Peltzman Revisited: Quantifying 21st Century Opportunity Costs of FDA Regulation

Casey B. Mulligan
DECEMBER 2021
PELTZMAN REVISITED:
QUANTIFYING 21ST CENTURY OPPORTUNITY COSTS OF FDA REGULATION

Casey B. Mulligan

December 2021

I appreciate discussions with Sam Peltzman, Tom Philipson, Andrei Shleifer, Alex Tabarrok, Cliff Winston, Kevin Murphy, Bob Topel, and Troy Durie and financial support from the University of Chicago’s Initiative on Enabling Choice and Competition in Healthcare.

© 2021 by Casey B. Mulligan. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.
Peltzman Revisited: Quantifying 21st Century Opportunity Costs of FDA Regulation
Casey B. Mulligan
December 2021
JEL No. I18,L51,L65,O31

ABSTRACT

This paper revisits Peltzman (1973) in light of two recent opportunities to quantitatively assess tradeoffs in drug regulation. First, reduced regulatory barriers to drug manufacturing associated with the 2017 reauthorization of Generic Drug User Fee Amendments were followed by significantly more entry and lower consumer prices for prescription drugs. Using a simple and versatile industry model and historical data on entry, I find that easing generic restrictions discourages innovation, but this welfare cost is more than offset by consumer benefits from enhanced competition, especially after 2016. Second, accelerated vaccine approval in 2020 had unprecedented net benefits as it not only improved health but substantially changed the trajectory of the wider economy. The evidence suggests that cost-benefit analysis of FDA regulation is incomplete without accounting for substitution toward potentially unsafe and ineffective treatments that are both outside FDA jurisdiction and heavily utilized prior to FDA approval. Moreover, the policy processes initiating these 21st century regulatory changes show a clear influence of Peltzman’s 1973 findings.

Casey B. Mulligan
University of Chicago
Department of Economics
1126 East 59th Street
Chicago, IL 60637
and NBER
c-mulligan@uchicago.edu
“...consumer losses from purchases of ineffective drugs or hastily-marketed unsafe drugs appear to have been trivial compared to their gains from innovation.” Peltzman (1973), describing drug markets before the Food and Drug Administration was tasked with certifying drug efficacy.

I. Introduction

An important fraction of improved living standards in the past, and likely the future, has originated from new medical products. Both markets and regulators have the potential to contribute to, or detract from, the innovative process. On the market side there are concerns that competition may erode financial rewards to innovation, or that large firms may be too bureaucratic to foster the consideration of new products and methods. Meanwhile government stands as a gatekeeper for new medical products for the stated purpose of protecting consumers.

A leading example of the regulation of new medical products is the 1962 “Drug Efficacy Amendment” (EA) to the Federal Food, Drug, and Cosmetic Act, which made proof of efficacy a requirement for the approval of new drugs by the Food and Drug Administration (FDA). Peltzman (1973) pioneered cost-benefit analysis of the EA by estimating the consumer benefit (if any) of curtailing the sale of ineffective drugs and comparing it to the opportunity cost of effective drugs that were not introduced into the U.S. market due to the additional approval costs created by the EA. Peltzman concluded that the EA imposed a net cost on consumers of magnitude similar to a “5-10 percent excise tax on all prescriptions sold.”

Given the sudden and obvious reduction in the rate at which new drug formulas were introduced into the market after 1962, perhaps the greatest challenge Peltzman faced was quantifying the degree to which the foregone drugs would have been ultimately deemed as ineffective by consumers and their physicians. Two drug market events between 2017 and 2021 offer fresh perspectives on the consumer costs and benefits of the entry barriers created by the FDA approval processes. One relates to the FDA regulation of the manufacturing of generic

---

1 Early analyses of these concerns appear in Schumpeter (1942), Arrow (1962), and Schmookler (1972).
drugs, for which there is little scope for protecting consumers from ineffective products because, as a group, consumers already have years of experience accumulated while the drug was produced under patent. The other relates to approval delays of COVID-19 vaccines, which have particularly concrete opportunity costs as well as further illustrating some of Peltzman’s influence on public policy.

Section II briefly reviews Peltzman’s methodology. Section III describes the U.S. generic drug market in recent years. A simple and versatile conceptual model of prices and entry is provided in Section IV for the purpose of quantifying the welfare benefits of the deregulation of generic entry that occurred since 2012, without restricting the values of the price elasticity of demand or the level of marginal cost. In this light, the generic entry data suggest that easing generic restrictions discourages innovation, but that this cost is more than offset by consumer benefits from enhanced competition, especially after 2016. Section V describes the timing of COVID-19 vaccine development and approval. An excess burden framework is applied to better measure the opportunity cost of regulatory delays, including substitution towards potentially harmful remedies that need not demonstrate safety or effectiveness because they are outside FDA jurisdiction. Section VI concludes.

II. Peltzman’s Methodology

Peltzman’s most memorable finding was a sharp and persistent decline in new chemical entities (NCEs) beginning in 1963, at the same time that drug market trends suggested that the number of NCEs would have been similar to what it was in the 1950s. His point estimate of the effect of the 1962 EA is a reduction of NCEs from 41 per year to 16 (about 60 percent). He also notes that U.S. NCEs fell 54 percent relative to British NCEs.

Peltzman acknowledges that the EA was intended to reduce the number of NCEs by eliminating the ineffective ones that prior to the EA consumers were purportedly duped into purchasing until experience revealed their true value. Indeed, his examination of drug evaluations by the American Medical Association (AMA) shows that (i) AMA-designated “ineffective drugs” were more common before 1963 and (ii) such drugs lost market share after introduction. However, this category of drugs was a somewhat greater percentage of drug sales
after the EA than before. Moreover, NCEs as a whole did not maintain sales over time any better after the EA, when the FDA was certifying them as effective.

A significant part of his paper, and this one, is dedicated to quantitatively assessing the tradeoffs of regulating drug-market entry. Peltzman (1973, p. 1090) concluded that the net effect of the EA on consumers was “comparable to their being taxed something between 5 and 10 percent on their … drug purchases.”

Peltzman’s paper influenced both academics and policymakers. Klein and Tabarrok (2002) cite a wide range of academics sharing Peltzman’s conclusions. For example, M.I.T. Professor Peter Temin (1980, p. 206) said that “whether or not people are capable of understanding the relevant information, I would still favor giving people more choice for their own well-being than the current [FDA] system allows.” Philipson and Sun (2008) and others followed Peltzman in evaluating the effects of new FDA regulation on drug approvals. Philipson et al. (2008), for example, found that 1992 FDA reforms significantly increased consumer surplus and the returns to innovation with a minimal cost in terms of drug safety, which influenced further FDA deregulation efforts under George W. Bush and later in the Trump Administration when Philipson was part of the senior White House staff.2

Pointing specifically to Peltzman’s paper, the 1975 Economic Report of the President (p. 159) concluded that “existing laws and institutions are imposing significant costs on the economy.” The 1987 Economic Report of the President also cited it as part of its discussion of recent FDA changes, including rules around generic entry, intended to relax “unnecessarily stringent regulatory requirements [that] can lead to more deaths and lower health levels” (Council of Economic Advisers January 1987, pp. 194-5). The 1989 Economic Report of the President listed several possible changes to FDA regulation, citing Peltzman’s results as to the agency’s dramatic effect on the rate of innovation (Council of Economic Advisers January 1989, pp. 218-20). Several reports from the Trump Administration focused on the fact that FDA

---

2 These deregulation efforts, and an emphasis on medical innovation, are described further in Mulligan (2020a). The 2018 Economic Report of the President had a full chapter about the health sector, half of which was about “Improving People’s Health through More Access to Medical Innovations” and “Encouraging Innovation, and Making It Affordable.” The 2019 Economic Report of the President looked at the possible negative innovation effects of a proposed federal ban on for-profit healthcare.
III. Generic-Drug Approvals and Prescription-Drug Prices

As Peltzman recognized, the challenge to signing the welfare effect of FDA entry barriers is that the barriers are often tied to the agency’s certification of effectiveness. Generic drug approvals are interesting in this regard because they involve drugs familiar to the market and that already had been certified as effective when the FDA approved them as NCEs.\(^3\) Although empirical relationships between drug prices and the number of generic manufacturers have been estimated in previous work, it is not easy to identify the reasons why some markets have more generics than others. Recent changes in generic drug regulation thereby offer a unique new opportunity to assess welfare effects of FDA entry barriers.

