Appendix

Rival Signals and Project Selection: Insights from the Drug Development Process

Appendix A: Data, Sample, and Variables

A.1. Definitions of Controls

Projects and molecular compounds that are licensed from other firms may have a different likelihood of being selected. Drugs that are in-licensed likely have a higher expectation of being developed, but there are also extra monetary and non-monetary costs to developing in-licensed drugs. For instance, royalties and milestone fees (Crama et al. 2008, Mason et al. 2008, Savva and Scholtes 2014) and the well-known "not-invented-here" syndrome (Katz and Allen 1982) may unfavorably impact the likelihood of selection. *Inlicensed* captures whether the drug was in-licensed (1) or originated in-house (0).

Similarly to in-licensed projects, drugs entering the pipeline as the result of an M&A may have different selection criteria applied to them. We code the binary variable *Acquired*, which captures whether a drug entered the development pipeline through an M&A (1) or began development in-house (0). (See also Appendix Appendix C: for more information on how we track the movement of firms throughout the industry.)

In many cases, once a firm receives regulatory approval for a certain drug-indication, it pursues approval for various follow-on indications. Neuberger et al. (2019) find that a new molecular entity (NME) with a successful primary indication has a 65% chance of approval for a secondary indication, compared to an 8.5% chance for an NME with a terminated primary indication. In other words, some compounds may belong to a separate, superior "class" of drugs in their ability to successfully treat multiple indications. To proxy for such compound "quality" (Danzon et al. 2005, Hoang and Rothaermel 2010), we create the binary variable $Multiple$. This variable takes on the value 1 (at the drug level) if a drug was successfully launched for two or more indications, and 0 otherwise.

Firms often pursue line extensions to previously approved drugs and also "cocktail" combinations of previously approved drugs. Because line extensions and cocktails can inflate the likelihood of firms successfully launching further drugs, we control for this by creating the variable *Extension*. In Cortellis, drug names allow us to distinguish between "new" drugs and extensions and/or cocktails. Performing a textual search on the name of the drug, we code *Extension* as 1 if the name of the drug includes any of the following phrases: "+" (which denotes drug cocktails), "release" (i.e., extended release), "formulation," "infusion," "patch," or "oral."

As returns from prior R&D investments may affect the selection decision (Loch and Kavadias 2002), we control for previous successes in developing an indication. We adapt Bennett and Snyder (2017) and creating rolling windows of cumulative ongoing and prior development successes. Specifically, we calculate: $Carryover_{i,y} = \sum_{x=1}^{3} PriorProj_{i,y-x} + \sum_{x=1}^{3} OngoingProj_{i,y-x}$. This captures carryovers from completed and ongoing projects in the same indication for a firm i in year y, over a sliding window of three years. The first term captures the number of projects that succeeded in the three years prior to the project being selected, and the second term

captures returns that may manifest from projects that were started in the years prior to the focal project being selected, but were still ongoing.

We control for a firm's R&D spend by first merging in annual R&D expenditures from Compustat (variable xrd).[1] We create *stdRD*, which is a standardized measure and rescales R&D expenditure to have a mean of 0 and a standard deviation of 1. We take a firm's R&D spend, subtract the sample mean R&D spend, and divide this by the sample standard deviation. We opt for a standardized measure instead of other transformations in order to avoid issues of collinearity between covariates in our models. For example, *ln*-transformed measures lead to variance inflation factors greater than 60 in our baseline estimations. Regardless of which measure we include, the results are not materially different.

Firms may select projects and be more likely to successfully develop them due to various development or downstream marketing capabilities they have gained. To proxy for such capabilities, we include firm- and indication-level measures of scale and scope. First, we capture the level of scope/diversification in a firm's pipeline by calculating the variable *IndScope*. This is a Herfindahl-Hirschman Index (HHI) of a firm's projects across all indications. Second, the variable *IndScale* captures economies of scale at the indication-level by measuring the number of ongoing projects for a clinical indication. Finally, *FirmScale* is the total number of ongoing projects across all clinical indications at the firm-level.

Firms may be more likely to pursue projects targeting a specific indication if there is a "hot" market or if they perceive the industry to be moving in that direction. To proxy for the possibility of herding (Banerjee 1992, Bikhchandani et al. 1992) in a domain, we control for the amount of deal-making activity surrounding an indication. We include the variable *Deals*, which is calculated as the year-on-year change in the number of in-licensing deals for development entered into by the Top 15 firms for a given indication.

Firms may be more likely to select projects for development if their pipelines are thinning and they become more desperate. We capture the strength of a firm's pipeline by calculating the variable *Strength* in two steps: First, we count the number of ongoing projects in a year at each phase of development, multiply each by its respective probability of being launched given its current phase (see Appendix Appendix C:), and add each measure together to create a cumulative, firm-year pipeline score. This score is indicative of the number of later-stage projects a firm has in its pipeline. As in Higgins and Rodriguez (2006), we then calculate the year-on-year change in firm-level scores to capture whether pipelines are strengthening or weakening.

We control for the proportion of pre-clinical projects in a firm's pipeline that are sent to clinical trials in a year. We calculate *SelectionRatio* as the ratio of the number of projects sent to clinical trials in a given year, over the number of projects sent to clinical trials plus those that remained in pre-clinical trials. In this sense, *SelectionRatio* captures how "wide" each firm's gate between pre-clinical and clinical trials is. As a firm's *SelectionRatio* increases, the likelihood of any one project being selected increases, so we expect this variable to be positively correlated with selection. In addition to firm fixed effects, this allows us to control for differing selection protocols among firms.

Unanticipated "bad news" may also affect selection. For example, drug withdrawals and late-stage clinical failures are likely to be accompanied by an acceleration of other compounds into the pipeline. We follow the intuition in Girotra et al. (2007) and rely on changes in market valuation to capture the magnitude of such bad news. We obtain daily closing stock price information from Thomson Reuters Datastream, and we create a variable that measures the most acute day-on-day percentage drop in a firm's stock price in a calendar year, taking care not to count stock splits. To ease interpretation, we take the absolute value so that more positive values of *PriceShock* correspond to larger drops in share price. Table A.I lists the top 10 price shocks in our data set. Because we employ *SelectionRatio* and *PriceShock* as instruments in Section 6, we measure these in

[1] For AbbVie, we take R&D expenditures from Abbott Laboratories from 1999 through 2013 and R&D information from AbbVie from 2014 through 2016 (after it was spun-off).

Table A.I: Top 10 Price Shocks

Notes: This table lists the most acute day-on-day stock price drops.

Firm	Date	Price Chg.	Event
Eli Lilly	Aug 9, 2000	-30.9%	Sooner-than-anticipated patent expiration for blockbuster drug Prozac
Merck & Co Inc	Sept 30, 2004	-26.8%	Withdrawal of blockbuster drug VIOXX due to safety concerns
Gilead	Nov 2, 1999	-22.5%	FDA rejection of adefovir dipivoxil application
Bristol Myers Squibb	Apr 19, 2000	-22.4%	Withdrawal of FDA filing for Vanlev due to safety concerns
Bayer	Aug 8, 2001	-16.8%	Withdrawal of Bayacol after links to adverse side effects and deaths
AstraZeneca	Aug 19, 2002	-16.3%	Failure of Iressa (gefitinib) in clinical trials
AbbVie (Abbott)	Jun 11, 2002	-16.1%	Earnings outlook cut due to manufacturing problems and slower sales of Meridia/Reductil
Bristol Myers Squibb	Aug 5, 2016	-16.0%	Phase III failure of Opdivo (nivolumab) to treat non-small cell lung cancer
Johnson & Johnson	Jul 19, 2002	-15.8%	FDA legal suit surrounding anemia drug Eprex
Bristol Myers Squibb	Mar 20, 2002	-15.6%	Poor late-stage clinical trial results for Vanlev

the year prior to the selection decision. This helps us ensure that both are *shaping* the selection decision, rather than the outcome thereof.

We include TA fixed effects in order to control for the TA the project targets. To generate categories for the TA, we enlisted the assistance of a licensed physician who went through indications in Cortellis and coded each indication as belonging to a certain TA. The final list of TAs was (i) Cardiovascular, (ii) Central Nervous System/Neurology/Psychiatry, (iii) Endocrinology/Metabolic, (iv) Gastrointestinal/Hepatology, (v) Immunology/Rheumatology/Dermatology, (vi) Infectious Diseases, (vii) Nephrology/Genitourinary, (viii) Oncology, (ix) Respiratory, (x) Otolaryngology/Ophthalmology, and (xi) Other. A complete mapping of indications to TAs is provided in Appendix Appendix D:.

A.2. Correlation Matrix

Table A.II: Correlation Matrix for all Variables

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1)	*Selected*										
(2)	*Success*	-									
(3)	*PreClinSignal*	**-0.221**	**-0.078**								
(4)	*TrialSignal*	**0.059**	**-0.093**	**0.639**							
(5)	*Biologic*	**0.140**	**0.096**	0.002	**0.085**						
(6)	*Rare*	**0.159**	0.002	**-0.126**	**-0.073**	**0.093**					
(7)	*Incidence_Reverse*	0.014	0.083	**-0.568**	**-0.628**	0.006	**0.417**				
(8)	*Inlicensed*	-0.009	0.033	**0.029**	0.022	**0.139**	0.026	-0.002			
(9)	*Acquired*	**-0.119**	0.029	-0.003	-0.017	**0.083**	0.020	0.045	**-0.058**		
(10)	*Multiple*	**0.196**	**0.362**	**-0.109**	**-0.070**	0.018	**0.088**	**0.157**	-0.022	**-0.029**	
(11)	*Extension*	**0.122**	**0.220**	**-0.056**	-0.016	**-0.078**	0.005	0.037	0.012	-0.01	**0.112**
(12)	*Carryover*	**0.063**	**0.158**	**-0.053**	**0.044**	**0.046**	**-0.045**	0.063	-0.015	-0.026	-0.001
(13)	*stdRD*	**0.123**	-0.027	0.004	**0.155**	**0.115**	**0.066**	0.014	**0.028**	**0.051**	-0.007
(14)	*IndScope*	**-0.054**	0.010	0.021	**-0.042**	-0.001	-0.008	-0.009	-0.019	-0.005	-0.024
(15)	*IndScale*	**-0.122**	**-0.077**	**0.744**	**0.685**	0.004	**-0.128**	**-0.529**	-0.016	**0.029**	**-0.113**
(16)	*FirmScale*	**0.140**	**-0.073**	-0.012	**0.131**	**0.074**	**0.053**	0.034	0.011	**0.036**	0.011
(17)	*Deals*	-0.005	-0.017	**0.170**	**0.147**	0.007	-0.002	-0.030	**0.047**	0.014	-0.003
(18)	*Strength*	**-0.076**	0.022	-0.014	**-0.044**	-0.027	-0.023	0.017	**-0.045**	**0.242**	-0.007
(19)	*SelectionRatio*	**0.210**	0.042	**-0.056**	-0.000	**0.033**	**0.035**	-0.014	0.017	**-0.069**	**0.079**
(20)	*PriceShock*	**0.063**	-0.016	**-0.050**	**-0.066**	**-0.045**	**-0.033**	-0.029	**-0.035**	**0.040**	0.014

		(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	
(12)	*Carryover*	**0.077**									
(13)	*stdRD*	0.019	0.020								
(14)	*IndScope*	-0.005	**-0.055**	**-0.371**							
(15)	*IndScale*	-0.027	**0.083**	**0.180**	**-0.039**						
(16)	*FirmScale*	**0.034**	**0.032**	**0.769**	**-0.424**	**0.222**					
(17)	*Deals*	0.011	-0.007	0.015	-0.019	**0.146**	**0.032**				
(18)	*Strength*	-0.002	-0.004	**-0.044**	-0.000	0.018	-0.023	**0.044**			
(19)	*SelectionRatio*	0.023	**0.043**	0.012	**-0.095**	**-0.046**	**0.029**	0.008	-0.020		
(20)	*PriceShock*	**0.035**	0.002	**-0.182**	**0.128**	**-0.060**	**-0.169**	0.005	**0.068**	**0.063**	

Notes: Correlations in bold are significant at the $\alpha = 0.01$ level.

Appendix B: Robustness Checks

We perform a series of tests in order to ensure that our results are not being driven by our econometric setup, and that they are robust to alternative model specifications. §B.1. redefines selection as entry into Phase II trials and also presents a model with selection at Phase I and Phase II. §B.2. estimates models that test H2 and H3 simultaneously. §B.3. compares our heckprobit model to models without instruments and shows that not accounting for selection when estimating success biases estimates. §B.4. shows robustness to the inclusion of indication-level fixed effects and §B.5. to concerns of right censoring. §B.6. shows robustness to an alternative definition of success for outlicensed projects, and §B.7. clusters standard errors at different levels. Finally, in §B.8. we address concerns about endogeneity due to the emergence of potential breakthrough technologies.

B.1. Phase II Selection & Selection at Multiple Stages

First, we test the robustness of our results when we change the definition of selection to the decision to send a drug into Phase II trials. Whereas Phase I is the first test in humans that takes the compound "out of the lab," Phase II is another milestone in the drug development process as it establishes the efficacy of the compound. Although we believe that the decision to advance a compound from pre-clinical to Phase I trials is a more valid definition of selection in this context, we nevertheless test whether the above analysis is robust to selection defined as entry into Phase II trials.

We define the binary variable $SelectedPII$ as 1 if a project was sent into Phase II clinical trials, and 0 if it was never sent to Phase II. We re-estimate our models and display the results in Table B.I. In Model I, we find that $PreClinSignal$ is negatively associated with selection and $TrialSignal$ is positively associated with selection, providing support for H1a and H1c. In Model II, we find support for H2 as the interactions $TrialSignal\_Bio \times Biologic$ and $TrialSignal\_NonBio \times Biologic$ are significant, whereas the interactions with pre-clinical signals are not. These results mirror those from Table 4 in the manuscript. Finally, Models III and IV show that market potential augments the effects of early-stage signals from rival projects substantially more than it affects late-stage signals, providing support for H3. Although we do obtain significance on $TrialSignal \times Incidence\_Reverse$, the effect size is several times smaller than that on $PreClinSignal \times Incidence\_Reverse$ and may be attributed to the particularities of the small subsample on which we are estimating the model.

