One product, many patents: Imperfect intellectual property rights in the pharmaceutical industry

Job Market Paper

CHARU N. GUPTA*
The Wharton School, University of Pennsylvania

November 25, 2020
Latest version here

Abstract

Economists’ standard notion of intellectual property rights considers a single patent per product, with a clearly defined scope, certain enforcement, and a fixed term of monopoly protection. Yet common across industries are “imperfect” intellectual property rights: More than one patent may cover a single product, with the scope and enforcement of each uncertain, contributing to an indeterminate period of monopoly protection. Using data on the pharmaceutical industry, I systematically document the presence of imperfect intellectual property rights and provide the first evidence on the extent to which they impact competition. In a sample of novel drugs, I show that roughly 70 percent of drugs are covered by multiple intellectual property rights. I offer evidence on two mechanisms by which the accumulation of such rights for a single drug may delay generic entry: by introducing a binding later patent expiration and by increasing uncertainty in the scope and enforceability of remaining patents. In an instrumental variables analysis, I determine that the accumulation of patents for a single drug product delays generic entry by over 2 years per drug (amounting to 17.5 percent of mean monopoly life), well beyond the expiration of the drug’s initial molecule patent. This research suggests large consequences for consumer welfare in terms of drug pricing and offers an important nuance for future work on optimal patent policy and innovation—that intellectual property rights are less rigid than we typically assume.

*guptac@wharton.upenn.edu. I am grateful to my advisors Matthew Grennan, Daniel Levinthal, and Noam Yuchtman for their continual guidance and support. I thank Francesco Agostinelli, Juan Pablo Atal, Aymeric Bellon, Laura Boudreau, Stuart Craig, Andre Victor Doherty Luduvice, Hanming Fang, Mallick Hossain, Kathleen Hui, Hyo Kang, Darius Lakdawalla, Aviv Nevo, Alex Olssen, Oren Reshef, Lindsey Patterson, Stefano Pietrosanti, Alejandro Sanchez Becerra, Carlos Segura-Rodriguez, Amanda Starc, Chad Stecher, Filipe Stona, and participants at the CCC Doctoral Conference, Mack Innovation Doctoral Association Seminar, NBER Aging and Health Trainee Seminar, Penn Empirical Micro Lunch, and Penn Health Economics Workshop for helpful comments and discussions. Financial support for this research was provided by the Mack Institute for Innovation Management. Any errors are my own.
1 Introduction

Standard economic theory views intellectual property (IP) rights as “well-defined” and “enforceable” (Coase, 1960; Mas-Colell et al., 1995), meaning a single patent covers a product, that patent has a certain scope, and there is certain enforcement of monopoly rights while the patent is in effect. Under these assumptions, an innovator enjoys a fixed period of monopoly protection meant to recoup its costs of research and development. Once the fixed term expires, competitors may enter the market with an imitation product. Existing theoretical literature on patent races and optimal IP policy relies on this “one-patent-one-product” model (e.g., Denicolo, 1996; Gilbert and Shapiro, 1990; Klemperer, 1990; Nordhaus, 1969, 1972; Scherer, 1972), and recent empirical evidence has highlighted the disincentives the fixed-term system creates for products requiring longer times to commercialization (Budish et al., 2015).

In reality, the assumptions of the standard model rarely hold (Bessen, 2009; Lemley and Shapiro, 2005). I define “imperfect” intellectual property rights as encompassing three possible characteristics: (1) Multiple patents may cover a single product and the (2) scope and (3) enforcement of each patent may be uncertain. These imperfections contribute to an indeterminate period of monopoly protection. For example, innovators in high-technology industries—such as software, semiconductors, information technology, and pharmaceuticals—often take out numerous patents on different aspects of the same product (Cohen et al., 2000; Noel and Schankerman, 2013). This practice can feasibly extend a product’s actual monopoly term well beyond its initial patent life. While oft-discussed among industry leaders, policy makers, and legal scholars, the prevalence of such imperfect intellectual property rights and their impact on competition and innovation has yet to be quantified.

This paper provides the first systematic evidence on the existence and importance of imperfect intellectual property rights using data from the pharmaceutical industry. In the first part of the paper, I document the presence of imperfect IP and of monopoly power beyond the initial patent term. In a sample of novel drugs, I find that 70 percent of drugs are covered by multiple intellectual property rights and that mean actual monopoly term (13.3 years) is more than 2 years longer than mean initial protection term (11.1 years); the final part of the paper determines the extent to which such extensions in monopoly life are attributable to the accumulation of intellectual property protection. In the second part, I discuss the mechanisms by which IP accumulation for a single drug may influence the timing of generic entry: by introducing a binding later IP expiration for generic entry and by increasing uncertainty in the scope and enforceability of remaining IP rights. Consistent with these mechanisms, I document non-monotonic generic entry patterns relative to IP expirations. In the third part, I use data on patent applications and an instrumental variables analysis to evaluate the causal impact such IP accumulation has on a drug’s monopoly life. This analysis corrects for endogeneity concerns stemming from branded firms’ patenting decisions and generic firms’ decisions on which markets to enter: Branded firms seek to obtain additional patents
on their most valuable drugs; these are the same drugs that may experience shorter monopoly terms due to their attractiveness to generic entrants. Note that by accumulation I am referring to the combination of the three imperfect IP characteristics (many patents per product and uncertain scope and enforcement of each). My results indicate that the accumulation of patents extends the average drug’s monopoly life by more than 2 years (equal to 17.5 percent of mean monopoly term), suggesting large potential consequences for consumer welfare in terms of drug pricing.

The pharmaceutical industry offers a noteworthy setting for the study of imperfect intellectual property rights. First, the size of the industry alone—$485 billion in the United States in 2018—makes it an important market for analysis (IQVIA, 2019). Second, debate is ongoing on the strategic use of patenting for branded drug products—referred to as “life-cycle management” within the industry and as “evergreening” by critics. Branded manufacturers often obtain additional patents beyond a drug’s initial patent for its active ingredient (the primary patent). These additional patents, called secondary patents, cover auxiliary features such as a drug’s formulation (i.e., physical form and route of administration), manufacturing process, or use in treating a particular condition. One view holds that, given the high costs of pharmaceutical research and development (R&D), these additional patents are necessary to incentivize innovation (Gaudry, 2011; Voet, 2020); critics, however, argue that this practice serves only to extend a drug’s monopoly life and maintain high drug prices (Angell, 2005; Tribble, 2019). Empirical evidence has shown that length of monopoly protection is an important determinant of innovation in the pharmaceutical sector (Budish et al., 2015), and thus, estimating the causal impact of an additional patent on monopoly life becomes a salient exercise.

Empirically, the pharmaceutical industry also offers advantages that simplify a study of imperfect intellectual property rights. First, there is a clear distinction between the primary and secondary patents, and the primary patent generally has certain enforcement. As such, the setting allows measurement of any extension in monopoly term beyond the initial (or primary) patent term. This may not be the case in other industries where a single product may not have an explicit primary patent. Further, the notion of a product (a given drug) is simpler than in other industries and there are limits to the possible scope of patents that can influence a drug’s monopoly term. The Food and Drug Administration (FDA) requires relevant patents on a drug’s substance, formulation, and approved uses to be listed in its Orange Book; these are the only patents that may extend a drug’s period of monopoly protection. With these simplifications, my estimates of imperfect intellectual property and its impact on competition are likely conservative; the imperfections may be even larger in industries where the definition of a product is more complex and involves patents on numerous component products with nebulous scope (consider, e.g., a smartphone).

Figure 1 offers a visual depiction of the differences between the standard and imperfect models of intellectual property rights under the simplifications of the pharmaceutical sector. Under the standard model, there is a fixed period of protection while the only patent is in effect. After
its expiration, competitors may choose to enter or not with an imitation product. Under the imperfect model, there is a fixed initial period of protection for the primary patent that is fully enforced and an uncertain monopoly extension afterwards, due to uncertain scope and enforcement of the secondary patents. This extension term becomes the empirical outcome of interest in the instrumental variables analysis.

As a motivating example, consider the antiretroviral drug Kaletra—used in the treatment of HIV and recently studied as a possible therapy for COVID-19 (Cao et al., 2020). Kaletra, a combination of the active ingredients lopinavir and ritonavir, is protected by 28 separate patents and 5 regulatory exclusivities—another form of IP protection granted to pharmaceuticals by the FDA, described in more detail in Section 2. The drug was launched in the U.S. in September 2002, and its primary patent expired in May 2014. Generic entry, however, did not occur until December 2016, after the expiration of several secondary patents and more than 2.5 years after the primary patent expiration. This paper aims to quantify the degree to which such extensions in a drug’s monopoly life are attributable to secondary patent accumulation.