About 90 percent of U.S. prescriptions are filled with generics (Pew Charitable Trusts 2019). A manufacturer aspiring to sell an off-patent generic drug must submit an Abbreviated New Drug Application (ANDA) for FDA approval, as distinct from the New Drug Applications (NDAs) required for introducing a new compound into the market. Although not requiring the safety and effectiveness studies that are part of NDAs, ANDAs are still costly. As of 2015, the FDA had thousands of ANDAs pending, resulting in a median approval time of 42 months (Pew Charitable Trusts 2019, Table 1).\(^4\) About a quarter of markets, and likely more, had only one generic manufacturer approved.\(^5\)

---

\(^3\) For this reason, Tabarrok (2015) and others have argued for international regulatory reciprocity for generic drugs. Namely, American consumers should be permitted to import generic drugs that have been approved abroad, are no longer protected by U.S. patents, and have FDA-approved NDA.

\(^4\) The Hatch-Waxman Act requires FDA approval within 180 days (21 U.S.C. \(355(j)(5)(A)\)), but provided no enforcement mechanism.

\(^5\) Berndt, Conti and Murphy (2017). For application purposes, the FDA defines a market as an active ingredient(s), dosage form, and route of administration, “ignoring differences in strength and packaging sizes” (Food and Drug Administration 2017b). According to the FDA *Orange
Reduced barriers to generic entry have the potential to significantly reduce drug prices by increasing competition. FDA (2017b) found that the generic was sold at only a six percent discount to the branded drug when the market had only one approved generic, but at 48 and 56 percent discounts when the market had two or three approved, respectively. With less than 3000 extant markets at a point in time (Food and Drug Administration 2021), these findings suggest that a policy increasing a year’s approvals by, say, 300, could reduce drug prices by two percent over a one-year time frame in addition to any affect they had on the branded price or on prices in the markets for therapeutic substitutes.

Using data on actual quantities purchased as well as measuring prices before and after ANDA approval, Conrad et al (2017) estimated that the ANDA approvals in 2017 alone would reduce annual consumer expenditures by $16 billion in a $324 billion prescription-drug market (IQVIA Institute for Human Data Science 2018). This includes changes in the branded price in the same market with ANDA approved, but not price changes for therapeutic substitutes.

The 1984 Hatch-Waxman Act required the FDA to approve or disapprove ANDAs within 180 days. Rarely meeting this requirement, the FDA asserted that it needed more funds to hire review staff (Woodcock 2016). With the 2012 Generic Drug User Fee Amendments (GDUFA I), Congress authorized user fees that add a cash barrier to entry but “dedicated toward expediting … the review of human drug applications.” Between 2012 and 2015, the median time from ANDA receipt to approval further increased from 32 months to 42 months (Pew Charitable Trusts 2019, Table 1). The number of final and tentative approvals failed to increase.

The number of foreign and domestic facilities manufacturing generic drugs declined. Book, 1447 markets in 2015 had at least one approved generic manufacturer and 323 of those had exactly one. Moreover, 1275 markets had zero generic manufacturers. Later the FDA would find that hundreds of such markets had no exclusivities or blocking patents (Food and Drug Administration 2017c). See also Berndt, Conti and Murphy (2017, Figure 11).

Caves et al (1991) find only modest changes in brand prices when generic entry occurs. However, their data was from 1976-1987 when brand quantity shares (measuring conditional on generic competition or for the sector as a whole) were triple or quadruple what they would be in the 2010s (Grabowski, et al. 2016).

See also Lichtenberg (2021) on the importance of competition between therapeutic substitutes.

Section 101(b) of Public Law 112-144.

The FDA Orange Book shows 433, 400, and 483 ANDA final approvals in fiscal years 2013-15, as compared to 413, 454, and 510 for the prior three fiscal years (before GDUFA I).
The apparent lack of success of GDUFA I may seem surprising given its apparent similarities to the 1992 Prescription Drug User Act (PDUFA), which required FDA to meet performance goals in its processing of NDAs in exchange for collecting fees from manufacturers. Perhaps PDUFA was successful in reducing approval times (Cantor 1997, Vernon, et al. 2009) because it simultaneously benefitted manufacturers, consumers, and FDA. In contrast, accelerated processing of ANDAs includes a significant amount of redistribution to consumers from expired patent holders (and perhaps also the generic manufacturers participating in markets with only one generic), which I quantify in Section IV.

Three related changes occurred in 2017. By May 9, 2017, Scott Gottlieb was nominated and confirmed to head the FDA. He had been an outspoken critic of the FDA’s slow approval process, which he described as “evading the law” (Gottlieb 2010). He immediately told Congress that his FDA would prioritize competition (U.S. House of Representatives, Committee on Appropriations, Agriculture Subcommittee 2017). In June, FDA (2017a) announced a Drug Competition Action Plan with procedural details published in November (Food and Drug Administration, Office of Generic Drugs 2017). It immediately promulgated, and subsequently maintained, a list of drugs with no blocking patents or exclusivities but still no approved generics (Food and Drug Administration 2017c). Section 801 of GDUFA II, which became law in August, instructed FDA to prioritize the review of drugs with no blocking patents or exclusivities that have three or fewer ANDAs or NDAs already approved. On paper at least, FDA appeared to be looking toward competition rather than purely bureaucratic metrics such as numbers of applications and approval times.

Table 1, based on the FDA’s Orange Book listing approved NDAs and ANDAs, shows approvals at a higher rate during Gottlieb’s tenure as compared to either the two years before or the two years after. Before Gottlieb, and therefore also before GDUFA II, FDA averaged 54 approvals per month (1,286 total for the 24 months). The average was 73 per month during Gottlieb’s tenure (through April 2019) and 61 in the subsequent 24 months. FDA approvals of reported similar results for the sum of final and tentative ANDA approvals. Food and Drug Administration (2016, slide 18) and Figure 2 of Berndt, Conti and Murphy (2018).

10 All but three of these 48 months were under GFUA II. The number of drugs coming off patent seems to be fairly constant over time after peaking between 2012 and 2014 (Council of Economic Advisers October 2018, Figure 3). The t-statistic for the hypothesis that Gottlieb’s
new drugs were also high during his tenure, as shown in the second row of the table. The third and fourth rows show the corresponding approvals for biologics reported in FDA’s Purple Book, which are a large share of expenditure on physician-administered drugs but a small share of retail prescription drugs.\textsuperscript{11}

**Table 1. Entry and price changes in drug markets**

by two-year period

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ANDAs approved</td>
<td>1,286</td>
<td>1,754</td>
<td>1,475</td>
</tr>
<tr>
<td>NDAs approved</td>
<td>184</td>
<td>229</td>
<td>202</td>
</tr>
<tr>
<td>Biosimilars approved</td>
<td>4</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>New biologics approved</td>
<td>73</td>
<td>82</td>
<td>62</td>
</tr>
<tr>
<td>Rx CPI change relative to all items, %</td>
<td>3.7%</td>
<td>-1.5%</td>
<td>-5.1%</td>
</tr>
</tbody>
</table>

Addendum: Inflation-adjusted changes in unit cost between calendar years, in %

-0.1%  -3.6%  N/A

Addendum: Medicaid/CHIP enrollment change, per capita, in %

3.4%  -5.3%  12.8%

Notes: CPI and Medicaid changes are from April to April. Unit cost is per-prescription consumer expenditure, including insurance plan expenditures net of rebates and discounts. Gottlieb was FDA commissioner May 2017 - April 2019. Biosimilars and new biologic approvals include Supplemental BLAs.

Sources: FDA Orange Book, FDA Purple Book, BLS CPI series CUSR0000SEMF01 and CUSR0000SA0, Express Scripts/Evernorth Drug Trend Reports

---

\textsuperscript{11} Biologics, which include vaccines, gene therapy, and Insulin, have applications akin to NDAs (Biological License Applications under 351(a)) and ANDAs (BLAs under 351(k). Both types of BLAs are approved by the FDA’s Center for Biologics Evaluation and Research and tabulated in the FDA’s Purple Book. Note that biologics are more than 70 percent of total drug spending under Medicare Part B (physician-administered drugs) and less than 10 percent of total drug spending under Medicare Part D (retail prescription drugs) (United States Department of Health and Human Services 2020, Anderson-Cook, Maeda and Nelson 2019).
Drug-market performance appears to reflect additional competition. Berndt, Conti and Murphy (2018) found that, as of April 30, 2017, Teva Pharmaceutical owned 1,611 ANDAs out of about 10,000 in existence. The second largest owner was Mylan Inc., with 668 ANDAs. Teva’s stock crashed in the summer of 2017, with Teva’s CEO reporting that the company would henceforth be less profitable owing to “greater competition as a result of an increase in generic drug approvals by the U.S. FDA” (Sheetz 2017). Real retail prescription drug prices fell 1.5 percent during Gottlieb’s tenure, as compared to a 3.7 percent increase in the prior two years and a 5.1 percent decrease in the subsequent two years.12 These findings are consistent with the hypothesis that the GDUFA II, or Gottlieb’s management, or some combination thereof, increased drug-market competition by reducing entry barriers.

