Do firms underinvest in long-term research?
Evidence from cancer clinical trials
Online appendix

Eric Budish
University of Chicago

Benjamin N. Roin
MIT

Heidi Williams
MIT and NBER

February 4, 2015

A Appendix: Proofs (not for publication)

A.1 Proof of Proposition 1

Part 1 follows immediately from (5) since conditions (a) and (b) together imply $EML = ETL$ and condition (c) states $\pi = v$.

For Part 2, the expected social return to A exceeding that to B can be written as:

$$\frac{ETL_A \cdot v_A}{c_A} \geq \frac{ETL_B \cdot v_B}{c_B}$$

Multiplying both sides by $\frac{EML_A}{ETL_A} \frac{\pi_A}{v_A}$ gives

$$\frac{EML_A \cdot \pi_A}{c_A} \geq \frac{ETL_B \cdot v_B}{c_B} \frac{EML_A}{ETL_A} \frac{\pi_A}{v_A}$$

Suppose that neither (a) nor (b) hold, i.e. $\frac{\pi_A}{v_A} \geq \frac{\pi_B}{v_B}$ and $\frac{EML_A}{ETL_A} \geq \frac{EML_B}{ETL_B}$. Then:

$$\frac{ETL_B \cdot v_B}{c_B} \frac{EML_A}{ETL_A} \frac{\pi_A}{v_A} \geq \frac{ETL_B \cdot v_B}{c_B} \frac{EML_B}{ETL_B} \frac{\pi_B}{v_B} = \frac{EML_B \cdot \pi_B}{c_B}$$

hence

$$\frac{EML_A \cdot \pi_A}{c_A} \geq \frac{EML_B \cdot \pi_B}{c_B}$$

Hence if invention B is pursued, so is invention A. A contradiction.

A.2 Proof of Proposition 2

Follows immediately from equation (6) as described in the text.

A.3 Proof of Proposition 3

Proof of Part 1. For this proof we work with the integral forms of EML and EPL: specifically, $EML = \int_{t_{comm}}^{t_{patent}} (\delta\eta)^t \, dt$ and $EPL = \int_{t_{comm}}^{t_{patent}} \delta^t dt$. Observe that, for any $t_{comm} \leq t_{patent}$, the ratio $\frac{(\delta\eta)^t}{\delta^t} = \eta^t$ is positive and strictly decreasing in t over the interval $[t_{comm}, t_{patent}]$. This immediately implies that $\frac{\partial\left(\frac{EML}{EPL}\right)}{\partial t_{comm}} < 0$.

Proof of Part 2. $\frac{EPL}{ETL} = 1 - \delta^{t_{patent} - t_{comm}}$. Hence $\frac{\partial \frac{EPL}{ETL}}{\partial_{t_{comm}}} = \log(\delta)\delta^{t_{patent} - t_{comm}} < 0$.

A.4 Proof of Proposition 4

Proof of Part 1. This follows immediately from the private investment condition (3). Since surrogate endpoints decrease t_{comm}, they increase EML, which might cause additional investments to occur. Formally, let t_{comm} and $\hat{t}_{comm}$ denote a drug's commercialization lag with and without surrogate endpoints, respectively. Consider a drug where surrogate endpoints strictly decrease commercialization lag, i.e.

$t_{comm} < \hat{t}_{comm}$. Let $EML^{Surrogate}$ and $EML^{NoSurrogate}$ denote EML with and without the surrogate endpoint; $t_{comm} < \hat{t}_{comm}$ implies that $EML^{Surrogate} > EML^{NoSurrogate}$. Now choose π, c and p such that

$$EML^{Surrogate} \cdot \pi \geq \frac{c}{p} > EML^{NoSurrogate} \cdot \pi$$

Such an invention will get commercialized with surrogate endpoints but not without. The second part of the statement follows from our assumption that surrogate endpoints always decrease commercialization lag, and hence always increase EML.

Proof of Part 2. The social welfare associated with a successfully commercialized invention is $EPL \cdot v^{monop} + (ETL - EPL) \cdot v$, where EPL, effective patent life, is defined according to $EPL = p \sum_{t_{comm}}^{t_{patent}-1} \delta^t$ (see discussion in Section 1.5.3). A reduction in t_{comm} strictly increases EPL, and has no effect on $ETL - EPL$, because both ETL and EPL increase by the expected number of additional years that the drug will be commercially available, in present value terms discounted at δ. Hence, the social welfare associated with any commercialized invention goes up, sometimes strictly. In combination with part (1) this yields that overall social welfare strictly increases.

Proof of Part 3. Follows immediately from the assumption that t_{comm} is independent of $\hat{t}_{comm}$.

A.5 Proof of Proposition 5

The result follows immediately from the fact that $\frac{EPL}{ETL} = 1 - \delta^{t_{patent} - t_{comm}}$ since $t_{patent} - t_{comm}$ is now constant.

A.6 Proof of Proposition 6

Define social welfare as a function of commercialization lag, t_{comm}, and post-commercialization patent length, denoted x (i.e., $t_{patent} = t_{comm} + x$), by

$$W(t_{comm}, x) \quad = \quad \int_{EML \cdot \pi \geq c} (EPL \cdot v^{monop} + (ETL - EPL) \cdot v - c) \, dF_{t_{comm}}(\cdot) \tag{13}$$

where $dF_{t_{comm}}(\cdot)$ denotes the distribution of invention parameters conditional on t_{comm}; EML, EPL and ETL are defined as in the text of Section 1 but using the notation $t_{patent} = t_{comm} + x$; and the integral is taken over all inventions that satisfy the private investment condition $EML \cdot \pi \geq c$ as defined in equation (3). Using (13), we can define the optimal choice of post-commercialization patent length x as an implicit function of commercialization lag t_{comm}:

$$x^*(t_{comm}) \in \arg\max_x W(t_{comm}, x)$$

Our goal is to show that $x^*(t_{comm})$ is increasing in t_{comm}. By Topkis's theorem, it is sufficient for us to show that the cross-partial $\frac{\partial^2 W(t_{comm},x)}{\partial x \partial t_{comm}}$ is strictly positive for all x and t_{comm}. (Topkis's theorem also tells us that a strictly positive cross-partial implies that the optimum $x^*(t_{comm})$ is unique for all t_{comm}).

To study the cross-partial $\frac{\partial^2 W(t_{comm},x)}{\partial x \partial t_{comm}}$, we first decompose the partial $\frac{\partial W}{\partial x}$ into two components: the benefit from eliciting more inventions at the extensive margin, and the costs of additional deadweight loss from inventions on the intensive margin, i.e., inventions that would have been elicited even without the increase in x. Write this as follows:

$$\begin{aligned}\frac{\partial W}{\partial x} &= BenefitsExtensive(t_{comm}, x) - CostsIntensive(t_{comm}, x)\\ BenefitsExtensive(t_{comm}, x) &= k \cdot \mathbb{E}_{(EML \cdot \pi = c)}\left(EPL \cdot v^{monop} + (ETL - EPL) \cdot v - c\right)\\ CostsIntensive(t_{comm}, x) &= \int_{EML \cdot \pi \geq c} p\delta^{t_{comm}+x}(v - v^{monop})dF_{t_{comm}}(\cdot)\end{aligned}$$

with k denoting the rate at which new inventions are elicited on the margin, which we have assumed is uniform.[54] We will show that $\frac{\partial^2 W(t_{comm},x)}{\partial x \partial t_{comm}}$ is strictly positive by showing that $\frac{\partial CostsIntensive(t_{comm},x)}{\partial t_{comm}}$ is strictly negative and $\frac{\partial BenefitsExtensive(t_{comm},x)}{\partial t_{comm}}$ is weakly positive.[55]

First, consider $\frac{\partial CostsIntensive(t_{comm},x)}{\partial t_{comm}}$. There are two effects. First, increasing t_{comm} reduces the deadweight loss cost associated with invention parameter tuples on the intensive margin, because these costs are pushed out further in time. Second, increasing t_{comm} reduces the set of inventions for which deadweight loss is suffered. Both effects are negative. Formally,

$$\begin{aligned}\frac{\partial CostsIntensive(t_{comm}, x)}{\partial t_{comm}} &= \int_{EML \cdot \pi \geq c} \frac{\partial}{\partial t_{comm}} p\delta^{t_{comm}+x}(v - v^{monop})dF_{t_{comm}}(\cdot)\\ &\quad - \mathbb{E}_{(EML \cdot \pi = c)}\left(p\delta^{t_{comm}+x}(v - v^{monop})\right) \cdot f_{t_{comm}}(EML \cdot \pi = c)\end{aligned}$$

The first term simplifies to $\ln(\delta)CostsIntensive(t_{comm}, x)$, which is strictly negative since $\delta < 1$. The second term is weakly negative since $v \geq v^{monop}$, and strictly negative if $v > v^{monop}$. Hence, $\frac{\partial CostsIntensive(t_{comm},x)}{\partial t_{comm}} < 0$.