My empirical analysis combines several sources of data for the pharmaceutical industry, including information on each drug patent and exclusivity from the FDA, approval dates for branded drugs, and corresponding launch dates for each drug’s first generic entrant. I focus on new chemical entities (NCEs) approved between 1985 and 2010 with a generic entrant. New chemical entities are those drugs with an entirely new active ingredient never before approved by the FDA (Thomas, 2017). For example, Prozac, launched in 1987 for the treatment of depression, is the first drug approved with active ingredient fluoxetine and an NCE; its product extensions with the same active ingredient—Prozac Weekly and Sarafem—are not. Focusing on NCEs with a generic entrant offers two main advantages: First, I capture drugs with a full history in the data (from branded launch through generic entry); empirically, this is necessary so that I have a measure of each drug’s actual monopoly
life.\(^1\) Second, my estimates will offer a lower bound on the extent to which secondary patenting lengthens monopoly life in that they do not include the effects of product extensions under new brand names.\(^2\)

From these sources, I construct three novel analytic datasets. The first is a drug-level dataset of all 370 NCEs meeting the above criteria, used in descriptive analyses. The second is a drug-month-level dataset following each NCE from branded launch through generic entry. Drugs are observed for 159 months on average (from branded launch through generic entry), yielding a total of 58,898 drug-month observations. I use this dataset to explore how secondary IP rights influence first generic entry timing. The third dataset is a drug-application-level dataset merging each NCE with its secondary patent applications (both granted and rejected) from the United States Patent and Trade Office (USPTO). The matched dataset, restricted to drugs that obtain secondary patent applications and have non-missing values of control variables, covers 241 NCEs and 3,021 drug-application observations and is used in the instrumental variables analysis.\(^3\)

In the first part of the paper, I provide descriptive evidence on the presence of imperfect intellectual property rights and the duration of monopoly protection. First, I show that multiple IP rights cover a single product. In particular, the average drug is covered by 4 patents and 3.4 regulatory exclusivities. Second, I show that actual monopoly terms last beyond initial protection terms. On average, NCEs have an initial monopoly life—meaning time from branded drug launch to primary IP expiration—of 11.1 years. Average effective monopoly life—meaning time from branded drug launch to first generic entry—is 13.3 years. Of course, this 2.2-year extension in monopoly life cannot be considered a causal result of secondary patents, due to systematic differences in drugs that have more patents relative to those that do not and the endogenous nature of generic firms’ decisions on which markets to enter. I determine a causal estimate in the final part of the paper. Third, I show that not all IP are enforced. The mean potential monopoly life—time from branded launch to the final secondary IP expiration—is 17.9 years. This is the monopoly term the average NCE would receive if all its secondary IP were fully enforced.

In the second part of the paper, I discuss the mechanisms through which secondary IP influences first generic entry timing. I present a simple conceptual framework suggesting that secondary IP delays generic entry in two ways: by introducing a binding later IP expiration and by increasing

---

1. NCEs which are not included in the sample are those approved by 2010 that have not yet experienced generic competition and those brought to market but then discontinued; this restriction has minimal impact on the sample.
2. That is, a branded manufacturer may take out additional patents on its planned product extensions and launch these as the NCE’s monopoly protection is set to expire. In recent work, Fowler (2019) confirms such launch timing of product extensions.
3. Specifically, I include as controls a drug’s market size and therapeutic area, discussed more fully in Section 6. I determine drug market size using the Medical Expenditure Panel Survey’s (MEPS) Prescribed Medicines Files and therapeutic area using the World Health Organization’s Anatomical Therapeutic Chemical (ATC) classification system. The restriction to NCEs with non-missing market size means the exclusion of (1) drugs with low market sizes and hence not captured by survey data and (2) drugs that are received in inpatient settings, e.g., chemotherapy drugs.
uncertainty in the scope or enforceability of remaining IP rights. I then provide a series of descriptive evidence consistent with these mechanisms. Visual bunching evidence confirms that IP extensions are binding in their delay of generic entry: While 20 percent of branded drugs see their first generic entrant at primary IP expiry, another 10 percent see generic entry at exactly the time of a later secondary IP expiration. For those drugs with non-binding IP expirations—whose first generic enters either after all IP have expired or between IP expirations—I exploit variation in the number of secondary IP on a drug to disentangle the sources of delayed entry. I separate uncertainty in market profitability from uncertainty in IP scope/enforceability and find that the latter is an important factor.

In the final part of the paper, I estimate the causal impact of an additional patent on a drug’s monopoly term. Underlying endogeneity concerns stem from the interrelatedness of branded manufacturers’ patenting decisions and generic manufacturers’ entry decisions: Branded firms want to protect their most valuable drugs with additional patents; these are the same markets generic firms want to enter earlier, putting the drugs at risk of shorter monopoly terms. I employ an instrumental variables analysis using secondary patent applications for a drug; I evaluate the impact of whether a secondary application is granted on a drug’s extra monopoly life (meaning time from primary IP expiry to generic entry). Looking at patent applications accounts for endogenous selection by branded firms into applying for a patent. Assuming that conditional on being included in an application, whether a drug is granted a secondary patent is not as good as random, it is necessary to find an instrument for whether the patent application is granted. For this purpose, I construct a measure of examiner leniency; such examiner designs have become increasingly popular in the economics literature across a variety of settings (e.g., Arnold et al., 2018; Farre-Mensa et al., 2020; Kling, 2006; Sampat and Williams, 2019). The relevance condition for a valid instrument is readily satisfied: A 10-percentage-point increase in an examiner’s average grant rate is associated with an 8.8-percentage-point increase in the likelihood a secondary drug patent application is granted. The exclusion restriction requires that examiner leniency only affects extra monopoly life through the likelihood a patent application is granted, which is a reasonable assumption. The IV analysis indicates that a randomly granted secondary patent extends monopoly life by 10.3 months. Given that the average drug has 2.7 secondary patents obtained prior to generic entry, this would suggest 28 months, or more than 2 years, of extra monopoly life per drug attributable to patent accumulation. Given a mean actual monopoly life of 13.3 years, this suggests secondary IP rights account for 17.5 percent of a drug’s monopoly term.

This paper contributes most directly to the literature on intellectual property protection and incentives for pharmaceutical innovation (Branstetter et al., 2016; Budish et al., 2015; Gaessler and Wagner, 2019; Gilchrist, 2016; Higgins et al., 2020; Mohapatra and Zhang, 2020). These papers evaluate how intellectual property duration affects outcomes such as new product introductions, R&D investment, or project continuation decisions but do not explore how firms can influence that IP duration. In particular, Budish et al. (2015) suggest that the fixed-term patent system reduces
incentives for longer-term research due to the lag between discovery (when a drug is first patented) and commercialization. I add to this research by showing that an innovating firm may extend its product’s monopoly term beyond the initial fixed patent term by accumulating patents on the product. If firms with products requiring longer development times offset reductions in the initial term by acquiring more patents, then the distortions of the fixed-term system may be less than previously assumed. Conversely, if the products acquiring monopoly extensions are those that do not require greater development costs or time, then patent accumulation on a single product may exacerbate the distortions of intellectual property design. My work thus offers an important nuance that intellectual property rights are less rigid than we typically assume and that these imperfections can alter our discussions on optimal patent policy and innovation.

The extent to which IP accumulation extends monopoly life is also important due to the trade-off inherent in intellectual property protection: A longer patent life encourages firms’ ex-ante incentives to innovate but also increases the deadweight loss resulting from monopoly pricing (Nordhaus, 1969). Quantifying the causal impact of an additional patent on monopoly extension is thus a requisite first step in any evaluation of intellectual property policy. Within the pharmaceutical industry, related papers have explored the patenting propensity of branded pharmaceutical firms and corresponding patent challenges by generics (Grabowski et al., 2017; Grabowski and Kyle, 2007; Hemphill and Sampat, 2011, 2012). None have estimated a causal impact of patenting on monopoly life. To my knowledge, I provide the first such causal estimate of the extent to which an incremental patent extends innovators’ monopoly protection in any industry.

My research on imperfect intellectual property rights relates to other work on non-standard intellectual property. Lemley and Shapiro (2005) provide an overview on the economics of “probabilistic patents”—including uncertainty in the commercial significance of a granted patent or in the scope and enforcement of the patent—and Bessen (2009) offers a theoretical treatment of contracting in the presence of imperfect IP rights. I offer a contribution by empirically documenting the extent and consequences of imperfect intellectual property rights in a key sector of the U.S. economy.4

The remainder of this paper proceeds as follows: Section 2 provides institutional details related to the pharmaceutical industry. Section 3 describes the data. Section 4 documents the existence of imperfect IP and of monopoly power beyond the initial patent term. Section 5 explores the mechanisms through which secondary IP impacts generic entry timing via a simple conceptual framework and a series of empirical evidence. Section 6 estimates a causal impact of secondary IP on monopoly term extension. Section 7 concludes.