Duggan and Morton (2006) argued that Medicaid enrollment is also an important driver of prescription-drug prices because Medicaid reimburses based on the average price in the commercial market. Drug companies setting their commercial prices know that raising them will reduce utilization among their commercial customers, but would automatically increase Medicaid revenue with hardly any effect on Medicaid utilization. The relative importance of these factors depends on the share of customers enrolled in Medicaid versus commercial plans. Between April 2015 and April 2017, per capita Medicaid enrollment increased 3.4 percent and then fell in the subsequent 24 months. Real drug prices show a similar qualitative pattern, except in the third period when real drug prices and Medicaid enrollment move sharply in opposite directions. More research is needed to determine how much of the pattern Duggan and Morton find across drug markets is expected to be observed in the national time series.

---

12 CEA (October 2019) explains why the CPI indicates the contribution of prescription drugs to the cost of living and how it relates to other drug-price-inflation measures. As a cost-of-living index, the CPI is based only on price changes within product and (unlike unit cost) does not reflect any increase from the introduction of a new expensive product because consumers can opt not to purchase the new product. Table 1’s first addendum row shows results for unit cost measured from the Express Scripts (now EverNorth) database of U.S. consumer costs for prescription drugs, which is one of the largest in the world.
IV. Welfare Consequences of Generic Entry Barriers

IV.A. Competition, Entry-Expenditure, and Innovation Costs

Entry barriers associated with ANDAs have three kinds of welfare consequences: a competition effect, an entry-expenditure effect, and an innovation effect. One is the “static” effect on ongoing competition for the supply of drugs after patent expiration. Social welfare increases with the quantity of drugs sold to consumers with marginal value exceeding the marginal cost of production. A second effect is that entry barriers are real costs, in the form of delay as well as labor effort at FDA and manufacturers, that are incurred to the extent that entry occurs. NDAs impose such costs too, except that the ANDA costs are paid by generic manufacturers and NDA costs paid by the first manufacturer of the drug. It is theoretically possible that barriers relaxed, but not eliminated, increase entry enough to increase aggregate expenditure on ANDAs and this added cost exceeds the benefit of additional competition. As I show below, much can be said about the sum of these two costs with minimal assumptions about industry demand and supply.

A third effect of ANDA barriers is that, by limiting entry, they indirectly encourage innovation by protecting incumbents, including the incumbent who brought the product to market in the first place. Quantifying this effect requires, in addition to “static” information about entry and the distribution of surplus among producers, information or assumptions about the effect of post-patent profits on the number of new drugs brought to market. Entry costs that are high enough to prevent all generic entry, and thereby not paid in equilibrium, would be the economic equivalent of an infinite patent life. Lesser generic entry barriers also resemble infinite patent life by indefinitely elevating incumbent profits, except that such barriers (i) elevate profits less than the infinite patent would and (ii) involve generic entry costs, which are real costs that would not be incurred if patent lives were infinite. Nevertheless, if actual patent lives were enough less than the social-surplus maximizing patent life, generic entry barriers could have a net social benefit through their oblique transfer of surplus to patent holders after their patent expires.

To focus issues, this paper evaluates the third welfare effect of generic entry assuming that the optimal patent duration has already been achieved and leaves it to future research to
assess whether actual patent lives are too short or too long.\textsuperscript{13} Specifically, the elasticity of new drugs with respect to innovator profits is assumed to be constant with a value that exactly offsets the static benefits (evaluated at pre-GDUFA entry) of marginally shortening the patent life. The optimal-duration assumption by itself implies a significant innovation benefit of a marginal increase in innovator profits while the constant-elasticity assumption dictates the rate at which the marginal innovation benefit diminishes.

Reducing generic entry costs enough to induce exactly one generic necessarily increases aggregate expenditure on entry. Because generic entry can reduce innovation and, by some estimates (Food and Drug Administration 2017b), have a limited competition effect, social welfare can be less with one generic entrant than with zero. But further entry-barrier reductions (enough to have more than one generic) would be associated with greater competition effects, smaller marginal innovation effects, and potentially less entry expenditure. This theoretical result is interesting to consider in light of GDUFA II’s emphasis on second and third generic entrants, in contrast with previous FDA metrics that weighted all ANDAs equally regardless of their market consequences.

\textbf{IV.B. Cournot Competition with Unit Pass-through}

Just a few market assumptions deliver a number of quantitative results about the relationship between entry barriers and market performance. First, I assume that market demand is the same before and after patent expiration with all manufacturers in the same market producing perfect substitutes. Second, I assume Cournot competition after the patent expires, which means that each entrant chooses its quantity taking as given the quantities produced by other manufacturers. The competition is symmetric, except that the innovator does not pay ANDA costs. Third, pass-through of common marginal costs is one-for-one regardless of the

number of entrants. This passthrough implies that all (infinite) potential entrants face the same ANDA and marginal production cost, which I denote \( a \) and \( c \), respectively. It also follows that market demand has a constant semi-elasticity with respect to price.\(^{14}\) Conversely, any market demand curve with constant (negative) semi-elasticity has unit pass-through in the Cournot equilibrium conditional on the number of entrants.

The equilibrium pricing function conditional on the number of manufacturers \( n \geq 1 \) is the sum of marginal cost \( c \) and a term inversely proportional to \( n \). Policies that change \( a \), which are the focus of this section, trace out this pricing function. In particular, both the monopoly and duopoly prices can be many times greater than the perfectly-competitive price \( c \). For many purposes, these theoretical pricing results are a good approximation of empirical findings for drug markets (Food and Drug Administration 2017b, Dave, et al. 2017, Conrad and Lutter 2019). In contrast, drug markets are poorly described by Bertrand competition (for a homogeneous product), which has trouble explaining costly entry as it predicts that price would be \( c \) regardless of whether the market has one generic or ten.\(^{15}\)

More surprising, the pass-through assumption delivers a number of quantitative static welfare conclusions – that is, conclusions about the sum of the competitive and entry-expenditure effects – without restricting the values of the price elasticity or semi-elasticity of demand or the level of marginal cost. The relationship between the ANDA cost \( a \) and aggregate expenditure on them, \((n-1)a\), follows a single-peak Laffer-curve shape. Aggregate ANDA expenditure is zero both in the monopoly case and in the limit as \( a \) goes to zero. Ignoring integer constraints, the peak ANDA expenditure is at the ANDA cost supporting \((3 + \sqrt{5})/2 \) (about 2.6) manufacturers, as shown in Appendix I.

\(^{14}\) Vives (1999, p. 104). A demand curve with a constant semi elasticity is consistent with any (negative) value for the equilibrium price elasticity of demand. Like many other demand curves, it is more price elastic at higher prices. Although the demand curve has no choke price, the area under the demand curve is finite.

\(^{15}\) The tension between Bertrand competition and costly market entry is known as the “Bertrand paradox” (Mukherjee 2005). Indeed, for explaining drug-market pricing, even the Cournot model may not be far enough from Bertrand because the former predicts that the first generic entrant reduces the markup by a factor of two, which appears to be more than occurs empirically. Also note that the crucial assumption of the Cournot model for my purposes is not that manufacturers “set quantities rather than prices” but rather that each manufacturer perceives that it must select an own quantity and price combination from a “demand” curve (inclusive of all competitors’ strategic reactions) that has the same slope as the market demand curve.
Ignoring integer constraints, the static marginal benefit of relaxing ANDA barriers is zero in the neighborhood of the entry barrier supporting $n = 1$ as an equilibrium. The innovation (opportunity) costs are strictly greater than zero, which means that the first generic entrant can have a net social cost. However, static social surplus strictly declines with ANDA costs $a$ at any value of $a$ supporting $n > 1$. Reducing $a$ in the amount that induces exactly one more entrant creates the most static welfare when it is adding a third manufacturer (i.e., second generic) and second-most when adding the fourth manufacturer.\footnote{In decreasing order of static social-welfare increments, the subsequent additions are fifth, second, sixth, seventh manufacturers, etc. Note that, even by the static criterion, the Cournot equilibrium has excessive entry at any given level of $a$ (Mankiw and Whinston 1986).}