Next, we sign $\frac{\partial BenefitsExtensive(t_{comm},x)}{\partial t_{comm}}$. Using the relationships $v^{monop} = a \cdot \pi$ and $v = b \cdot \pi$, and the fact that $EML \cdot \pi = c$ on the extensive margin, we can rewrite $BenefitsExtensive(t_{comm}, x)$ as

$$BenefitsExtensive(t_{comm}, x) = k \cdot \mathbb{E}_{(EML \cdot \pi = c)}\left(\left(\frac{EPL}{EML} \cdot a - 1\right) \cdot c + \left(\frac{ETL - EPL}{EML} \cdot b\right) \cdot c\right) \quad (14)$$

Hence we need to sign

$$\frac{\partial}{\partial t_{comm}} \mathbb{E}_{(EML \cdot \pi = c)}\left(\left(\frac{EPL}{EML} \cdot a - 1\right) \cdot c + \left(\frac{ETL - EPL}{EML} \cdot b\right) \cdot c\right) \quad (15)$$

We will show that all of the terms in the main parenthetical of (15) are positive and weakly increasing in t_{comm} along the extensive margin $EML \cdot \pi = c$. First, the ratio $\frac{EPL}{EML}$ can be simplified to $\frac{1-\eta\delta}{1-\delta}\frac{(1-\delta^x)}{(1-(\eta\delta)^x)}\eta^{-t_{comm}}$. Since we are holding η and δ fixed they do not vary with t_{comm} at the extensive margin; hence the ratio is increasing in t_{comm} since $\eta^{-t_{comm}}$ is increasing in t_{comm} and all other terms stay constant. Since, in addition $\frac{EPL}{EML} \geq 1$ and $a \geq 1$, we have that the object $\left(\frac{EPL}{EML} \cdot a - 1\right)$ is positive and weakly increasing. The ratio $\frac{ETL-EPL}{EML}$ simplifies to $\frac{1-\eta\delta}{1-\delta}\frac{\delta^x}{1-(\eta\delta)^x}\eta^{-t_{comm}}$. As with $\frac{EPL}{EML}$, the only

[54] Formally, what we call the density of inventions on the extensive margin – the rate at which additional inventions are elicited as x is increased – is $f(EML \cdot \pi = c)\frac{\partial EML}{\partial x}$ which we assume is constant in t_{comm} up to finite upper bounds on t_{comm} and x. The proof works equivalently if this density is weakly increasing in t_{comm}. The proof works with some modification if this density is decreasing in t_{comm} but not too rapidly; see the next footnote.

[55] As mentioned in the body of the text, it is not necessary for the result that $\frac{\partial BenefitsExtensive(x;t_{comm})}{\partial t_{comm}}$ is weakly positive; it can be negative, so long as it is less negative than $\frac{\partial CostsIntensive(x;t_{comm})}{\partial t_{comm}}$. For this reason, several of our assumptions can be slightly relaxed. For example, the density of inventions on the extensive margin, described in the previous footnote, could be decreasing in t_{comm} so long as the decline is not too rapid. Also, as mentioned in the body of the text, we have a numerical counterexample in which the density declines rapidly within a region; intuitively, in such a region, the benefit of eliciting additional inventions at the extensive margin does not justify the additional deadweight loss costs from the inventions on the intensive margin.

factor that varies with t_{comm} at the extensive margin is $\eta^{-t_{comm}}$, so the ratio is increasing in t_{comm}. Since $ETL \geq EPL$ and $b \geq 1$ we have that $\frac{ETL-EPL}{EML} \cdot b$ is positive as well. Last, expected costs c are weakly increasing in t_{comm} on the extensive margin by assumption. Hence, the parenthetical of (15) consists of positive terms that are all weakly increasing in t_{comm}, hence the sign of $\frac{\partial BenefitsExtensive(t_{comm},x)}{\partial t_{comm}}$ is positive.

For intuition, rewrite the rightmost parenthetical of (15) as $[(EPL \cdot a - 1) + (ETL - EPL) \cdot b] \cdot \pi$. Increasing t_{comm} (while holding fixed x) decreases both EPL and $ETL - EPL$, because it pushes both the period of patent protection and the period post-patent protection out into the future. However, in the other direction, increasing t_{comm} improves the quality of inventions at the extensive margin, as measured by π, which by assumption also improves quality as measured by social value v^{monop} and v. The sign of (15) thus tells us that the benefits from higher quality inventions exceed the costs from additional time discounting. Note as well the role of excess impatience. If $\eta = 1$ then the ratios $\frac{EPL}{EML}$ and $\frac{ETL-EPL}{EML}$ are constant in t_{comm}, whereas if $\eta < 1$ these ratios increase exponentially in t_{comm} at rate $\Omega(\eta^{-t_{comm}})$, which directly causes $\frac{\partial BenefitsExtensive(t_{comm},x)}{\partial t_{comm}}$ to grow exponentially in t_{comm}.

Putting this all together, we have

$$\begin{aligned} sign\left(\frac{dx^*(t_{comm})}{dt_{comm}}\right) &= sign\left(\frac{\partial^2 W(t_{comm},x)}{\partial x \partial t_{comm}}\right) \\ &= sign\left(\frac{\partial BenefitsExtensive(t_{comm},x)}{\partial t_{comm}} - \frac{\partial CostsIntensive(t_{comm},x)}{\partial t_{comm}}\right) \\ &= sign\left([\geq 0] - [< 0]\right) \end{aligned}$$

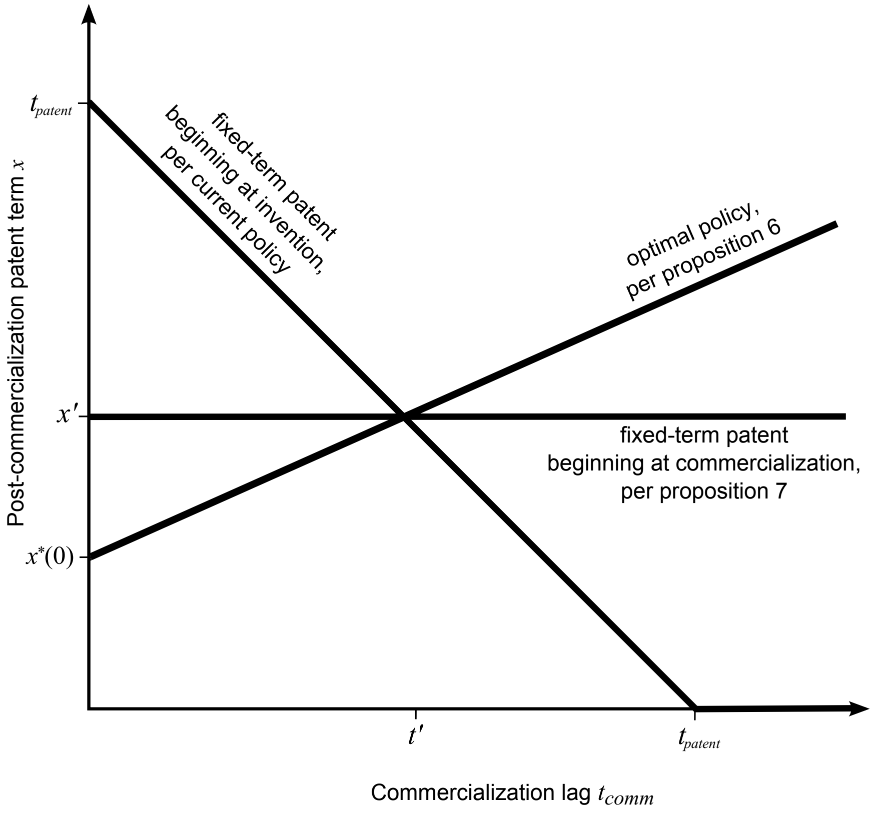
so $\frac{dx^*(t_{comm})}{dt_{comm}} > 0$, as required.

A.7 Proof of Proposition 7

Take as given a fixed patent term running from the date of invention, t_{patent}. This patent term induces a strictly negative relationship between commercialization lag and years of post-commercialization patent life. Letting x_{invent} denote the number of years of post-commercialization patent life when the clock starts at invention, we have $x_{invent}(t_{comm}) = \max(t_{patent} - t_{comm}, 0)$ which is strictly downward sloping in t_{comm} while $t_{comm} < t_{patent}$ and then flat at zero while $t_{comm} \geq t_{patent}$.

In contrast, Proposition 6 shows that the optimal post-commercialization patent life, x^*, is strictly *increasing* in t_{comm}. Suppose that the strictly increasing curve $x^*(t_{comm})$ and the strictly decreasing (then flat) curve $x_{invent}(t_{comm})$ intersect, which will occur iff $x^*(0) < x_{invent}(0)$; that is, if optimal policy involves awarding less than t_{patent} years of post-commercialization protection to inventions that have commercialization lag of 0. Let (t', x') denote the point of intersection. Construct a fixed patent term running from the date of commercialization using $x_{comm} = x'$. This policy, illustrated in Figure 7, strictly increases welfare relative to an t_{patent} term starting at invention. First, take an invention with $t_{comm} < t'$. For such an invention, we have $x^*(t_{comm}) < x_{comm} < x_{invent}(t_{comm})$, that is, optimal post-commercialization protection is smaller than our constructed policy x_{comm}, which itself is smaller than the given policy $x_{invent}(\cdot)$. By the same argument as in the proof of Proposition 6, reducing post-commercialization protection from $x_{invent}(t_{comm})$ to x_{comm} is welfare improving for these inventions (and we would be better off reducing further to $x^*(t_{comm})$): awarding more protection than $x^*(t_{comm})$ increases deadweight loss faster than it increases the gains from non-elicited inventions. Similarly, now take an invention with $t_{comm} > t'$. For such an invention we have $x^*(t_{comm}) > x_{comm} > x_{invent}(t_{comm})$, and our increase of patent protection from $x_{invent}(t_{comm})$ to x_{comm} increases the gains from eliciting more inventions faster than it increases deadweight loss.