4Other empirical studies have examined patent thickets—the fragmentation of intellectual property rights across firms within an industry (Galasso and Schankerman, 2010; Hall and Ziedonis, 2001; Noel and Schankerman, 2013; Ziedonis, 2004)—and the within-firm strategic use of patents (Abrams et al., 2018; Cockburn and MacGarvie, 2011; Hegde et al., 2009; Kurakina, 2020; Noel and Schankerman, 2013; Righi and Simcoe, 2020). My paper considers imperfections at a more granular level of analysis (the product level) and corresponding impacts to monopoly protection.
2 Background: The pharmaceutical industry

The Drug Price Competition and Patent Term Restoration Act of 1984, more commonly known as the Hatch-Waxman Act, established today’s system of both drug regulation and intellectual property protection (Eisenberg, 2001). Meant to balance the competing interests of generic and big pharma manufacturers, the act was a “complex legislative compromise” that had dual goals of simultaneously fostering the generic drug industry while providing additional protections to branded drugs (Angell, 2005; Eisenberg, 2001, p. 121). This section briefly describes branded drug development and approval, the intellectual property protection available for small-molecule drugs, the generic entry process, and proposed reforms to pharmaceutical IP protection. For an overview of the pharmaceutical industry more generally, see Lakdawalla (2018).

2.1 Branded drug development and approval

Drug development begins with discovery of a new molecule. Once a molecule is designed, the firm will patent it (see more on drug patenting below), and the molecule begins preclinical, nonhuman testing. Successful drug candidates will then move through successive phases of clinical trials in human volunteers to establish safety and efficacy. Preclinical and clinical trials for a new molecule can take 10 years on average, upon the conclusion of which a manufacturer will submit a New Drug Application (NDA) to the FDA for approval, including the results of all phases of testing (PhRMA, 2015). Only after FDA approval can a manufacturer market a drug to consumers. While estimates vary, development of a brand new molecule is costly, ranging from $500 million to $2 billion per drug, depending on the therapy and manufacturer (Adams and Brantner, 2006). As such, intellectual property protection is considered necessary to incentivize firms to undertake these high costs of development.

2.2 Intellectual property protection

The pharmaceutical industry is unique in that it receives intellectual property protection not only via patents granted by the USPTO but also via regulatory exclusivities by the FDA (Eisenberg, 2001). I describe here the two types of IP protection provided for pharmaceutical manufacturers.

---

5Development costs for new drugs that make use of an already approved molecule are substantially reduced (according to Chong and Sullivan (2007), by about 40 percent). As existing molecules have already demonstrated safety in clinical trials, producers can typically obtain approvals for new uses of those drugs with fewer resources, i.e. time, money, and trial subjects (Hernandez et al., 2017).
2.2.1 Drug patents

Drug patents extend 20 years from the date of USPTO application and cover one of four characteristics: (1) the drug substance, or active ingredient; (2) its use for treating a certain indication; (3) its formulation, including both the physical form (e.g. liquid or capsule) and administration route (e.g. by mouth or injection); and (4) the manufacturing process to make the drug (Angell, 2005). Once a molecule is developed, the branded firm will patent its active ingredient; this patent offers the strongest form of protection and is considered the primary patent. We think of primary patents as having certain enforcement. Secondary patents, covering one of the other three characteristics, are filed over the course of a drug’s development and even after its launch. Multiple patents covering the same product can extend effective patent life, and Angell (2005) notes that even the pill colors and coatings can be patented. Hatch-Waxman also provides for a single USPTO patent term extension per new active ingredient of up to 5 years to account for time in which a drug is in development or under FDA review; branded manufacturers typically choose to apply this extension to the primary, or strongest, patent. Note that while patents last from 20 years after their initial filing, the effective patent term is typically smaller, due to the lag between when a patent is obtained (during the drug discovery/development process) and when a product is launched.

Hatch-Waxman requires branded manufacturers to list relevant patents (covering either the drug substance, formulation, or approved uses) in the FDA’s Orange Book (Angell, 2005). When a generic firm wants to enter a market, it must account for those patents listed in the Orange Book. As a drug substance patent is the first filed, it often expires before the secondary ones. The scope of secondary patents are described in the Orange Book under use codes or descriptions provided by manufacturers; the FDA relies on these descriptions rather than an actual reading of the patent claims when making generic approval determinations (Rai, 2012). As such, branded firms have previously employed a tactic of listing overly broad use codes in the Orange Book that exceed the actual scope of the patent (Rai, 2012). Under Hatch-Waxman, generic manufacturers, meanwhile, can file a “Paragraph IV certification” or challenge to a branded drug’s remaining patents (described in more detail below).

2.2.2 FDA regulatory exclusivities

The Hatch-Waxman Act established a second pathway for IP protection for branded drugs—regulatory exclusivities from the FDA. They are granted once a drug is approved for marketing, and during periods of exclusivity, generic manufacturers cannot enter the market. These exclusivities actually take two different forms: (1) data exclusivity and (2) marketing exclusivity. Data exclusivity covers a firm’s clinical trial data on safety and efficacy. During periods of data exclusivity, generic manufacturers may not rely on the original firm’s data to submit an abbreviated new drug application (ANDA, described in Section 2.3; however, generic firms are free to develop and submit their own full NDA. Marketing
are submitted to the USPTO for approval, exclusivities are submitted to the FDA. In applying for exclusivities, applicants may submit studies that were conducted by another entity—meaning manufacturers can rely on (1) clinical data or literature from other companies and/or (2) the FDA’s prior safety and efficacy findings from an approved drug—thus lowering their burden of proof. Patent terms and exclusivity periods may or may not run concurrently, depending on the type of exclusivity (FDA, 2018).

Numerous different exclusivities exist to protect branded drugs. New chemical entity (NCE) exclusivity is a 5-year protection starting from drug launch for a new active ingredient never before approved by the FDA (Thomas, 2017). Exclusivities may also be given to existing drug products. Changes to approved drugs result in a three-year clinical investigation exclusivity (CIE), including changes in dosage form or dosing regimen, administration route, strength, disease indication, and change from a prescription (Rx) to an over-the-counter (OTC) indication, among others (FDA, 1999; Thomas, 2017). Other exclusivities include those for orphan drugs, for approved medicines that undergo subsequent pediatric clinical trials, and for infectious disease products. Appendix Table 6 provides more detail on these key exclusivities; see Thomas (2017) for a description of all 16 FDA exclusivities that currently exist.

### 2.3 Generic entry

The Hatch-Waxman Act also provided a pathway for generic entry—the Abbreviated New Drug Application (ANDA)—which lowered the burden of evidence required for generic drug approval. Prior to the act, generic manufacturers were required to submit full NDAs proving safety and efficacy of their drugs. An ANDA allows the generic manufacturer to rely on clinical trial data of the original manufacturer and instead show only bioequivalence to the original branded drug. Note that generic drugs contain the same active ingredient, in the same dosage form, same strength, and route of administration as the original drug. The time and costs of generic development are nontrivial. Copying a branded drug and conducting bioequivalence studies can take a few years, and the ANDA approval process requires both factory inspections by the FDA and laboratory tests of the generic product. Thus, although substantially lower than the costs required for branded drug development, there are still sizeable sunk costs of entry for a generic manufacturer, estimated to be between $250,000 and $20 million per ANDA (Scott Morton, 1999).

A generic firm may submit an ANDA to the FDA four years after NCE launch, although actual entry cannot occur until at least one year later, after the NCE exclusivity has expired. At the time of application, a generic firm must account for a drug’s patents listed in the Orange Book and all remaining exclusivities. That is, the generic entrant must specify its intended market entry date and show that, by that date, any unexpired IP on a drug will either be invalid or that the generic exclusivity prevents generic firms from submitting either application to the FDA irrespective of where the clinical trial data come from (Thomas, 2017). Effectively, exclusivities—whether data or market—forestall generic entry.
can enter without infringing upon those remaining IP. The FDA may then tentatively approve a generic application, with official approval coming only after the scope/enforceability of remaining IP has been fully determined.

Hatch-Waxman created a specific process by which such determinations occur. Specifically, generic firms may submit a Paragraph-IV challenge of a branded drug's remaining patents on the grounds of either non-infringement or invalidity. When a generic manufacturer files an ANDA with a Paragraph-IV certification, the branded firm can either (1) do nothing or (2) sue (within 45 days) the generic manufacturer for patent infringement. If the branded firm does nothing, then the FDA may approve the generic drug.

If the branded manufacturer sues, then there is an automatic stay on FDA approval to allow for ensuing litigation or a settlement. If the courts decide in favor of the branded firm, then the patent challenge fails and the FDA will not approve generic entry. For the generic firm that chooses to submit an ANDA while patents remain on a drug, they thus risk potentially large litigation expenses (and the previously incurred development expenses) without guarantee of approval. Whenever a first court ruling decides in favor of the generic (that either all or some remaining secondary patents are unenforceable), the stay ends and the FDA will officially approve generic entry. Generic entry can occur immediately if all remaining patents are deemed unenforceable. If there is partial enforcement, i.e. at least one remaining secondary patent is fully enforced while others are deemed unenforceable, then generic entry can occur at the point at which those fully enforced patents expire. The first generic firm to win a patent challenge receives 180 days of generic exclusivity, during which time the branded firm and generic challenger constitute a duopoly. After the 180 days, other generic manufacturers may submit ANDAs and enter the market. (Branstetter et al., 2016). Thus, a trade-off exists for the first generic entrant for a drug: Earlier entry gives it the opportunity to earn duopoly profits but comes with greater uncertainty about IP enforceability, raising potential litigation costs.