Second, the pharmaceutical drug development process can be modeled as one comprised of multiple stages and gates rather than a one-shot selection into clinical trials. We consider the "nestedness" of the selection process by augmenting our heckprobit model. This trivariate probit model with selection simultaneously estimates a probit model of selection into Phase I, a probit model of selection into Phase II, and a probit model of success, and it captures any additional selection bias that may be occurring between (i) Phase I and Phase II, and (ii) Phase II and launch. This model estimates the following three equations simultaneously.

$$Selected^* = \kappa_0 + \kappa \mathbf{A} + \varepsilon_1 \tag{B1}$$

$$Selected = \mathbb{1}[Selected^* > 0]$$

$$SelectedPII^* = \lambda_0 + \lambda \mathbf{B} + \varepsilon_2 \tag{B2}$$

$$SelectedPII = \mathbb{1}[SelectedPII^* > 0]$$

$$Success^* = \nu_0 + \nu \mathbf{C} + \varepsilon_3 \tag{B3}$$

$$Success = \mathbb{1}[Selected^* > 0]$$

In order to estimate the trivariate probit with selection model, we also specify the variables $SelectionRatioPII$ (the ratio of the number of projects sent to Phase II trials in a given year, over the number of

Table B.I: Phase II Selection

Notes: Probit model estimating selection into Phase II. Reported estimates for Model I are average marginal effects, for Models II–IV are probit coefficients. Robust standard errors clustered by drug are in parentheses. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $+p < 0.10$

	(I) Base H1	(II) Biologic H2	(III) Rare Diseases H3	(IV) Incidence H3
PreClinSignal	-0.0040***		-0.0148***	-0.5507***
	(0.0003)		(0.0011)	(0.1106)
TrialSignal	0.0021***		0.0075***	0.1087***
	(0.0004)		(0.0016)	(0.0213)
Biologic	0.0877***	0.3926***	0.3149***	0.1031
	(0.0169)	(0.0682)	(0.0586)	(0.1311)
PreClinSignal_Bio		0.0012		
		(0.0048)		
PreClinSignal_Bio × *Biologic*		0.0031		
		(0.0121)		
TrialSignal_Bio		0.0008		
		(0.0063)		
TrialSignal_Bio × *Biologic*		0.0433***		
		(0.0114)		
PreClinSignal_NonBio		-0.0194***		
		(0.0020)		
PreClinSignal_NonBio × *Biologic*		-0.0002		
		(0.0043)		
TrialSignal_NonBio		0.0120***		
		(0.0022)		
TrialSignal_NonBio × *Biologic*		-0.0269***		
		(0.0048)		
Rare	0.0720**	0.2459**	0.4953***	-0.1203
	(0.0231)	(0.0797)	(0.1049)	(0.1230)
PreClinSignal × *Rare*			-0.2210***	
			(0.0471)	
TrialSignal × *Rare*			-0.0031	
			(0.0097)	
Incidence_Reverse				0.0105
				(0.0399)
PreClinSignal × *Incidence_Reverse*				-0.0381***
				(0.0090)
TrialSignal × *Incidence_Reverse*				0.0087***
				(0.0017)
All Controls	Yes	Yes	Yes	Yes
TA Indicators	Yes	Yes	Yes	Yes
Firm Indicators	Yes	Yes	Yes	Yes
Year Indicators	Yes	Yes	Yes	Yes
# Projects	7,418	7,418	7,418	1,266
Pseudo R^2	0.272	0.278	0.276	0.235
Log-likelihood	-3,555	-3,523	-3,535	-601.7

Table B.II: Trivariate Probit with Selection

Notes: Estimates from the simultaneous estimation of Equations B1, B2, and B3. Reported estimates are average marginal effects. Robust standard errors clustered by drug are in parentheses. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $+p < 0.10$

	Eqn. B1 *Selected*	Eqn. B2 *SelectedPII*	Eqn. B3 *Success*
PreClinSignal	-0.0041***	-0.0004	-0.0006
	(0.0003)	(0.0008)	(0.0010)
PISignal	0.0050***	-0.0107***	-0.0047+
	(0.0011)	(0.0012)	(0.0027)
PIISignal	0.0075***	0.0100***	-0.0062*
	(0.0012)	(0.0020)	(0.0026)
PIIISignal	0.0090**	0.0041	0.0100*
	(0.0031)	(0.0041)	(0.0042)
RegSignal	0.0028	0.0337	0.0518*
	(0.0211)	(0.0268)	(0.0244)
SelectionRatioPII	0.3110***	0.6748***	
	(0.0592)	(0.1139)	
PriceShockPII	-0.5891*	0.6941*	
	(0.2599)	(0.2758)	
SelectionRatio	0.2553***		
	(0.0637)		
PriceShock	1.9111***		
	(0.2585)		
$\rho_{B1,B2}$		-0.416*	
$\rho_{B2,B3}$		0.241	
$\rho_{B1,B3}$		-0.496+	
All Controls		Yes	
TA Indicators		Yes	
Firm Indicators		Yes	
Year Indicators		Yes	
# Projects		6,502	
Log pseudolikelihood		-5,438	

projects sent to Phase II plus those that remained in Phase I) and $PriceShockPII$, which we insert into Eqns. B1 and B2, but not into Eqn. B3. These are meant to serve as instruments that predict selection into Phase II, but not success. In this manner, the model corrects for the possibility that $\rho_{B1,B2} \neq 0$, $\rho_{B2,B3} \neq 0$, and $\rho_{B1,B3} \neq 0$.

Table B.II reports estimates from these models. We first note that we cannot reject the null hypothesis that the correlation between the stage two and three errors is different from zero (that is, $\rho_{B2,B3} = 0$). Contrast this to Phase I, where $\rho_{B1,B2} = -0.416$ ($p < 0.05$), and we argue that managers must decide whether or not to send a project to clinical trials. This suggests that (to the extent that estimating models of success is concerned) the only selection bias needing to be considered is that arising from the decision to send a project into Phase I trials in the first place.

The results for the most part mirror our main findings from Section 6. Regarding Phase I selection, we find that pre-clinical signals are associated with a lower likelihood of selection, whereas signals in Phase I and beyond are associated with a higher likelihood of success. Regarding Phase II selection, we find that, conditional on already being selected into Phase I, signals from rival projects in pre-clinical and Phase I are associated with a lower likelihood of selection, whereas Phase II and later signals are associated with a higher likelihood of success. Taken together with the Phase I selection results, this suggests more generally that signals from rival projects in the same phase and earlier are associated with stronger competitive effects and a lower likelihood of selection, whereas technological signals dominate when firms receive signals from later-stage rival projects. Finally, regarding Success, we document a pattern similar to our main heckprobit results from Table 6.

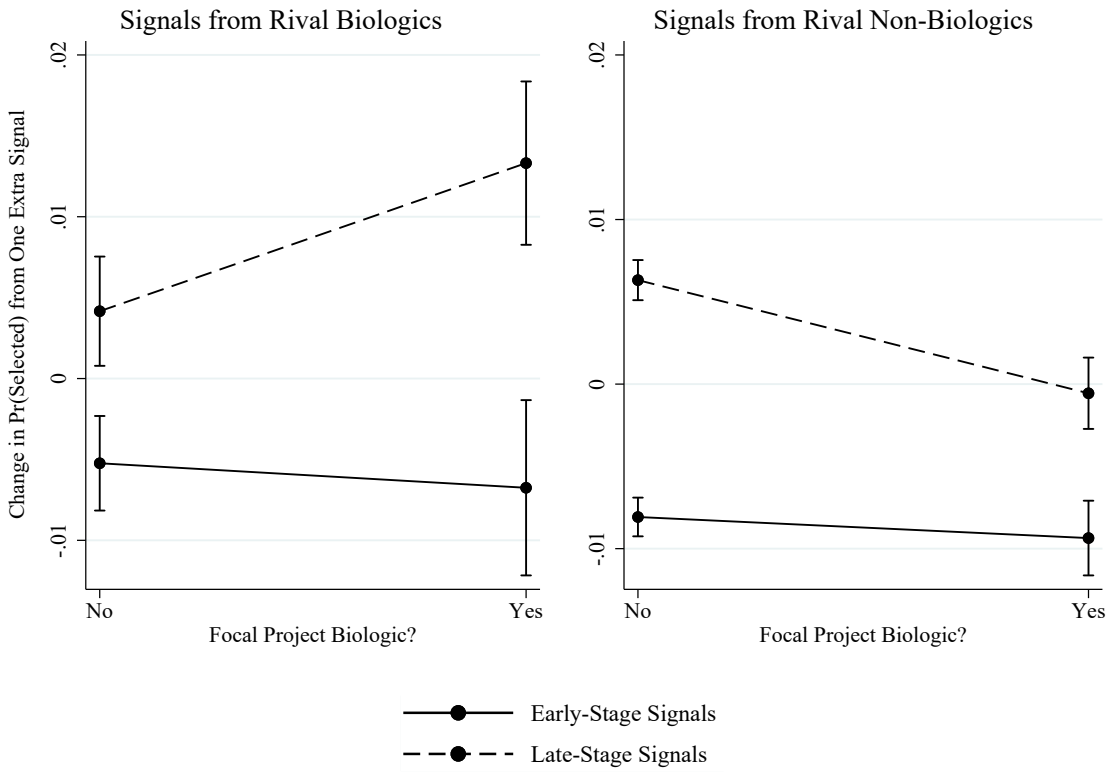

Figure B1: Signals from Rival Biologic and Non-Biologic Projects

B.2. A "Full" Model of Selection

In Sections 5.2.1 and 5.2.2, we use two different models to test H2 and H3. The reasons for doing so are twofold. First, a full-fledged model estimating all these coefficients becomes quite complex, with many interactions. Second, there is the issue of perfect multicollinearity. In testing market potential (H3), for example, we include the variable $PreClinSignal$ in one model, whereas when testing technological relatedness (H2), we decompose this into $PreClinSignalBio$ and $PreClinSignalNonBio$ in the other model.[2] Since $PreClinSignal = PreClinSignalBio + PreClinSignalNonBio$, we cannot include all three variables in the same model.

One remedy is to estimate one "full" model that tests H3 using the decomposed variables, although this also suffers from the "complexity" noted above. Nevertheless, to assess the robustness of our results to the inclusion of both sets of variables simultaneously, we estimate a full model and display the results in Table B.III and Figures B1 and B2. Our results are robust to the inclusion of both sets of variables, and our conclusions do not change. In particular:

1. Pre-clinical (clinical) signals are negatively (positively) correlated with Phase I selection, corroborating H1a and H1c,

2. Clinical-phase signals from rival projects are stronger when the focal project's technology matches the rival project's technology, corroborating H2, and

3. The effect of pre-clinical rival signals on selection are stronger (more negative) in low-market-potential regimes than in high-market-potential regimes, corroborating H3.

[2]The analogous line of thinking holds for $TrialSignal$ as well.

Table B.III: Probits Testing H2 and H3

Notes: All models include fixed effects at the level of the indication. Reported estimates for Models I and II are probit coefficients. Robust standard errors clustered by drug are in parentheses. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{+}p < 0.10$

	(I)	(II)
$Biologic$	0.3679***	0.1987
	(0.0648)	(0.1753)
$PreClinSignal\_Bio$	-0.0087**	-0.6307+
	(0.0031)	(0.3239)
$PreClinSignal\_Bio \times Biologic$	0.0016	-0.0911
	(0.0073)	(0.0671)
$TrialSignal\_Bio$	0.0136*	0.0940
	(0.0058)	(0.0797)
$TrialSignal\_Bio \times Biologic$	0.0367**	0.0444*
	(0.0116)	(0.0221)
$PreClinSignal\_NonBio$	-0.0167***	-0.9019***
	(0.0012)	(0.1473)
$PreClinSignal\_NonBio \times Biologic$	0.0030	0.0207
	(0.0026)	(0.0391)
$TrialSignal\_NonBio$	0.0215***	0.0959**
	(0.0022)	(0.0358)
$TrialSignal\_NonBio \times Biologic$	-0.0236***	-0.0193+
	(0.0045)	(0.0107)
$Rare$	0.6423***	0.3384*
	(0.1131)	(0.1586)
$PreClinSignalBio \times Rare$	-0.2586*	
	(0.1166)	
$PreClinSignalNonBio \times Rare$	-0.3053***	
	(0.0541)	
$TrialSignalBio \times Rare$	0.0152	
	(0.0322)	
$TrialSignalNonBio \times Rare$	-0.0007	
	(0.0144)	
$Incidence\_Reverse$		-0.0639
		(0.0443)
$PreClinSignalBio \times Incidence\_Reverse$		-0.0474+
		(0.0255)
$PreClinSignalNonBio \times Incidence\_Reverse$		-0.0648***
		(0.0118)
$TrialSignalBio \times Incidence\_Reverse$		0.0065
		(0.0064)
$TrialSignalNonBio \times Incidence\_Reverse$		0.0065*
		(0.0029)
All Controls	Yes	Yes
TA, Firm, Year Indicators	Yes	Yes
# Projects	8,024	1,424
Pseudo R^2	0.267	0.224
Log-likelihood	-4,020	-402.6

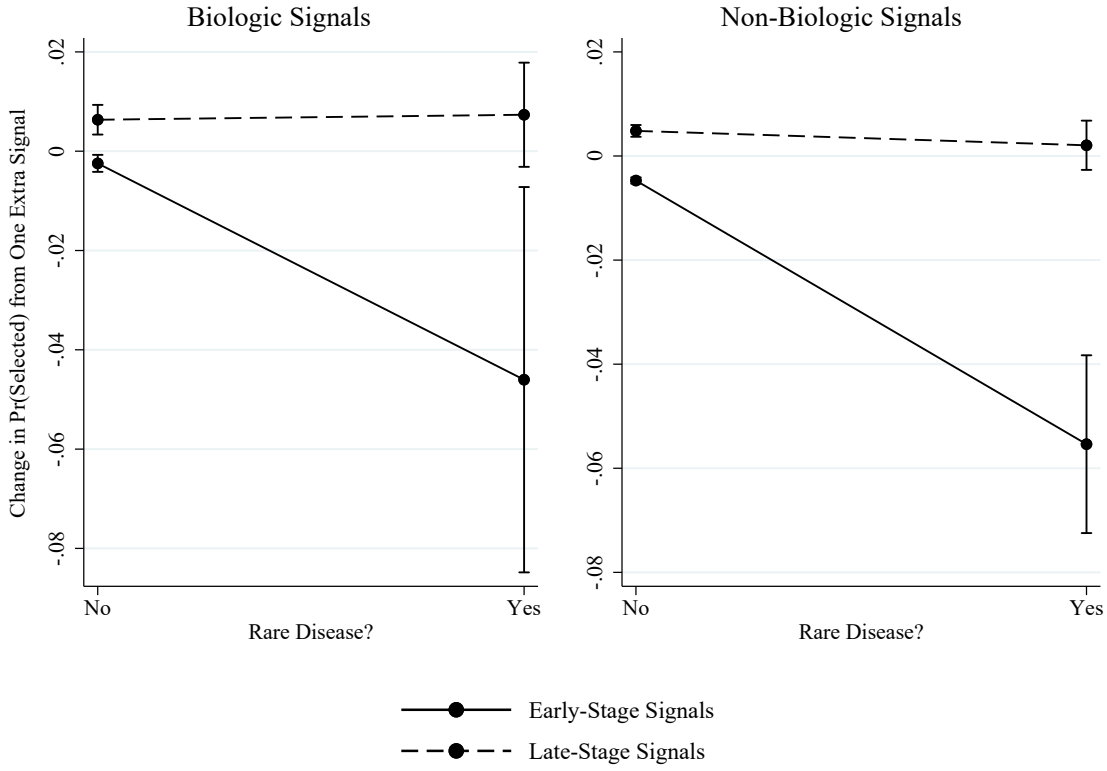

Figure B2: Rival Project Signals and Market Potential

B.3. Heckprobit Specification

Our objective in this section is twofold. First, heckprobit models are known to suffer if exclusion restrictions are incorrectly specified, and so we first make sure that our models are not adversely affected by our two instruments. Second, we compare our heckprobit model to a "naive" probit model of success (which does not correct for selection) and show that not correcting for selection may lead to biased estimates. Table B.IV displays estimates from several different models, which we explain in turn.