Figure 7: **Illustration of the Proof of Proposition 7**



Notes: The post-commercialization patent term function under current policy, denoted $x_{invent}(\cdot)$ in the text, has a negative 45-degree slope until it reaches zero, then it is flat. The post-commercialization patent term under optimal policy, denoted $x^*(\cdot)$ in the text, is strictly increasing, but we do not know its shape or location.

Last, suppose that the curves $x^*(t_{comm})$ and $x_{invent}(t_{comm})$ do not intersect; this occurs iff $x^*(0) > x_{invent}(0)$. In this case, construct our post-commercialization patent term according to $x_{comm} = x^*(0)$. For all inventions, we have $x^*(t_{comm}) \geq x_{comm} > x_{invent}(t_{comm})$, and the argument in the preceding paragraph implies that this policy strictly increases welfare.

A.8 Proof of Proposition 8

Choose arbitrary commercialization lags t_{comm}, t'_{comm}, with $t_{comm} < t'_{comm}$. Initially, suppose that inventions with these commercialization lags receive the same per-commercialization-effort subsidy, of $s > 0$ dollars. We will argue that the marginal social welfare benefit of an increase in s for t'_{comm} inventions is strictly larger than that for t_{comm} inventions.

Consider a marginal increase in the invention subsidy for commercialization lags t_{comm} and t'_{comm}. By the same logic as in the proof of Proposition 6, the marginal invention with lag t'_{comm} is weakly more socially valuable than the marginal invention with lag t_{comm}. Also by the same logic as in the proof of Proposition 6, the cost associated with providing a subsidy to inframarginal inventions (i.e., inventions we would have received anyway) is smaller the longer is the commercialization lag, since any set of invention parameters that leads to commercialization with lag t'_{comm} also leads to commercialization when the lag is the shorter t_{comm}.

Hence, at the margin, additional subsidy to t'_{comm} inventions is more valuable than to t_{comm} inventions. This implies that optimal subsidies are increasing in commercialization lag, as required.

A.9 Power calculation derivation

Our conceptual framework is based on the idea that inventions which require long commercialization lags may be under-incentivized. Empirically, we focus on patient survival as a determinant of clinical trial length: because clinical trials must generally show evidence that treatments improve mortality-related outcomes, clinical trials tend to be longer when enrolling patients with longer survival times. In this section, we outline one example of a power calculation of the type used to guide the design of clinical trials in order to fix ideas on this point.

Approval of a drug compound by the US FDA requires evidence of efficacy and safety. Traditionally, "evidence of effectiveness" has been interpreted as evidence from controlled clinical trials. While most FDA approvals are based on placebo control groups, oncology trials instead compare the new drug compound to a non-placebo control of existing therapy. When testing the null hypothesis of no difference in mortality outcomes between the treatment and control groups, the traditional threshold for statistical evidence in oncology trials allows for a 1-in-20 chance of a false positive conclusion, or a p-value of 0.05.[56] This type of bar for statistical evidence motivates a calculation of what clinical trial design will be needed to achieve adequate statistical power to detect a statistically significant difference between the treatment and control groups.

An enormous literature exists on the design of clinical trials. Here, we simply focus on one type of calculation as an example. Collett (2003)'s *Modelling Survival Data in Medical Research* textbook includes a chapter on clinical trial design when survival is the outcome. Collett frames the design problem as a calculation of the required number of total deaths that must be observed.[57] Following this approach, at a given follow-up time k after treatment is administered, we can express the total number of deaths as $D = \frac{N}{2}(1-\mu^k) + \frac{N}{2}[1-(1-R(1-\mu))^k]$, where μ is the per-period survival rate of untreated individuals, k is the number of periods of patient follow-up, N is the sample size (equally divided between the treatment group and the control group), and R is a constant per-period multiplicative treatment effect. This expression can be derived as follows:

$$\begin{aligned}
Pr(\text{die at time } t|\text{survival to time } t-1) &= \begin{cases} 1-\mu & \text{for Control} \\ R(1-\mu) & \text{for Treatment} \end{cases} \\
Pr(\text{survive at time } t|\text{survival to time } t-1) &= \begin{cases} \mu & \text{for Control} \\ 1-R(1-\mu) & \text{for Treatment} \end{cases}
\end{aligned}$$

where μ is bounded by 0 and 1 and R is constrained such that $R(1-\mu)$ also is bounded by 0 and 1.

Consider first the control group. In the initial period there are $\frac{N}{2}$ individuals. In the subsequent period, there are $\frac{N}{2}\cdot\mu$, and in the kth period there are $\frac{N}{2}\cdot\mu^k$. Thus in the kth period there are $\frac{N}{2} - \frac{N}{2}\cdot\mu^k = \frac{N}{2}(1-\mu^k)$

[56] See, for example, http://www.fda.gov/downloads/AboutFDA/CentersOffices/CDER/ucm103366.pdf.

[57] See the discussion in Chapter 10.

deaths in the control group. Similarly, in the treatment group, at time k there are $\frac{N}{2} \cdot [1 - R(1-\mu)]^k$ survivors and $\frac{N}{2} \cdot \{1 - [1 - R(1-\mu)]^k\}$ deaths. Thus the total number of deaths in the sample at time k is:

$$\begin{aligned} D &= \frac{N}{2}(1-\mu^k) + \frac{N}{2} \cdot [1 - (1 - R(1-\mu))^k] \\ &= \frac{N}{2}[2 - \mu^k - (1 - R(1-\mu))^k] \end{aligned}$$

Applying the implicit function theorem, we can derive the following two comparative statics:

$$\frac{\partial N}{\partial \mu} = \frac{2D(k\mu^{k-1} + Rk(1 - R(1-\mu))^{k-1})}{[2 - \mu^k - (1 - R(1-\mu))^k]^2}$$

$$\frac{\partial k}{\partial \mu} = \frac{k\mu^{k-1} + Rk(1 - R(1-\mu))^{k-1}}{-[\mu^k \ln \mu + (1 - R(1-\mu))^k \ln(1 - (R(1-\mu))]}.$$

Since $0 < \mu < 1$ and $0 < R(1-\mu) < 1$, both of these partial derivatives are positive. Thus, we have two results. First, the required follow-up period k is increasing in the per-period survival rate μ: $\frac{\partial k}{\partial \mu} > 0$. Second, the required sample size N is increasing in μ: $\frac{\partial N}{\partial \mu} > 0$.

The first comparative static is the focus of our conceptual framework: clinical trials enrolling patients with longer expected survival times will - all else equal - require longer follow-up periods. The second comparative static is related to our conceptual framework in a more nuanced way. In the absence of detailed data on clinical trial costs (which are confidential), it is difficult to know whether the financial cost of enrolling an additional patient is higher or lower than an equivalently effective lengthening of the trial. However, in addition to the financial cost of enrolling additional patients, there is also a *time cost* of an increase in sample size because of the time required to recruit patients.

A variety of sources have stressed the time required to recruit patients as a barrier to clinical development; for example, Bartfai and Lees (2006) argue: "[m]any trials take a long time because the rate of enrollment is low. It is not uncommon that a 90-day drug trial takes 18 months to complete for all enrolled patients; it might take 90 days for each patient, but by the time the selected centers reach the required numbers 1.5 years have flown by."[58] A book on clinical trial management notes, "access to patients remains critical for the success of clinical development programs" because "[s]low patient recruitment can delay product launch with revenue loss during the precious product patent life" (Chin and Bairu (2011)). Thus, although at first blush clinical trial size might seem to be a mechanism for increasing the statistical power of clinical trials that is independent of trial length, this margin of adjustment also fits into our conceptual framework.

A.10 Elasticity calculation

In this section, we outline our rough estimate of the elasticity of R&D investment with respect to an additional year of commercialization lag.

From our empirical work in Section 3, we have an estimate of how R&D investment responds to the 5-year survival rate, $\frac{\partial(\text{R\&D investment})}{\partial(\text{5-year survival rate})}$. To translate this estimate into our elasticity of interest, we would like to scale $\frac{\partial(\text{R\&D investment})}{\partial(\text{5-year survival rate})}$ by an estimate of how commercialization lag varies with the 5-year survival

[58] See also Hovde (2006), Goffin (2009), Malani and Philipson (2011), and Allison (2012).

rate. By combining these estimates, we could then estimate the elasticity of interest:

$$\frac{\frac{\partial(\text{R\&D investment})}{\partial(\text{5-year survival rate})}}{\frac{\partial(\text{commercialization lag})}{\partial(\text{5-year survival rate})}} = \frac{\partial\,(\text{R\&D investment})}{\partial\,(\text{commercialization lag})}$$

The conceptual problem with estimating $\frac{\partial(\text{commercialization lag})}{\partial(\text{5-year survival rate})}$ is that - by construction - we only observe clinical trial length *conditional* on a drug compound being placed in clinical trials. Because - consistent with our model - we document that fewer drug compounds are placed in clinical trials for patients with longer survival times, we expect selection into clinical trials to bias the relationship between patient survival and clinical trial length in the set of observed clinical trials. Perhaps the most natural selection story is that firms are only willing to place a drug compound in clinical trials for patients with long expected survival times if they receive permission to use a surrogate endpoint in place of survival as an endpoint; in this case, the relationship between patient survival and clinical trial length would be biased towards zero.[59] Given this selection, we cannot obtain an unbiased empirical estimate of $\frac{\partial(\text{commercialization lag})}{\partial(\text{5-year survival rate})}$. To overcome this selection problem, we instead calibrate the relationship between commercialization lag and the 5-year survival rate using the power calculation outlined in Appendix A.9.

