P-hacking in clinical trials and how incentives shape the distribution of results across phases

Jérôme Adda[a,b,c,1], Christian Decker[d,e,1], and Marco Ottaviani[a,b,c,1,2]

[a]Department of Economics, Bocconi University, 20136 Milan, Italy; [b]Bocconi Institute for Data Science and Analytics, Bocconi University, 20136 Milan, Italy; [c]Innocenzo Gasparini Institute for Economic Research, Bocconi University, 20136 Milan, Italy; [d]Department of Economics, University of Zurich, 8001 Zurich, Switzerland; and [e]UBS Center for Economics in Society, University of Zurich, 8001 Zurich, Switzerland

Edited by Jose A. Scheinkman, Columbia University, New York, NY, and approved April 24, 2020 (received for review November 15, 2019)

Clinical research should conform to high standards of ethical and scientific integrity, given that human lives are at stake. However, economic incentives can generate conflicts of interest for investigators, who may be inclined to withhold unfavorable results or even tamper with data in order to achieve desired outcomes. To shed light on the integrity of clinical trial results, this paper systematically analyzes the distribution of *P* values of primary outcomes for phase II and phase III drug trials reported to the ClinicalTrials.gov registry. First, we detect no bunching of results just above the classical 5% threshold for statistical significance. Second, a density-discontinuity test reveals an upward jump at the 5% threshold for phase III results by small industry sponsors. Third, we document a larger fraction of significant results in phase III compared to phase II. Linking trials across phases, we find that early favorable results increase the likelihood of continuing into the next phase. Once we take into account this selective continuation, we can explain almost completely the excess of significant results in phase III for trials conducted by large industry sponsors. For small industry sponsors, instead, part of the excess remains unexplained.

clinical trials | drug development | selective reporting | p-hacking | economic incentives in research

The evidence produced in clinical trials is susceptible to many kinds of bias (1–3). While some such biases can occur accidently, even unbeknownst to the study investigators, other biases may result from strategic behavior of investigators and sponsors. In addition to the public value of improving medical treatments, the information obtained through clinical trials is privately valuable for the sponsoring pharmaceutical companies that aim to demonstrate the safety and efficacy of newly developed drugs—the prerequisite for marketing approval by authorities such as the US Food and Drug Administration (FDA). Given the sizeable research and development costs involved (4) and the lure of large potential profits, investigators can suffer from conflicts of interest (5–8) and pressure to withhold or "beautify" unfavorable results (9, 10) or even fabricate and falsify data (11).

In the 1990s and 2000s, many medical scholars began calling for more transparency in clinical research (12), following public outcry over alarming evidence of selective publication of trial results (13–15), cases of premature drug approvals (16), and allegations of data withholding (17). As a response to these concerns, policymakers established publicly accessible registries and result databases (18, 19), such as ClinicalTrials.gov (20, 21) (see *SI Appendix* for more details on the ClinicalTrials.gov registry and the legal requirements for reporting trial results).

ClinicalTrials.gov now contains sufficient data to allow for a systematic evaluation of the distribution of reported *P* values. Our analysis builds on and develops the methods proposed in the literature that investigates "p-hacking," publication bias, and the "file-drawer problem" (22, 23) for academic journal publications in a number of fields, ranging from life sciences (24) to psychology (25, 26), political science (27, 28), and economics (29–31).

Given the escalation of stakes as research progresses through phases, clinical trials are particularly well suited to detect how economic incentives of sponsoring parties drive research activity (32–34) and reporting bias. Economic incentives in clinical trials may depend on the size of the sponsoring firm (32). Compared to larger companies, smaller firms may have more to gain by misreporting results—and less reputation to lose if they are exposed. In other contexts, such reputational concerns have been found to vary by firm size (35, 36) or by academic prominence (37).

While the previous literature focused mostly on scientific publications in academic journals for which prepublication research results are typically not observable, ClinicalTrials.gov allows us to observe results from clinical trials in earlier research phases. Thus, we are able to follow the evolution of research results over time and construct counterfactuals not available in previous work. By linking trials across different phases of clinical research, we are able to quantify the effect of the incentives to selectively continue experimental research depending on early stage results.