This model also says a lot about how the social benefits of innovation are allocated between the innovator and consumers. It predicts innovator profits and social surplus for the $T$ years that the patent is effective as well as the years after patent expiration. The ratio of the present value of innovator profits to the present value of social surplus is (1)

$$\text{innovator share} = \frac{1}{2} \left[ 1 - \frac{n_1^2 - 1}{2} \right] = \frac{1}{1 + n_1^2 + (e^{\rho T} - 1)2n_1^2 e^{\frac{n_1-1}{n_1}}} \in \left(0, \frac{1}{2}\right)$$

where $\rho$ is the annual discount rate and $n_1 > 1$ is the number of manufacturers after patent expiration. The discount rate reflects not only the time value of money but also any exponential trends in market demand, such as decay that may occur as therapeutic substitutes come on the market (Lichtenberg and Philipson 2002). The fact that the share is significantly less than one means that the social benefit of innovation significantly exceeds what the innovator spends on it. Without any assumption about the relative costs and benefits of a $T$-year patent life, Figure 1 shows level curves of the innovator share as a function of $\rho T$ and $n_1$, suggesting that innovators capture about a quarter of the social surplus.\footnote{The relationship between innovation share and $n_1$ in equation (1) and Figure 1 is a parametric relationship traced out by changing the ANDA cost $a$. Using a linear demand assumption, Philipson and Jena (2006) estimate a smaller share for HIV drugs. Nordhaus (2004) estimates an even smaller share for the economy as a whole, which includes innovations without legal intellectual property rights.} Not surprisingly, this share increases with the patent duration and the discount rate but decreases with the number of post-patent competitors.
Figure 2 shows level curves for the elasticity of the quantity of new drugs with respect to innovator profits that justifies $T$ years as the optimal patent life. This is the critical elasticity I use in what follows to weight the innovation benefits of protecting incumbents against the static social benefits of enhanced competition.

**Figure 1. Level curves of the innovator's share of surplus**

**Figure 2. Level curves of the elasticity of innovation with respect to profit**

that justifies the actual patent length.
Table 2 shows the welfare consequences of increasing generic entry by one manufacturer, achieved through reduced ANDA barriers. The first column refers to the case in which one generic enters a post-patent market that had none, the second column a second generic entering the market, etc. Each column’s entry is simulated by reducing the ANDA cost $a$ from the value supporting $n_1$ equilibrium generics to the value supporting $n_1 + 1$. The first three rows of the table refer to impacts on post-patent flows of surpluses and costs, with units normalized so that the profit flow under patent is one. Row (1) is the competition effect, which unsurprisingly is greatest for the first generic entry because of the large gap between marginal cost and marginal social value under monopoly. Even though the innovator loses profit, combined producer and consumer surplus increase by an amount that is 47.3 percent of the profit a monopoly manufacturer would earn. The increments to surplus fall with the second, third, etc., generic entrants. On the other hand, ANDA costs are incurred with the first generic that would not be incurred under monopoly. This expenditure amounts to 41.2 percent of the profit a monopoly manufacturer would earn. Reducing per-entrant costs enough to induce a second entrant also increases aggregate entry costs, although much less than the first entrant did. Further reductions in per-entrant costs reduce aggregate costs. Although not shown in the Table, in the limit aggregate entry costs go to zero even though the number of entrants is unbounded.
Table 2. Welfare consequences of reducing generic entry barriers

<table>
<thead>
<tr>
<th>Reduced barriers increase number of generic manufacturers from:</th>
<th>0 to 1</th>
<th>1 to 2</th>
<th>2 to 3</th>
<th>3 to 4</th>
<th>4 to 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impact on post-patent flows, expressed relative to profit flow under patent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) Consumer and producer surplus, ignoring entry costs</td>
<td>0.473</td>
<td>0.124</td>
<td>0.049</td>
<td>0.024</td>
<td>0.014</td>
</tr>
<tr>
<td>(2) Aggregate generic entry costs</td>
<td>0.412</td>
<td>0.021</td>
<td>-0.036</td>
<td>-0.041</td>
<td>-0.037</td>
</tr>
<tr>
<td>(3) Static surplus after expiration = (1) - (2)</td>
<td>0.061</td>
<td>0.103</td>
<td>0.085</td>
<td>0.065</td>
<td>0.050</td>
</tr>
<tr>
<td>Impact on annuitized values, discount rate 6% and patent life 11.5, from the perspective of patent beginning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4) Static surplus after expiration</td>
<td>0.031</td>
<td>0.052</td>
<td>0.043</td>
<td>0.033</td>
<td>0.025</td>
</tr>
<tr>
<td>(5) Innovation benefit (- is a cost), with elasticity 0.076 justifying patent life</td>
<td>-0.053</td>
<td>-0.024</td>
<td>-0.012</td>
<td>-0.006</td>
<td>-0.004</td>
</tr>
<tr>
<td>(6) All three welfare components = (4) + (5)</td>
<td>-0.023</td>
<td>0.028</td>
<td>0.031</td>
<td>0.026</td>
<td>0.021</td>
</tr>
<tr>
<td>(7) Cumulative from 0 generics</td>
<td>-0.023</td>
<td>0.005</td>
<td>0.036</td>
<td>0.063</td>
<td>0.084</td>
</tr>
<tr>
<td>Impact on annuitized values, discount rate 4% and patent life 11.5, from the perspective of patent beginning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(8) Static surplus after expiration</td>
<td>0.038</td>
<td>0.065</td>
<td>0.054</td>
<td>0.041</td>
<td>0.032</td>
</tr>
<tr>
<td>(9) Innovation benefit (- is a cost), with elasticity 0.060 justifying patent life</td>
<td>-0.056</td>
<td>-0.028</td>
<td>-0.014</td>
<td>-0.008</td>
<td>-0.005</td>
</tr>
<tr>
<td>(10) All three welfare components = (8) + (9)</td>
<td>-0.018</td>
<td>0.038</td>
<td>0.039</td>
<td>0.033</td>
<td>0.027</td>
</tr>
<tr>
<td>(11) Cumulative from 0 generics</td>
<td>-0.018</td>
<td>0.020</td>
<td>0.059</td>
<td>0.092</td>
<td>0.119</td>
</tr>
</tbody>
</table>

Note: Present values are from the perspective of the effective beginning of the patent and are then annuitized (over a perpetual horizon) in order to compare to the profit flow under patent. This comparison profit flow does not vary across columns or rows. Discount rates and patent lives matter only as a product.
Table 2’s row (3) shows the net of the previous two lines. This static surplus is positive for all increments although, as noted previously, is greater for the second, third, and fourth generic than for the first. As with results shown in Figures 1 and 2, this part of the Table does not depend on the magnitude of the semi-elasticity of demand with respect to price. Moreover, because it refers only to post-patent flows, it does not depend on the discount rate or patent duration.

Although not shown in the table, the redistribution from innovator to consumers due to the first generic is almost ten times the addition to static surplus shown in the table’s row (3). The amount of redistribution not only speaks to the political economy of generic-entry policy, but also raises concerns that generic entry may discourage innovation. The second and third panels of Table 2 account for the innovation effects under two alternative assumptions about the product $\rho T$ of the discount rate and patent length. The product is six (four) percent of 11.5 years for the second and third panels, respectively.

Generic entry policy affects surplus flows, such as those shown in the table’s rows (1) and (3), after patent expiration. Using the two alternative discount factors, rows (4) and (8) discount row (3) back to the date when the innovator’s product entered the market and then converts into a perpetual annuity. As a result, row (4)’s annuitized values of static surplus are about half of the corresponding entry in row (3). The discount factor used to calculate row (8) is further from zero due to the lower discount rate.

Generic entry reduces innovator profit, which reduces the supply of new drugs to the marketplace (Vernon, et al. 2009). The actual patent life would have been optimal on average prior to GDUFA if the elasticity of new drugs with respect to innovator profit were 0.076 (0.060), which are the values assumed for the second (third) panel, respectively.20 The present

---

18 See also Appendix I.
19 The Congressional Research Service (Schacht and Thomas 2012) estimates that the average effective patent life for drugs to be 11.5 years, including the extensions provided by the Hatch-Waxman Act. The effective life is less than the statutory life (20 years) due to the patent time used during the FDA approval process. Table 2’s results would be identical if instead, for example, the products were taken as three and four percent of 23 years.
20 For the pre-GDUFA optimal-patent-life analysis, an average of three generic entrants is assumed after patent expiration. Each of the two panel’s assumptions therefore corresponds to its own point in Figure 2 at a horizontal position of four manufacturers. See Appendix II for the
value of the foregone social surplus associated with the foregone new drugs is shown in rows (5) and (9), as negative numbers because the foregone drugs are a cost. Rows (6) and (10) show the sum of the innovation benefit and the post-patent static surplus. For the first generic entrant, the foregone innovation dominates but the static surplus dominates for subsequent entrants. After the second entrant, the static surplus is a good approximation of the total welfare consequences.