We can approximate our estimate of $\frac{\partial(\text{commercialization lag})}{\partial(\text{5-year survival rate})}$ with an estimate of $\frac{\partial(\text{clinical trial length})}{\partial(\text{5-year survival rate})}$, given that we expect commercialization lag to scale one-for-one with clinical trial length. In the language of the power calculation outlined in Appendix A.9, we can re-write this elasticity as:

$$\frac{\partial\,(\text{commercialization lag})}{\partial\,(\text{5-year survival rate})} = \frac{\partial\,(\text{clinical trial length})}{\partial\,(\text{5-year survival rate})} = \frac{\partial k}{\partial \mu} = \frac{k\mu^{k-1} + Rk(1-R(1-\mu))^{k-1}}{-[\mu^k \ln\mu + (1-R(1-\mu))^k \ln(1-(R(1-\mu))]}$$

where μ is the per-period survival rate of untreated individuals, k is the number of periods of patient follow-up, and R is a constant per-period multiplicative treatment effect.

Intuitively, μ and k come in pairs - not all μ and k are feasible conditional on a given technology (R). Here, we take the two (μ, k) pairs from the examples in the introduction given that by construction these are feasible pairs (given that the trials were completed), and that we know these trials looked at survival outcomes (rather than surrogate endpoints). We assume a technology of $R = 0.8$, which translates to a 20 percent improvement in the five-year survival rate; this choice of R is arbitrary but we explore robustness to alternative values of R below. Given the assumed value of R, our two introduction examples can be written as:

1. Metastatic prostate cancer: 5-year survival rate of 20 percent ($\mu = 0.2$)

   (a) Follow-up time of 12.8 months ((12.8/12)/5 $=>$ $k = 0.213$ units in 5-year increments)

   (b) Total trial length of 3 years (3/5 $=>$ $k = 0.6$ in 5-year increments)

2. Localized prostate cancer: 5-year survival rate of 80 percent ($\mu = 0.8$)

[59] If we estimate this relationship in our data, we do estimate a statistically significant relationship; however, the magnitude is implausibly small, consistent with our prior that this relationship would be biased towards zero (a ten percentage point increase in the five-year survival rate is associated with a 1.5 percent increase in average clinical trial length - an increase on the order of one month).

(a) Follow-up time of 9.1 years (9.1/5 => $k = 1.82$ units in 5-year increments)

(b) Total trial length of 18 years (18/5 => $k = 3.6$ units in 5-year increments)

Plugging in these values for μ, k, and R into the above formula for $\frac{\partial k}{\partial \mu}$ gives estimates of 2.234 for metastatic prostate cancer, and 0.766 for localized prostate cancer. Those estimates are in units of 5-year increments, and multiplying them by 5 to translate them into a 1-year unit gives 11.170 and 3.827. In words, a change from 0 to 1 in the 5-year survival rate translates to between a 3.827-11.170 year increase in patient follow-up time.

Our estimate from Section 3 implies that a change from 0 to 1 in the 5-year survival rate translates into an 86.9% reduction in R&D investment. Scaling this estimate by our estimates of $\frac{\partial(\text{commercialization lag})}{\partial(\text{5-year survival rate})}$ implies an estimated semi-elasticity of R&D investment with respect to a one-year change in commercialization lag of between 7.779% (based on metastatic prostate cancer; $86.9/11.170 = 7.779$) and 22.707% (based on localized prostate cancer; $86.9/3.827 = 22.707$).

Alternatively, we can do the same calculation using total trial length (3 and 18 years) rather than follow-up times (12.8 months and 9.1 years). Reassuringly, we obtain nearly identical estimates: 7.993% (based on metastatic prostate cancer; $86.9/10.872 = 7.993$) and 23.416% (based on localized prostate cancer; $86.9/3.711 = 23.416$).

We can investigate sensitivity of our estimates to different assumed values of R, the quality of the technology. A 'reasonable' range of R might be between 0.15-0.95, in which case our estimated elasticities fall between 6-54%.[60]

[60] Our metastatic prostate cancer example from the introduction - where the treatment resulted in a gain of 3.9 months on average - corresponds to $R = 0.961$, which implies elasticity estimates between 6-20%. On the other extreme Gleevec, often referenced as a "miracle" drug, is estimated to have increased the five-year survival rate from 30% to 89% - implying $R = 0.157$, and elasticity estimates between 17-54%.

B Appendix: Data (not for publication)

B.1 Description of SEER cancer registry data

The Surveillance, Epidemiology, and End Results (SEER) data is compiled by the National Cancer Institute (NCI), and is considered the authoritative source of information on cancer incidence and survival in the US.[61]

SEER collects data from population-based cancer registries covering approximately 28 percent of the US population. Specifically, the SEER data aims to be a comprehensive census of all cancer cases diagnosed among residents of geographic areas covered by SEER cancer registries. In order to focus on a geographically consistent sample over time, we analyze data from the seven original SEER registries that joined in 1973: the states of Connecticut, Iowa, New Mexico, Utah, and Hawaii and the metropolitan areas of Detroit and San Francisco-Oakland. Funding for the data collection varies by state, and is a mix of funding from the NCI, the Centers for Disease Control and Prevention (CDC), and state funding.

The SEER registries collect detailed information on cancer patients near the time of diagnosis, including data on patient demographics, primary tumor site, tumor morphology and stage at diagnosis, and first course of treatment.[62] This data is administratively linked to follow-up mortality data from the National Center for Health Statistics (NCHS).

We use the 1973-2009 SEER Research Data (ASCII text format) as downloaded on 25 June 2012, which includes patients diagnosed from 1973-2009.[63] The follow-up mortality data cutoff date is 31 December 2009. The key variables that we obtain from the SEER data are the following:

- **Cancer information**. We use the SEER site recode with kaposi sarcoma and mesothelioma variable to identify the cancer type for each individual in our sample. For example, a value of 20010 for this variable corresponds to a diagnosis of lip cancer. There are 80 unique cancer categories, as listed here: http://seer.cancer.gov/siterecode/icdo3_d01272003/. This variable is non-missing for all observations.

- **Stage information**. We use the SEER historic stage A variable to identify the stage of cancer for each individual in our sample: in situ, localized, regional, metastatic, or unknown. As described on the SEER website, stage information is not available for all observations for three reasons.[64] First, some cancers are not staged by SEER: for example, brain cancers are not staged.[65] Second, some cancers are not staged in a subset of years: for example, between 1973-1982 nose, nasal cavity, and middle ear cancers were not staged.[66] Third, some individual observations that should be staged

[61] For more details, see http://www.seer.cancer.gov.

[62] Importantly for our purposes, the SEER website notes that the SEER data is the only comprehensive source of population-based information in the US that includes stage of cancer at the time of diagnosis and patient survival data.

[63] This data is available via a research data agreement; see https://www.seer.cancer.gov/seertrack/data/request/ for details. The citation for this data is: Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2009), National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2012, based on the November 2011 submission.

[64] For more information, see http://seer.cancer.gov/seerstat/variables/seer/yr1973_2009/lrd_stage/index.html.

[65] The SEER site recodes that are not staged by SEER are: brain (31010); cranial nerves and other nervous system (31040); pleura (22050); hodgkin and non-hodgkin lymphoma (both nodal and extranodal; 33011, 33012, 33041, and 33042); myeloma (34000); acute lymphocytic leukemia (35011); chronic lymphocytic leukemia (35012); other lymphocytic leukemia (35013); acute myeloid leukemia (35021); acute monocytic leukemia (35031); chronic myeloid leukemia (35022); other myeloid/monocytic leukemia (35023); other acute leukemia (35041); aleukemic, subleukemic, and not otherwise specified leukemias (35043); kaposi sarcoma (36020); and miscellaneous (37000).

[66] The SEER site recodes that are staged by SEER for a subset of years are: nasopharynx (not staged 2004 and later; 20060); peritoneum, omentum, and mesentery (not staged 1973-1987; 21120); nose, nasal cavity, and middle ear (not staged

have missing stage data. The first two categories - for which missing stage data are "expected" - result in stage data missing for 19 percent of the SEER sample; the third category - for which missing stage data is "unexpected" - results in stage data missing for an additional 5 percent of the SEER sample. The exceptions to the standard staging categories are as follows:

  - Prostate cancer. Prostate cancer is staged by SEER starting in 1995, but uses a combined localized/regional category rather than separate localized and regional stages. We code the localized/regional prostate cases as regional cancers.
  - Bladder cancer. All in situ cases of bladder cancer (29010) in the SEER data were re-coded by SEER to appear as localized cancers.

For consistency, we code all unstaged cancers and cancers that utilize only one stage into an "unstaged" stage classification in our analysis.