Methods and Results

Our focus is on preapproval interventional superiority studies on drugs carried out as phase II and phase III trials. Trials in phase II investigate drug safety and efficacy, typically with a small sample of experimental subjects. Phase III trials investigate

Significance

Statistical significance in clinical trials is a key prerequisite for marketing approval of new drugs. The large economic payoffs at stake might undermine investigators' ethical obligations and incentivize manipulation of results. This study systematically evaluates the integrity of results reported to the largest registry of clinical trials, ClinicalTrials.gov. Contrary to what has been documented in previous studies of academic publications across a number of disciplines, our analysis does not detect evidence for widespread manipulation of results to clear the 5% threshold for statistical significance. However, we find that the increase in the share of significant results from phase II to phase III can be explained only partially by investigators' incentives to selectively continue experimentation following favorable early results.

Author contributions: J.A., C.D., and M.O. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Data deposition: A complete replication package is available at the Harvard Dataverse (https://doi.org/10.7910/DVN/NBLYSW).

[1]J.A., C.D., and M.O. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: marco.ottaviani@unibocconi.it.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1919906117/-/DCSupplemental.

First published June 2, 2020.

efficacy, while monitoring adverse effects on a larger sample of individuals, and play a central role in obtaining approval to market the drug from regulators such as the FDA. To facilitate the analysis, we transformed the *P* values into test statistics, supposing that they would all originate from a two-sided Z test of a null hypothesis that the drug has the same effect as the comparison. This transformation allowed us to investigate both the overall shape of the distribution and the region around the thresholds for statistical significance more easily (see *Materials and Methods* and *SI Appendix* for further information on the data and the *P*–Z transformation).

The Distribution of Z Scores: Irregularity Tests. Fig. 1 displays density estimates of the constructed z statistics for tests performed for primary outcomes of phase II and phase III trials. We present results for all trials in Fig. 1*A* and subsequently provide the breakdown by affiliation of the lead sponsor: nonindustry (NIH, US federal agencies, universities, etc.) in Fig. 1*B*, top-10 industry (the 10 pharmaceutical companies in the sample with the largest revenues in 2018; *SI Appendix*, Table S1) in Fig. 1*C*, and small industry (the remaining smaller pharmaceutical companies) in Fig. 1*D*.

Next, we diagnosed three possible irregularities in the distribution of z statistics of trials, at or above the 5% significance threshold, corresponding to a z statistic of 1.96. Further technical details and robustness checks are gathered in *SI Appendix*.

Spike in the Density Function Just Above 1.96. First, we detected no spikes in the densities (or discontinuities in the distribution functions) just above 1.96, the salient significance threshold. Such spikes, indicating that results are inflated to clear the significance hurdle, have been documented in previous studies of z distributions for tests in academic publications across life sciences (24), as well as economics (31) and business studies (39). Thus, the more natural distribution of z scores from ClinicalTrials.gov displays more integrity compared to results reported for publications in scientific journals. This difference may partially be explained by the absence of the additional layer of editorial selection, which may be based also on the statistical significance of presented results. This first finding suggests that registered results are not inflated at the margin just to clear the significance threshold.

Discontinuity of the Density Function at 1.96. Second, we investigated the presence of a discontinuity in the density of z statistics with a test that relies on a simple local polynomial density estimator (38). The densities for phase II trials were smooth and did not show a noteworthy upward shift at the 1.96 threshold in all cases. In contrast, the densities of z statistics for industry-sponsored (both small and top 10) phase III trials displayed a break at 1.96. The break was statistically significant only for phase III trials undertaken by small pharmaceutical companies (Fig. 1*D*), with a persistent upward shift to the right of the threshold, indicating an abnormal amount of significant results. This pattern is suggestive of "selective reporting" i.e., strategic concealment of some nonsignificant results.

The different patterns observed between large and small industry sponsors (Fig. 1 *C* and *D*) were robust across a wide range of alternative ways to define "large" sponsors (*SI Appendix*, Fig. S1). Moreover, we found a similar discontinuity for phase III trials by small industry sponsors when transforming *P* values to test statistics of a one-sided instead of a two-sided test (*SI Appendix*, Fig. S2).