Rows (7) and (11) show the cumulative welfare consequences – that is, the net benefit of having 1, 2, 3, etc., generics compared to zero. Contrary to what is suggested by the static surplus, social welfare is less with one generic than with zero. But the zero-generic surplus is still less than the surplus from two or more generics. Comparing the second and third panel shows how these results are fairly insensitive to the discount term \( \rho T \). The priority given by GDUFA II to second and third entrants may have an economic justification in Table 2.

Table 3 applies Table 2’s results to the empirical distribution of drug markets as measured in the FDA Orange Book. Following FDA protocol for NDAs and ANDAs, I define a drug market to be the cross between active ingredient(s), dosage form, and route of administration, “ignoring differences in strength and packaging sizes” (Food and Drug Administration 2017b). For each market and month, I used the Orange Book to track the number of NDAs and ANDAs so far approved for that market. The total number of approved manufacturers for each cell is then merged with the cumulative benefits and costs shown in Table 2, extended to include results for more than five generic manufacturers. Each cumulative benefit and cost is then averaged across markets at a point in time, with the results shown in the left half of Table 3. The right half of the table derives changes over time.

Rows (1) and (2) of Table 3 show how first (and perhaps second) generic approvals were comparatively important between 2012 and 2016 because, on average, competition benefits algebra deriving the innovation elasticity and then using it to calculate Table 2’s rows (5) and (9).

\(^{21}\) Applicants with more than one approved NDA or ANDA are counted only once. A weakness of the Orange Book data is that “…both FDA and industry personnel believe a substantial number of approved ANDAs are no longer marketed, but it is unknown how large is their number” (Berndt, Conti and Murphy 2018). To the extent that ANDA exit occurs in low-volume markets, it may not be important for the industry surplus results of this paper.

\(^{22}\) If none of the markets at a point in time had a generic, then the first two rows of the corresponding column in the left half of Table 3 would be zero. If every market had exactly one generic, then those entries would be the same 0.473 and 0.412 entries shown in Table 2, etc.
increased more during that period (0.018) than 2016-20 (0.014) but so did entry costs. Combined static surplus increased more during the latter period: 0.019 compared to 0.013. Rows (5) and (9) suggest that innovation opportunity costs increased about the same amount during the two periods, so that the combination of all three welfare components increased almost twice as much (about 0.008) between 2016 and 2020 than it did 2012-16. Over the eight years the combined benefits are, amortized over time, over one percent of the profit flow received by the patent holder prior to the patent expiration. Assuming that a significant majority of drug revenue prior to expiration goes toward patent-holder profit, then the net social benefits are an amount equivalent to about one percent of revenue prior to expiration. This appears to be of the same order of magnitude, although likely smaller than, Peltzman’s estimate of the net costs of the 1962 EA, which he found to be about eight percent of industry revenue.23

---

23 Peltzman’s revenue base included revenues from drugs both on and off patent.
Table 3. Welfare consequences of generic entry under GDUFA I & II

Assuming all changes in the empirical distributions of number of generics derived from entry costs

| Impact on post-patent flows, expressed relative to profit flow under patent | Empirical distributions compared with no generics | Changes over time |
|---|---|---|---|
| | 2012 | 2016 | 2020 | 2012-16 | 2016-20 | 2012-20 |
| (1) Consumer and producer surplus, ignoring entry costs | 0.320 | 0.337 | 0.351 | 0.018 | 0.014 | 0.031 |
| (2) Aggregate generic entry costs | 0.174 | 0.178 | 0.173 | 0.004 | -0.005 | -0.001 |
| (3) Static surplus after expiration = (1) - (2) | 0.146 | 0.160 | 0.179 | 0.013 | 0.019 | 0.032 |

| Impact on annuitized values, discount rate 6% and patent life 11.5, from the perspective of patent beginning | Empirical distributions compared with no generics | Changes over time |
|---|---|---|---|
| | 2012 | 2016 | 2020 | 2012-16 | 2016-20 | 2012-20 |
| (4) Static surplus after expiration | 0.073 | 0.080 | 0.090 | 0.007 | 0.009 | 0.016 |
| (5) Innovation benefit (- is a cost), with elasticity 0.076 justifying patent life | -0.044 | -0.047 | -0.049 | -0.003 | -0.002 | -0.005 |
| (6) All three welfare components = (4) + (5) | 0.029 | 0.033 | 0.040 | 0.004 | 0.007 | 0.011 |

| Impact on annuitized values, discount rate 4% and patent life 11.5, from the perspective of patent beginning | Empirical distributions compared with no generics | Changes over time |
|---|---|---|---|
| | 2012 | 2016 | 2020 | 2012-16 | 2016-20 | 2012-20 |
| (8) Static surplus after expiration | 0.092 | 0.101 | 0.113 | 0.008 | 0.012 | 0.020 |
| (9) Innovation benefit (- is a cost), with elasticity 0.060 justifying patent life | -0.049 | -0.052 | -0.055 | -0.003 | -0.003 | -0.006 |
| (10) All three welfare components = (8) + (9) | 0.044 | 0.049 | 0.058 | 0.005 | 0.009 | 0.014 |

Note: Row numbering is the same as for Table 2. Rows (7) and (11) are not relevant here. The various costs and benefits from the constant-semi-elasticity model are averaged across numbers of producers using the Orange Book cross-market empirical distribution as weights. For each market, the total number of extant manufacturers with approved ANDA is measured in December 2012, 2016, and 2020. A market is the cross of active ingredient(s), dosage form, and route of administration as reported in the FDA Orange Book.
V. Barriers to the Marketing of Pandemic Vaccines

Pandemic vaccines provide a unique opportunity to measure the consequences of regulatory barriers to drug approval. The opportunity costs of pandemic vaccines and treatments are less speculative than those for most other drugs. For example, Peltzman found 25 fewer new drugs per year as a result of the 1962 EA. Because those drugs were not produced, it is difficult to know which diseases they would have treated and what benefits they would have conferred. In contrast, a pandemic vaccine treats a specific disease. The opportunities foregone during a pandemic without an available vaccine are primarily familiar and normal activities that are curtailed as individuals and organization seek to avoid infection.

In addition, the experience with pandemic vaccines, tests, and treatments suggest that the opportunity costs of approval barriers may have a fat right tail. Especially dangerous pandemics occur just once or twice per century and did not occur during Peltzman’s sample.

The COVID-19 pandemic also shows how FDA approval barriers can, in some instances, have the opposite of the effect intended since the 1962 EA, which is to reduce the number of unsafe and ineffective treatments used by consumers. The demand for treatments during a pandemic is high enough that many treatments are supplied outside FDA jurisdiction, especially while treatments under FDA jurisdiction are delayed by its approval process.

V.A. Changes in federal pandemic vaccine approval policy

In addition to the large economic literature critical of FDA approval barriers generally, a couple of strands of research influenced federal pandemic policy. One is a series of papers by Tomas Philipson, Pierre-Yves Geoffard, Michael Kremer, Rachel Glennerster and coauthors (Berndt, Glennerster, et al. 2007, Geoffard and Philipson 1996, 1997, Glennerster, Kremer and Williams 2006, Kremer 2000, Philipson 1996) on vaccine demand functions, incentivizing innovation, and whether effective vaccines can be expected to eradicate an infectious disease. Another includes work by Murphy and Topel (2003), Cutler and Kadiyala (2003), Lichtenberg
(2003), Nordhaus (2003), Becker, Philipson and Soares (2005) and others estimating the value of life-saving medical innovations. Another is work on the application of Bayesian decision theory to new-drug approval by Berry (1985) and Isakov, Lo and Montazerhodjat (2019) noting that the tradeoff between speed and safety or speed and efficacy is disease specific. Counterterrorism experts concerned with the use of viruses as bioweapons brought urgency to vaccine innovation (Borio, et al. 2002, Clark and Pazdernik 2016).

In 2018, Borio was working in the White House as National Security Director for Medical & Biodefense Preparedness and asked Council of Economic Advisers member Philipson that CEA undertake an economic analysis of pandemic vaccine innovation. CEA concluded that “…improving the speed of vaccine production is more important for decreasing the number of infections than improving vaccine efficacy” and emphasized the need for large-scale manufacturing and the possible advantages of “public-private partnerships” (Council of Economic Advisers 2019). Borio and the CEA vaccine report prompted a 2019 President’s Executive Order, also before the pandemic, noting that “viruses emerge from animals … that can spread efficiently and have sustained transmission among humans” and concluding that “vaccination is the most effective defense…” (U.S. President 2019).