- **Survival time**. Because the SEER data are linked to follow-up mortality data from the National Center for Health Statistics, for each individual in our sample we know survival time in months as calculated using the date of diagnosis and one of the following: date of death, date last known to be alive, or follow-up cutoff date of 31 December 2009. This variable is non-missing for all observations.
- **Year of diagnosis**. The SEER data record the year of diagnosis for each patient, defined as the year the tumor was first diagnosed by a recognized medical practitioner, whether clinically or microscopically confirmed. The year of diagnosis varies from 1973 to 2009, and is non-missing for all observations. We use the year of diagnosis together with information on patient sex and age at diagnosis to calculate life expectancy at the time of diagnosis (in the absence of cancer) for each individual in the SEER sample.
- **Age at diagnosis**. The SEER data record the patient's age in years at diagnosis. This variable is missing for 692 of 3,245,656 individuals (0.02 percent of the sample). Because we need information on age at diagnosis in order to calculate life expectancy at the time of diagnosis, we drop these 692 individuals from the sample.
- **Sex**. The SEER data record the sex of the patient at diagnosis. This variable is non-missing for all observations. We use this variable together with information on year of diagnosis and patient age at diagnosis to calculate life expectancy at the time of diagnosis for each individual in the SEER sample.

Between 1973 and 2009, 3,245,656 individuals were diagnosed in catchment areas of the seven original SEER registries. Our only sample restriction is to exclude the 692 individuals missing data on age at diagnosis (0.02 percent of the sample), leaving us with a final SEER sample of 3,244,964 individuals.

SEER also produced population data which can be used to normalize the cancer incidence data into rates per population. We use the 1969-2009 SEER population data (ASCII text format) for the catchment areas of the seven original SEER registries as downloaded on 28 June 2012.[67]

1973-1982; 22010); larynx (not staged 2004 and later; 22020); lung and bronchus (not staged 1973-1987; 22030); trachea (not staged 2004 forward; 22060); vagina (not staged 2004 forward; 27050); prostate (not staged 1973-1994; 28010); other endocrine including thymus (not staged 2004 and later; 32020); and mesothelioma (not staged 2004 and later; 36010).

[67] The citation for this data is: Surveillance, Epidemiology, and End Results (SEER) Program Populations (1969-2009) (www.seer.cancer.gov/popdata), National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released January 2011.

B.2 Life expectancy data

We use year-age-gender-specific period life expectancy data for 1973-2006 from the National Center for Health Statistics (NCHS) files posted at http://www.cdc.gov/nchs/products/life_tables.htm. For 2000-2006, digitized files are available from NCHS. For 1973-1999, the data was entered by the firm Digital Divide Data (http://www.digitaldividedata.org/) and was funded by NIA Grant Number T32-AG000186 to the NBER. No data is available for 1979 nor 1981.

Based on the year of diagnosis, age at diagnosis, and gender of each patient, we use this NCHS data to construct year-age-gender-specific life expectancy for each patient - in the absence of cancer - at the time of diagnosis. Because in most years the life tables end at age 85, we apply the life expectancy numbers for 85-year-old individuals to all individuals age 85 and older. Because data is not available for years 1979, 1981, 2007, 2008, and 2009, we fill in the data as follows: apply the 1978 life expectancy data in 1979; apply the 1980 life expectancy data in 1981; and apply the 2006 life expectancy data in 2007, 2008, and 2009. Using this life expectancy data, we calculate the life lost for each individual as their life expectancy at the time of diagnosis minus their survival time in years.

We focus on measuring life lost among patients diagnosed between 1973-1983 to minimize censoring. In this sample, median survival time by cancer-stage is almost never censored; in the handful of cases where censoring is an issue, we top code survival time at 25 years.

B.3 Description of classification method for research investment datasets

Unlike the SEER data, our data measuring research investments in cancer treatments were not originally developed as research data, and hence required a large amount of restructuring to be converted to a format useable for our analysis. We detail this restructuring below, but first describe the method we use to classify the cancer and stages to which a given clinical trial is relevant.

We start by compiling a classification system which can consistently code observations that vary in the aggregation level at which the cancer type was identified in the original data. For example, while some cancer clinical trials enroll stage III breast cancer patients, others are open to all patients with "solid tumors" and we need a way of classifying which cancer types are solid tumors. We base our classification around the SEER site recode ICD-O-3 (1/27/2003) definition.[68] These SEER site records define the major cancer sites (e.g. breast, stomach, prostate) and are the standard set of cancer classifications used by cancer researchers. Importantly, the SEER cancer registry data include SEER site recodes, so using the SEER site recodes as the basis of our classification of the research investment datasets is what enables a cross-walk between the SEER cancer registry data and the research investment data.

For each research investment observation, we search the textual description of the cancer type for which the observation is relevant in order to match the observation to one or more of the SEER site recodes.[69] Most of the search words are drawn directly from the SEER site recode title (e.g. "lip" for lip cancer, SEER site recode 20010), but we also search for variations on cancer names that are frequently observed in the data (e.g. searching both "pancreas" and "pancreatic" for cancer of the pancreas, SEER site recode 21100). We allow a given observation to be labeled as relevant to multiple cancer types (e.g. an observation labeled as being relevant for hematologic/blood cancers is classified as relevant to all hematologic/blood cancers, such as both acute myeloid leukemia and chronic myeloid leukemia). While there are surely imperfections in this classification system, it allows for a consistent coding of our data.

One additional issue that deserves discussion is off-label use of drugs. Off-label prescription of a drug refers to use of the drug outside of what is prescribed on its FDA-approved label (Leveque (2008)). Off-label use of drugs is generally thought to be widespread, particularly in cancer, although very few studies have actually measured the extent of off-label drug use in representative populations. An exception is

[68] Available at http://seer.cancer.gov/siterecode/icdo3_d01272003/.

[69] Detailed documentation on the precise search words used in this classification system are available on request.

a recent study by Agha and Molitor (2012), who estimate that 22 percent of cancer drug prescriptions were off-label in the Medicare population between 1998-2008. Why could off-label drug use be important in this context? If off-label use were completely unrestricted, firms would face an incentive to always approve drugs for the group of patients for which trials were the least expensive, and then *ex post* have the drugs be prescribed for all patients. In practice, although in the US physicians are free to prescribe drugs for off-label uses, it is illegal for pharmaceutical firms to actively advertise/promote those uses, and additional constraints are often imposed by insurers for reimbursement. Perhaps the clearest evidence that restrictions on off-label use appear to be binding comes from the fact that we very frequently observe firms making large R&D investments to re-approve a given drug compound for an additional indication (see, e.g., Eisenberg (2005)), which they would have no need to do if off-label use restrictions were not binding.

B.4 Description of NCI cancer clinical trial registry data

The Physician Data Query (PDQ) Cancer Clinical Trials Registry is the National Cancer Institute (NCI)'s cancer clinical trials database. The NCI registry was created via the National Cancer Act of 1971, and claims to be the world's most comprehensive cancer clinical trial registry. The intended purpose of this registry is to allow cancer patients and physicians to search for clinical trials now accepting participants, and to allow them to access information and results from closed trials.

We use the 12 July 2011 version of the NCI registry data (XML format), which includes all clinical trials entered into the registry prior to that date.[70] The registration of clinical trials in the NCI registry is strictly voluntary but strongly encouraged. The NCI registry is thought to include most clinical trials sponsored by the NCI, as well as a substantial share of clinical trials sponsored by pharmaceutical companies, medical centers, and other groups. For example, the NCI registry includes all cancer clinical trials registered under the requirements specified by Section 113 of the Food and Drug Administration Modernization Act of 1997 (phase II and higher drug treatment trials), all cancer clinical trials registered under the requirements of the International Committee of Medical Journal Editors (phase II and higher trials that have a comparison or control group), and all cancer clinical trials that are included in the US National Institutes of Health (NIH) `Clinicaltrials.gov` database.

Many trials in our sample enroll multiple patient "types" as measured by the cancer-stages eligible for participation in the trial, and we expand the data to unique trial-cancer-stage observations. We make three sample restrictions. First, cancer stages are sometimes reported in the NCI data at a finer level of granularity than we observe in the SEER data: for example, a given trial may list breast cancer stage IIA and breast cancer IIB patients as eligible for enrollment, but we do not consistently observe cancer stage at that level of detail in the SEER data. To avoid double-counting, we remove duplicate observations at the trial-cancer-stage level. Second, because we do not observe remission in the SEER data, we are unable to construct measures of the patient population eligible for these trials and thus drop trials enrolling only remission cases from the sample. Third, because we do not observe recurrent cases in the SEER data, we are again unable to construct measures of the patient population eligible for these trials and thus drop trials enrolling only recurrent patients from the sample.