Scholars disagree on the impact of additional IP to generic entry. One view holds that secondary patents and exclusivities would have little effect in forestalling generic entry since generic manufacturers theoretically can enter after the primary IP expiration (i.e., the latter expiration of either the active ingredient patent or NCE exclusivity). For instance, if a brand new molecule is initially approved by the FDA for the treatment of depression and then subsequently obtains secondary IP for the treatment of bipolar disorder, a generic manufacturer theoretically should be able to enter the market for depression at the primary IP expiration. Yet other scholars note that strategic patenting practices by branded manufacturers still create a considerable barrier to generic manufacturers for entry for any use, as they must navigate an “infringement minefield” of all potential infringement cases that could result (Furrow, 2008; Tribble, 2019). That is, there is considerable

---

8Under non-infringement, generic firms seek approval for the generic with a narrower labeling that excludes on-patent uses.
uncertainty in the scope and enforceability of secondary IP that must be resolved. I settle this debate by providing the first systematic evidence on when generic entry occurs relative to a drug’s primary and secondary IP expirations.

2.4 Proposed reforms to pharmaceutical IP

Various legislative proposals have suggested reforms to IP policy in the pharmaceutical industry to limit the extent to which IP accumulation extends a drug’s monopoly life. These include reducing exclusivity periods and restricting CIE to drugs showing significant clinical benefit relative to existing therapies (Thomas, 2017). Other proposals similarly suggest raising the threshold for patentability, as is the case in other countries. India, for example, requires higher standards for nonobviousness, and any patent applications that claim improvements over existing drugs must show evidence of clinical benefit (Amin and Kesselheim, 2012). A first step to evaluating any such change to IP policy is to quantify the impact of secondary intellectual property protections for pharmaceuticals on generic entry.

3 Data

This project draws from several data sources. The main source of information on drug approvals, patents, and exclusivities is the FDA. Historical FDA Orange Books list all relevant patents and exclusivities and their expiration dates for each small molecule drug in the US. I supplement these data with information from the Drugs@FDA database, including drug approval dates. More details on each granted patent as well as all patent applications (both granted and rejected) come from research datasets at the USPTO.9 I determine drug market size using the Medical Expenditure Panel Survey’s (MEPS) Prescribed Medicines Files and therapeutic area using the World Health Organization’s Anatomical Therapeutic Chemical (ATC) classification system.

From these sources, I create three novel analytic datasets. The first is a drug-level dataset of the 370 new chemical entities (NCEs) approved between 1985 and 2010 that have a generic entrant as of July 2020. Focusing on NCEs with a generic entrant offers two main advantages: First, I capture drugs with a full history in the data (from branded launch through generic entry); empirically, this is necessary so that I have a measure of each drug’s actual monopoly life.10 Second, my estimates will offer a lower bound on the extent to which secondary patenting lengthens monopoly life (by excluding the effects of product extensions under new brand names). Key variables include branded launch date, generic entry date, and the start and expiry of each intellectual property right. One

10NCEs which are not included in the sample are those approved by 2010 that have not yet experienced generic competition and those brought to market but then discontinued; this restriction has minimal impact on the sample.
challenge is determining which of a drug’s IP is primary and which are secondary. I consider a drug’s molecule patent to be the one receiving a USPTO patent term extension and in the case of no such extension, the earliest patent filed. I take the primary IP expiry to then be the latter of either the molecule patent or NCE exclusivity expiry. All subsequent patents and exclusivities are secondary.

The second is a drug-month-level dataset of these NCEs from branded launch through the month of first generic entry. Drugs are observed for 159 months on average, yielding a total of 58,898 drug-month observations. This dataset tracks the periods during which each patent or exclusivity is in effect, and I use it to explore how secondary IP influence first generic entry timing.

The third dataset is a drug-application-level dataset merging each NCE with its secondary patent applications, both granted and rejected. Construction of this dataset requires restricting the universe of all USPTO patent applications to the set of drug applications and determining which of these applications correspond to each NCE. I attach a patent application to a drug if its claim text includes its active ingredient and if its assignee is one of the assignees for the drug’s granted patents listed in the Orange Book. I restrict to drugs that apply for secondary patents and have non-missing values of control variables. The matched dataset includes 241 NCEs and 3,021 drug-application observations. I use this final dataset in the OLS and IV analyses of the impact of secondary patents on extra monopoly life.

4 Existence of imperfect IP

I present in this section a series of descriptive statistics invalidating the standard “one-patent-one-product” model. I show that pharmaceutical firms obtain multiple IP rights per drug product, that not all IP rights are enforced, and that monopoly terms across drugs vary, lasting beyond the initial term indicated by the primary IP.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patents</td>
<td>4</td>
<td>3.8</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>Number of FDA exclusivities</td>
<td>3.4</td>
<td>2.7</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>Initial monopoly term (years)</td>
<td>11.1</td>
<td>3.8</td>
<td>1.1</td>
<td>19.4</td>
</tr>
<tr>
<td>Potential monopoly term (years)</td>
<td>17.9</td>
<td>6.7</td>
<td>3.3</td>
<td>44.9</td>
</tr>
<tr>
<td>Actual monopoly term (years)</td>
<td>13.3</td>
<td>4.3</td>
<td>3.4</td>
<td>32.6</td>
</tr>
</tbody>
</table>

Total number of drugs is 370. All monopoly terms are relative to branded launch.

The reduction in sample size from 370 to 241 NCEs is primarily due to the restrictions to drugs with secondary patent applications and to drugs with non-missing market size. Due to the survey nature of the MEPS data, drugs with low market sizes or received in inpatient settings will have missing values for market size.
Table 1 provides summary statistics on IP accumulation and monopoly life. The first key takeaway from this table is that multiple IP may cover a single drug. The average drug is protected by 4 patents and 3.4 FDA exclusivities, with substantial cross-drug heterogeneity in that IP acquisition.\textsuperscript{12} For example, HIV treatment Kaletra, discussed previously, is covered by a maximum 28 patents, and Gleevec (imatinib), an oral chemotherapy medication used for leukemia, is covered by a maximum 17 FDA exclusivities.

The bottom three rows of Table 1 look at different measures of IP duration. If drugs were restricted to solely their primary (or initial) IP, they would experience an average monopoly life of 11.1 years on the market.\textsuperscript{13} (Note that this measure restricts to time on the market post-launch—from branded launch to primary IP expiry—and does not include time when the drug may have been protected by a primary patent but still in development). The fourth row gives statistics on potential monopoly term—time from branded launch to final IP expiry. This represents the monopoly life a branded firm would experience if all secondary IP were fully enforced, 17.9 years on average. Comparing this row to the final row indicates that for at least some drugs, not all IP are enforced. That is, when generic entry occurs, additional patents or exclusivities may still be in place but are deemed unenforceable in terms of preventing generic entry. We infer this because mean actual effective monopoly life—time from branded launch to generic entry—is 13.3 years, more than 4 years shorter than the mean potential term. Note that this actual monopoly term, or time to generic entry, is more than two years longer than the initial monopoly term. Subsequent analyses will evaluate the extent to which such extensions in monopoly life represent a causal impact of secondary IP rights.

Figure 2 compares the distributions of initial and actual monopoly terms. Two noteworthy facts emerge from this figure. First, there is substantial heterogeneity across drugs in both time to primary IP expiry and in time to generic entry. Second, the distribution of monopoly life shifts right once secondary IP is included, i.e. the distribution of time to generic entry is to the right of that of time to primary IP expiry. This figure suggests that considering only the fixed initial term of the primary IP may result in an incomplete measure, or underestimate, of a product’s monopoly term. Future research on optimal IP design and innovation should incorporate this potential for products to experience monopoly extensions.

The section confirms that intellectual property rights in the pharmaceutical industry are imperfect: multiple IP cover a single drug, not all of those IP are enforced, and drugs experience varying

\textsuperscript{12}The patent numbers focus on patents specifically listed in the Orange Book; these patents cover drug substance, formulation, or approved uses and are the ones generic firms must account for when they seek to enter a market. A branded firm may also take out other patents that are not listed in the Orange Book, such as patents on manufacturing process.