Excess of Significant Results in Phase III Compared to Phase II. Third, Fig.1 indicates an excess of favorable results over the 1.96 threshold in phase III compared to phase II. More favorable results were more likely to be observed in phase III than in phase II. The phase III distribution of z statistics stochastically dominates the phase II distribution. Dominance is particularly strong for industry-sponsored trials (Fig. 1 *C* and *D*). This pattern appears suspicious, but it is not as alarming as a spike at the significance threshold. While only 34.7% of phase II trial results by nonindustry sponsors fell above 1.96 (and 34.8%, respectively, for phase III, a difference that is not statistically significant), the fraction of significant results rose to 45.7% in phase II and 70.6% in phase III for industry-sponsored trials.

Recall that the analysis above considered only *P* values associated to primary outcomes of trials. These results constitute

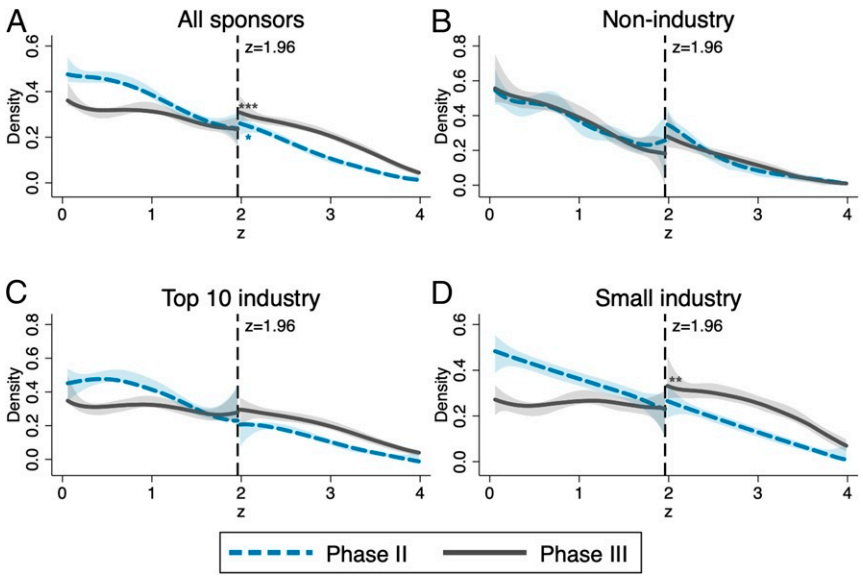

Fig. 1. Comparison of phase II and phase III densities of the z score and tests for discontinuity at $z = 1.96$, depending on the affiliation of the lead sponsor. Density estimates of the constructed z statistics for primary outcomes of phase II (dashed blue lines) and phase III (solid gray lines) trials are shown. The shaded areas are 95% confidence bands, and the vertical lines at 1.96 correspond to the threshold for statistical significance at the 0.05 level. Sample sizes: $n = 3,953$ (phase II), $n = 3,664$ (phase III) (*A*); $n = 1,171$ (phase II), $n = 720$ (phase III) (*B*); $n = 1,332$ (phase II), $n = 1,424$ (phase III) (*C*); and $n = 1,450$ (phase II), $n = 1,520$ (phase III) (*D*). Significance levels for discontinuity tests (38) are shown. $^{**}P < 0.05$; $^{***}P < 0.01$. Exact *P* values are reported in *SI Appendix*, Table S2.

the main measure for success of the treatment being trialed, for both the investigators themselves and the evaluating authorities. The densities of z scores from lower-stake secondary outcomes for all groups of sponsors and both phases did not display any meaningful discontinuity at the significance threshold (*SI Appendix*, Fig. S3 and Table S5). Moreover, for secondary outcomes, the excess of significant results from industry-sponsored trials in phase III relative to phase II was much smaller compared to the distribution for primary outcomes. We found irregularities only for higher-stake primary outcomes, suggesting that incentives of reporting parties play a role.

Linking Trials across Phases: Controlling for Selective Continuation. The FDA focuses mainly on phase III results when deciding about marketing approval, a decision with major financial consequences for pharmaceutical companies. Given these incentives, the observed excess of significant results, particularly in the group of industry-sponsored phase III trials, could be interpreted as evidence of tampering (p-hacking) or nondisclosure of negative results (selective reporting). However, this conclusion would be premature without first carefully examining the dynamic incentives underlying clinical research, as we set out to do.