The key variables that we obtain from the NCI registry data are the following:

- **Cancer information**. We identify the types of cancers eligible for enrollment in each clinical trial, coded by the SEER site recodes. By construction, this variable is non-missing for all observations.
- **Stage information**. In the NCI registry data, the cancer stages eligible for enrollment in each clinical trial are most frequently identified in the following categories: stage 0, stage 1, stage 2,

[70] This data is available via a research licensing agreement; see `http://www.cancer.gov/licensing` for details. The scripts used to extract the XML files are available on request.

stage 3, stage 4, recurrent cancers, cancers in remission, and localized cancers. As discussed above, we drop trials enrolling only remission or recurrent patients from the sample. We then need a crosswalk which maps the remaining stage categories in the NCI registry data - stage 0, stage 1, stage 2, stage 3, stage 4, and localized cancers - to the SEER historic stage A categories in the SEER data (in situ, localized, regional, and metastatic). We follow the *AJCC Cancer Staging Manual* (American Joint Committee on Cancer (2010)) and use the following mapping: stage 0 maps into in situ; stages I, II, and localized map into localized; stage III maps into regional; and stage IV maps into metastatic.[71] In addition, to harmonize the NCI registry stage coding with the SEER stage coding we make the following revisions:

  - Prostate cancer. In the SEER data, the localized and regional prostate cases (28010) are coded into a joint localized/regional category which as described above we code as regional cancers. Analogously, in the NCI registry data we code trials for either localized or regional prostate cancer as being for regional prostate cancer.
  - Bladder cancer. In the SEER data, all in situ cases of bladder cancer (29010) are coded as localized cancers. Analogously, in the NCI registry data we code trials for in situ bladder cancers as being for localized bladder cancers.

For consistency, we code all cancers which SEER codes as either unstaged or utilizing only one stage into an "unstaged" stage classification in our analysis.

- **Clinical trial sponsorship.** Approximately 50 percent of clinical trials in the NCI registry data are listed as being either publicly sponsored or privately sponsored. We define publicly-sponsored trials as trials that are solely publicly sponsored, and define privately-sponsored trials as trials that are solely privately sponsored; in our sponsorship analysis, we treat the approximately 1 percent of trials that are listed as being both publicly sponsored and privately sponsored as missing sponsorship data.

- **Clinical trial length.** Length of clinical trials is very rarely reported in the NCI registry data. Our understanding is that this is because the NCI registry is primarily oriented towards the recruitment of patients into clinical trials, whereas trial lengths are typically reported at the time of trial completion. In order to obtain data on clinical trial length, we take advantage of the fact that the NCI registry includes - where available - a `Clinicaltrials.gov` trial ID number. `Clinicaltrials.gov` is a registry *and results* database of clinical trials: likely because `Clinicaltrials.gov` includes clinical trial results, trial length is much better reported relative to the NCI registry data. The NCI registry claims to include all cancer clinical trials listed in the `Clinicaltrials.gov` registry, and approximately 70 percent of the NCI trials are included in the `Clinicaltrials.gov` registry. While we rely on the more complete NCI registry for the main analysis, we use the `Clinicaltrials.gov` subsample in order to examine data on trial length. Approximately 60 percent of trials in the NCI registry which appear in the `Clinicaltrials.gov` registry have non-missing data on trial length. Much of the missing data appears to be explained by trial length being more frequently reported in more recent years (even given that we would expect missing data for ongoing trials to bias upwards the share of trials with missing data in more recent years): on the order of 80 percent of trials

[71] The exact language on page 12 is as follows: "Stage I is usually assigned to tumors confined to the primary site with a better prognosis, stages II and III for tumors with increasing local and regional nodal involvement, and stage IV to cases with distant metastatic disease. In addition, a group termed stage 0 is assigned to cases of carcinoma in situ (CIS)."

starting in 1997 have missing data on trial length, compared to 50 percent in 2003 and 25 percent in 2011. This increased reporting over time likely in part reflects increased incentives for reporting: for example, there was a tightening of reporting requirements affecting clinical trials initiated in or ongoing as of September 2007; see http://clinicaltrials.gov/ct2/info/results for details.

C Appendix: Industry interviews (not for publication)

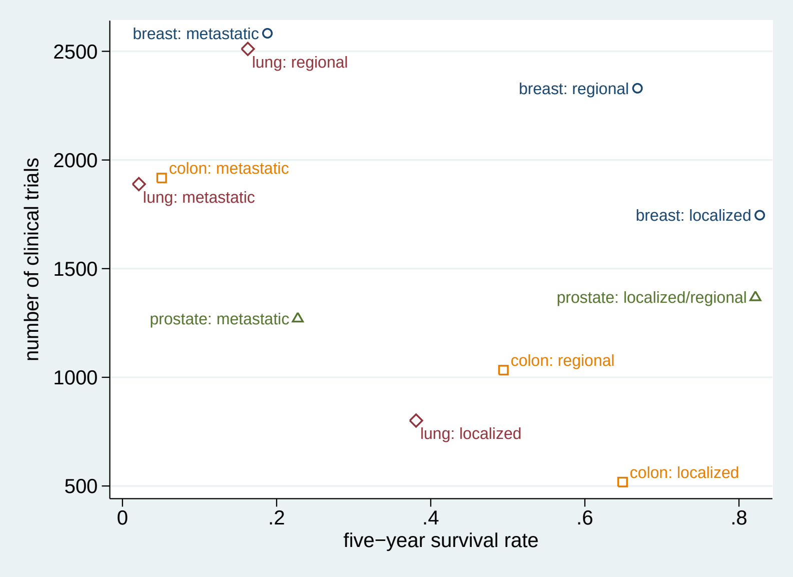
As discussed in the introduction, while the importance of patents has been debated in many industries, given our empirical focus on the pharmaceutical industry it is worth noting that a variety of evidence suggests that patents play a key role in motivating innovation in the pharmaceutical industry, including industry interviews (Edwin Mansfield, Mark Schwartz and Samuel Wagner 1981; Mansfield 1986; Levin et al. 1987; Wesley Cohen, Richard Nelson and John Walsh 2000), the cost structure of new drug development relative to the generic production (Joseph DiMasi, Ronald Hansen and Henry Grabowski 2003; Adams and Brantner 2006; Wroblewski et al. 2009), and the fact that standard investment models used by pharmaceutical firms pay close attention to effective patent length (Mayer Brown 2009). In this appendix, we document some additional evidence on this point including summarizing some informal interviews that we conducted for this paper.

Academic clinicians, clinical researchers, and firms all report a reluctance to invest in drugs that - by nature of requiring lengthy clinical trials - receive short effective patent terms. For example, a medicinal chemistry textbook notes: "...patents normally run for 20 years from the date of application, ...some compounds are never developed because the patent protected production time available to recoup the cost of development is too short" (Thomas, 2003). We interviewed several venture capitalists for this paper, and while their confidentiality agreements with the companies they evaluate prevented them from naming any specific examples of drugs that failed to reach the market due to short expected patent terms, they each claimed that it happens "all the time." Below are two excerpts from their comments:

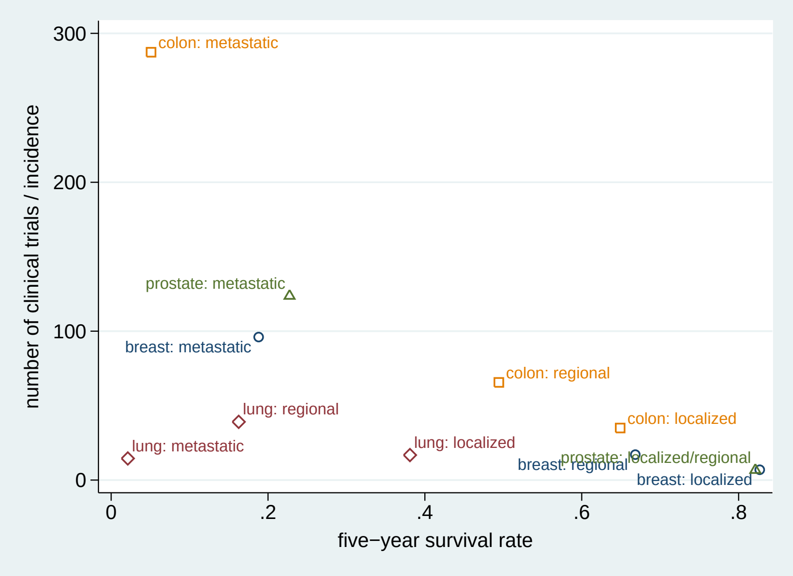
- *The running clock on these patents is a huge deal. Companies absolutely choose not to go forward with drugs because their remaining patent life isn't sufficient. In the models we use to decide whether to in-license a drug, we measure the time and cost of doing the trials against the period of exclusivity and time until peak market share. You have a pretty good sense of how long it will take to get to approval, at least by the time you're in the Phase 2A trials, so these things happen pretty early in development. Companies de-prioritize those drugs. Quite often we've declined to take advantage of an opportunity because we thought there wouldn't be enough time under the patent term to earn a return on the investment.* (Venture Capitalist A)
- *The shorter the remaining patent term, the more certainty you need that the drug will work, and the more it needs to have a large market. Also, the ramp is important. You want at least a couple years of peak sales. It happens all the time that we pass on a drug, one we think would probably work, because there wouldn't be enough life left on its patent by the time it reached the market.* (Venture Capitalist B)

D Appendix: Additional figures and tables (not for publication)

Figure D.1: **Survival time and R&D investments: Breast, colon, lung, and prostate cancers**