Moderna finalized its mRNA sequence on January 13, 2020 and manufactured its first clinical vaccine batch by February 7, only one day after the first American died (in Wuhan China) from COVID-19. At this time, there were only 12 known cases in the U.S. (Centers for Disease Control and Prevention 2020). Nevertheless, experts were pessimistic as to the duration of vaccine development. The director of the National Institute of Allergy and Infectious Diseases told the U.S. Senate in March 2020 that “a vaccine … will take at least a year or year and a half” largely because, he said, approval by the FDA necessarily requires a year or more (NBC News 2020). The New York Times put the vaccine-development timeline out to the year 2034 (Thompson 2020). Even PhRMA, the trade association for pharmaceutical manufacturers, said on March 13 that “12 to 18 months … is a best case estimate” of the amount of time it would take for a vaccine to be available to the public.

24 Moderna (2021) and Zhong and Wong (2020).
Contrary to the above assertions, “when COVID-19 emerged, the White House was ready and expeditiously applied the [CEA] report's deregulatory and fiscal lessons to streamline FDA approval for vaccines and their parallel manufacturing on a large scale” (Grogan and Philipson 2020). The $20 billion federal program Operation Warp Speed launched in April 2020 to encourage and accelerate the development and mass manufacturing of COVID-19 vaccines, streamline Federal approval for vaccines and their manufacture, and provide Federal funds for private vaccine research and advance-purchase orders. Pfizer’s vaccine was given emergency use authorization by the FDA on December 11 and Moderna’s on December 18. By the end of that month, at least 14 million vaccine doses had been manufactured and distributed in the U.S.

V.B. Excess burden as an analytical framework for measuring opportunity costs

Let $E$ denote exposure to the disease per unit time and $t$ the health cost of exposure. For example, $t$ could be the value of a statistical life times the fatality rate per exposure. The arrival of a vaccine is a change in $t$: $dt < 0$. Let $U(E)$ denote the “economic benefits” of exposure, without regard for the health costs $tE$. Therefore, reducing exposure by $dE$ involves an opportunity cost $U'(E)dE$. The net benefit of the arrival of a vaccine is:

$$d[U(E) - tE] = d[U(E)] - d[tE]$$

In this formulation, one way to measure the opportunity-cost of vaccine delay is as the effect of delay on the value of economic activity, which is the $d[U(E)]$ term, net of the effect on equilibrium health costs $d[tE]$. I refer to equation (2) as the “excess burden method.”

Rearranging terms in equation (2), we have:

$$d[U(E) - tE] = [U'(E) - t]dE - Edt$$

If the amount of exposure equated benefits and costs at the margin, then the first term on the RHS of (3) would disappear and we would have a second measurement method, which is the

---

26 FDA (2020) and Moderna (2020).
28 This is a simplified version of Philipson’s (1995) of a disease as a “tax” on at-risk activities $E$, levied at rate $t$. 
product of equilibrium exposure $E$ and the effect $dt$ of the vaccine on health costs conditional on exposure. In principle, $dt$ would be revealed in double-blind clinical trials by comparing the vaccinated and the unvaccinated. For example, if the health costs of exposure were entirely deaths valued at the VSL and the vaccine reduces the disease’s mortality by ninety percent, then $Edt$ would be 90 percent of the mortality cost that would be incurred without the vaccine. I refer to $Edt$ as the “envelope method.”

Regardless of whether exposure equates marginal costs and benefits, it is likely that people reduce their exposure in response to additional severity. Formally, $dE/dt < 0$ so that the arrival of a vaccine has a larger effect $Edt$ on conditional health costs than its effect $d[tE]$ on equilibrium health costs. Indeed, nothing in this model rules out the possibility that exposure is elastic to severity in the relevant range, so that the arrival of a vaccine that is effective at preventing death from the disease, but less than 100 percent effective, increases equilibrium deaths. This possibility highlights the more general result that proper measurement of vaccine benefits either holds exposure constant (equation (3)’s final term) or accounts for the economic benefits or reduced excess burden $u'(E)dE$ of the additional equilibrium exposure (equation (2)).

The model (2) is a static model, but in reality the distribution of a vaccine requires a prolonged period of time. Mass manufacturing is required and a distribution network must be built and staffed. More individuals become convinced to take a valuable vaccine as they observe others experiencing it. Even though vaccine approval does not in reality create an immediate transition to the static equilibrium associated with the vaccine, the static model’s comparative statics are useful because an $x$-month approval delay postpones each stage of the prolonged processes of manufacturing, distribution, etc., by approximately $x$ months.

\[ \text{In practice, those treated in clinical trials may suspect that they were treated because they experience different symptoms than members of the placebo arm. If the treated respond by increasing their exposure, then the health gap between the two arms involves a } tdE \text{ term and not just the } Edt \text{ term.} \]
V.C. Examples of substitution outside FDA jurisdiction to therapeutic substitutes that were hardly safe or effective

The simple excess burden framework reveals that medical innovation such as a vaccine reducing $t$ has therapeutic substitutes, broadly construed. Namely, any action $dE < 0$ that individuals and organizations take to reduce exposure is a prevention substitute for vaccination. Many of these therapeutic substitutes, such as remote work, closing schools, and canceling normal medical appointments, are beyond the jurisdiction of the FDA and can be utilized without any attempt to demonstrate their safety or efficacy. Indeed, rational individuals may use such substitutes merely because they suspect them to be sufficiently safe and effective, especially during periods of time when more proven treatments are still awaiting FDA approval and therefore prohibited.

Closing schools to in-person learning is an important example of prevention activity – hardly safe or effective – that was available and applied to tens of millions of children in the U.S. because it was outside the FDA’s jurisdiction, while vaccines were inside.30 Closed schools proved to significantly harm student learning and create extraordinary household stress. Halloran et al. (2021) find, for example, “that offering full in-person instruction rather than full hybrid or virtual instruction reduces test score losses in math by 10.1 percentage points (on the base of 14.2 percentage points).”31 In many public school districts, the resumption of in-person learning was conditioned on teacher vaccination (Shapiro and Hubler 2020, Blume and Esquivel 2021).32

30 The potential timing of school closings and vaccine distribution are not as different as they might appear. The first day of closed schools in most states was Monday March 16, 2020 which is the same day that Moderna first administered its vaccine to humans. Of course, the two prevention measures are different in terms of the speed at which they can be scaled and command public acceptance. COVID-19 tests were created by commercial and academic labs already in February, but they too were delayed from the market by FDA approval processes (Baird 2020).
31 See also Mulligan (2021a) and the references cited therein.
32 The tie between vaccine availability and school openings might be “blamed” on school districts rather than the FDA, but point here is that a proper cost-benefit analysis of FDA decisions must acknowledge the real world in which the FDA is embedded. That real world
If closing schools prevented infection, the effect has proven difficult to detect. The COVID-19 risks of teaching and learning in person during the pandemic, without a vaccine and including secondary transmission to teacher and student families, was comparable to the risk of driving a car a few miles (Mulligan 2021a). Indeed, closing schools may have contributed to spreading the disease. On either an absolute or hourly basis, students and teachers appear to have been more likely to be infected outside of school than in school, where prevention protocols were more likely implemented and followed (Mulligan 2021b). Obviously the FDA effectiveness standard for vaccines differs from the effectiveness standard (if any) that school districts applied in deciding to close schools.

During the pandemic, 20 percent of adults skipped and delayed medical appointments for serious health problems in order to avoid infection (Findling, Blendon and Benson 2020). Because these choices had other health consequences and already by April 2020 medical facilities proved to be more effective at slowing the spread than the wider community (Mulligan 2021b), this seems to be another instance in which vaccine delay encouraged unsafe and ineffective substitutes. FDA delays in approving tests had a similar effect. As John Cochrane (2021) put it, “someone could sell you a thermometer to detect a COVID-19 fever, but if someone tried to sell you any [test] more effective, the FDA would stop them. … Sure, the test might not be perfect, but it would be a lot better than relying on thermometers.”

The pandemic was not the first important instance that FDA barriers steered consumers toward unsafe treatments outside FDA jurisdiction.33 In 2010, the FDA pursued a “reformulation” policy in which OxyContin, a leading prescription opioid, would be removed from the market and replaced with a new “abuse-deterrent formulation” that would not be crushed or dissolved as easily (a common recreational practice, contrary to the prescribed method). Several studies have concluded that this change pushed many consumers from opioid prescriptions to heroin and then illicit fentanyl, both of which are manufactured and sold illegally includes actors seeking potentially ineffective therapeutic substitutes in the absence of FDA approval.