\textsuperscript{13}All drugs in the sample are new chemical entities (NCEs) and thus receive a 5-year NCE exclusivity. The minimum initial monopoly term in Table 1 (1.1 years) is below 5 years because in a handful of cases, the branded manufacturer discontinued its original product and transferred the remaining exclusivity period to another product with the same active ingredient. In other cases, a branded manufacturer licensed rights to its drug to different manufacturer who then received the remaining exclusivity.
terms of monopoly protection that can last beyond the initial term. The next section explores the channels through which secondary IP rights influence generic entry timing, and Section 6 estimates the causal effect of those rights on monopoly extension.

5 Mechanisms: How secondary IP impacts generic entry timing

This section discusses the mechanisms through which the accumulation of IP may delay generic entry. I present a simple conceptual framework that produces testable implications I can take to my data. In particular, I posit that secondary IP rights work in two ways: by introducing a binding later expiration for some drugs and by increasing uncertainty in overall IP scope and enforceability for others. I then present a series of empirical evidence that is consistent with the conceptual framework.

5.1 Conceptual framework

I consider a generic firm’s decision to enter a pharmaceutical market (where each market corresponds to a branded drug), taking the level of intellectual property protection on the drug as given. The timing of first generic entry in a market exhibits a trade-off between incentives to enter early and preempt competitors on the one hand, and uncertainty in market profitability and/or IP enforceability on the other hand. These two types of uncertainty have different implications. Uncertainty in market conditions leads the generic to actively delay entry until some signal is
received that reduces or resolves that uncertainty (as in real options theory). Uncertainty in IP enforceability delays entry (even if the generic manufacturer wants to enter) until that uncertainty is resolved via established legal institutions. I discuss this distinction in the conceptual framework that follows. I distinguish between two separate scenarios: when a branded drug is protected only by primary IP (as in the standard model) and when it is protected by both primary and secondary IP (in the imperfect model).

Suppose first that the branded drug is covered only by primary IP. We generally think of the primary IP as a strong type of protection covering the drug’s molecule with certain enforcement. For a generic firm deciding when to enter, attempting entry prior to that expiration will likely lead to huge costs in litigation that outweigh the benefits. Thus, the generic firm will delay its entry until that primary IP expires. Indeed, while generic firms may prepare and submit applications to the FDA in advance of the actual primary IP expiration, they will specify their anticipated entry date as when the primary IP expires. The FDA may then tentatively approve the generic with an official approval date occurring upon primary IP expiration. We would thus expect the primary IP to be binding for many generic entrants, particularly in those markets with large potential revenues, with first generic entry occurring as soon as that primary IP expires.

**Implication 1 (standard model): There will exist a spike in entry at the time of primary IP expiration, due to binding primary IP.**

Note that not all generic firms will enter immediately upon primary IP expiration. The firm may wait to learn more information that resolves some uncertainty in market profitability. For example, the generic firm may delay entry in markets offering less revenue or where there may be concern about new treatments in development that would make the drug in question obsolete. In such cases, the generic firm may delay undertaking the costs of bioequivalence trials and its application to the FDA until receiving some positive signal on market profitability.

**Implication 2 (standard model): There will exist gradual entry delayed past primary IP expiration, due to learning about market conditions.**

Now let us consider the imperfect model where a branded drug has both primary and secondary IP protection. For such drugs, binding primary IP and uncertainty in market profitability remain factors that the generic entrant must consider. However, the accumulation of IP on a single drug can also lengthen the branded manufacturer’s protection term in two other ways—by introducing a binding secondary IP expiration and by increasing uncertainty in overall scope/enforceability of remaining IP rights. Due to such uncertainty, generic entry for many drugs, as described in Section 2.3, occurs following patent challenges in which a generic entrant seeks to enter the market prior to the expiration of all secondary patents. The generic firm will submit an application to the FDA specifying that remaining patents are either invalid or that the generic can enter without infringing upon those patents. The uncertainty is then resolved via litigation in the courts, with final
FDA approval for the generic occurring only following a court decision in its favor or a settlement agreement between the branded and generic manufacturers.

Suppose the court rules one or more remaining patents are fully enforceable. In this case, the FDA will approve generic entry can occur at the expiration of the final enforceable patent, and we would expect a binding secondary IP for these entrants. As in the case of binding primary IP, we would anticipate such a scenario in more attractive markets, e.g. that treat chronic conditions or generate higher revenues, and where the generic manufacturer is willing to incur litigation expenses for earlier entry. Thus, if one or more secondary IP rights are perfectly enforced, then IP accumulation can extend a branded drug’s monopoly term by changing the binding expiration for generic entry—from the primary IP to a secondary one.

**Implication 3 (imperfect model):** There will exist a spike in entry delayed to the time of a secondary IP expiration, due to binding secondary IP.

We may also see intermediate entry occurring between IP expirations, i.e. after the expiration of at least one (primary or secondary) IP but before the expiration of all IP. This would occur if the courts rule that all remaining patents are unenforceable. Note that in this case, the delayed or non-binding entry arises due to the need to wait for legal institutions to resolve questions on patent enforceability rather than from the generic firm’s active decision to wait until some additional learning occurs (as suggested by Implication 2). When entry does occur, we would expect it to be faster than in the case of delayed entry under market uncertainty. Here, as in the case under binding expirations, the generic manufacturers expect the market to be profitable. They have completed their bioequivalence studies and submitted ANDAs to the FDA but must now wait for a court decision. As soon as a favorable decision to the generic is made not to enforce remaining patents, they enter the market.

**Implication 4 (imperfect model):** There will exist intermediate entry occurring between two IP expirations, due to nonenforcement of remaining secondary IP. All else equal, this delayed entry will occur at a faster rate than the delayed entry that occurs with learning about market conditions.

In summary, this conceptual framework suggests that under the standard model, we would expect to see either immediate entry at expiration of the primary patent (Implication 1) or delayed entry under learning about market conditions (Implication 2). Under the imperfect model, two additional sources of delay are possible. We will see entry at the expiration of a binding later secondary patent (Implication 3) or intermediate entry occurring between expirations but at a faster rate than occurs under market learning (Implication 4).
5.2 Empirical evidence

I present a series of empirical evidence consistent with the above conceptual framework. First, I test for Implications 1 and 3. I explore the extent to which IP rights—both primary and secondary—are binding in blocking generic entry. Figure 3 displays the timing of generic entry in months for each drug relative to its IP expirations. Panel (a) gives generic entry relative to the primary IP expiration, i.e. a drug’s extra monopoly life relative to the standard model. The spike at 0 months confirms Implication 1: A drug’s primary IP is binding for 74 out of 370 drugs, or 20 percent of the sample. But looking at a drug’s primary IP does not give a complete picture. Panel (b) thus considers generic entry relative to the branded drug’s most recent (primary or secondary) IP expiration prior to generic entry. This panel confirms Implication 3: With secondary IP included, the spike at month 0 rises dramatically, with another 37 drugs, or 10 percent of the sample, experiencing generic entry at exactly when a secondary IP expires. Thus, binding secondary IP rights are important, accounting for half as much binding entry as primary IP.

Two questions emerge from this figure: First, which secondary IP matters? Is it the first secondary IP that is binding, or do later secondary IP similarly block entry? Second, we see that for many drugs, IP expirations are non-binding—for these drugs, the first generic enters either after the drug’s final IP expiration or between expirations, as indicated by Implications 2 and 4. How do we separate the sources of non-binding entry? I address each of these questions in turn.

14Note that a small portion of drugs (11 percent) see generic entry prior to primary IP expiration. A manual review of these drugs suggests these entrants are authorized generics, marketed by either a separate company that has received permission by the branded firm to enter early in exchange for a portion of its profits or by the branded manufacturer itself (the branded manufacturer may choose to launch its own generic to preempt future outside generic competition). I exclude these drugs from Figure 3.
Figure 3: Timing of generic entry (in months)

(a) Relative to primary IP expiration

(b) Relative to most recent IP expiration

Restricted to NCEs with generic entry at or after primary IP expiration, i.e. 329 drugs out of total sample of 370. Panel (a) gives timing of generic entry relative to primary expiration while panel (b) gives timing relative to most recent (primary or secondary) IP expiration prior to generic entry.

To answer the first question as to which secondary IP is binding, I estimate a linear probability model of first generic entry and include as explanatory variables indicators for the month of primary and secondary IP expirations and the periods between those expirations. I use the drug-month-level dataset tracking drugs from branded launch through initial generic entry. There are 370 drugs observed for 159 months on average, for 58,898 drug-month observations total. I estimate the following regression:

\[ 1 \{ \text{Entry} \}_{it} = X_{it}\beta + \delta_i + \varepsilon_{it} \]  

The outcome of interest \( 1 \{ \text{Entry} \}_{it} \) is a binary variable indicating whether or not first generic entry for drug \( i \) occurs in month \( t \) following branded launch. \( X_{it} \) is a vector of indicators for month of primary IP expiration, time between primary and first secondary IP expirations, month of first secondary IP expiration, time between first secondary and second secondary IP expirations, and so forth. For fourth and subsequent secondary IP, I collapse the periods during which they are in effect into a single indicator and their expirations into another. A final indicator specifies when all IP has expired. Note that as multiple patents or exclusivities may expire in the same month (or even on the same day), I form these indicators based not on individual patents and exclusivities but rather on periods of IP protection. That is, multiple secondary patents may expire in the “month of first secondary IP expiration.” With the inclusion of these indicators, the excluded reference period is the time after branded launch and before the primary IP expires. I include drug fixed effects and robust standard errors clustered at the drug level.