An alternative explanation for the excess of significant results in phase III relative to phase II is the selective continuation of drug testing to the next phase only when initial results are sufficiently encouraging. Selective continuation saves on costly clinical research and can thus even be socially desirable, as long as such economic considerations do not distort research activity away from important, but costly, projects (8). Also, from an ethical viewpoint, no further trials with volunteer patients should be conducted when a drug is highly unlikely to have a positive impact. Time and resources should be devoted to more promising projects instead. We outline a model of the sponsor's continuation decision in *Materials and Methods*.

To identify the impact of selective continuation, we developed a procedure to link phase II and phase III trials in our dataset based on the main intervention (i.e., the tested drug or combination of drugs), the medical condition to be treated, and the timing. This procedure is illustrated in Fig. 2. A given phase II trial may either 1) have no corresponding phase III trial with the same intervention and same condition; or 2) have one or multiple matches in phase III. In the latter case, we considered the phase II trial as continued into phase III. The resulting linked data, which we make available to the research community (40), is a key input in the methodology we developed to estimate a selection function capturing selective continuation for industry-sponsored trials.

Following our model of the firm's continuation decision, we estimated the selection function with a logistic regression of a dummy variable indicating if there is at least one match among the phase III trials in the database (regardless of whether phase III results are reported or not) on the phase II z score. We controlled for adjustment for multiple hypothesis testing, a flexible time trend, and other covariates that might influence the perceived persuasiveness of phase II results (square root of overall enrollment to each trial as proxy for power of the statistical tests and active comparator vs. placebo) or the economic incentives to undertake research (fixed effects for the treated condition) on top of the z score; see *Materials and Methods* for the exact specification. The predicted values of this selection function can be interpreted as the probability that a drug progresses to phase III, conditional on the information available at the end of phase II, consisting of the phase II z score and other covariates.

In most cases, very low P values are no longer reported precisely, but only as being below the thresholds 0.001 or 0.0001 (e.g., $P < 0.001$ instead of $P = 0.0008$). Therefore, we estimated the continuation probability separately for those two cases by including dummies for "$z > 3.29$" (corresponding to the P value being reported as $P < 0.001$) and "$z > 3.89$" (corresponding to $P < 0.0001$) in the specification of the selection function.

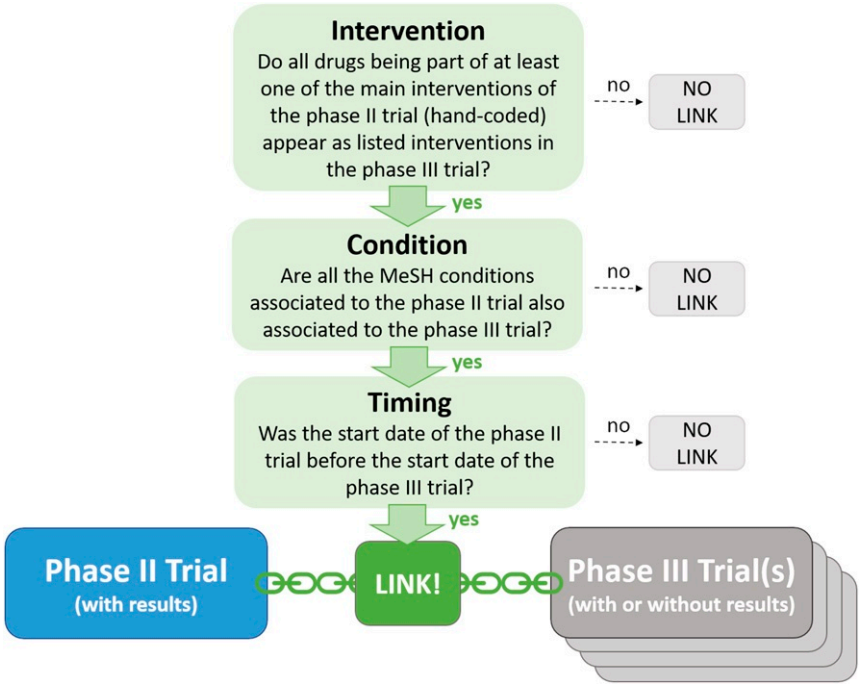

Fig. 2. Linking phase II and phase III trials. We considered a phase II trial as continued if we found at least one phase III trial registered in the database (regardless of whether associated results are reported or not) fulfilling all three criteria (intervention, condition, and timing). See *SI Appendix* for a more detailed description of the linking procedure.

