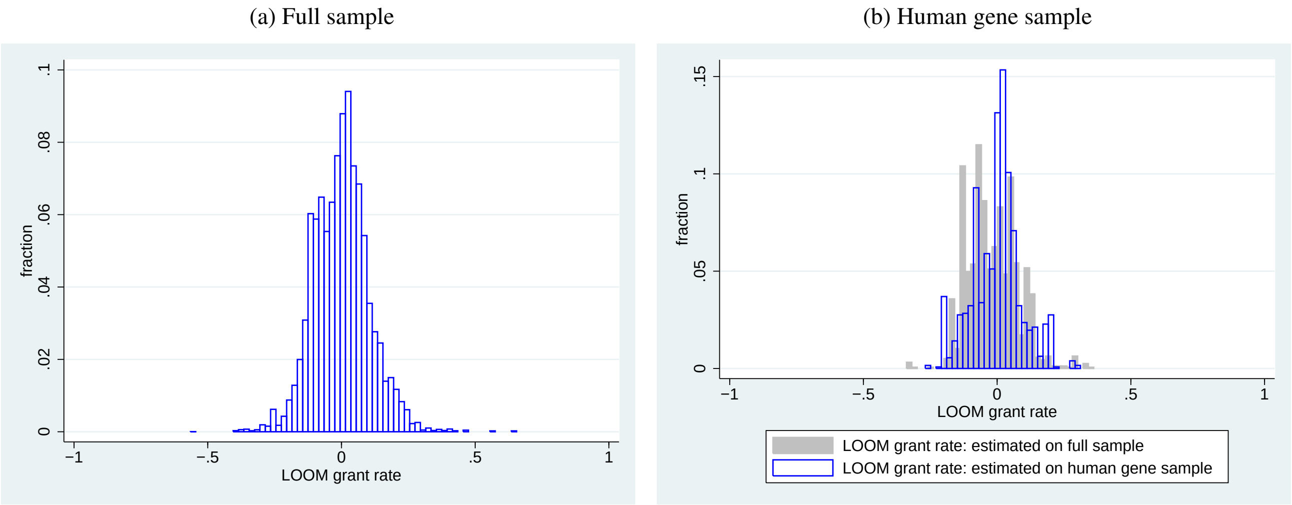
(b) Gene-level clinical trials: Binary

(c) Gene-level scientific publications: Count

(d) Gene-level clinical trials: Count

*Notes*: This figure plots trends in patenting and follow-on innovation by year separately for two groups of genes: genes claimed in patent applications that were assigned to above-median leniency patent examiners, and genes claimed in patent applications that were assigned to below-leniency patent examiners. The figure is constructed from patent application-gene level data. The right-hand side panels use gene-level scientific publications as a measure of follow-on innovation, and the left-hand-side panel uses gene-level clinical trials as a measure of follow-on innovation. The first row of figures plots the average of indicator variables for any follow-on innovation by year, and the second row of figures plots the average number of each follow-on measure by year.

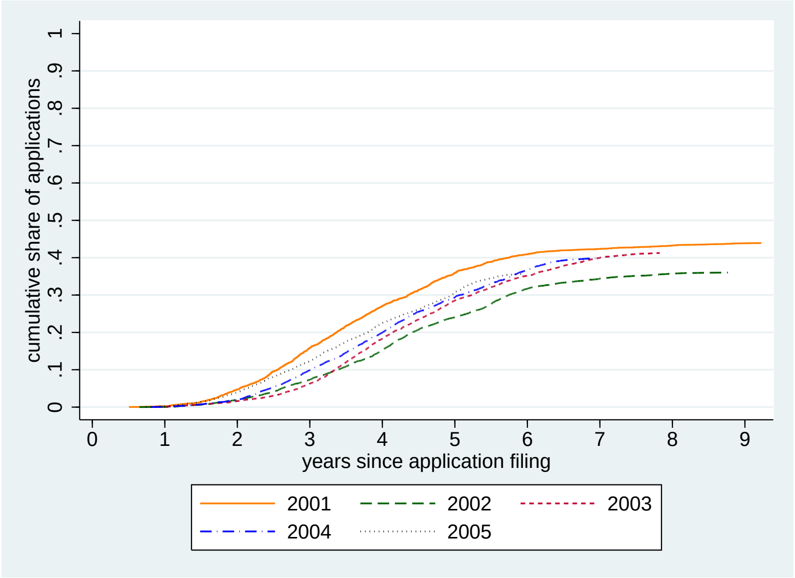
Figure D.4: **Histogram of examiner leniency instrument** ($Z_{itja}$)


(a) Full sample

(b) Human gene sample

*Notes*: This figure plots a histogram of our patent application-level observations by our leave-one-out-mean (LOOM) grant rate instrument, residualized by Art Unit-by-application year fixed effects. Panel (a) plots this variable for the set of patent applications in our full sample. Panel (b) plots two versions of this variable — one estimated on the full sample, and one estimated on the human gene sample — for the set of patent applications in our human gene sample.

Figure D.5: **Cumulative share of applications granted a patent, by years since filing**



*Notes*: This figure plots — for our first five cohorts of patent applications, filed in 2001-2005 — the distribution of time between the application filing and the patent grant date, for the set of granted patents in our full sample.