Table 4 plots the coefficients of the linear probability model; all are significant at the 1-percent level except the coefficient on the indicator on the “month of third secondary IP expiration,” which...
is significant at the 5-percent level. Generic entry is 23.2 percentage points higher in the month of a drug’s primary IP expiration than in the months prior. After the primary IP expiry, the relative probability of generic entry then drops dramatically (to merely 2.6 percentage points higher) until the first secondary IP expiry, at which point it rises again (to 17.1 percentage points higher). This non-monotonic pattern continues with subsequent secondary IP, suggesting the binding nature of these IP in delaying generic entry.

A key takeaway from this figure is that not only does the first secondary IP matter but later secondary IP also introduce binding entry points. It is also important to note that this figure plots relative entry in a *given* month. As the periods between two IP expirations for a drug may last for many months, when we consider total entry, there is more non-binding entry (occurring between or after all expirations) than binding entry (occurring exactly when an IP expires). I next turn to answering the second question of what accounts for this non-binding entry.

To separately test for Implications 2 and 4, I exploit variation across drugs in their number of secondary IP. Figure 5 repeats the analyses of Figure 3 on timing of generic entry but separated by whether or not a drug has any secondary IP. For drugs without secondary IP, entry timing is determined by either binding primary IP (Implication 1) or uncertainty in market profitability (Implication 2). Panel (a) of Figure 5 shows the generic entry timing for these drugs relative to primary IP expiration. We see, as would be expected, that primary IP is binding for a much larger portion of these drugs (39 percent) relative to the entire sample (20 percent in Figure 3). Nonetheless, many drugs in this sample still have delayed entry (despite no secondary IP to forestall it), meaning that for these drugs it is the uncertainty in market profitability that matters, as
suggested by Implication 2.

Figure 5: Timing of generic entry (in months)

(a) Drugs with no secondary IP; relative to primary IP expiration
(b) Drugs with secondary IP; relative to most recent IP expiration

Restricted to NCEs with generic entry at or after primary IP expiration. Panel (a) gives timing of generic entry relative to primary expiration for drugs with no secondary IP; number of drugs is 75. Panel (b) gives timing relative to most recent (primary or secondary) IP expiration prior to generic entry for drugs with secondary IP; number of drugs is 254.

Panel (b) of Figure 5 restricts to drugs with secondary IP and depicts generic entry timing relative to the branded drug's most recent (primary or secondary) IP expiration prior to entry. For this set of drugs, intellectual property protection is binding in blocking generic entry for 33.1 percent (or 84 out of 254 drugs). Separating this binding IP expiry by type, for 18.5 percent the primary IP is binding and for the other 14.6 percent, secondary IP is binding. For the remaining drugs with non-binding or delayed entry, two sources of uncertainty are possible, in either market profitability or IP scope/enforceability.

To disentangle these sources of uncertainty, I narrow in on exactly the set of drugs with non-binding generic entry; Figure 6 presents their generic entry timing. Panel (a) depicts drugs without secondary IP. We know that delay for these drugs must reflect learning in market conditions (Implication 2). For drugs with secondary IP, I further divide the sample into those where the generic enters after all secondary IP have expired and those with intermediate entry between IP expirations. Panel (b) looks at this first group. For these drugs, IP enforceability is no longer an issue and market profitability is the only source of uncertainty. We would expect that these drugs to have similar entry patterns as those in panel (a). Indeed, the entry patterns look visually similar. Out of the drugs in panel (a) without secondary IP and delayed entry, 30 percent enter by 10 months after their primary IP expiration, 52 percent by 25 months, and 72 percent by 50 months. For the drugs in panel (b) with secondary IP and delayed entry after all secondary IP have expired, the comparable statistics are similar: 29 percent by 10 months post-final secondary IP expiration, 46
Figure 6: Timing of delayed generic entry (in months)

(a) Drugs with no secondary IP; relative to primary IP expiration
(b) Drugs with secondary IP & all have expired; relative to final IP expiration
(c) Drugs with secondary IP & at least one still unexpired; relative to most recent IP expiration

Restricted to NCEs with non-binding IP expirations, i.e., delayed generic entry. Panel (a) gives timing of delayed generic entry relative to primary expiration for drugs with no secondary IP; number of drugs is 46. Panel (b) gives delayed timing relative to final IP expiration for drugs with secondary IP and all secondary IP expired; number of drugs is 28. Panel (c) gives delayed timing relative to most recent (primary or secondary) IP expiration prior to generic entry for drugs with at least one secondary IP still unexpired; number of drugs is 142.
percent by 25 months, and 68 percent by 50 months. To further assess their similarity, I apply the non-parametric Kolmogorov-Smirnov test of the equality of two distributions; with a \( p \)-value of 0.897, the test fails to reject the null hypothesis that these two entry distributions are the same.

Panel (c) looks at entry timing for generics with secondary IP that enter while at least one secondary IP still remains on a drug; entry is graphed relative to the most recent (primary or secondary) IP expiration. By Implication 4, uncertainty in IP scope/enforceability is an important contributor, with intermediate entry occurring once the nonenforcement of remaining IP is determined. Consider the counterargument: There is no uncertainty about which secondary patents are enforced (i.e., there is certain unenforcement of these patents) and the only source of uncertainty is in market profitability. If this were the case, we would expect the entry timing distribution for this group to match those in panels (a) and (b). But this is not what we see—the entry distribution looks remarkably different. As suggested by Implication 4, entry is faster: 40 percent enter by 10 months post-expiration; 70 percent by 25 months; and 92 percent by 50 months. This makes sense if we consider the different types of drugs in these samples. Drugs that have delayed entry after all IP expired are likely to be drugs less attractive to generic entry, e.g., that may be in lower revenue markets or at risk of becoming obsolete pending results of clinical trials on other drugs in development. Rather than a situation where the generic firm is waiting to enter as soon as the binding IP expires, entry occurs only once a positive signal has been received on market profitability—e.g., a medical study that suggests a new use for an older off-patent drug.

In contrast, a generic with intermediate entry between IP expirations is likely a more attractive drug, e.g., in a higher revenue market, and the generic must wait for a determination on the remaining secondary IP scope and enforceability. Under such a situation, we would expect faster entry for these drugs, occurring as soon as nonenforcement of remaining IP is established. To confirm, I again apply the Kolmogorov-Smirnov test to assess the equality of the distribution in panel (c) with those in panels (a) and (b). These tests reject the hypotheses that the generic entry distribution for drugs with secondary IP still remaining is the same as that of drugs with no secondary IP (with a \( p \)-value of 0.001) or that of drugs with all secondary IP expired (\( p \)-value of 0.009). I also conduct a Cox regression of entry timing in which I compare the hazard or probability of generic entry for drugs with secondary IP still unexpired to those with all IP expired.\(^{15}\) The regression indicates that the probability a drug sees its first generic entrant in any given month after its final enforced IP expiration (conditional on not having had a generic entrant prior) is twice as likely for drugs with at least one IP still unexpired than for those with all IP expired.\(^{16}\)

\(^{15}\)Specifically, I run the following regression: \( h_{it} = h_0(t) \exp(\beta \mathbb{1}\{IP \text{ unexpired}\}_{it}) \), where the hazard rate \( h_{it} \) is the conditional probability a drug sees its first generic entrant in month \( t + \delta \) post-final enforced IP expiration given that it has not had a generic entrant through month \( t \); \( h_0(t) \) gives the baseline hazard, i.e. the probability of generic entry for any drug \( t \) months post-final enforced IP expiration; and \( \mathbb{1}\{IP \text{ unexpired}\}_{it} \) is an indicator for having at least one secondary IP still unexpired. I include robust standard errors clustered at the drug level. The sample includes 170 drugs and 4,480 drug-month observations.