33 Other licensing requirements have at least occasionally led consumers to pursue dangerous alternatives outside the licensing jurisdiction. Carroll and Gaston (1981) found that licensing “that reduce the density of electricians are significantly associated with a rise in the rate of death from accidental electrocution,” apparently as consumers did their own electrical work.
without FDA oversight. Any reduction in OxyContin overdose deaths was dwarfed by the increase in deaths from fentanyl overdose, which were enough to reduce nationwide life expectancy two years consecutively for the first time in at least 50 years.

V.D. The opportunity costs of COVID-19-vaccine delays

For both estimation methods, the opportunity cost of delay depends on the duration of delay and the value of a statistical life. I use a six-month delay in what follows; readers may rescale results to approximately consider longer or shorter delays. I use a VSL of $2.1 million to cost each death from COVID-19, which is an estimate accounting for the age and morbidity of those who died from the disease.

The simplest envelope method, based on equation (3)’s final term at taking the primary health costs of the disease to be mortality rather than morbidity, is simply the product of the duration of delay, the number of COVID deaths per unit time, the VSL, and the factor reflecting the percentage effectiveness of the vaccine in terms of preventing death. Sensitivity analysis is therefore straightforward rescaling. I take the number of U.S. COVID deaths as 500,000 annually and vaccine effectiveness as 90 percent. Therefore, according to the envelope

---

34 See Alpert, Powell and Pacula (2018); Schnell (2018); Mallatt (2018); Evans, Lieber and Power (2019) and Ruhm (2019).
35 Mulligan (2020b). Law enforcement and technological changes also contributed to an expansion of heroin and illicit fentanyl markets.
36 To arrive at the VSL estimate, I begin with Kniesner and Viscusi’s (2019) $10 million estimate for 2017. Following the convention in this field, I then adjust to 2019 according to the increase in nominal GDP per capita. Because I am comparing the mortality costs to costs incurred during the pandemic, I then adjust for the ratio of pandemic consumption to 2019 consumption among the affected population, which I take to be 0.9. The result of these two adjustments is $9.8 million. The more important adjustment is for differences between the age and comorbidities of the general population and the population dying from COVID-19. The mid estimate from Briggs (2020) is that those who died from COVID-19 had a discounted average of 4.1 remaining quality-adjusted life years (QALYs), as compared to 19.1 for the general population. My $2.1 million VSL is 9.8*4.1/19.1. Somewhat lower VSL would be obtained by using Briggs’ undiscounted QALY or undiscounted life expectancy.
37 Institute for Health Metrics and Evaluation (2021). Their estimate assumes that the clinical trials are double-blind but in fact vaccination has obvious side effects. These effects may induce
method, the dollar cost of a six-month delay is $0.5 \times 500,000 \times 2,100,000 \times 0.9 = 0.47$ trillion in the U.S.

Mulligan (2021c) estimates the opportunity costs from a specific change in exposure $E$, namely, from normal levels to a complete shutdown of non-essential activities. The costs include foregone market and nonmarket production, including human capital opportunity costs, as well as deadweight costs of relief policies. During much of the pandemic, and in parts of the country, exposure was not reduced this much, which Mulligan (2021c) associates with a 25.4 percent reduction in the rate of goods and service production in the private sector and an annualized opportunity cost of $8.6$ trillion for the U.S. Because this paper’s term $d[U(E)]$ reflects the opportunity costs of exposure from vaccine delays, it rescales the $8.6$ trillion for various plausible effects of vaccine delay on the flow of economic activity. These results are shown in Table 4’s column for the cost “of reducing exposure,” which can be rescaled proportionally to accommodate alternative estimates of the costs of reducing exposure while a vaccine is delayed.

---

38 Behavioral changes offsetting some of the vaccine’s effect with behavior constant, which is what is needed for the envelope method.

38 See also the discussion by Castillo et al (2021) of both the health and opportunity costs of limits on vaccine capacity.
Table 4. Vaccine-delay opportunity costs according to the excess burden method
Assumed six-month delay

<table>
<thead>
<tr>
<th>Private production</th>
<th>Mortality cost</th>
<th>Assumed impact of vaccine arrival, %</th>
<th>Delay costs in $ trillions</th>
<th>Equilibrium mortality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Of reducing exposure</td>
<td></td>
<td>Sum</td>
</tr>
<tr>
<td>3</td>
<td>-33</td>
<td>3</td>
<td>0.51</td>
<td>0.17</td>
</tr>
<tr>
<td>3</td>
<td>-50</td>
<td>3</td>
<td>0.51</td>
<td>0.26</td>
</tr>
<tr>
<td>3</td>
<td>-67</td>
<td>3</td>
<td>0.51</td>
<td>0.35</td>
</tr>
<tr>
<td>5</td>
<td>-33</td>
<td>5</td>
<td>0.85</td>
<td>0.17</td>
</tr>
<tr>
<td>5</td>
<td>-50</td>
<td>5</td>
<td>0.85</td>
<td>0.26</td>
</tr>
<tr>
<td>5</td>
<td>-67</td>
<td>5</td>
<td>0.85</td>
<td>0.35</td>
</tr>
<tr>
<td>10</td>
<td>-33</td>
<td>10</td>
<td>1.70</td>
<td>0.17</td>
</tr>
<tr>
<td>10</td>
<td>-50</td>
<td>10</td>
<td>1.70</td>
<td>0.26</td>
</tr>
<tr>
<td>10</td>
<td>-67</td>
<td>10</td>
<td>1.70</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Notes: A full shutdown of nonessential activities is assumed to cost $4.3 trillion per 6 months, including nonmarket opportunity costs and the deadweight costs of relief packages, while reducing GDP by 25.4 percent. The "reducing exposure" column of the table rescales the $4.3 trillion. The full (100%) mortality cost is 250,000 deaths per six months valued at $2.1 million each.

The fourth column of Table 4 shows estimates of the $d[tE]$ term. This term is less in magnitude than the $Edt$ term used for the envelope method ($0.47 trillion) as long as exposure $E$ falls with severity $t$. Nevertheless, the sum of the two columns, which is the opportunity cost according to the excess-burden method, exceeds $0.47 trillion. This discrepancy has two explanations. One is that my implementation of at least one of the methods has incorrectly calibrated one of the parameters. Another explanation is that exposure during the pandemic was reduced excessively relative to a $2.1 million VSL, so that the $[U'(E)−t]$ term in equation (3) is positive. Indeed, in hindsight this was the case with at least some prevention measures such as closing schools.

The costs of distributing or administering the vaccine are real, as evidenced by the fact that vaccine uptake was much less than 100 percent and prolonged over time. Such costs are not included in the table because it quantifies the costs of delay rather than the cost of forgoing a vaccine indefinitely. Even at a 10 percent annual interest rate at 250 million Americans
vaccinated, distribution/administration costs less than $800 per vaccinated would be overwhelmed by the table’s rounding error.

VI. Conclusions

Decades ago, Peltzman concluded that the FDA was not stopping enough ineffective drugs to justify the consumer costs of its barriers to valuable medical innovation. Recent drug market events reinforce his conclusion. Drug prices are greater, and quantities less, because FDA approval barriers limit competition. Generic-entry policy changes initiated in 2012 began to alleviate some of those costs, but had little effect on entry and social surplus until after 2016 when FDA approvals accelerated and prioritized second and third generics (Table 3). I estimate that these changes reduced the value created by innovation at about the same rate before and after 2016, but due to the time lag between innovation and generic entry this opportunity cost is dwarfed by the “static” benefits of generic competition. The pandemic vaccine approval process, although surprisingly accelerated during COVID-19, still has large and obvious opportunity costs on the order of a trillion dollars in the U.S. for just a half-year delay (Table 4), and even more costs worldwide.

One of Peltzman’s approaches to assessing the benefits of FDA screening is whether it stops drugs from making it to market that would be deemed “ineffective” by experts such as the American Medical Association. However, the alternatives against which the AMA judges effectiveness or costs may differ substantially from the practical alternatives employed by patients. Approval delays for pandemic tests and vaccines pushed tens of millions of individuals and businesses into preventions and treatments that were both outside FDA jurisdiction and hardly safe or effective. The pandemic experience raises the question of whether, on the whole, consumers engage in more unsafe and ineffective practices than they would if FDA approval were not a prerequisite for pharmaceutical sales. It also highlights another pitfall in the use of
government licensing and approval requirements as consumer protection against imperfect information.39

My welfare analysis of generic entry barriers relies on a simple model of competition and entry. Future research needs to extend the model to allow for “non-price” competition such as advertising or negotiated discounts, which are quantitatively important differences in the conduct of brand names and generics.40 Such an extension may further reduce the static social benefit of the first generic entrant relative to the benefit of the second and third, thereby strengthening the conclusion from Table 3 that generic-barrier reductions after 2016 were more beneficial than those before. As long as the optimal patent length hypothesis is maintained, it is unclear how this extension would affect the relative magnitude of the innovation cost of reduced generic-entry barriers.