Comparing the first stage of our main heckprobit (Column I) to a probit model of selection (Col. III), we see that, in absolute terms, the coefficients correspond quite well with one another. The relative bias between the first-stage heckprobit and the probit is also relatively low, with the one notable exception being $IndScope$, which anyway is statistically non-significant. Small deviations between the two models likely arise as the heckprobit model simultaneously estimates the first and second stages, and so is maximizing a likelihood function that includes information on project successes as well (i.e., because the errors are correlated, the estimation of the success equation affects the estimates of the selection equation). Importantly, there are no sign flips, and significance levels are quite similar across the board. The inclusion of our two instruments—at the least—does not seem to be making the estimation of selection worse.

Col. IV then estimates a probit model of success that does not correct for possible selection bias. In contrast to the comparison between Cols. I and III, a failure to account for sample selection leads to significant differences in coefficient estimates between the probit (IV) and the second stage of the heckprobit (II), both in absolute and relative terms. These differences are quite pronounced. Moreover, assuming that the heckprobit model in Col. II is valid, the probit coefficients on the signals from rival projects are consistently biased away from zero (i.e., the effect sizes are stronger). In other words, estimation of a heckprobit model with our instruments seems to result in a much better specified model.

Next, we compare the first stage of a heckprobit without these two instruments (Col. V) to our baseline probit of selection (Col. III). However, having established that $SelectionRatio$ and $PriceShock$ both significantly load on selection, any model that does not include these two variables is likely to be misspecified. This can be seen by examining the bias between these two models in Col. VIII, which is generally worse than the bias from our first-stage heckprobit with the two instruments in Col. IV.

Taken together, the results suggest that including these two variables leads to a better heckprobit model that closely agrees with a probit model of selection and is better able to address sample selection concerns.

Table B.IV: Heckprobit and Probit Comparisons

Notes: This table compares estimates from heckprobit and probit models of selection and success. Reported estimates are average marginal effects. Robust standard errors clustered by drug are in parentheses. Log-P(L) refers to either log-pseudolikelihood (heckprobit) or log-likelihood (probit). $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{+}p < 0.10$

	(I) Heckprobit Model	(II)	(III) Probit Model		(IV) Probit Model		(V) Heckprobit w/o ExcR		(VI) Heckprobit w/o ExcR	
	Selected	*Success*	*Selected*	Bias (III vs. I)	*Success*	Bias (IV vs. II)	*Selected*	Bias (V vs. III)	*Success*	(VI vs. II)
PreClinSignal	-0.0042***	-0.0009*	-0.0042***	-0.05%	-0.0013**	42.60%	-0.0044***	3.80%	-0.0007*	-138.36%
	(0.0003)	(0.0004)	(0.0003)		(0.0005)		(0.0003)		(0.0003)	
PISignal	0.0050***	-0.0025**	0.0051***	-0.76%	-0.0031**	24.56%	0.0051***	0.77%	-0.0027**	8.29%
	(0.0011)	(0.0010)	(0.0011)		(0.0012)		(0.0011)		(0.0010)	
PIISignal	0.0075***	-0.0010	0.0074***	0.40%	-0.0017+	68.41%	0.0079***	6.49%	-0.0007	-29.98%
	(0.0012)	(0.0008)	(0.0012)		(0.0010)		(0.0012)		(0.0007)	
PIIISignal	0.0092**	0.0047**	0.0088**	4.29%	0.0056*	19.39%	0.0095**	7.74%	0.0050**	6.45%
	(0.0032)	(0.0018)	(0.0032)		(0.0023)		(0.0032)		(0.0018)	
RegSignal	0.0016	0.0232**	0.0018	-12.75%	0.0297**	28.14%	0.0007	-60.90%	0.0229**	-1.12%
	(0.0211)	(0.0086)	(0.0211)		(0.0110)		(0.0212)		(0.0087)	
Biologic	0.0575**	0.0160	0.0604***	-4.88%	0.0235+	46.77%	0.0583**	-3.54%	0.0123	-23.27%
	(0.0176)	(0.0102)	(0.0177)		(0.0129)		(0.0181)		(0.0103)	
Rare	0.0243	-0.0120	0.0256	-4.84%	-0.0161	34.27%	0.0249	-2.55%	-0.0128	6.96%
	(0.0293)	(0.0163)	(0.0291)		(0.0216)		(0.0302)		(0.0154)	
Multiple	0.3819***	0.2487***	0.3766***	1.38%	0.2319***	-6.74%	0.3863***	2.58%	0.2841***	14.25%
	(0.0285)	(0.0310)	(0.0289)		(0.0295)		(0.0279)		(0.0356)	
Extension	0.1658***	0.1275***	0.1600***	3.67%	0.1377***	7.97%	0.1744***	9.05%	0.1440***	12.95%
	(0.0280)	(0.0230)	(0.0285)		(0.0252)		(0.0269)		(0.0250)	
Inlicensed	-0.0641***	0.0247*	-0.0631***	1.64%	0.0342**	38.55%	-0.0671***	6.37%	0.0207*	-16.22%
	(0.0172)	(0.0104)	(0.0173)		(0.0130)		(0.0172)		(0.0101)	
Acquired	-0.2204***	-0.0012	-0.2212***	-0.38%	0.0047	-480.63%	-0.2255***	1.95%	-0.0038	202.82%
	(0.0230)	(0.0204)	(0.0234)		(0.0300)		(0.0237)		(0.0179)	
Carryover	0.0266**	0.0119**	0.0263**	0.87%	0.0160***	34.24%	0.0278**	5.67%	0.0123**	2.83%
	(0.0096)	(0.0037)	(0.0094)		(0.0048)		(0.0098)		(0.0038)	
stdRD	-0.0256+	0.0113	-0.0242	5.97%	0.0179	57.93%	-0.0153	-36.49%	0.0112	-0.90%
	(0.0149)	(0.0105)	(0.0150)		(0.0134)		(0.0153)		(0.0098)	
IndScope	0.1910	0.0951	0.1979	-3.49%	0.0430	-54.81%	-0.1206	-160.96%	0.0323	-66.07%
	(0.5172)	(0.2611)	(0.5343)		(0.3643)		(0.4861)		(0.2405)	
IndScale	-0.0010	0.0031*	-0.0010	5.66%	0.0041*	32.85%	-0.0012	19.97%	0.0027*	-12.61%
	(0.0013)	(0.0013)	(0.0013)		(0.0017)		(0.0014)		(0.0012)	
FirmScale	0.0001	-0.0001	0.0001	-0.92%	-0.0001	44.36%	0.0001	-44.01%	-0.0001	-7.95%
	(0.0002)	(0.0001)	(0.0002)		(0.0001)		(0.0002)		(0.0001)	
Deals	0.0020	-0.0032	0.0019	0.08%	-0.0043	35.90%	0.0018	-5.51%	-0.0026	-18.37%
	(0.0027)	(0.0023)	(0.0027)		(0.0030)		(0.0027)		(0.0023)	
Strength	-0.0012	-0.0000	-0.0012	-6.34%	-0.0001	790.98%	-0.0011	-13.03%	0.0000	-346.72%
	(0.0011)	(0.0006)	(0.0011)		(0.0008)		(0.0011)		(0.0006)	
SelectionRatio	0.3928***		0.4051***	-3.04%						

	(0.0554)			(0.0560)								
PriceShock	1.5226***			1.5879***		-4.11%						
	(0.1962)			(0.1915)								
ρ	-0.593**									-0.881***		
$Wald\chi^2(1)\rho = 0$	10.71									16.63		
TA Indicators	Yes			Yes			Yes			Yes		
Firm Indicators	Yes			Yes			Yes			Yes		
Year Indicators	Yes			Yes			Yes			Yes		
# Projects	6,506	3,040		6,506		3,020			6,506			3,040
Log-(pseudo)likelihood	-3,963			-3,353.0			-615.1			-4,052.0		
Pseudo-R^2				0.254		0.331						

B.4. Indication Effects

We test the robustness of our results to the inclusion of indication-level fixed effects. The intuition is that there may be some indication-specific factors that affect selection and may not be getting picked up by our TA-level fixed effects. We note that this approach has three main drawbacks. First, there are 485 unique indications in our data set, which introduces several hundred variables into the models (thus drastically reducing our degrees of freedom). Second, there are many observations for which indication-fixed effects perfectly predict selection (i.e., there is no variation within indication regarding selection), and these observations drop out of our models, reducing our estimating sample. Third, since *Rare* is at the level of the indication, it is perfectly collinear with the indication-fixed effects, and we cannot test H3 with this variable. Because of the above limitations and because including dummies does not provide much explanatory power (the vast majority of these controls are statistically non-significant; not reported), we provide these models as robustness checks. These results are presented in Table B.V and show that all our results hold.

Table B.V: Probits Controlling for Indication Effects

Notes: All models include fixed effects at the level of the indication. Reported estimates for Model I are average marginal effects, for Models II and III are probit coefficients. Robust standard errors clustered by drug are in parentheses. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $+p < 0.10$

	(I) Base	(II) Biologic	(III) Incidence
$PreClinSignal$	-0.0035***		-0.8836***
	(0.0005)		(0.1535)
$TrialSignal$	0.0052***		0.0695+
	(0.0007)		(0.0361)
$Biologic$	0.0816***	0.3863***	0.1177
	(0.0159)	(0.0747)	(0.1472)
$PreClinSignal\_Bio$		-0.0093*	
		(0.0037)	
$PreClinSignal\_Bio \times Biologic$		0.0027	
		(0.0071)	
$TrialSignal\_Bio$		0.0176**	
		(0.0058)	
$TrialSignal\_Bio \times Biologic$		0.0283**	
		(0.0102)	
$PreClinSignal\_NonBio$		-0.0146***	
		(0.0022)	
$PreClinSignal\_NonBio \times Biologic$		0.0018	
		(0.0026)	
$TrialSignal\_NonBio$		0.0208***	
		(0.0028)	
$TrialSignal\_NonBio \times Biologic$		-0.0193***	
		(0.0041)	
$Incidence\_Reverse$			0.2198
			(0.7813)
$Incidence\_Reverse \times PreClinSignal$			-0.0659***
			(0.0124)
$Incidence\_Reverse \times TrialSignal$			0.0038
			(0.0029)
All Controls	Yes	Yes	Yes
Indication Effects	Yes	Yes	Yes
Firm Effects	Yes	Yes	Yes
Year Effects	Yes	Yes	Yes
# Projects	7,416	7,416	1,186
# Indications	273	273	30
Pseudo R^2	0.294	0.297	0.268
Log-likelihood	-3,607	-3,591	-354.3

Table B.VI: Probits Considering Censoring

Notes: Probit models of selection excluding the five most recent years of projects in our sample (i.e., from 2012 to 2016). Reported estimates for Model I are average marginal effects, for Models II–IV are probit coefficients. Robust standard errors clustered by drug are in parentheses. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{+}p < 0.10$

	(I) Base	(II) Biologic	(III) Rare Diseases	(IV) Incidence
PreClinSignal	-0.0043***		-0.0146***	-0.8500***
	(0.0002)		(0.0009)	(0.1305)
TrialSignal	0.0052***		0.0174***	0.0766*
	(0.0006)		(0.0021)	(0.0306)
Biologic	0.0797***	0.3476***	0.2729***	-0.0656
	(0.0171)	(0.0704)	(0.0583)	(0.1535)
PreClinSignal_Bio		-0.0082*		
		(0.0042)		
PreClinSignal_Bio × *Biologic*		0.0021		
		(0.0113)		
TrialSignal_Bio		0.0138+		
		(0.0071)		
TrialSignal_Bio × *Biologic*		0.0286*		
		(0.0143)		
PreClinSignal_NonBio		-0.0167***		
		(0.0013)		
PreClinSignal_NonBio × *Biologic*		0.0024		
		(0.0034)		
TrialSignal_NonBio		0.0216***		
		(0.0023)		
TrialSignal_NonBio × *Biologic*		-0.0208***		
		(0.0050)		
Rare	0.1118***	0.3725***	0.6469***	0.3295+
	(0.0285)	(0.0990)	(0.1331)	(0.1777)
Rare × *PreClinSignal*			-0.3358***	
			(0.0711)	
Rare × *TrialSignal*			0.0205	
			(0.0159)	
Incidence_Reverse				-0.0940+
				(0.0500)
PreClinSignal × *Incidence_Reverse*				-0.0602***
				(0.0107)
TrialSignal × *Incidence_Reverse*				0.0053*
				(0.0025)
All Controls	Yes	Yes	Yes	Yes
TA Effects	Yes	Yes	Yes	Yes
Firm Effects	Yes	Yes	Yes	Yes
Year Effects	Yes	Yes	Yes	Yes
# Projects	6,651	6,651	6,651	1,127
Pseudo R^2	0.258	0.261	0.262	0.225
Log-likelihood	-3,413	-3,398	-3,393	-327.7

B.5. Censored Projects

Right-censoring may be an issue in estimating selection and success, especially because projects in pharmaceutical pipelines remain in development for many years before they are selected and end up succeeding. As a robustness check, we rerun our models by excluding projects in the last five years of our data set (i.e., projects from 2012 and after). Tables B.VI and B.VII show that our results are robust.