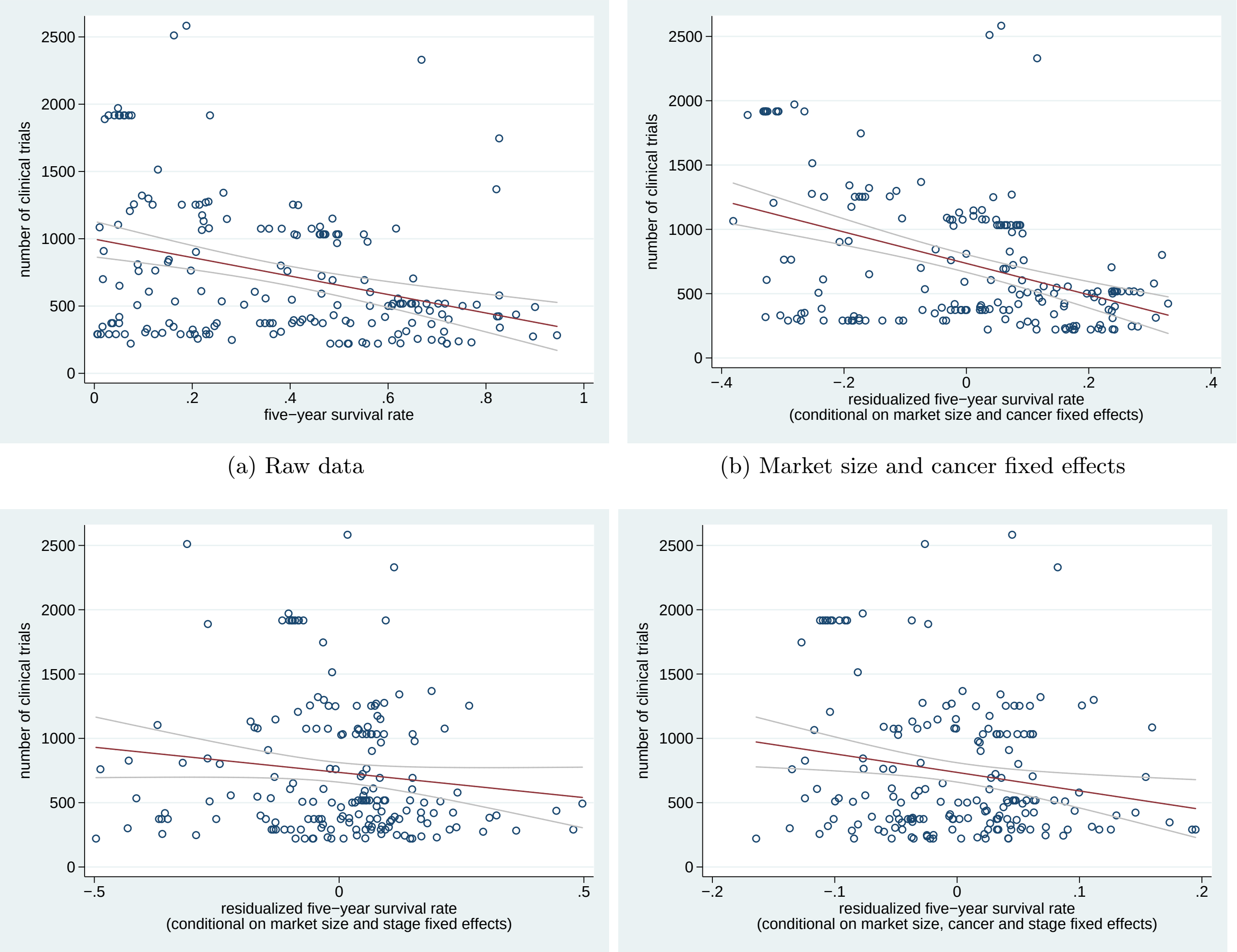
(a) Survival time and R&D investments



(b) Survival time and market-size adjusted R&D investments

Notes: This figure shows the relationship between the five-year survival rate among patients diagnosed with a given cancer-stage between 1973-2004 (the cohorts for which five-year survival is uncensored), and two measures of clinical trial activity for that cancer-stage from 1973-2011, for the "big four" cancers: breast (26000), colon (specifically, ascending colon; 21043), lung (22030), and prostate (28010). The level of observation is the cancer-stage. Panel (a) plots the number of clinical trials enrolling patients of each cancer-stage from 1973-2011; Panel (b) plots the number of clinical trials enrolling patients of each cancer-stage from 1973-2011 divided by the number of patients diagnosed with that cancer-stage from 1973-2009, as a rough adjustment for market size. For details on the sample, see the text and data appendix.

Figure D.2: **Survival time and R&D investments: Residualized cancer-stage data**



(a) Raw data

(b) Market size and cancer fixed effects

(c) Market size and stage fixed effects

(d) Market size, cancer and stage fixed effects

Notes: This figure shows the relationship between the five-year survival rate among patients diagnosed with each cancer-stage between 1973-2004 (the cohorts for which five-year survival is uncensored), and the number of clinical trials enrolling patients of each cancer-stage from 1973-2011. The level of observation is the cancer-stage. Panel (a) shows the raw data; Panel (b) residualizes market size and cancer fixed effects; Panel (c) residualizes market size and stage fixed effects; and Panel (d) residualizes market size, cancer fixed effects, and stage fixed effects. *Market size* denotes the inclusion of a covariate measuring the number of patients diagnosed with that cancer-stage between 1973-2009. As explained in the text, unstaged cancers are omitted from these figures since these observations do not identify the relationship of interest once we include cancer fixed effects and by definition unstaged cancers do not correspond to localized, regional, or metastatic stage definitions; Figure 2 shows an analogous scatterplot which includes unstaged cancers. For details on the sample, see the text and data appendix.

Table D.1: **Survival time and R&D investments: Robustness to cancer and stage fixed effects**

	Dependent variable: Number of clinical trials (mean = 945)				
	(1)	(2)	(3)	(4)	(5)
five-year survival rate	-0.963 ***	-1.151 ***	-1.588 ***	-0.339	-1.360 ***
	(0.236)	(0.188)	(0.132)	(0.305)	(0.315)
log(market size)	-	0.189 ***	0.098 **	0.193 ***	0.059
		(0.040)	(0.045)	(0.036)	(0.037)
cancer fixed effects	no	no	yes	no	yes
stage fixed effects	no	no	no	yes	yes

Notes: This table shows the relationship between the five-year survival rate among patients diagnosed with each cancer-stage between 1973-2004 (the cohorts for which five-year survival is uncensored), and the number of clinical trials enrolling patients of that cancer-stage from 1973-2011. The level of observation is the cancer-stage. Estimates are from quasi-maximum likelihood Poisson models. Standard errors are clustered at the cancer level. *: p<0.10; **: p<0.05; ***: p<0.01. *Market size* denotes the number of patients diagnosed with that cancer-stage between 1973-2009. As explained in the text, unstaged cancers are omitted from these regressions since these observations do not identify the relationship of interest once we include cancer fixed effects and by definition unstaged cancers do not correspond to localized, regional, or metastatic stage definitions; n=182. For details on the sample, see the text and data appendix.

Table D.2: **Survival time and R&D investments: Robustness to alternative survival measures**

	Dependent variable: Number of clinical trials (mean = 945)				
	(1)	(2)	(3)	(4)	(5)
one-year survival rate	-0.781 **				
	(0.325)				
five-year survival rate		-0.868 ***			
		(0.319)			
1973 survival (years)			-0.034 ***		
			(0.013)		
1973 one-year survival rate				-0.597 **	
				(0.297)	
1973 five-year survival rate					-0.731 **
					(0.309)

Notes: This table shows the relationship between various measures of the survival rate among patients diagnosed with each cancer-stage and the number of clinical trials enrolling patients of that cancer-stage from 1973-2011. The level of observation is the cancer-stage. Estimates are from quasi-maximum likelihood Poisson models. Standard errors are clustered at the cancer level. *: p<0.10; **: p<0.05; ***: p<0.01. The number of observations is 201 in Columns (1) and (2), and 187 in Column (3), because 14 cancer-stages had no patients diagnosed in 1973. For details on the sample, see the text and data appendix.

Table D.3: **Survival time and R&D investments: Robustness across samples**

	Dependent variable: Number of clinical trials (mean in Columns (1), (2) = 945)					
	(1)	(2)	(3)	(4)	(5)	(6)
five-year survival rate	-0.868 ***	-1.113 ***	-1.241 **	-1.498 ***	-0.963 ***	-1.151 ***
	(0.319)	(0.286)	(0.529)	(0.434)	(0.236)	(0.188)
log(market size)	-	0.243 ***	-	0.275 ***	-	0.189 ***
		(0.055)		(0.072)		(0.040)
excluding metastatic cancers	no	no	yes	yes	no	no
excluding unstaged cancers	no	no	no	no	yes	yes

Notes: This table shows the relationship between the five-year survival rate among patients diagnosed with each cancer-stage between 1973-2004 (the cohorts for which five-year survival is uncensored), and the number of clinical trials enrolling patients of that cancer-stage from 1973-2011. The level of observation is the cancer-stage. Estimates are from quasi-maximum likelihood Poisson models. Standard errors are clustered at the cancer level. *: p<0.10; **: p<0.05; ***: p<0.01. *Market size* denotes the number of patients diagnosed with that cancer-stage between 1973-2009. N = 201 in Columns (1) and (2); given the sample restrictions noted in the table, n=140 in Columns (3) and (4), and n=182 in Columns (5) and (6). For details on the sample, see the text and data appendix.