\(^{16}\)The unexponentiated coefficient on \( \mathbb{1}\{IP \text{ unexpired}\}_{it} \) is 0.686, significant at the 1-percent level. The exponentiated coefficient, or hazard ratio, is 1.985—indicating a hazard (probability of generic entry) twice as large for drugs
Taken together, this evidence can be summarized as follows: Secondary IP are important for influencing first generic entry timing and they work in two ways. First, secondary IP rights introduce later binding expirations—secondary IP account for half as many binding expirations (10 percent) as primary IP (20 percent). Second, they also influence generic entry by increasing uncertainty in the scope or enforceability of the remaining IP on a drug. Nearly 40 percent of drugs in the sample see their first generic enter between two IP expirations. Comparing the distribution of generic entry timing for these drugs relative to those whose first generic enters after all IP have expired suggests that market profitability alone is not driving delayed entry for these drugs. Rather, uncertainty in IP scope/enforceability must also be a factor.\footnote{Considering the results of the Cox regression—which shows a hazard twice as large for this group as for those with all IP expired—if we suppose that certain unenforcement is the case for one-third of this group (and hence uncertainty in market profitability the contributing factor behind their delayed entry), that would suggest for $\frac{2}{3} \times 40 = 27$ percent of the sample, uncertainty in IP scope/enforceability is the limiting factor.}

6 Causal impacts to monopoly term

6.1 Naive analysis: Comparing drugs by number of secondary patents

Estimating a causal impact of secondary IP on monopoly life is difficult due to endogeneity concerns stemming from both the branded firm’s decision to patent and the generic firm’s decision to enter a market. The naive analysis would simply compare drugs with secondary IP to those without and observe resulting monopoly terms. If intellectual property rights were as good as randomly assigned, this would be enough to estimate a causal effect of secondary IP. Two main endogeneity concerns make such an analysis problematic. First, there is the potential for omitted variable bias: Characteristics of a drug that are unobservable to the researcher may influence both the branded firm’s decision to patent and the generic firm’s decision to enter (and hence, the monopoly life of a drug). Stated differently, there is selection into patenting in that drugs with secondary IP are systematically different from those without. Second, there is a reverse causality concern: The branded firm may anticipate the end of its monopoly life and as a result, attempt to obtain more intellectual property to forestall entry.

Intuitively, branded firms want to protect their most valuable or promising drugs with additional IP, and these are exactly the drugs that generics want to copy (which would shorten a drug’s effective monopoly life). By not accounting for this selection into patenting or unobserved attractiveness to generic entry, then the estimated coefficient on a simple regression of monopoly life on number of secondary IP would be too small (a negative bias). That is, such a regression would underestimate the causal effect of secondary IP. I present the naive analysis in Table 2. For comparison with subsequent analyses, I focus specifically on secondary patents.\footnote{Appendix Table 7 gives the naive analysis using number of all secondary IP (both patents and exclusivities) as}

\begin{footnotesize}
\footnotesize
   with secondary IP still unexpired as those with all IP expired.
\end{footnotesize}
I regress a drug’s extra monopoly life in months (meaning time between initial IP expiry and generic entry) on its number of secondary patents obtained prior to generic entry. Column (1) shows this basic regression; the coefficient of interest is insignificant. A first solution to improve the naive analysis would be to add controls to proxy for a drug’s unobserved characteristics. Column (2) adds fixed effects for branded drug approval year, Column (3) adds fixed effects for therapeutic area, and Column (4) adds a control for market size. Scott Morton (1999) shows that therapeutic area and market size are important for generic entry decisions; their inclusion helps control for attractiveness to a generic firm. Moving left to right across the table we see that the addition of these controls does help explain the total variation in extra monopoly life and that their inclusion further raises the coefficient on the number of secondary patents. Finally, as expected, drugs with larger markets have shorter extra monopoly terms and controlling for this raises the estimated coefficient on number of secondary patents. Looking at the specification in Column (4), with both sets of fixed effects and the control for market size, suggests that each secondary patent is associated with an additional 1.8 months of monopoly life.

Table 2: Secondary patents and extra monopoly life: Drug-level analysis

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td># secondary patents</td>
<td>0.327</td>
<td>0.890</td>
<td>1.328</td>
<td>1.823**</td>
</tr>
<tr>
<td></td>
<td>(0.857)</td>
<td>(0.894)</td>
<td>(0.871)</td>
<td>(0.874)</td>
</tr>
<tr>
<td>ln(market size)</td>
<td>-3.813**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.568)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>288</td>
<td>288</td>
<td>288</td>
<td>288</td>
</tr>
<tr>
<td>Adj. $R^2$</td>
<td>-0.003</td>
<td>0.127</td>
<td>0.176</td>
<td>0.202</td>
</tr>
<tr>
<td>Approval year FE</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Therapeutic area FE</td>
<td>yes</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dependent variable in all regressions is extra monopoly life in months (i.e., from a drug’s initial IP expiry to generic entry). Analysis at the drug level. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, and * p<0.1.

6.2 Comparing granted versus rejected secondary patent applications

Even with the inclusion of these controls, it is plausible that observed differences in monopoly life across drugs may still reflect selection of which drugs are patented rather than a causal effect of secondary patents on delaying generic entry. Thus, a next step would be to compare monopoly terms for drugs in accepted versus rejected secondary patent applications. If conditional on being included in an application, whether a patent is granted is as good as random, then such an analysis

---

19 Requiring drugs with non-missing market size reduces the sample size because the MEPS survey data do not capture (1) drugs with low market sizes and (2) drugs that are received in inpatient settings, e.g., chemotherapy drugs.
would imply a causal effect. I match drugs to their patent applications and carry out a drug-application-level analysis. The dependent variable is again a drug’s extra monopoly life and the main explanatory variable is an indicator for whether or not each secondary patent application for that drug is granted. Table 3 presents these results, with Columns (1) through (4) adding the same controls as before. The increase in the coefficient on $1_{\{\text{granted}\}}$ relative to the estimates of naive analysis confirms a selection into patent applications. Moving across specifications from left to right, the increase in the adjusted $R^2$-squared indicates that these controls soak up variation in the model, but the stability of the coefficient on $1_{\{\text{granted}\}}$ suggests that they not correlated with whether or not a given patent application is granted, conditional on an application being submitted to the USPTO. The results of Table 3 suggest that each secondary patent application granted is associated with 6.7 months of additional monopoly life.

Table 3: Secondary patents and extra monopoly life for drugs in granted versus rejected applications: Application-level analysis

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1_{\text{granted}}$</td>
<td>6.835***</td>
<td>6.703***</td>
<td>6.838***</td>
<td>6.704***</td>
</tr>
<tr>
<td></td>
<td>(1.689)</td>
<td>(1.560)</td>
<td>(1.503)</td>
<td>(1.496)</td>
</tr>
<tr>
<td>$\ln(\text{market size})$</td>
<td>-1.664***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.365)</td>
</tr>
<tr>
<td>Observations</td>
<td>3,021</td>
<td>3,021</td>
<td>3,021</td>
<td>3,021</td>
</tr>
<tr>
<td>Adj. $R^2$</td>
<td>0.005</td>
<td>0.225</td>
<td>0.357</td>
<td>0.361</td>
</tr>
<tr>
<td>Approval year FE</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Therapeutic area FE</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dependent variable in all regressions is extra monopoly life in months (i.e., from a drug’s initial IP expiry to generic entry). Analysis at the drug-application level. Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, and * $p<0.1$.

6.3 Exploiting variation in examiner leniency

Comparing granted and rejected patent applications above controlled for selection into obtaining a patent application. It is plausible that there still remains a selection issue into whether or not a patent is granted. This may be the case, e.g., if branded firms devote more resources to revising to a secondary patent application for certain drugs that are more valuable or more at risk of earlier generic entry. Assuming that conditional on being included in an application, whether a drug is granted a secondary patent is not as good as random, the natural solution is to find an instrument for which patent applications are granted patents. I follow previous research and use the “leniency” of the assigned patent examiner as an instrument (Farre-Mensa et al., 2020; Sampat and Williams, 2019). Prior work confirms that patent examiners have considerable discretion in whether or not to grant a patent (Cockburn et al., 2003; Lemley and Sampat, 2012; Sampat and Lemley, 2010) and that the assignment of patent applications is random conditional on application
year and technology area (Sampat and Williams, 2019). This type of analysis thus exploits both the variation in cross-examiner leniency and the quasi-random assignment of patent applications to examiners.

I construct a measure of examiner leniency—the share of applications granted by an examiner—using a separate first-stage sample of applications for the NBER drugs & medical class that are not attached to the NCEs in my analytic sample. Three criteria must be satisfied for a valid instrument: the relevance condition, exclusion restriction, and monotonicity assumption. The relevance condition requires that this measure of examiner leniency is relevant to explaining whether a patent application is granted. Table 4 gives first-stage results regressing whether a patent application is granted on the leniency measure; the table demonstrates the relevance condition is satisfied. Across specifications, the 0.88 point estimate suggests that a 10-percentage-point increase in an examiner’s average grant rate is associated with a 8.8-percentage-point increase in the likelihood a secondary patent application is granted. The $F$-statistic of these regressions is roughly 1600, substantially above the threshold for weak instruments. The exclusion restriction requires that examiner leniency only affects extra monopoly life through the likelihood a drug’s secondary patent application is granted, which is a reasonable assumption. Finally, the monotonicity assumption requires that patents granted by a stricter examiner would also be granted by a more lenient examiner, and vice versa.