Table B.VII: Heckprobits Considering Censoring

Notes: Heckprobit models excluding the five most recent years of projects in our sample (i.e., from 2012 to 2016). Model I measures signals within the same indication as the focal project (e.g., *PreClinSignal*), and Model II within the same TA (e.g., *PreClinSignalTA*). Reported estimates are average marginal effects. Robust standard errors clustered by drug are in parentheses. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $+p < 0.10$

	(I) Indication-Level Signals		(II) TA-Level Signals	
	Selected	*Success*	*Selected*	*Success*
$PreClinSignal(TA)$	-0.0043***	-0.0010**	-0.0009**	-0.0000
	(0.0003)	(0.0004)	(0.0003)	(0.0002)
$PISignal(TA)$	0.0047***	-0.0025*	0.0001	-0.0001
	(0.0011)	(0.0010)	(0.0003)	(0.0003)
$PIISignal(TA)$	0.0074***	-0.0007	0.0001	0.0001
	(0.0013)	(0.0008)	(0.0002)	(0.0002)
$PIIISignal(TA)$	0.0107**	0.0051**	0.0011+	0.0001
	(0.0033)	(0.0018)	(0.0006)	(0.0004)
$RegSignal(TA)$	0.0115	0.0216*	-0.0009	0.0004
	(0.0252)	(0.0089)	(0.0012)	(0.0007)
$Multiple$	0.3602***	0.2414***	0.3942***	0.2633***
	(0.0299)	(0.0320)	(0.0285)	(0.0348)
$Extension$	0.1619***	0.1368***	0.1760***	0.1500***
	(0.0298)	(0.0248)	(0.0326)	(0.0275)
$Inlicensed$	-0.0652***	0.0288**	-0.0742***	0.0243*
	(0.0178)	(0.0109)	(0.0190)	(0.0111)
$Acquired$	-0.2308***	-0.0018	-0.2413***	-0.0036
	(0.0230)	(0.0212)	(0.0242)	(0.0234)
$Carryover$	0.0292**	0.0119**	0.0697***	0.0215***
	(0.0103)	(0.0037)	(0.0109)	(0.0039)
$stdRD$	-0.0327*	0.0131	-0.0308+	0.0148
	(0.0152)	(0.0110)	(0.0160)	(0.0121)
$IndScope$	-0.1642	0.1694	0.1184	0.2386
	(0.5897)	(0.2521)	(0.6446)	(0.2860)
$IndScale$	-0.0008	0.0030*	-0.0123***	-0.0010
	(0.0014)	(0.0014)	(0.0010)	(0.0009)
$FirmScale$	0.0001	-0.0001	0.0003	-0.0001
	(0.0002)	(0.0001)	(0.0002)	(0.0001)
$Deals$	-0.0009	-0.0024	-0.0046+	-0.0020
	(0.0029)	(0.0024)	(0.0027)	(0.0023)
$Strength$	-0.0008	-0.0003	-0.0011	-0.0003
	(0.0011)	(0.0006)	(0.0012)	(0.0007)
$SelectionRatio$	0.3938***		0.4323***	
	(0.0563)		(0.0603)	
$PriceShock$	1.6971***		1.9405***	
	(0.2069)		(0.2163)	
ρ	-0.630***		-0.519**	
$Wald\ \chi^2(1) : \rho = 0$	13.6		8.6	
TA Indicators	Yes		Yes	
Firm Indicators	Yes		Yes	
Year Indicators	Yes		Yes	
# Projects	5,961	2,843	5,961	2,843
L-(P)L	-3611.2		-3854.8	

Table B.VIII: Outlicensed Projects

Notes: This table reports estimates models with alternative specifications of outlicensing. Robust standard errors clustered by drug are in parentheses. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $+p < 0.10$

	Outlicensed Recoded		Outlicensed Excluded	
	Selected	*Success*	*Selected*	*Success*
PreClinSignal	-0.0042***	-0.0009*	-0.0045***	-0.0010*
	(0.0003)	(0.0004)	(0.0003)	(0.0004)
PISignal	0.0050***	-0.0026**	0.0057***	-0.0024*
	(0.0011)	(0.0010)	(0.0012)	(0.0010)
PIISignal	0.0075***	-0.0010	0.0069***	-0.0010
	(0.0012)	(0.0008)	(0.0012)	(0.0008)
PIIISignal	0.0092**	0.0048**	0.0094**	0.0049**
	(0.0032)	(0.0018)	(0.0032)	(0.0018)
RegSignal	0.0018	0.0228**	0.0103	0.0236**
	(0.0211)	(0.0086)	(0.0216)	(0.0088)
Biologic	0.0573**	0.0146	0.0540**	0.0176+
	(0.0176)	(0.0105)	(0.0179)	(0.0107)
Rare	0.0244	-0.0125	0.0232	-0.0116
	(0.0293)	(0.0167)	(0.0300)	(0.0170)
Multiple	0.3820***	0.2557***	0.3648***	0.2363***
	(0.0284)	(0.0332)	(0.0291)	(0.0308)
Extension	0.1672***	0.1351***	0.1718***	0.1372***
	(0.0281)	(0.0247)	(0.0291)	(0.0243)
Inlicensed	-0.0643***	0.0245*	-0.0636***	0.0226*
	(0.0171)	(0.0106)	(0.0171)	(0.0104)
Acquired	-0.2204***	-0.0034	-0.2232***	-0.0023
	(0.0230)	(0.0203)	(0.0230)	(0.0207)
Carryover	0.0264**	0.0120**	0.0256**	0.0118**
	(0.0096)	(0.0037)	(0.0096)	(0.0038)
stdRD	-0.0256+	0.0085	-0.0258+	0.0111
	(0.0149)	(0.0107)	(0.0151)	(0.0112)
IndScope	0.1917	0.1206	0.2313	0.1183
	(0.5163)	(0.2607)	(0.5238)	(0.2550)
IndScale	-0.0010	0.0028*	-0.0012	0.0031*
	(0.0013)	(0.0014)	(0.0014)	(0.0014)
FirmScale	0.0001	-0.0000	0.0001	-0.0001
	(0.0002)	(0.0001)	(0.0002)	(0.0001)
Deals	0.0020	-0.0028	0.0024	-0.0036
	(0.0027)	(0.0024)	(0.0027)	(0.0024)
Strength	-0.0012	0.0001	-0.0014	-0.0001
	(0.0011)	(0.0006)	(0.0011)	(0.0006)
SelectionRatio	0.3932***		0.3961***	
	(0.0551)		(0.0562)	
PriceShock	1.5131***		1.5807***	
	(0.1970)		(0.2021)	
ρ	-0.625***		-0.574**	
$\chi^2(1)$	11.61		9.80	
TA Indicators	Yes		Yes	
Firm Indicators	Yes		Yes	
Year Indicators	Yes		Yes	
# Projects	6,506	3,040	6,196	2,900
Log-likelihood	-3974.1		-3,747.3	

B.6. Outlicensed Projects

In the main analysis, we defined projects that were outlicensed to another company as $Success = 0$. Here, we perform two robustness checks. First, we recode projects that were outlicensed but eventually achieved regulatory approval under the development of another company as $Success = 1$. Second, we also exclude any outlicensed projects from the analysis. We perform these checks in order to ensure that our results are not being driven due to our definition of what constitutes a success in the context of an outlicensed project. Table B.VIII displays the results from these two models. Our results hold and are robust to these alternative specifications.

B.7. Clustering of Errors

In our main models, we estimate models that allow standard errors to correlate within drugs. It is plausible, however, that an alternative error structure is present, i.e., due to drugs relying on "common science." We test the robustness of our models to alternate levels of clustering standard errors. Tables B.IX and B.XI cluster on Firm-Indication and Indication, and Tables B.X and B.XII cluster on Technology (i.e., biologic or chemical molecule). All our models are robust to these alternative classifications, but we do urge caution interpreting the results in Tables B.X and B.XII given the small number of clusters.

Table B.IX: Probit Models Clustering Errors at Firm-Indication and Indication

Notes: This table reports estimates from probit models of selection with standard errors clustered at the Firm-Indication (Models I–IV) and Indication (Models V–VIII) levels. Reported estimates for Models I and V are average marginal effects, for Models II–IV and VI–VIII are probit coefficients. Robust standard errors are in parentheses.
$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{+}p < 0.10$

	Clustering at Firm-Indication				Clustering at Indication			
	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)
	Base	Biologic	Rare Diseases	Incidence	Base	Biologic	Rare Diseases	Incidence
PreClinSignal	-0.0042***		-0.0146***	-0.8389***	-0.0042***		-0.0146***	-0.8389***
	(0.0003)		(0.0010)	(0.1243)	(0.0005)		(0.0020)	(0.1284)
TrialSignal	0.0053***		0.0181***	0.0843**	0.0053***		0.0181***	0.0843**
	(0.0006)		(0.0022)	(0.0280)	(0.0011)		(0.0040)	(0.0301)
Biologic	0.0902***	0.3684***	0.3112***	0.1164	0.0902***	0.3684***	0.3112***	0.1164
	(0.0133)	(0.0561)	(0.0459)	(0.1019)	(0.0122)	(0.0554)	(0.0440)	(0.0738)
PreClinSignal_Bio		-0.0080*				-0.0080*		
		(0.0035)				(0.0036)		
PreClinSignal_Bio × *Biologic*		0.0014				0.0014		
		(0.0067)				(0.0079)		
TrialSignalBio		0.0135+				0.0135		
		(0.0078)				(0.0144)		
TrialSignal_Bio × *Biologic*		0.0371**				0.0371		
		(0.0135)				(0.0241)		
PreClinSignalNonBio		-0.0168***				-0.0168***		
		(0.0013)				(0.0027)		
PreClinSignal_NonBio × *Biologic*		0.0033				0.0033		
		(0.0026)				(0.0023)		
TrialSignalNonBio		0.0217***				0.0217***		
		(0.0026)				(0.0055)		
TrialSignal_NonBio × *Biologic*		-0.0241***				-0.0241***		
		(0.0054)				(0.0064)		
Rare	0.0886**	0.2975**	0.6593***	0.3508*	0.0886*	0.2975*	0.6593***	0.3508*
	(0.0275)	(0.0964)	(0.1102)	(0.1547)	(0.0396)	(0.1390)	(0.1228)	(0.1639)
Rare × *PreClinSignal*			-0.2956***				-0.2956***	
			(0.0557)				(0.0395)	
Rare × *TrialSignal*			0.0014				0.0014	
			(0.0133)				(0.0135)	
Incidence_Reverse				-0.0553				-0.0553
				(0.0500)				(0.0586)
PreClinSignal × *Incidence_Reverse*				-0.0597***				-0.0597***
				(0.0101)				(0.0100)
TrialSignal × *Incidence_Reverse*				0.0056*				0.0056*
				(0.0023)				(0.0023)
Multiple	0.2735***	1.0374***	1.0580***	0.6953***	0.2735***	1.0374***	1.0580***	0.6953***
	(0.0243)	(0.1212)	(0.1153)	(0.1934)	(0.0233)	(0.1161)	(0.1125)	(0.2024)

Extension	0.1586***	0.5735***	0.5626***	0.1894	0.1586***	0.5735***	0.5626***	0.1894
	(0.0203)	(0.0784)	(0.0781)	(0.2067)	(0.0212)	(0.0820)	(0.0816)	(0.1520)
Inlicensed	-0.0370**	-0.1073*	-0.1223**	0.2217+	-0.0370**	-0.1073*	-0.1223*	0.2217+
	(0.0126)	(0.0443)	(0.0439)	(0.1331)	(0.0137)	(0.0450)	(0.0479)	(0.1135)
Acquired	-0.1717***	-0.5912***	-0.5960***	0.1285	-0.1717***	-0.5912***	-0.5960***	0.1285
	(0.0175)	(0.0631)	(0.0640)	(0.1743)	(0.0203)	(0.0729)	(0.0750)	(0.2156)
Carryover	0.0190*	0.0592+	0.0702*	-0.4027**	0.0190*	0.0592*	0.0702**	-0.4027**
	(0.0090)	(0.0315)	(0.0312)	(0.1385)	(0.0076)	(0.0273)	(0.0253)	(0.1447)
stdRD	-0.0268*	-0.0880+	-0.0904*	-0.1819	-0.0268**	-0.0880**	-0.0904**	-0.1819+
	(0.0130)	(0.0450)	(0.0453)	(0.1602)	(0.0097)	(0.0328)	(0.0338)	(0.1100)
IndScope	-0.0463	-0.4090	-0.2121	8.4397	-0.0463	-0.4090	-0.2121	8.4397
	(0.4737)	(1.5660)	(1.6196)	(9.0987)	(0.4435)	(1.4881)	(1.5195)	(8.6960)
IndScale	-0.0018	-0.0080	-0.0061	-0.0187	-0.0018	-0.0080	-0.0061	-0.0187
	(0.0017)	(0.0059)	(0.0060)	(0.0266)	(0.0017)	(0.0069)	(0.0059)	(0.0185)
FirmScale	0.0002	0.0005	0.0005	0.0033*	0.0002	0.0005	0.0005	0.0033**
	(0.0001)	(0.0005)	(0.0005)	(0.0013)	(0.0001)	(0.0004)	(0.0004)	(0.0012)
Deals	0.0004	-0.0009	0.0016	0.0046	0.0004	-0.0009	0.0016	0.0046
	(0.0019)	(0.0065)	(0.0066)	(0.0313)	(0.0019)	(0.0064)	(0.0066)	(0.0289)
Strength	-0.0017+	-0.0064*	-0.0065*	-0.0009	-0.0017*	-0.0064*	-0.0065*	-0.0009
	(0.0009)	(0.0032)	(0.0032)	(0.0114)	(0.0009)	(0.0030)	(0.0030)	(0.0135)
SelectionRatio	0.3684***	1.2533***	1.2774***	1.4623**	0.3684***	1.2533***	1.2774***	1.4623**
	(0.0483)	(0.1719)	(0.1704)	(0.5444)	(0.0452)	(0.1627)	(0.1593)	(0.5228)
PriceShock	1.4306***	4.9231***	4.9742***	1.8858	1.4306***	4.9231***	4.9742***	1.8858
	(0.1679)	(0.5852)	(0.5820)	(1.6157)	(0.1687)	(0.5886)	(0.5783)	(1.6226)
TA Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# Projects	8,024	8,024	8,024	1,424	8,024	8,024	8,024	1,424
Pseudo R^2	0.257	0.262	0.263	0.215	0.257	0.262	0.263	0.215
Log-likelihood	-4,078	-4,050	-4,047	-407.1	-4,078	-4,050	-4,047	-407.1

Table B.X: Probit Models Clustering Errors on Technology

Notes: This table reports estimates from probit models of selection with standard errors clustered at the Technology (i.e., biologic vs. non-biologic) level. Reported estimates for Model I are average marginal effects, for Models II–IV are probit coefficients. Robust standard errors are in parentheses. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $+p < 0.10$

	Clustering on Technology			
	H1	H2	H3	H3
PreClinSignal	-0.0042***		-0.0146***	-0.8388***
	(0.0004)		(0.0013)	(0.1695)
TrialSignal	0.0053***		0.0181***	0.0843**
	(0.0009)		(0.0033)	(0.0287)
Biologic	0.0901***	0.3683***	0.3112***	0.1164
	(0.0100)	(0.0347)	(0.0422)	(0.0720)
PreClinSignal_Bio		-0.0080***		
		(0.0005)		
PreClinSignal_Bio × *Biologic*		0.0014***		
		(0.0001)		
TrialSignalBio		0.0135***		
		(0.0015)		
TrialSignal_Bio × *Biologic*		0.0371***		
		(0.0015)		
PreClinSignalNonBio		-0.0168***		
		(0.0007)		
PreClinSignal_NonBio × *Biologic*		0.0033***		
		(0.0000)		
TrialSignalNonBio		0.0217***		
		(0.0010)		
TrialSignal_NonBio × *Biologic*		-0.0241***		
		(0.0010)		
Rare	0.0885***	0.2974***	0.6593***	0.3505***
	(0.0045)	(0.0063)	(0.0858)	(0.0437)
Rare × *PreClinSignal*			-0.2957***	
			(0.0224)	
Rare × *TrialSignal*			0.0014	
			(0.0072)	
Incidence_Reverse				-0.0552***
				(0.0100)
Incidence_Reverse × *PreClinSignal*				-0.0597***
				(0.0138)
Incidence_Reverse × *TrialSignal*				0.0056*
				(0.0026)
All Controls	Yes	Yes	Yes	Yes
TA Effects	Yes	Yes	Yes	Yes
Firm Effects	Yes	Yes	Yes	Yes
Year Effects	Yes	Yes	Yes	Yes
# Projects	8,024	8,024	8,024	1,424
Pseudo R^2	0.257	0.262	0.263	0.215
Log-likelihood	-4,078	-4,050	-4,047	-407.1

Table B.XI: Heckprobit Models Clustering Errors at Firm-Indication and Indication

Notes: This table reports estimates from models with standard errors clustered at the Firm-Indication (Columns I and II) and Indication (Columns III and IV) levels. Reported coefficients are average marginal effects. Robust standard errors clustered by drug are in parentheses. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $+p < 0.10$