Table D.4: **Survival time and FDA drug approvals: Cancer-stage data**

	Dependent variable: Number of approved drugs (mean = 0.507)		
	(1)	(2)	(3)
five-year survival rate	-2.306 **	-2.719 ***	-2.341 ***
	(0.912)	(0.798)	(0.823)
log(market size)	-	0.393 ***	-
		(0.101)	
log(life-years lost)	-	-	0.438 ***
			(0.133)

Notes: This table shows the relationship between the five-year survival rate among patients diagnosed with each cancer-stage between 1973-2004 (the cohorts for which five-year survival is uncensored), and the number of drugs approved by the US FDA for that cancer-stage from 1990-2002. The level of observation is the cancer-stage. Estimates are from quasi-maximum likelihood Poisson models. Standard errors are clustered at the cancer level. *: p<0.10; **: p<0.05; ***: p<0.01. *Market size* denotes the number of patients diagnosed with that cancer-stage between 1973-2009. *Life-years lost* denotes age-gender-year specific life expectancy (in the absence of cancer) in the year of diagnosis, less observed survival time in years, averaged over patients diagnosed with that cancer-stage between 1973-1983 (to minimize censoring) multiplied times market size. The number of observations is 201 in Columns (1) and (2), and 192 in Column (3), because 9 cancer-stages had no patients diagnosed between 1973-1983. For details on the sample, see the text and data appendix.

Table D.5: **Surrogate endpoints, survival time, and drug approvals**

Panel (A): Level of R&D, Dependent variable: Number of approved drugs (mean = 0.507)						
	(1)		(2)		(3)	
five-year survival rate	-2.327	***	-2.815	***	-2.405	***
	(0.902)		(0.785)		(0.814)	
(0/1: *hematologic*)	1.250	***	1.178	***	1.032	**
	(0.458)		(0.393)		(0.432)	
log(market size)	-		0.398	***	-	
			(0.104)			
log(life-years lost)	-		-		0.413	***
					(0.141)	
Panel (B): Composition of R&D, Dependent variable: Number of approved drugs (mean = 0.507)						
	(1)		(2)		(3)	
(five-year survival rate)*(0/1: *hematologic*)	6.632	***	6.543	***	6.075	***
	(1.668)		(1.622)		(1.622)	
five-year survival rate	-3.743	***	-3.925	***	-3.539	***
	(1.273)		(1.054)		(1.111)	
(0/1: *hematologic*)	-1.032		-1.190	*	-1.164	*
	(0.725)		(0.639)		(0.605)	
log(market size)	-		0.376	***	-	
			(0.109)			
log(life-years lost)	-		-		0.386	**
					(0.153)	

Notes: This table shows two analyses of how cancer R&D differs on hematologic malignancies relative to other cancers, as a way of shedding light on how surrogate endpoints - which are more commonly used for hematologic malignancies - affect R&D investments. Panel (A) regresses the number of drugs approved by the US FDA for that cancer-stage from 1990-2002 on the five-year survival rate among patients diagnosed with each cancer-stage between 1973-2004 (the cohorts for which five-year survival is uncensored) and an indicator for hematological malignancies. Panel (B) regresses the number of drugs approved by the US FDA for that cancer-stage from 1990-2002 on the five-year survival rate among patients diagnosed with each cancer-stage between 1973-2004, an indicator for hematological malignancies, and an interaction between these two variables. The level of observation is the cancer-stage. Estimates are from quasi-maximum likelihood Poisson models. Standard errors are clustered at the cancer level. *: $p<0.10$; **: $p<0.05$; ***: $p<0.01$. *Market size* denotes the number of patients diagnosed with that cancer-stage between 1973-2009. *Life-years lost* denotes age-gender-year specific life expectancy (in the absence of cancer) in the year of diagnosis, less observed survival time in years, averaged over patients diagnosed with that cancer-stage between 1973-1983 (to minimize censoring) multiplied times market size. The number of observations is 201 in Columns (1) and (2), and 192 in Column (3), because 9 cancer-stages had no patients diagnosed between 1973-1983. For details on the sample, see the text and data appendix.

E Appendix: Development of chemoprevention drugs (not for publication)

In a review article on chemoprevention drugs in the journal *Cancer Prevention Research*, Meyskens et al. (2011) compile a list of the FDA approved drugs which prevent human cancers: BCG for bladder carcinoma in situ, Diclofenac for actinic keratoses, Celecoxib for familial adenomatous polyposis (FAP)-polyps, Photofrin for Barrett's esophagus, Tamoxifen/Raloxifene for breast cancer, and vaccines (Gardasil and Cervarix) to prevent cervical cancer. As summarized in Section 4.3, our qualitative investigation of the history of these FDA drug approvals suggests that each of these six approvals was either financed by the public sector (Tamoxifen and BCG) or relied on the use of surrogate endpoints (Diclofenac, Celecoxib, Photofrin, and cervical cancer vaccines). In this appendix, we provide documentation for this assertion.

E.1 BCG

The 1990 FDA approval for bladder carcinoma in situ was supported by clinical trials funded by the National Cancer Institute (NCI).[72] A popular press citation in the *New York Times* (Leary (1990)) noted the approval was supported by *"controlled, multi-center trials sponsored by the National Cancer Institute."* Lippman and Hawk (2009) cite the importance of one particular trial by Lamm et al. (1991) as supporting this approval, which was NCI-funded. Lippman and Hawk (2009) note: *"The FDA approved BCG for preventing recurrence of superficial bladder cancer in 1990 based on several clinical trials including one by the Southwest Oncology Group (Lamm et al. (1991))."* The acknowledgements in the Lamm et al. (1991) paper note: *"Conducted by the Southwest Oncology Group and supported in part by Public Health Service Cooperative Agreement grants from the National Cancer Institute (CA-04915, CA-37429, CA-42777, CA-04919, CA-27057, CA-13512, CA-1238, CA-36020, CA-22433, CA-16385, CA-20319, CA-37918, CA-13238, CA-35109, CA-12213, CA-12644, CA-35090, CA-461433, CA-35996, CA-35261, CA-14028, CA-03096, CA-35274, CA-22411, CA-35178, CA-35117, CA-35176, CA-35281, CA-28862, CA-03389, and CA-32102)."*

E.2 Diclofenac

Diclofenac is a topical treatment for actinic keratoses, which is clinically recommended for treatment to prevent disease progression to squamous cell carcinomas.[73] See, for example, the FDA approval letter and medical review for Diclofenac: http://www.accessdata.fda.gov/drugsatfda_docs/appletter/2000/21005ltr.pdf and http://www.accessdata.fda.gov/drugsatfda_docs/nda/2000/21005_Solaraze_medr_P1.pdf.

E.3 Celecoxib

The clinical trial endpoint for Celecoxib was a reduction in the number of adenomatous colorectal polyps, as a surrogate endpoint for gastrointestinal and other familial adenomatous polyposis (FAP)-related cancers. The medical review for Celecoxib's FDA approval notes: *"The sponsor has submitted clinical efficacy*

[72] Note that there seems to be a typo in the FDA approval date (1978) listed by Meyskens et al. (2011), because the FDA approval seems to have been in 1990.

[73] See, for example, Mcintyre, Downs and Bedwell (2007), who note: *"Actinic keratoses should be treated because of their potential to progress to squamous cell carcinomas."*

and safety data in support of the following new indication for Celebrex [celecoxib]: reduction in the number of adenomatous colorectal polyps in familial adenomatous polyposis patients...based on improvement in a surrogate endpoint" (http://www.accessdata.fda.gov/drugsatfda_docs/nda/99/21156-S007_Celebrex_medr.pdf).

E.4 Photofrin

The clinical trial endpoint for Photofrin was "complete ablation of high-grade dysplasia in patients with Barrett's esophagus," as a surrogate endpoint for the incidence of esophageal carcinoma. See, for example, the label for Photofrin: http://www.accessdata.fda.gov/drugsatfda_docs/label/2011/020451s020lbl.pdf.

E.5 Tamoxifen

Lippman and Brown (1999) note that the 1998 FDA approval of Tamoxifen as a chemoprevention agent was supported by the National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention Trial (Fisher et al. (1998)), which was funded by the National Cancer Institute and the National Institutes of Health. Lippman and Brown (1999) note: *"Tamoxifen as a chemopreventive agent has produced a fundamental change in the outlook for controlling breast cancer. Tamoxifen in the National Surgical Adjuvant Breast and Bowel Project (NSABP) P-1 Breast Cancer Prevention Trial (BCPT) achieved a striking 49% reduction in the incidence of invasive breast disease in women at increased risk of breast cancer (Fisher et al. (1998)). With this finding, the Food and Drug Administration (FDA) approved tamoxifen for risk reduction in this setting, marking the historic first FDA approval of any agent for cancer risk reduction."* The acknowledgements in the Fisher et al. (1998) paper note: *"This investigation was supported by Public Health Service grants U10-CA- 37377 and U10-CA-69974 from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services."*

E.6 Cervical cancer vaccines

The clinical trial endpoints for cervical cancer vaccines (Gardasil and Cervarix) were the incidence of cervical CIN 2/3 (cervical intraepithelial neoplasia grade 2/3) and cervical AIS (cervical adenocarcinoma in situ), as surrogate endpoints for the incidence of cervical cancer. See, for example, the label for the Gardasil vaccine: http://www.fda.gov/downloads/biologicsbloodvaccines/vaccines/approvedproducts/ucm111263.pdf, which states: *"CIN 2/3 and AIS are the immediate and necessary precursors of squamous cell carcinoma and adenocarcinoma of the cervix, respectively. Their detection and removal has been shown to prevent cancer; thus, they serve as surrogate markers for prevention of cervical cancer."*