Table 4: First-stage results

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>leniency</td>
<td>0.885***</td>
<td>0.887***</td>
<td>0.885***</td>
<td>0.884***</td>
</tr>
<tr>
<td></td>
<td>(0.0214)</td>
<td>(0.0214)</td>
<td>(0.0218)</td>
<td>(0.0218)</td>
</tr>
<tr>
<td>ln(market size)</td>
<td>-0.00875**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00404)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>2,868</td>
<td>2,868</td>
<td>2,868</td>
<td>2,868</td>
</tr>
<tr>
<td>Approval year FE</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Therapeutic area FE</td>
<td>yes</td>
<td></td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>$F$-statistic</td>
<td>1563.0</td>
<td>1611.6</td>
<td>1590.9</td>
<td>1595.6</td>
</tr>
</tbody>
</table>

Dependent variable in all regressions is an indicator for whether a patent application is granted. Analysis at the drug-application level. Robust standard errors in parentheses. *** p < 0.01, ** p < 0.05, and * p < 0.1.

Table 5 presents the IV results. The estimates in all 4 specifications are substantially larger than the OLS estimates, suggesting that, as expected, selection into whether a patent is granted indeed biases the OLS estimates downward. The preferred specification in Column (4) indicates that a randomly granted secondary patent extends monopoly life by 10.3 months. Given that the average drug has 2.7 secondary patents obtained prior to generic entry, this would suggest nearly 28 months, or more than 2 years, of extra monopoly life attributable to secondary patent acquisition. It is important to note that the IV estimate is not giving the average treatment effect but rather the local average
treatment effect (or LATE)—the effect of an additional patent granted on extra monopoly life for those drugs where changing examiner leniency would change whether the patent application was granted. It does not apply to those applications that would always be granted irrespective of examiner leniency or never granted. That is, we can consider the 10.3 months as an average treatment effect specifically for those patents that are granted by a more lenient examiner but not by a stricter one. If we think that such patents are less likely to be enforced later by the courts (and hence less likely to contribute to delays in generic entry), then the LATE will underestimate the average effect of a secondary patent and this per-drug extension in monopoly life of 28 months will be a conservative estimate of the extent to which secondary patents delay generic entry.

Table 5: Secondary patents and extra monopoly life using variation in examiner leniency: IV analysis

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1_Granted</td>
<td>11.18***</td>
<td>10.33***</td>
<td>10.13***</td>
<td>10.31***</td>
</tr>
<tr>
<td></td>
<td>(2.873)</td>
<td>(2.745)</td>
<td>(2.578)</td>
<td>(2.568)</td>
</tr>
<tr>
<td>ln(market size)</td>
<td></td>
<td></td>
<td></td>
<td>-1.690***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.359)</td>
</tr>
<tr>
<td>Observations</td>
<td>2,868</td>
<td>2,868</td>
<td>2,868</td>
<td>2,868</td>
</tr>
<tr>
<td>Approval year FE</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Therapeutic area FE</td>
<td>yes</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dependent variable in all regressions is extra monopoly life in months (i.e., from a drug’s initial IP expiry to generic entry). Examiner leniency instruments for whether a secondary patent application is granted or not. Analysis at the drug-application level. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, and * p<0.1.

7 Conclusion

Despite the standard “one-patent-one-product” model in intellectual property rights discourse, a common practice by innovators is to acquire numerous patents on different features of the same product to extend its monopoly life. In this paper, I systematically document the presence of imperfect intellectual property rights in the pharmaceutical industry and provide the first evidence on how they extend monopoly life. By exploiting the quasi-random assignment of patent applications to examiners and variation in cross-examiner leniency, I determine that the average drug receives more than 2 years of additional monopoly life due to the accumulation of secondary patents.

This research has important implications for the design of intellectual property rights, innovation incentives, and antitrust concerns both broadly and within the pharmaceutical industry. Some observers argue that the ability of branded firms to make minor tweaks to existing drugs in dosage or formulation disproportionately extends their effective patent terms relative to the R&D costs...
of these improvements.\textsuperscript{20} Others point out that, given the high costs of drug development, these extensions are necessary to incentivize a drug’s initial development. While the average 2-year extension I estimate obviously has sizable impacts for drug pricing and, in turn, consumer welfare, I cannot make a statement on the innovation incentives for pharmaceutical manufacturers and thus overall dynamic welfare impacts. In separate preliminary work, I find that these extensions do not vary with scientific novelty and vary only minimally with drug efficacy, suggesting a potential misalignment between resulting monopoly terms and a drug’s degree of innovation. I leave such questions for future research.

\textsuperscript{20}This relates to the antitrust issue of “product hopping” whereby a manufacturer will cease production of prior dosages or formulations and switch consumers to the newer one at higher monopoly prices.
References


IQVIA (2019). The global use of medicine in 2019 and outlook to 2023: Forecasts and areas to watch.


Lakdawalla, D. N. (2018). Economics of the pharmaceutical industry. Journal of Economic Literature 56(2): 397–449.


A Appendix

A.1 FDA regulatory exclusivities

Table 6 provides term lengths for several key exclusivities; see Thomas (2017) for a description of all 16 FDA exclusivities that currently exist.

Table 6: Key FDA exclusivities by type and term length

<table>
<thead>
<tr>
<th>Exclusivity type</th>
<th>Term length</th>
</tr>
</thead>
<tbody>
<tr>
<td>New chemical entity (NCE)</td>
<td>5 years</td>
</tr>
<tr>
<td>Clinical investigation exclusivity (CIE)</td>
<td>3 years</td>
</tr>
<tr>
<td>– New dosage form or dosing regimen</td>
<td></td>
</tr>
<tr>
<td>– New administration</td>
<td></td>
</tr>
<tr>
<td>– New strength</td>
<td></td>
</tr>
<tr>
<td>– New indication</td>
<td></td>
</tr>
<tr>
<td>– Switch from Rx to OTC</td>
<td></td>
</tr>
<tr>
<td>Orphan drug exclusivity (ODE)</td>
<td>7 years</td>
</tr>
<tr>
<td>Pediatric (PED)</td>
<td>6 months</td>
</tr>
<tr>
<td>Qualified infectious disease products (QIDP)</td>
<td>5 years</td>
</tr>
</tbody>
</table>

A.2 Naive analysis: Comparing drugs by number of secondary IP

Table 7 repeats the analysis of Table 2 using the number of all secondary IP (both patents and exclusivities) as the main explanatory variable.

Table 7: Secondary IP and extra monopoly life: Drug-level analysis

<table>
<thead>
<tr>
<th># secondary IP</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln(market size)</td>
<td>-0.0433</td>
<td>0.209</td>
<td>0.456</td>
<td>0.826</td>
</tr>
</tbody>
</table>
|                | (0.519) | (0.551)| (0.551)| (0.554)| 33

Observations 288 288 288 288
Adj. $R^2$ -0.003 0.123 0.168 0.191
Approval year FE yes yes yes
Therapeutic area FE yes yes

Dependent variable in all regressions is extra monopoly life in months (i.e., from a drug’s initial IP expiry to generic entry). Analysis at the drug level. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, and * p<0.1.
A.3 Application-level analyses with clustered standard errors

Tables 8 and 9 repeat the analyses of Tables 3 and 5 using robust standard errors clustered at the drug level.

Table 8: Secondary patents and extra monopoly life for drugs in granted versus rejected applications: Application-level analysis

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Granted</td>
<td>6.835</td>
<td>6.703</td>
<td>6.838*</td>
</tr>
<tr>
<td></td>
<td>(5.286)</td>
<td>(4.670)</td>
<td>(3.917)</td>
</tr>
<tr>
<td>ln(market size)</td>
<td>-1.664</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.297)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Observations: 3,021 3,021 3,021 3,021
Adj. $R^2$: 0.005 0.225 0.357 0.361
Num. of (drug) clusters: 241 241 241 241
Approval year FE: yes yes yes
Therapeutic area FE: yes yes

Dependent variable in all regressions is extra monopoly life in months (i.e., from a drug’s initial IP expiry to generic entry). Analysis at the drug-application level. Robust standard errors clustered at the drug level in parentheses. *** p<0.01, ** p<0.05, and * p<0.1.

Table 9: Secondary patents and extra monopoly life using variation in examiner leniency: IV analysis

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Granted</td>
<td>11.18</td>
<td>10.33</td>
<td>10.13*</td>
</tr>
<tr>
<td></td>
<td>(7.374)</td>
<td>(6.749)</td>
<td>(5.687)</td>
</tr>
<tr>
<td>ln(market size)</td>
<td>-1.690</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.283)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Observations: 2,868 2,868 2,868 2,868
Num. of (drug) clusters: 237 237 237 237
Approval year FE: yes yes yes
Therapeutic area FE: yes yes

Dependent variable in all regressions is extra monopoly life in months (i.e., from a drug’s initial IP expiry to generic entry). Examiner leniency instruments for whether a secondary patent application is granted or not. Analysis at the drug-application level. Robust standard errors clustered at the drug level in parentheses. *** p<0.01, ** p<0.05, and * p<0.1.