	Clustering at Firm-Indication		Clustering at Indication	
	(I)	(II)	(III)	(IV)
	Selected	*Success*	*Selected*	*Success*
PreClinSignal	-0.0042***	-0.0009**	-0.0042***	-0.0009**
	(0.0004)	(0.0003)	(0.0007)	(0.0003)
PISignal	0.0050***	-0.0025**	0.0050*	-0.0025*
	(0.0012)	(0.0009)	(0.0021)	(0.0010)
PIISignal	0.0075***	-0.0010	0.0075***	-0.0010
	(0.0014)	(0.0007)	(0.0020)	(0.0007)
PIIISignal	0.0092**	0.0047**	0.0092*	0.0047**
	(0.0033)	(0.0016)	(0.0039)	(0.0016)
RegSignal	0.0016	0.0232**	0.0016	0.0232***
	(0.0211)	(0.0079)	(0.0184)	(0.0059)
Biologic	0.0575***	0.0160+	0.0575***	0.0160+
	(0.0158)	(0.0097)	(0.0161)	(0.0097)
Rare	0.0243	-0.0120	0.0243	-0.0120
	(0.0331)	(0.0119)	(0.0417)	(0.0172)
Multiple	0.3819***	0.2487***	0.3819***	0.2487***
	(0.0249)	(0.0252)	(0.0249)	(0.0254)
Extension	0.1658***	0.1275***	0.1658***	0.1275***
	(0.0248)	(0.0227)	(0.0259)	(0.0233)
Inlicensed	-0.0641***	0.0247*	-0.0641***	0.0247*
	(0.0142)	(0.0102)	(0.0152)	(0.0110)
Acquired	-0.2204***	-0.0012	-0.2204***	-0.0012
	(0.0179)	(0.0149)	(0.0195)	(0.0138)
Carryover	0.0266*	0.0119**	0.0266**	0.0119*
	(0.0108)	(0.0042)	(0.0086)	(0.0053)
stdRD	-0.0256+	0.0113	-0.0256*	0.0113
	(0.0138)	(0.0105)	(0.0105)	(0.0116)
IndScope	0.1910	0.0951	0.1910	0.0951
	(0.4672)	(0.2032)	(0.4637)	(0.1873)
IndScale	-0.0010	0.0031*	-0.0010	0.0031*
	(0.0016)	(0.0013)	(0.0016)	(0.0014)
FirmScale	0.0001	-0.0001	0.0001	-0.0001
	(0.0002)	(0.0001)	(0.0002)	(0.0001)
Deals	0.0020	-0.0032	0.0020	-0.0032
	(0.0024)	(0.0023)	(0.0034)	(0.0021)
Strength	-0.0012	-0.0000	-0.0012	-0.0000
	(0.0010)	(0.0006)	(0.0010)	(0.0006)
SelectionRatio	0.3928***		0.3928***	
	(0.0523)		(0.0501)	
PriceShock	1.5226***		1.5226***	
	(0.1844)		(0.1664)	
ρ	-0.593***		-0.593***	
$Wald\chi^2(1)\rho = 0$	13.40		14.25	
TA Indicators	Yes		Yes	
Firm Indicators	Yes		Yes	
Year Indicators	Yes		Yes	
# Projects	6,506	3,040	6,506	3,040
Log-(pseudo)likelihood	-3,963		-3,963	

Table B.XII: Heckprobit Models Clustering Errors on Technology

Notes: This table reports estimates from models with standard errors clustered at the Technology (i.e., biologic vs. non-biologic) level. Reported coefficients are average marginal effects. Robust standard errors clustered by drug are in parentheses. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $+p < 0.10$

	Clustering on Technology	
	(I)	(II)
	Selection	*Success*
PreClinSignal	-0.0042***	-0.0009*
	(0.0004)	(0.0004)
PISignal	0.0050***	-0.0025***
	(0.0015)	(0.0007)
PIISignal	0.0075***	-0.0010
	(0.0004)	(0.0006)
PIIISignal	0.0092***	0.0047***
	(0.0027)	(0.0004)
RegSignal	0.0016	0.0232***
	(0.0105)	(0.0047)
Biologic	0.0575***	0.0160*
	(0.0033)	(0.0070)
Rare	0.0243+	-0.0120
	(0.0133)	(0.0248)
Multiple	0.3819***	0.2487***
	(0.0163)	(0.0727)
Extension	0.1658***	0.1275***
	(0.0054)	(0.0187)
Inlicensed	-0.0641	0.0247*
	(0.0600)	(0.0100)
Acquired	-0.2204***	-0.0012
	(0.0038)	(0.0073)
Carryover	0.0266***	0.0119**
	(0.0067)	(0.0042)
stdRD	-0.0256*	0.0113**
	(0.0103)	(0.0034)
IndScope	0.1910**	0.0951
	(0.0716)	(0.1723)
IndScale	-0.0010**	0.0031***
	(0.0003)	(0.0005)
FirmScale	0.0001***	-0.0001***
	(0.0000)	(0.0000)
Deals	0.0020	-0.0032***
	(0.0032)	(0.0003)
Strength	-0.0012**	-0.0000
	(0.0004)	(0.0004)
SelectionRatio	0.3928***	
	(0.0701)	
PriceShock	1.5226***	
	(0.1267)	
ρ	-0.593+	
$Wald\chi^2(1)\rho = 0$	3.49	
TA Indicators	Yes	
Firm Indicators	Yes	
Year Indicators	Yes	
# Projects	6,506	3,040
Log-(pseudo)likelihood	-3,963	

B.8. Endogeneity from Breakthrough Technologies

In many industries, the emergence of new, breakthrough technologies may render older technological solutions obsolete. Such technological breakthroughs could conceivably lead to a simultaneous increase in new rival projects. In our context, this would mean that pre-clinical projects would be endogenous to the emergence of a technology. Similarly, a technology breakthrough could create a number of drug-indication projects, leading a focal project to merely be part of a "cohort" that moves into clinical trials together. Such a phenomenon would raise endogeneity concerns in our current study.

However, although there may indeed be industries whereby a new, superior technology renders an older one obsolete, such dynamics do not accurately characterize the pharmaceutical industry. For example, conventional therapies based on chemical formulations (which have been around for hundreds of years) are an important and growing technology in the industry (see, e.g., Berggren et al. 2018, pg. 5, Exh. 4). They are nowhere near to being supplanted by newer biologic technologies such as monoclonal antibodies and gene therapy. Different technologies are more effective for different indications, and even within the same indication there are multiple technologies that are more or less effective for different groups of patients depending on their medical history, genetic information, etc.

Moreover, even if such a superior technology existed, it is incredibly difficult to evaluate its feasibility and potential. Consider the case of gene therapy, arguably one of the "hottest" technologies in the industry at the moment: in 2014-15, M&A deal values were $5bn, and in 2018-19, this surged 980% to a combined two-year total of $49bn.[3] Interestingly, gene therapy as a technology is far from new. In fact, as noted in another article: "In September 1999, gene therapy looked to be on the cusp of a breakthrough in medicine. By the end of 2000, it seemed like a cautionary tale of scientific overreach."[4] Another example is m-RNA technology- originally developed in the 1990s. As a recent article notes "before messenger RNA was a multibillion-dollar idea, it was a scientific backwater. And for the Hungarian-born scientist behind a key mRNA discovery, it was a career dead-end."[5]

For the purposes of our study, the key point from those stories is that before reliable clinical signals are generated (the success of CAR-T therapy by Novartis in the case of gene therapy and the Covid-19 vaccines by BioNTech/Pfizer and Moderna in the case of m-RNA), the feasibility of such technologies remains unresolved. As such, we argue it is specifically through learning by rivals' actions that the potential of those new technologies is resolved. In the absence of such a reliable learning mechanism, we are not aware of other factors that would lead firms to exhibit such correlated behavior and select projects merely as a "cohort" of projects that implement a certain technology.

The above conceptual arguments notwithstanding, we also provide below empirical reasons and robustness tests that, taken together, lead us to believe that phenomenona such as the ones described above are not driving our results.

First, in all our models, we aim to control for herding or "faddish" behavior (variable: *Deals*): for example, moving into a technology or market simply because that domain is "hot." In all our models, we find that the coefficient on this term is statistically zero. This suggests that, at least in the search for novel drugs within an indication, selection decisions are not driven by some kind of herding.

Second, in line with the idea that selection decisions may be correlated due to, e.g., common science/technology, we assess the robustness of our results by varying the level at which we cluster model standard errors (see Appendix B.7.). Regardless of the level of clustering, we find that all our results hold.

[3] `https://www.pmlive.com/pharma_intelligence/After_years_of_potential,_cell_and_gene_therapy_is_ready_for_the_pharmaceutical_mainstream_1360242`, accessed March 30, 2021.

[4] `https://www.wired.com/2013/08/the-fall-and-rise-of-gene-therapy-2/`, accessed March 30, 2021.

[5] `https://www.statnews.com/2020/11/10/the-story-of-mrna-how-a-once-dismissed-idea-became-a-leading-technology-in-the-covid-vaccine-race/`, accessed March 20, 2021.

Third, we also looked to try to further control for the "technological quality" of a drug. If we could explicitly control for a "superior" technology, this would eliminate concerns about endogeneity. For this, we went back to the Cortellis database and looked for drugs that were designated as a "Breakthrough Therapy" (BT). The FDA assigns BT status to a drug if it "demonstrate[s] substantial improvement over existing therapies."[6]

Admittedly, this approach is imperfect: drugs with a BT designation do not necessarily use a superior technology. However, it is likely that firms will apply for BT designation if their drug indeed were to use a superior technology. The point is this: BT designation should be strongly correlated with and proxy for the existence of a superior technology. As such, although we present the following analysis merely as a robustness check, it should nonetheless be interpreted with caution.

We proceed by creating the binary variable $Breakthrough$, which denotes whether a drug obtained the FDA designation of "Breakthrough Therapy." On the one hand, if a project utilizes an inferior technology that leads to non-selection, then inclusion of $Breakthrough$ in our models should weaken $PreClinSignal$'s effect size and significance. On the other hand, if there is some kind of herding activity, then $Breakthrough$ should weaken the coefficient and significance on $TrialSignal$.

Table B.XIII Models I—IV show that, even after controlling for $Breakthrough$, all our results hold. Moreover, the effect sizes on $PreClinSignal$ and $TrialSignal$ in our base specification (Model I) are statistically identical to those from our main model (Table 3, Model II). Taken together, our results do not seem to be driven by technological considerations.

Fourth, as a final robustness check, in addition to $Year$- and TA-level effects, we also include a set of $Year \times TA$ effects. By including a TA-specific time trend in such a manner, we can again get closer to proxying for the emergence of a superior technology and its specific effect for a TA. As above, this is not a perfect approach; however, robustness of our results to this further specification would help to strengthen the evidence against an alternative, superior technology explanation. Indeed, in Table B.XIII, Models V—VIII, we find that our results are robust. All results retain statistical significance in the hypothesized direction.

[6] See `https://www.govinfo.gov/content/pkg/BILLS-112s3187enr/pdf/BILLS-112s3187enr.pdf`, Section 902, accessed March 30, 2021.

Table B.XIII: Robustness Checks around Superior Technologies

Notes: Models I—IV control for Breakthrough Therapy designation, and models V—VIII include firm × year indicator effects. Robust standard errors clustered by drug are in parentheses. ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$; +$p < 0.10$

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)
	Controlling for Breakthrough Therapy Designation				Controlling for Year × TA			
	H1	H2	H3	H3	H1	H2	H3	H3
PreClinSignal	-0.0041***		-0.0144***	-0.8427***	-0.0042***		-0.0149***	-0.8388***
	(0.0002)		(0.0009)	(0.1201)	(0.0002)		(0.0009)	(0.1180)
TrialSignal	0.0053***		0.0185***	0.0797**	0.0056***		0.0192***	0.0843**
	(0.0005)		(0.0019)	(0.0278)	(0.0006)		(0.0021)	(0.0280)
Biologic	0.0849***	0.3466***	0.2994***	0.0981	0.0924***	0.3912***	0.3256***	0.1164
	(0.0149)	(0.0636)	(0.0521)	(0.1338)	(0.0152)	(0.0667)	(0.0535)	(0.1348)
PreClinSignal_Bio		-0.0071*				-0.0060+		
		(0.0032)				(0.0035)		
PreClinSignal_Bio × *Biologic*		0.0026				0.0061		
		(0.0073)				(0.0074)		
TrialSignal_Bio		0.0143*				0.0185**		
		(0.0058)				(0.0059)		
TrialSignal_Bio × *Biologic*		0.0365**				0.0348**		
		(0.0115)				(0.0118)		
PreClinSignal_NonBio		-0.0168***				-0.0179***		
		(0.0012)				(0.0013)		
PreClinSignal_NonBio × *Biologic*		0.0032				0.0019		
		(0.0026)				(0.0026)		
TrialSignal_NonBio		0.0217***				0.0223***		
		(0.0022)				(0.0023)		
TrialSignal_NonBio × *Biologic*		-0.0239***				-0.0248***		
		(0.0045)				(0.0046)		
Rare	0.0800***	0.2712**	0.6544***	0.3256*	0.0925***	0.3171***	0.6615***	0.3505*
	(0.0241)	(0.0856)	(0.1161)	(0.1633)	(0.0240)	(0.0856)	(0.1143)	(0.1549)
PreClinSignal × *Rare*			-0.3061***				-0.2967***	
			(0.0500)				(0.0509)	
TrialSignal × *Rare*			-0.0015				0.0052	
			(0.0125)				(0.0124)	
Incidence_Reverse				-0.0643				-0.0552
				(0.0453)				(0.0442)
PreClinSignal × *Incidence_Rev*				-0.0597***				-0.0597***
				(0.0098)				(0.0097)
TrialSignal × *Incidence_Rev*				0.0053*				0.0056*
				(0.0022)				(0.0023)
Breakthrough	0.3507***	1.5503***	1.5680***	1.3520***				
	(0.0306)	(0.2329)	(0.2440)	(0.3928)				
All Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm Indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

TA Indicators	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Year Indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
TA × Year Indicators	No	No	No	No	Yes	Yes	Yes	Yes
# Projects	8,024	8,024	8,024	1,424	7,946	7,946	7,946	1,424
Pseudo R^2	0.268	0.273	0.274	0.245	0.269	0.275	0.275	0.215
Log-likelihood	-4018	-3989	-3986	-391.5	-3978	-3945	-3948	-407.1

Appendix C: Cortellis Database

C.1. Main Data Set and Sample Creation

A key step in understanding the drivers of project selection is to capture the timely development pipeline for each firm. The Cortellis database is substantially fine-grained so as to denote how a project moves from pre-clinical to clinical trials, and through regulatory approval. We restrict the data set and our analysis to the time period from 1999 through 2016 for two reasons. First, prior to 1999, data quality is poor and there is only very limited information on drug development activities for the industry. Second, the industry is constantly undergoing transformational M&As and in-licensing deals. The next several years from 1999 onward saw Pfizer purchasing Warner-Lambert in 1999, Astra and Zeneca merging in 1999, the merger between Glaxo Wellcome and SmithKline Beecham in 2000, and Sanofi acquiring Aventis in 2004. Therefore, we make sure to capture these years. Although this makes manually tracking drugs cumbersome, not taking these into account could potentially lead to questionable results.