The value of an institution’s reputation is another interesting area for future research. Accelerating approvals may reduce the expected costs of a present disease but risk complicating future information in the contingency that FDA suffers a reputational loss from an accelerated approval. However, such an analysis should also consider the reputational costs of other institutions such as school districts that serve the demand for less effective therapeutic substitutes, created by the FDA’s delay.

39 Winston (2021) reviews pre-pandemic studies of consumer-protection regulation. Tabarrok (2017, 2020) notes the irony that FDA test-approval policy increasingly has the effect of “limiting the information that patients may discover about their own bodies….” See also Bourne (2021). A related question is whether a drug should be made available over the counter or by prescription (Temin 1992).

40 See Lakdawalla and Philipson (2012) on pharmaceutical advertising before and after patent expiration. Murphy et al (2014) discuss the economics of negotiated discounts, which are a significant share of gross prescription-drug expenditure (Roehrig 2018).
VII. Appendix I: Cournot Competition Algebra with Unit Pass-Through

This appendix shows the algebra of the unit pass-through Cournot model, with emphasis on the quantitative results that are independent of the level of demand, the elasticity of demand, and the level of marginal cost. It considers a market with $Q < A$ aggregate units sold by $n$ sellers at price $p > 0$ and constant marginal cost $c$. The market demand curve is exponential:

$$Q = Ae^{-p/\alpha} \quad (4)$$

I impose no restrictions on the demand constants $a$ and $\alpha$ except that they are positive. The area under this demand curve is

$$u(Q) = \left(1 + \ln \frac{A}{Q}\right) \alpha Q \quad (5)$$

For a given number of producers, the Cournot equilibrium price $p$, aggregate quantity $Q$, and per-seller quantity $q$ are:

$$p = \frac{\alpha}{n} + c \quad (6)$$

$$nq = Q = Ae^{-\frac{1}{n} \frac{c}{\alpha}} \quad (7)$$

It is straightforward to verify that these equations are consistent with symmetry, the market demand curve (4), and unit passthrough of $c$ conditional on $n$. The profit maximization problem for a Cournot competitor is:

$$\max_q \left(\alpha \ln \frac{A}{q + (\text{others})} - c\right)q \quad (8)$$

where the first term in parentheses is the market price as a function of own sales $q$ and the sales of others. The first order condition of (8) is satisfied with symmetry only if and only if $q$ satisfies (7).
While the patent is in force, \( n = 1 \). Ignoring integer constraints on the number \( n \) of competitors, the equilibrium number after patent expiration is implicitly defined by:

\[
a = \frac{\alpha A}{n^2} e^{\frac{-1}{n} \frac{c}{\alpha}}
\]

where \( a \) is the ANDA cost paid by every seller except the one with the original patent. The number of competitors implied by equation (9) puts equilibrium profits at zero for all sellers except the one, whose profits are \( a \). It follows that profits prior to patent expiration are the RHS of equation (9) with \( n \) set to one, which is \( \alpha A e^{1-c/a} \). After expiration, the owner of the expired patent earns \( a \).

The two “static” flows represented in rows (1) and (2) of Tables 2 and 3 are \( u(Q) - cQ \) and \( (n-1)a \), respectively, expressed as ratios to \( \alpha A e^{1-c/a} \).

\[
\frac{u(Q) - cQ}{\alpha A e^{-1-c/a}} = \frac{n + 1}{n} e^{1-\frac{1}{n}}
\]

(10)

\[
\frac{(n - 1)a}{\alpha A e^{-1-c/a}} = \frac{n - 1}{n^2} e^{1-\frac{1}{n}}
\]

(11)

Note that these normalized flows depend only on \( n \geq 1 \) (generic manufacturers \( \geq 0 \)) and are thereby reported in Tables 2 and 3 without any quantitative assumptions about the demand parameters \( (A \) and \( \alpha \)) or the level of marginal cost \( (c) \). Equation (11) is the single-peak “Laffer-curve shape” referenced in the main text with peak at \( n = (3 + \sqrt{5})/2 \).

The present value of profits of the owner of a \( T \)-year patent are, from the perspective of the effective beginning of the patent,

\[
\frac{1 - e^{-\rho T}}{\rho} \alpha A e^{-1-c/a} + \frac{e^{-\rho T}}{\rho} a
\]

(12)

For reporting in Tables 2 and 3, these present values are annuitized (multiplied by \( \rho \)) and then divided by the flow of profits under patent, \( \alpha A e^{1-c/a} \). After this normalization, the expression (12) depends only on \( n \) and the product \( \rho T \).
The static social surplus (the difference between (10) and (11)) is greater with one generic \( (n = 2) \) than none \( (n = 1) \), but the increment is dwarfed by the redistribution from innovator to consumers. The share of the innovator’s monopoly profit flow lost from the first generic is

\[
1 - \frac{a}{\alpha A e^{-1-c/\alpha}} = 1 - \frac{\sqrt{e}}{4}
\]  

(13)

By comparison, as a ratio to the monopoly profit, the addition to social surplus is

\[
5 \frac{\sqrt{e}}{4} - 2
\]

(14)

The redistribution (13) is almost ten times (14).
VIII. Appendix II: Estimation of the Value of Foregone Innovation

Let $p$ denote the probability of creating a new product, $v$ the present social value of a new product, and $m$ the present value accruing to the innovator conditional on discovering the new product. Under an alternative entry policy, their values would be $p'$, $v'$, and $m'$, respectively. Consider the increment to expected present social value, normalized by $p$:

$$\frac{p'v' - pv}{p} = \left[\frac{p'}{p} - 1\right]v' + v' - v$$

(15)

The second term is effect of policy on the present value of static welfare. Converted to an annuity value, it is shown in rows (4) and (8) of Table 2. The first term in (15) can therefore be interpreted as the innovation-value term, which is shown as annuity values in rows (5) and (9) of Table 2. Each of the values $m$, $m'$, $v$, and $v'$ depends on the discounting term $\rho T$ and the number of producers $n$ after patent expiration, which is why the tables show multiple columns and panels. The ratio term in equation (15)'s square brackets is $\left(\frac{m'}{m}\right)^\eta$, where a single elasticity $\eta$ (regardless of $\rho T$ and $n$) is calibrated as follows.

There is only one patent policy even while generic barriers vary over time and drugs vary in terms of generic entry. If the actual patent life (about 11.5 years) maximized $pv$ for products that will have 4 producers after patents expire, then a marginal reduction in the patent life from 11.5 years would increase $v$ by the same percentage that it reduces the probability of innovation $p$:

$$\eta \frac{d \ln m}{dT} + \frac{d \ln v}{dT} = 0$$

(16)

$$\eta = \frac{m \frac{v_3 - v_0}{v} \frac{m_0 - m_3}{m}}$$

(17)

where (17) is derived from (16) using the fact that $v$ is the weighted sum of $v_0$ (surplus flow under patent) and $v_3$ (surplus flow with four producers, three of which are generics), and similarly for $m$. The second ratio in (17) features differences because the patent length changes

---

41 In 2012, the average number of producers indicated by the Orange Book was 3.6. It was 3.9 in 2016.
the weights in the sums without affecting the flows conditional on patent status. The first ratio is the innovator share of the present value of social surplus, shown in Figure 1 conditional on $\rho T$ and $n$; a value of $n = 4$ is used for the purpose applying (17). Application of (17) yields values $\eta = 0.076$ and $\eta = 0.060$ for $\rho = 0.06$ and $\rho = 0.04$, respectively.

Recall that $\eta$ is the elasticity of new drugs with respect to innovator reward, which has been the subject of previous studies such as Cerda (2003) and Acemoglu and Lin (2004). The latter, for example, estimates an elasticity of new drugs with respect to market size of about 4 to 6 by comparing drugs serving larger patient populations to those serving smaller ones. The discrepancy between their estimates and the values used in my Table 2 suggests that either (i) the actual patent length is far short of maximizing expected social surplus, (ii) the (positive) elasticity of the innovator’s return with respect to market size is close to zero, (iii) cross-sectional comparisons exaggerate the elasticity that is relevant for determining the optimal patent length, or (iv) some combination of (i)-(iii). As noted in the main text, to the extent that (i) is correct, generic entry barriers can enhance social surplus by serving as a backdoor method of extending patent life.
IX. Bibliography


Klein, Daniel B., and Alexander Tabarrok. "Is the FDA safe and effective?" *FDAreview.org* (The Independent Institute), 2002.