In tracking drug projects as they move into and out of the portfolio and through the pipeline, we require knowledge of where the project originated, whether and how it progressed from pre-clinical to clinical trials, and whether it was successfully developed or not. The Cortellis database contains detailed information for thousands of companies and institutions involved in the discovery and development of drugs. Cortellis is unique in that it is a near-exhaustive compilation of drug discovery and development activities in the industry. Analysts compile the data set using information from quarterly and annual reports, clinical trial registrations, scientific journal publications, patents, and even conference presentations. Although this is a significant advantage of the database, it also makes the preparation and exploitation of the data especially cumbersome. Because the data preparation outlined below was manually performed to ensure accuracy, for reasons of feasibility, the sample of firms considered in this paper was restricted to the top 15 pharmaceuticals by sales according to Palmer (2017).

For the current purpose of preparing the data set, the main variables in the Cortellis Drugs data set are the drug name, indication, development status, date, and country. Each observation (row) in Cortellis Drugs is a unique combination of these five variables. Our goal with the data preparation task was to establish the history of each drug-indication project. As mentioned in Section 4, we denote the starting point of our data set in 1999 and consider only those drug development projects that entered pre-clinical trials during and after 1999.

We then identify the first time a drug-indication project entered a given phase. After identifying this point, we remove all subsequent, similar drug-indication-phase observations. In Cortellis, such drug-indication-phase duplications occur because, for example, a drug-indication project can enter Phase I in the US in 2005 and then also enter Phase I in Europe in 2007. Because we are not concerned with where the project entered a phase—only that it did so—such duplicates are removed. Thus, we end up with each observation being a unique combination of drug, indication, and phase.

The next step is to obtain a list of all the merger and acquisitions (M&As) and subsidiaries of each pharmaceutical, the names of which are available within Cortellis. Compiling this list gives us the ability to track those drugs and projects that originated outside the focal pharmaceutical but were subsequently brought into the pipeline through an M&A. After obtaining this list, we triangulated its accuracy using various sources such as annual reports, public relations reports, industry news clippings, Wikipedia, and other online publications. We then obtained the exact dates of the M&As from similar sources.

Using the list of M&As and subsidiaries mentioned above, we then capture whether a drug originated within one of the 15 pharmaceuticals (in-house) or was externally acquired through an M&A or a licensing deal. This is captured at the drug level. Therefore, if a pharmaceutical obtained a drug from outside and then subsequently began researching a new indication, we consider this project as originating outside the pharmaceutical. Thus,

any subsequent drug developments post-M&A or post-inlicensing are considered as derivatives from the external purchase. Because we are concerned with projects from the moment they enter the focal pharmaceuticals' pipelines and after, we do not consider the history of externally acquired projects prior to their acquisition. Moreover, for drug indications that were developed and launched by one company and then in-licensed and marketed by a pharmaceutical, such projects are not included in the analysis since we are interested in drug discovery and development projects only.

All Phase II trials and beyond must be registered with the FDA (U.S. Food and Drug Administration 2007), and we have near-complete data from Phase II onward. Yet, although the Cortellis database is quite thorough and generally comprehensive, the dates and phases are missing for some projects; for some observations, there is no date that the project entered into the pre-clinical phase. In such cases, where possible, we manually recreate the history of a project in a "path-by-path" approach similar to Wong et al. (2019). When there are multiple indications for a single drug and where an indication is missing phase information, we impute dates for the indication with missing information by copying the date on which another indication in the same TA for the same drug entered into the phase. Where it is not possible to impute dates from other drug-indication projects, we leave the project history as is. Finally, when a project is missing information on the pre-clinical phase, we set the date of pre-clinical trials as January 1 of the same year as the next available time point.

We focus on the top 15 pharmaceutical firms by sales in 2016. These are: AbbVie, Amgen, AstraZeneca, Bayer, Bristol-Myers Squibb, Eli Lilly, Gilead Sciences, GlaxoSmithKline, Janssen (Johnson & Johnson's pharmaceutical division), Merck & Co. Inc., Novartis, Pfizer, Roche, Sanofi, and Teva. Together, these firms accounted for more than $524 billion in revenues in 2016. We focus on these firms for two reasons. A first reason is the feasibility of tracking projects as they historically move across companies. We do so by manually compiling all the M&As a company underwent, and then we individually assess each project's origins (i.e., in-house, in-licensed, acquired) to ensure completeness and accuracy. Second, focusing on the top 15 firms allows us to examine a very similar group of firms in terms of capabilities. Beyond "Big Pharma" are "mid-size Pharma" and smaller biotech companies that function differently in terms of organizational structure, portfolio focus, and R&D budgets, to name a few. We restrict our sample to these 15 to ensure that we draw conclusions from as homogeneous a sample as possible.

Finally, we briefly comment on the accuracy of our *raw* data set by comparing it to general trends and specific statistics from other publications. First, one of the key measures of productivity in the industry is the ability to progress a project through the various phases of development. The more projects that a firm is able to successfully push through different phases of clinical testing, the greater the likelihood of achieving regulatory approval. We compare the rates from our data to data from the Pharmaceutical Benchmarking Forum (KMR Group 2011), which analyzes a comparable sample of 12 of the then-largest pharmaceuticals in the industry (Abbott, AstraZeneca, Bayer, Boehringer Ingelheim, Bristol-Myers Squibb, Eli Lilly, Johnson & Johnson, Merck Research Labs, Novartis, Novo Nordisk, Pfizer, Roche, and Sanofi). In our raw data set, 58.4% of projects (vs. 63% in KMR Group (2011)) in the pre-clinical phase progress to Phase I trials, 44.8% (vs. 47%) in Phase I progress to Phase II, 21.3% (vs. 23%) in Phase II progress to Phase III, 53.0% (vs. 59%) in Phase III progress to pre-registration, and 83.1% (vs. 79%) in pre-registration eventually are launched.

Second, the distribution of projects with respect to TA seems to be consistent with recent publications. Using the Pharmaprojects database, Lloyd (2017, pg. 11) examines the distribution of R&D projects across the industry. Although assessing a much broader sample from ours, he finds that anticancer projects make up approximately 32.6% of firms' pipelines; in our raw data, approximately 33.4% of projects at any given time target cancer. Moreover, the report suggests that the second- and third-largest categories of drugs (neurological and anti-infectives) make up approximately 17.1% and 16.6% of pipelines, respectively.[7] In our raw data, projects

[7] Although these percentages for neurologicals and anti-infectives are not explicitly stated in Lloyd (2017), we calculate them as follows (see also pg. 11 and Figure 7 of that report). First, since the author explicitly mentions the proportion of anticancer drugs

in the central nervous system TA make up 16.2% of pipelines, and infectious diseases make up 11.7% of pipelines. Finally, examining the drug technology, Lloyd (2017, pg. 16) suggests that approximately 35% of drugs in the industry in 2016 were biological; in our raw data, this figure is 38.9%.

C.2. Cortellis Deals

To create our *Deals* control in Section A.1., we subsample the Cortellis database and focus only on data in the Cortellis universe concerning in-licensing deals. Similar to the methodology of the Cortellis Drugs data set, we begin by removing all deals which were entered into prior to 1999. We then constrain the sample to those deals entered into by the top 15 that were concerned with drug development and not patents, technologies, manufacturing, etc. These are denoted in Cortellis as "Drug - Asset Divestment," "Drug - Development/Commercialization License," and "Drug - Early Research/Development." Once more, we use the list of pharmaceuticals and subsidiaries to consolidate in-licensing companies (licensees) at the level of the parent organization.

as 32.6%, we divide the number of anticancer drugs (4,845) by this proportion to estimate a total of 14,862 drugs in the author's sample. We then divide the number of neurologicals (2,542) and anti-infectives (2,468) by the sample size to calculate the proportions of each. Also, although the author reports "biotechnology" in the same figure, it is unclear what this refers to. We assume that this references the technology of the drugs, and not a TA as per our definition.

Appendix D: Mapping of Indications to TAs

Table D.I provides a mapping of indication labels in Cortellis to TAs used as controls in our models.

Table D.I: Mapping of Indications to TAs

Notes: This table maps indications in Cortellis to TAs. CNS = Central nervous system/neurology/psychiatry, CVS = Cardiovascular, ENDO = Endocrinology/metabolic, ENT/OPH = Otolaryngology/ophthalmology, GI = Gastrointestinal/hepatology, GU = Nephrology/genitourinary, ID = Infectious diseases, IM/D = Immunology/rheumatology/dermatology, ONC = Oncology, RESP = Respiratory, and OTHER = Other.

Indications	Th. Area	Indications	Th. Area
Acne	IM/D	Acne vulgaris	IM/D
Acquired immune deficiency syndrome	ID	Acromegaly	ENDO
Acute coronary syndrome	CVS	Acute leukemia	ONC
Acute lung injury	RESP	Acute lymphoblastic leukemia	ONC
Acute myelogenous leukemia	ONC	Acute promyelocytic leukemia	ONC
Adenosine deaminase deficiency	ENDO	Adenovirus infection	ID
Adult varicella zoster virus infection	ID	Advanced solid tumor	ONC
Age related macular degeneration	ENT/OPH	Aging	OTHER
Alcoholism	CNS	Allergic conjunctivitis	ENT/OPH
Allergic rhinitis	ENT/OPH	Allergy	IM/D
Alopecia	IM/D	Alpha-1 antitrypsin deficiency	ID
Alzheimers disease	CNS	Amnesia	CNS
Amphetamine dependence	CNS	Amyloidosis	OTHER
Anaplastic thyroid cancer	ONC	Anemia	OTHER
Anesthesia	CNS	Angina	CVS
Angiogenesis disorder	CVS	Ankylosing spondylitis	IM/D
Anorexia nervosa	CNS	Anxiety disorder	CNS
Aplastic anemia	OTHER	Appetite loss	OTHER
Arenavirus infection	ID	Arteriosclerosis	CVS
Arthralgia	IM/D	Arthritis	IM/D
Aspergillus infection	ID	Asthma	RESP
Atherosclerosis	CVS	Atopic dermatitis	IM/D
Atrial fibrillation	CVS	Attention deficit hyperactivity disorder	CNS
Autism	CNS	Autoimmune disease	IM/D
B-cell lymphoma	ONC	Bacillus anthracis infection	ID
Bacterial infection	ID	Bacterial meningitis	ID
Bacterial pneumonia	ID	Bacterial respiratory tract infection	ID
Bacterial skin infection	ID	Bacterial urinary tract infection	ID
Basal cell carcinoma	ONC	Beta thalassemia	OTHER
Biliary cancer	ONC	Binge eating disorder	CNS
Bipolar disorder	CNS	Bladder cancer	ONC
Bladder tumor	ONC	Bleeding	OTHER
Blindness	ENT/OPH	Blood clotting disorder	OTHER
Blood transfusion	OTHER	Bone disease	OTHER
Bone injury	OTHER	Bone marrow disease	OTHER
Bone marrow transplantation	OTHER	Bone metastases	ONC
Bone tumor	ONC	Bordetella pertussis infection	ID
Brain disease	CNS	Brain hemorrhage	CNS
Brain injury	CNS	Brain tumor	ONC
Breast tumor	ONC	Bronchiectasis	RESP
Bronchitis	RESP	Burkholderia infection	ID
CMV retinitis	ID	Cachexia	OTHER
Cancer	ONC	Cancer pain	ONC
Candida albicans infection	ID	Candida infection	ID
Carcinoma	ONC	Cardiac failure	CVS
Cardiovascular disease	CVS	Cartilage disease	OTHER
Cataract	ENT/OPH	Celiac disease	GI
Central nervous system disease	CNS	Cerebral infarction	CNS
Cerebrovascular disease	CNS	Cervical dysplasia	GU
Chemotherapy induced nausea and vomiting	ONC	Chemotherapy-induced emesis	ONC

Chikungunya virus infection	ID	Chlamydia infection	ID
Chlamydia trachomatis infection	ID	Cholangiocarcinoma	ONC
Chronic bronchitis	RESP	Chronic fatigue syndrome	CNS
Chronic lymphocytic leukemia	ONC	Chronic myelocytic leukemia	ONC
Chronic obstructive pulmonary disease	RESP	Clostridium botulinum infection	ID
Clostridium difficile infection	ID	Clostridium tetani infection	ID
Cocaine addiction	CNS	Cognitive disorder	CNS
Colon tumor	ONC	Colorectal tumor	ONC
Common cold	ID	Condyloma	ONC
Congestive heart failure	CVS	Conjunctivitis	ENT/OPH
Connective tissue disease	OTHER	Constipation	OTHER
Contraception	GU	Corneal ulcer	ENT/OPH
Coronary artery disease	CVS	Coronavirus infection	ID
Corynebacterium diphtheriae infection	ID	Cough	RESP
Crohns disease	GI	Cryptococcus neoformans meningitis	ID
Cutaneous T-cell lymphoma	ONC	Cystic fibrosis	RESP
Cystitis	GU	Cytomegalovirus infection	ID
Deep vein thrombosis	CVS	Dementia	CNS
Dengue virus infection	ID	Depression	CNS
Dermatitis	IM/D	Dermatological disease	IM/D
Diabetes mellitus	ENDO	Diabetic complication	ENDO
Diabetic foot ulcer	ENDO	Diabetic macular edema	ENDO
Diabetic nephropathy	ENDO	Diabetic neuropathy	ENDO
Diabetic retinopathy	ENDO	Diarrhea	OTHER
Diffuse large B-cell lymphoma	ID	Down syndrome	CNS
Drug dependence	CNS	Dry age related macular degeneration	ENT/OPH
Duchenne dystrophy	CNS	Duodenal ulcer	ID
Dysmenorrhea	GU	Dyspepsia	GI
Dystonia	CNS	Eating disorder	CNS
Ebola virus infection	ID	Eczema	IM/D
Edema	OTHER	Elephantiasis	ID
Emesis	OTHER	Emphysema	RESP
End stage renal disease	GU	Endocrine disease	ENDO
Endometrioid carcinoma	ONC	Endometriosis	GU
Epidermolysis bullosa	IM/D	Epilepsy	CNS
Epstein Barr virus infection	ID	Erectile dysfunction	GU
Escherichia coli infection	ID	Esophageal disease	GI
Esophagitis	GI	Esophagus tumor	ONC
Estrogen deficiency	ENDO	Ewing sarcoma	ONC
Extrapyramidal syndrome	CNS	Fabry disease	ENDO
Factor IX deficiency	OTHER	Factor VIII deficiency	OTHER
Fallopian tube cancer	ONC	Familial hypercholesterolemia	ENDO
Fatigue	CNS	Fatty liver disease	GI
Female contraception	GU	Female genital tract infection	ID
Female infertility	GU	Female sexual dysfunction	GU
Fever	OTHER	Fibromyalgia	IM/D
Fibrosis	OTHER	Filovirus infection	ID
Flavivirus infection	ID	Follicle center lymphoma	ONC
Fragile X syndrome	CNS	Francisella tularensis infection	ID
Friedreich ataxia	CNS	Fungal infection	ID
Gastric motility disorder	GI	Gastritis	GI
Gastroesophageal reflux	GI	Gastrointestinal disease	GI
Gastrointestinal function disorder	GI	Gastrointestinal tumor	ONC
Gastroparesis	GI	Gaucher disease	ENDO
Generalized anxiety disorder	CNS	Genetic disorder	OTHER
Genitourinary disease	GU	Glaucoma	ENT/OPH
Glioblastoma	ONC	Glioma	ONC
Glomerulonephritis	GU	Gout	IM/D
Graft versus host disease	OTHER	Gram negative bacterium infection	ID
Gram positive bacterium infection	ID	Growth disorder	ENDO
Growth hormone deficiency	ENDO	Gynecological disorder	GU
HIV associated dementia	ID	HIV infection	ID
HIV-1 infection	ID	HSV-1 infection	ID
HSV-2 infection	ID	Haemophilus influenzae infection	ID
Hairy cell leukemia	ONC	Head and neck tumor	ONC
Head injury	CNS	Headache	CNS
Hearing disorder	ENT/OPH	Hearing loss	ENT/OPH

Heart arrhythmia	CVS	Heart disease	CVS
Helicobacter pylori infection	ID	Hematological disease	OTHER
Hematological neoplasm	ONC	Hemolytic uremic syndrome	OTHER
Hemophilia	OTHER	Hemorrhagic shock	OTHER
Hemorrhoids	GI	Hendra virus infection	ID
Heparin induced thrombocytopenia	OTHER	Hepatic encephalopathy	CNS
Hepatitis	GI	Hepatitis A virus infection	ID
Hepatitis B virus infection	ID	Hepatitis C virus infection	ID
Hepatitis virus infection	ID	Hepatocellular carcinoma	ONC
Hereditary angioedema	CVS	Herpes simplex virus infection	ID
Herpesvirus infection	ID	Hodgkins disease	ONC
Hormone deficiency	ENDO	Hormone dependent prostate cancer	ONC
Hormone refractory prostate cancer	ONC	Hot flashes	OTHER
Hunter syndrome	ENDO	Huntingtons chorea	CNS
Hypercalcemia	ENDO	Hypercholesterolemia	ENDO
Hyperlipidemia	ENDO	Hyperoxaluria	ENDO
Hyperparathyroidism	ENDO	Hyperphosphatemia	ENDO
Hypertension	CVS	Hypertriglyceridemia	ENDO
Hyperuricemia	ENDO	Hypoglycemia	ENDO
Hypogonadism	ENDO	Idiopathic pulmonary fibrosis	RESP
Immediate type hypersensitivity	IM/D	Immune deficiency	ID
Immune disorder	IM/D	Infectious disease	ID
Infertility	GU	Inflammatory bowel disease	GI
Inflammatory disease	IM/D	Influenza virus A infection	ID
Influenza virus infection	ID	Injury	OTHER
Insomnia	CNS	Insulin dependent diabetes	ENDO
Intermittent claudication	CVS	Intoxication	CNS
Iron deficiency anemia	OTHER	Iron overload	OTHER
Irritable bowel syndrome	GI	Ischemia	OTHER
Ischemic heart disease	CVS	Ischemic stroke	CNS
Islet cell transplant rejection	OTHER	Japanese encephalitis virus infection	ID
Kaposis sarcoma	ONC	Keratoconjunctivitis	ENT/OPH
Keratosis	IM/D	Kidney transplant rejection	OTHER
Lacrimal gland disease	ENT/OPH	Leishmania donovani infection	ID
Leishmania infection	ID	Leishmania tropica infection	ID
Leukemia	ONC	Leukopenia	OTHER
Leukopenia drug induced	OTHER	Lewy body dementia	CNS
Lipid metabolism disorder	ENDO	Lipoprotein lipase deficiency	ENDO
Liver cirrhosis	GI	Liver disease	GI
Liver fibrosis	GI	Liver tumor	ONC
Lung disease	RESP	Lung embolism	RESP
Lung infection	ID	Lung inflammation	RESP
Lung injury	RESP	Lung malformation	ONC
Lung tumor	ONC	Lupus nephritis	IM/D
Lyme disease	ID	Lymphoma	ONC
Lysosome storage disease	ENDO	MRSA infection	ID
Macroglobulinemia	ENDO	Macular degeneration	ENT/OPH
Major depressive disorder	CNS	Male contraception	GU
Male sexual dysfunction	GU	Mantle cell lymphoma	ONC
Marburg virus infection	ID	Measles virus infection	ID
Medullary thyroid cancer	ONC	Melanoma	ONC
Melioidosis	ID	Menopause	ENDO
Mesothelioma	ONC	Metabolic disorder	ENDO
Metabolic syndrome X	ENDO	Metastasis	ONC
Metastatic brain cancer	ONC	Metastatic breast cancer	ONC
Metastatic colon cancer	ONC	Metastatic colorectal cancer	ONC
Metastatic esophageal cancer	ONC	Metastatic head and neck cancer	ONC
Metastatic liver cancer	ONC	Metastatic lung cancer	ONC
Metastatic non small cell lung cancer	ONC	Metastatic ovary cancer	ONC
Metastatic pancreas cancer	ONC	Metastatic prostate cancer	ONC
Metastatic rectal cancer	ONC	Metastatic renal cancer	ONC
Metastatic stomach cancer	ONC	Migraine	CNS
Mild cognitive impairment	CNS	Mitochondrial disease	OTHER
Mood disorder	CNS	Motor neurone disease	CNS
Movement disorder	CNS	Mucopolysaccharidosis type I	ENDO
Mucositis	ONC	Multidrug resistant infection	ID
Multiple myeloma	ONC	Multiple sclerosis	CNS

Mumps virus infection	ID	Muscle disease	OTHER
Muscle hypertonia	CNS	Muscle spasm	CNS
Muscle wasting disease	CNS	Muscular dystrophy	CNS
Musculoskeletal disease	OTHER	Myasthenia gravis	CNS
Mycobacterium infection	ID	Mycobacterium leprae infection	ID
Mycobacterium tuberculosis infection	ID	Myelodysplastic syndrome	ONC
Myeloid leukemia	ONC	Myeloproliferative disorder	ONC
Myocardial disease	CVS	Myocardial infarction	CVS
Myopathy	CNS	Myotonic dystrophy type 1	CNS
Narcolepsy	CNS	Nasopharyngeal carcinoma	ONC
Nausea	OTHER	Neisseria gonorrhoeae infection	ID
Neisseria meningitidis infection	ID	Neisseria meningitidis meningitis	ID
Neonatal respiratory distress syndrome	RESP	Neoplasm	ONC
Nephritis	GU	Neuroblastoma	ONC
Neurodegenerative disease	CNS	Neuroendocrine tumor	ONC
Neurological disease	CNS	Neuromuscular disease	CNS
Neuropathic pain	CNS	Neuropathy	CNS
Neutropenia	OTHER	Nicotine dependence	CNS
Non-Hodgkin lymphoma	ONC	Non-alcoholic steatohepatitis	GI
Non-insulin dependent diabetes	ENDO	Non-small-cell lung cancer	ONC
Obesity	ENDO	Obsessive compulsive disorder	CNS
Ocular disease	ENT/OPH	Ocular hypertension	ENT/OPH
Ocular infection	ID	Ocular inflammation	ENT/OPH
Onchocerciasis	ID	Onychomycosis	ID
Open angle glaucoma	ENT/OPH	Opiate dependence	CNS
Oral mucositis	ONC	Organ transplantation	OTHER
Osteoarthritis	IM/D	Osteoporosis	OTHER
Osteosarcoma	ONC	Otitis media	ENT/OPH
Ovary tumor	ONC	Overactive bladder	GU
Pagets bone disease	IM/D	Pain	CNS
Pancreas disease	GI	Pancreas tumor	ONC
Pancreatitis	GI	Panic disorder	CNS
Papillomavirus infection	ID	Parasitic infection	ID
Parkinsons disease	CNS	Parturition	OTHER
Peptic ulcer	GI	Periodontal disease	ENT/OPH
Peripheral T-cell lymphoma	ONC	Peripheral arterial occlusive disease	CVS
Peripheral neuropathy	CNS	Peripheral vascular disease	CVS
Peritoneal tumor	ONC	Phenylketonuria	ENDO
Plasmodium falciparum infection	ID	Plasmodium infection	ID
Plasmodium vivax infection	ID	Pneumocystis carinii infection	ID
Pneumonia	ID	Poison intoxication	OTHER
Poliovirus infection	ID	Pollakiuria	ENDO
Polycystic kidney disease	GU	Pompes disease	ENDO
Post traumatic stress disorder	CNS	Postherpetic neuralgia	CNS
Postmenopausal osteoporosis	ENDO	Pox virus infection	ID
Pre-eclampsia	OTHER	Precancer	ONC
Premature ejaculation	GU	Premature labor	OTHER
Premenstrual syndrome	GU	Primary biliary cirrhosis	GI
Primary sclerosing cholangitis	GI	Prostate hyperplasia	GU
Prostate tumor	ONC	Pruritus	OTHER
Pseudomonas aeruginosa infection	ID	Pseudomonas infection	ID
Psoriasis	IM/D	Psoriatic arthritis	IM/D
Psychiatric disorder	CNS	Psychotic disorder	CNS
Pulmonary artery hypertension	RESP	Pulmonary fibrosis	RESP
Pulmonary hypertension	RESP	Rabies virus infection	ID
Radiation sickness	OTHER	Renal cell carcinoma	ONC
Renal disease	GU	Renal failure	GU
Renal fibrosis	GU	Renal injury	GU
Renal tumor	ONC	Renovascular hypertension	GU
Reperfusion injury	OTHER	Respiratory disease	RESP
Respiratory disorder	RESP	Respiratory distress syndrome	RESP
Respiratory syncytial virus infection	RESP	Respiratory tract infection	ID
Respiratory tract inflammation	ID	Restenosis	CVS
Restless legs syndrome	CNS	Retinitis pigmentosa	ENT/OPH
Retinopathy	ENT/OPH	Rett syndrome	OTHER
Rheumatoid arthritis	IM/D	Rhinitis	ENT/OPH
Rosacea	IM/D	Rotavirus infection	ID

Rubella virus infection	ID	SARS coronavirus infection	ID
Salmonella typhi infection	ID	Sanfilippo syndrome	ENDO
Sarcoma	ONC	Sarcopenia	OTHER
Scar tissue	IM/D	Schistosomiasis	ID
Schizophrenia	CNS	Scleroderma	IM/D
Secondary hyperparathyroidism	ENDO	Seizure disorder	CNS
Sepsis	ID	Septic shock	ID
Shigella infection	ID	Shock	OTHER
Sickle cell anemia	OTHER	Sinusitis	ENT/OPH
Skin burns	IM/D	Skin cancer	ONC
Skin infection	ID	Skin tumor	ONC
Skin ulcer	IM/D	Sleep apnea	RESP
Sleep disorder	CNS	Small-cell lung cancer	ONC
Soft tissue sarcoma	ONC	Solid tumor	ONC
Spinal cord injury	CNS	Spinal muscular atrophy	CNS
Squamous cell carcinoma	ONC	Stage IV melanoma	ONC
Staphylococcus aureus infection	ID	Staphylococcus infection	ID
Stem cell transplantation	OTHER	Stomach tumor	ONC
Stomach ulcer	GI	Streptococcus infection	ID
Streptococcus pneumoniae infection	ID	Stroke	CNS
Syndrome X	OTHER	Systemic lupus erythematosus	IM/D
T-cell lymphoma	ID	Tardive dyskinesia	CNS
Testosterone deficiency	ENDO	Thrombocytopenia	OTHER
Thrombocytopenic purpura	OTHER	Thromboembolism	CVS
Thrombosis	CVS	Thyroid tumor	ONC
Tinnitus	ENT/OPH	Topical anesthesia	CNS
Toxicity	OTHER	Transplant rejection	OTHER
Traumatic brain injury	CNS	Trypanosoma brucei infection	ID
Trypanosoma cruzi infection	ID	Trypanosomiasis	ID
Turners syndrome	OTHER	Ulcer	OTHER
Ulcerative colitis	GI	Unidentified indication	OTHER
Urinary dysfunction	GU	Urinary incontinence	GU
Urinary tract disease	GU	Urinary tract infection	GU
Urticaria	IM/D	Uterine cervix tumor	ONC
Uterine fibroids	GU	Uveitis	ENT/OPH
Vaccination	ID	Varicella zoster virus infection	ID
Variola virus infection	ID	Vascular disease	CVS
Venezuelan equine encephalitis virus infection	ID	Verruca vulgaris	ID
Vibrio cholerae infection	ID	Viral hemorrhagic fever	ID
Viral infection	ID	West Nile virus infection	ID
Wet age related macular degeneration	ENT/OPH	Wound healing	OTHER
Xerophthalmia	ENT/OPH	Xerostomia	ENT/OPH
Yellow fever virus infection	ID	Yersinia pestis infection	ID
Zika virus infection	ID	Zollinger-Ellison syndrome	GI

Appendix E: A Simple Model of Drug Selection

In this Appendix, we develop a simple model of drug selection decisions given the state of competitive R&D pipelines. Our goal is to derive some directional results on how the selection decision is affected by our primary factors of interest: the number of rival projects, the magnitude of learning, etc. To that end, we begin by estimating the expected value from continuation (selection) of a focal compound, given the stage of rivals' compounds (namely $\mathbb{E}[V|\omega]$)[8].

We use the following nomenclature definitions:

- the *rival pipeline state* $\omega : n_0, n_1$ where n_0 and n_1 denote the number of rival compounds in the early (pre-clinical), or late (clinical) stages of development respectively, which aim to address the same indication as the focal drug. For the purposes of this document, we assume that n_0 and n_1 can take values 0, 1, 2.
- p_0, p_1 are transition probabilities representing the likelihood that the compound meets the respective efficacy criteria to be entered in the clinical phase, or into the market respectively; also we define $\pi_0 = p_0 \cdot p_1$ and $\pi_1 = p_1$ as the likelihoods of successful launch given the stage of development of a compound. Those probabilities can be the same or different across rivals.
- V represents the total market revenue in each of the market periods as seen in the current moment. Upon the coexistence of two (or three) competing drugs in the same market, we assume that companies equally share the total revenues.[9] We also assume here that each compound serves a patient population only for two periods, to capture patent expiration considerations.
- δ is a standard per-period discount factor.
- The function $S(n_0, n_1|V)$ represents the **selection decision** of the focal company such that $S(n_0, n_1|V) = 1$ when $\mathbb{E}[V|\omega] > t$ and 0 otherwise. The threshold t captures the minimum hurdle that the project needs to overcome to be selected by the focal firm.

Early-stage Rival Efforts

Given the above definitions, we can calculate the expected value of the focal compound in the case where $n_0 > 0$ and $n_1 = 0$. We do this to keep the derivations as simple as possible. Then:

For $n_0 = 1$:

$$\begin{aligned}\mathbb{E}[V|n_0 = 1, n_1 = 0] &= -\mathbb{E}[C] + \pi_0(1-\pi_0)(V+\delta V) + \frac{1}{2}\pi_0^2(V+\delta V) \\ &= -\mathbb{E}[C] + \pi_0(V+\delta V)\left[1 - \pi_0 + \frac{\pi_0}{2}\right] \\ &= -\mathbb{E}[C] + \pi_0\left(1 - \frac{\pi_0}{2}\right)(1+\delta)V\end{aligned}$$

where we assume that the expected development cost is $\mathbb{E}[C] = c_0 + \pi_0 c_1$.

For $n_0 = 2$:

[8] We recognise the fact that a selection decision is a much more complex decision as opposed to a simple calculation of the expected value of a focal drug. However, we posit, reasonably, that a value $\mathbb{E}[V|\omega]$ higher than a threshold t implies a positive selection decision.

[9] Our results do not change even when we introduce more complex revenue structures both across time and between monopoly, duopoly, and oligopoly settings.

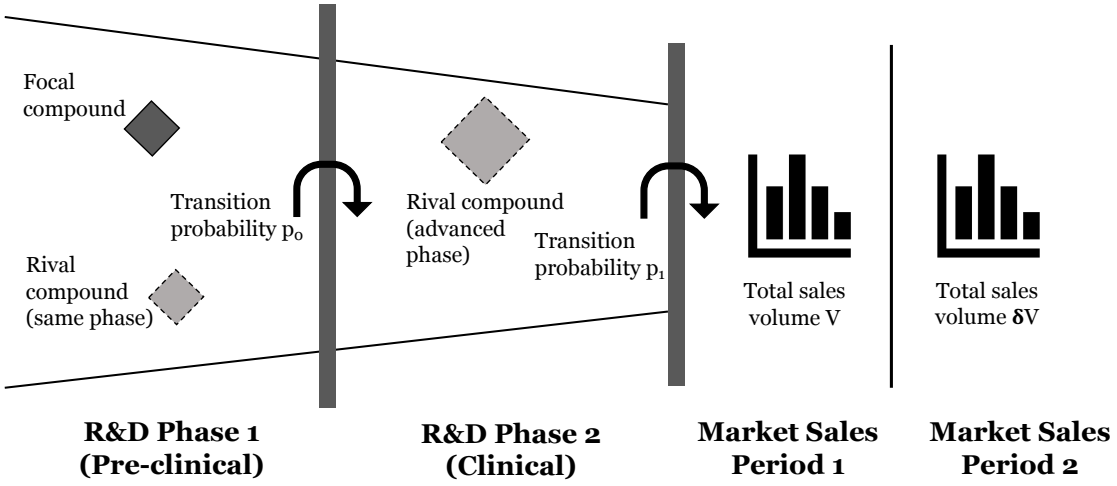

Figure E3: A Simple Representation of a Drug Development Pipeline

$$\begin{aligned}\mathbb{E}[V|n_0=2,n_1=0] &= -\mathbb{E}[C]+\pi_0(1-\pi_0)^2(V+\delta V)+2\pi_0^2(1-\pi_0)\left(\frac{\delta V+\delta V}{2}\right)+\pi_0^3\left(\frac{V+\delta V}{3}\right)\\ &= -\mathbb{E}[C]+\pi_0(1-\pi_0)^2(V+\delta V)+\pi_0^2(1-\pi_0)(V+\delta V)+\pi_0^3\left(\frac{V+\delta V}{3}\right)\\ &= -\mathbb{E}[C]+\pi_0\left[1-\pi_0+\frac{\pi_0^2}{3}\right](1+\delta)V\end{aligned}$$

Based on those derivations we can establish that higher competition (higher n_0) decreases the expected value of the focal compound:

$$\begin{aligned}\mathbb{E}[V|n_0=2,n_1=0]-\mathbb{E}[V|n_0=1,n_1=0] &= \pi_0 V(1+\delta)\left[\frac{\pi_0^2}{3}-\frac{\pi_0}{2}\right]\\ &= \pi_0^2 V(1+\delta)\left(\frac{\pi_0}{3}-\frac{1}{2}\right)<0\end{aligned}$$

Notice that the difference is negative for all values of π_0.

Our derivations thus far allow us to identify the following **selection effects** for the early-stage rival efforts. Consider first the case where $\mathbb{E}[V|n_0=1,n_1=0]<t$, so that $S(n_0=1,n_1=0)=0$ (the focal drug is not selected for $n_0=1$). Then, as shown in the previous paragraph, $\mathbb{E}[V|n_0=2,n_1=0]<\mathbb{E}[V|n_0=1,n_1=0]<t$, and therefore $S(n_0=2,n_1=0)=0$. However, if $\mathbb{E}[V|n_0=1,n_1=0]>t$, so that $S(n_0=1,n_1=0)=1$ (the focal drug is selected for $n_0=1$), then we may get either $S(n_0=2,n_1=0)=1$ (if the drop due to the second rival in the $\mathbb{E}[V|\omega]$ is not big enough to cross under the t threshold), or $S(n_0=2,n_1=0)=0$ (if the drop brings the value below the threshold, i.e., $\mathbb{E}[V|n_0=2,n_1=0]<t$). Taken together, the above results show that the *marginal effect* of early-stage rival efforts on selection, $D(n_0)=S(n_0=2,n_1=0)-S(n_0=1,n_1=0)$ will always be non-positive.

Moreover, given that $\mathbb{E}[V|\omega]$ increases in V, one can expect that for higher values ($V = V^H$), $\mathbb{E}[V|n_0 = 2, n_1 = 0]$ remains above t, and therefore $D(n_0/V^H) = S(n_0 = 2, n_1 = 0/V^H) - S(n_0 = 1, n_1 = 0/V^H) = 1 - 1 = 0$. In contrast, for smaller values ($V = V^L$), the additional rival could push a project below the selection threshold, i.e., $\mathbb{E}[V|n_0 = 2, n_1 = 0] < t < \mathbb{E}[V|n_0 = 1, n_1 = 0]$. In that case, $D(n_0/V^L) = S(n_0 = 2, n_1 = 0/V^L) - S(n_0 = 1, n_1 = 0/V^L) = 0 - 1 = -1$. These directional effects are consistent with the empirical results presented in Figure 2.

Late-stage Rival Efforts

In a similar way, we can derive the effects of $\omega : n_0 = 0, n_1 > 0$, which capture how late-stage rival projects shape the expected value of the focal drug project.

$$\begin{aligned}\mathbb{E}[V|n_0 = 0, n_1 = 1] &= -\mathbb{E}[C] + \pi_0(1-\pi_1)(V + \delta V) + \pi_0\pi_1\left(\frac{V}{2} + \delta V\right) \\ &= -\mathbb{E}[C] + \pi_0 V\left\{(1-\pi_1)(1+\delta) + \pi_1\left(\frac{1}{2} + \delta\right)\right\}\end{aligned}$$

$$\begin{aligned}\mathbb{E}[V|n_0 = 0, n_1 = 2] &= -\mathbb{E}[C] + \pi_0'(1-\pi_1)^2(V + \delta V) + 2\pi_0'(1-\pi_1)\pi_1\left(\frac{V}{2} + \delta V\right) + \pi_0'\pi_1^2\left(\frac{V}{3} + \delta V\right) \\ &\quad [...] \\ &= -\mathbb{E}[C] + \pi_0' V\left\{(1-\pi_1)(1+\delta) + \pi_1\delta + \frac{\pi_1^2}{3}\right\}\end{aligned}$$

Two important points are worth highlighting here: (i) we assume that the focal firm faces a $\pi_0' > \pi_0$ likelihood of successful regulatory approval to capture the knowledge spillovers due to late-stage rival efforts, i.e., the fact that more firms ($n_1 = 2$) progressed to that stage within the same target indication has generated more knowledge regarding what works and what does not. Such knowledge enables the focal firm to adjust their effort and increase their chances of successful regulatory approval. In other words, under the state $n_0 = 0, n_1 = 2$, there are *ceteris paribus* higher chances of success for the focal firm compared to the state $n_0 = 0, n_1 = 1$, and that is why $\pi_0' > \pi_0$. (ii) The per-period total market value is split *only* in the first sales period (as viewed by the focal firm) because of the assumed two-period patent effect. Thus followers face lower rents during the period of overlap with leader firms, but given the patent expiration they may enjoy discounted monopoly rents as well (captured by the term δV).

Let k denote the learning parameter, so that $\pi_0' = k\pi_0$ with $k > 1$. Note that, in the absence of any learning from late-stage efforts, i.e., $k = 1$, the change in the expected value is:

$$\begin{aligned}\mathbb{E}[V|n_0 = 0, n_1 = 2] - \mathbb{E}[V|n_0 = 0, n_1 = 1] &= \pi_0 V\left\{(1-\pi_1)(1+\delta) + \pi_1\delta + \frac{\pi_1^2}{3}\right\} \\ &\quad - \pi_0 V\left\{(1-\pi_1)(1+\delta) + \pi_1\left(\frac{1}{2} + \delta\right)\right\} \\ &= \pi_0\pi_1 V\left\{\frac{\pi_1}{3} - \frac{1}{2}\right\} < 0\end{aligned}$$

In other words, *absent the possibility to learn* from rival efforts, more late-stage rival efforts will decrease the expected value of the focal project.

However, in the presence of learning ($k > 1$),

$$\mathbb{E}[V|n_0=0,n_1=2]-\mathbb{E}[V|n_0=0,n_1=1]=\pi_0(k-1)V\left\{(1-\pi_1)(1+\delta)+\pi_1\delta+\frac{\pi_1^2}{3}\right\}-\pi_0\pi_1V\left\{\frac{1}{2}-\frac{\pi_1}{3}\right\},$$

which increases in k. Specifically, the difference becomes positive for *sufficient* learning, i.e.,

$$k>\overline{k}=1+\frac{\pi_1\left(\frac{1}{2}-\frac{\pi_1^2}{3}\right)}{1-\pi_1+\delta+\frac{\pi_1^2}{3}}.$$

As a result, we find that more late-stage rival efforts (targeting the same indication) increase the expected value of the focal project due to learning effects. We now examine the effect of late-stage signals on the *selection decision*.

First, consider the case where $\mathbb{E}[V|n_0=0,n_1=1]>t$, such that $S(n_0=0,n_1=1)=1$ (the focal project is selected for $n_1=1$). Then, in the presence of "strong" learning, we can safely derive that $\mathbb{E}[V|n_0=0,n_1=2]>\mathbb{E}[V|n_0=0,n_1=1]>t$, so that $S(n_0=0,n_1=2)=1$. In other words, if a project were to be selected under $n_1=1$, then it will also be selected under $n_1=2$.

Alternatively, if $\mathbb{E}[V|n_0=0,n_1=1]<t$, so that $S(n_0=0,n_1=1)=0$ (the focal project is not selected for $n_1=1$), then we may get either $S(n_0=0,n_1=2)=0$ (if the increase in the $\mathbb{E}[V|\omega]$ is not big enough to cross above the t threshold), or $S(n_0=0,n_1=2)=1$ (if the increase brings the value above the threshold $\mathbb{E}[V|n_0=0,n_1=2]>t$). In other words, if a project were to not be selected under $n_1=1$, then because of the presence of learning, an extra late-stage signal $n_1=2$ may actually "bump" the expected value of the project over the threshold, leading to selection. Taken together, the above results show that the *marginal effect* of late-stage rival efforts on selection $D(n_1)=S(n_0=0,n_1=2)-S(n_0=0,n_1=1)$ will always be non-negative, for any t value. This is consistent with H1c.

Lastly, when the difference between low (V^L) and high (V^H) market is substantial, so that the focal project is selected only in the latter case, then $D(n_1/V^L)=S(n_0=0,n_1=2/V^L)-S(n_0=0,n_1=1/V^L)=0-0=0$ and $D(n_1/V^H)=S(n_0=0,n_1=2/V^H)-S(n_0=0,n_1=1/V^H)=1-1=0$. Alternatively, when V^L is sufficiently high so that the project is selected even for low rival efforts (i.e., $\mathbb{E}[V|n_0=0,n_1=1]>t$), then it will also be selected for high rival efforts as well as high market potential. Again, $D(n_1/V^L)=D(n_1/V^H)=1-1=0$. These directional effects are consistent with the empirical results presented in Figure 2.

Appendix References

Banerjee AV (1992) A simple model of herd behavior. *The Quarterly Journal of Economics* 107(3):797–817.

Bennett VM, Snyder J (2017) The empirics of learning from failure. *Strategy Science* 2(1):1–12.

Berggren R, Fleming E, Keane H, Moss R (2018) R&D in the 'age of agile'. Technical report, McKinsey & Company.

Bikhchandani S, Hirshleifer D, Welch I (1992) A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy* 100(5):992–1026.

Crama P, Reyck BD, Degraeve Z (2008) Milestone payments or royalties? contract design for R&D licensing. *Operations Research* 56(6):1539–1552.

Danzon PM, Nicholson S, Pereira NS (2005) Productivity in pharmaceutical-biotechnology R&D: The role of experience and alliances. *Journal of Health Economics* 24(2):317–339.

Girotra K, Terwiesch C, Ulrich KT (2007) Valuing R&D projects in a portfolio: Evidence from the pharmaceutical industry. *Management Science* 53(9):1452–1466.

Higgins MJ, Rodriguez D (2006) The outsourcing of R&D through acquisitions in the pharmaceutical industry. *Journal of Financial Economics* 80(2):351–383.

Hoang H, Rothaermel FT (2010) Leveraging internal and external experience: Exploration, exploitation, and R&D project performance. *Strategic Management Journal* 31(7):734–758.

Katz R, Allen TJ (1982) Investigating the Not Invented Here (NIH) syndrome: A look at the performance, tenure, and communication patterns of 50 R&D project groups. *R&D Management* 12(1):7–20.

KMR Group (2011) Annual R&D general metrics study highlights new success rate and cycle time data. Press release, accessed April 6, 2018.

Lloyd I (2017) Pharma R&D annual review 2017. , URL `https://pharmaintelligence.informa.com/~/media/Informa-Shop-Window/Pharma/Files/PDFs/whitepapers/RD-Review-2017.pdf`, accessed April 6, 2018.

Loch CH, Kavadias S (2002) Dynamic portfolio selection of npd programs using marginal returns. *Management Science* 48(10):1227–1241.

Mason R, Savva N, Scholtes S (2008) The economics of licensing contracts. *Nature Biotechnology* 26(8):1–3.

Neuberger A, Oraiopoulos N, Drakeman DL (2019) Renovation as innovation: Is repurposing the future of drug discovery research? *Drug discovery today* 24(1):1.

Palmer E (2017) Top 15 pharma companies by 2016 revenue. *FiercePharma* URL `https://www.fiercepharma.com/special-report/top-15-pharma-companies-by-2016-revenues`, Accessed June 6, 2017.

Savva N, Scholtes S (2014) Opt-out options in new product co-development partnerships. *Production and Operations Management* 23(8):1370–1386.

US Food and Drug Administration (2007) *Food and Drug Administration Amendments Act of 2007*. `https://www.gpo.gov/fdsys/pkg/PLAW-110publ85/pdf/PLAW-110publ85.pdf`, Accessed July 9, 2018.

Wong CH, Siah KW, Lo AW (2019) Estimation of clinical trial success rates and related parameters. *Biostatistics* 20(2):273–286.