Table 1 displays the estimated logit coefficients for all industry sponsors (column 1) and for small and top-10 industry sponsors separately (columns 2 and 3, respectively). Fig. 3 illustrates the estimated selection functions graphically. The solid green line shows the predicted continuation probability as function of the phase II z score. A higher z score in phase II significantly increases the probability of continuation to phase III. The lighter dotted and darker dashed lines show the predictions when considering only trials conducted by small sponsors or, respectively, the 10 largest industry sponsors. The estimated continuation probabilities suggest that larger companies continue research projects more selectively. The overall share of matched trials is lower for large industry sponsors, captured by the downward shift of the selection function.

In the context of our model of the firm's continuation decision, the continuation probability is negatively associated with the opportunity cost of continuing a specific project. On average, this cost can be expected to be greater for large sponsors with many alternative projects. This interpretation is in line with findings from previous studies arguing that managers of larger firms with multiple products in development have less private costs attached to terminating unpromising research projects and, thus, are more efficient (32).

In *SI Appendix*, Table S6, we report estimates of the same logistic model when considering the phase II z scores associated to secondary outcomes instead of primary outcomes. The coefficients related to the z score are much smaller in magnitude, and most of the coefficients are not statistically significant, notwithstanding the much larger sample size. This finding confirms that the evaluation of a trial's success, and therefore also selective continuation, is based predominantly on primary outcomes.

Decomposition of the Difference in Significant Results between Phase II and Phase III. Under the assumption that, conditional on our control variables, the expected z statistic in phase III equals the z of a similar phase II trial, we can construct a hypothetical phase III distribution for primary outcomes accounting for selective continuation. To do so, we estimated the kernel density of phase II statistics (for now, disregarding $z > 3.29$ and $z > 3.89$) reweighting each observation by the continuation probability predicted by our selection function, given the characteristics of the phase II trial. The resulting counterfactual density can be compared to the actual phase II and phase III densities, which we estimated using a standard unweighted kernel estimator.

Since the selection function is increasing in the phase II z score, the counterfactual z density rotates counter-clockwise, increasing the share of significant results (*SI Appendix*, Fig. S4). To calculate the overall share of significant results under the hypothetical regime, we combined the estimated densities with the number of $z > 3.29$ and $z > 3.89$ results predicted from the selection functions and renormalize to one.

Based on this construction, we decomposed the difference in the share of significant results in phase II and phase III into two parts: selective continuation and an unexplained residual. As illustrated in Fig. 4*A* and *SI Appendix*, Table S7, when we considered all industry-sponsored trials, selective continuation, i.e., economizing on the cost of trials that are not promising enough, accounted for more than half of the difference, leaving 48.5% of the difference unexplained.

Next, we repeated the estimation procedure separately for trials sponsored by large and small industry. The difference in the share of significant results between phase II and phase III was slightly larger for trials by small sponsors (21.9 percentage points for top-10 industry vs. 25.8 percentage points for small industry). For trials sponsored by the 10 largest companies, the difference between the actual share of significant phase III results and the share predicted by selective continuation from phase II shrank to 3.4 percentage points and was no longer statistically significant. Thus, for top-10 industry sponsors, our methodology suggests no indication of selective reporting or potential tampering: Selective continuation can explain almost the entire excess share of significant results in phase III trials compared to phase II trials.

A different picture emerged for small industry sponsors. According to the selection function estimated in Table 1 and displayed in Fig. 3, small sponsors were much more likely to proceed to phase III than large sponsors, especially following phase II trials with relatively low z statistics. Hence, for small sponsors, selective continuation was less pronounced and can only account for less than one-third of the excess share of significant results in phase III trials compared to phase II trials. Phase III results actually reported by small sponsors appeared to be much more favorable than predicted by the selection function; for these sponsors, we are left with a statistically significant unexplained residual of 18.4 percentage points, as displayed in Fig. 4*A*.

As illustrated by Fig. 4 *B* and *C*, these different patterns between large and small industry sponsors are robust across a wide range of alternative ways to define "large" sponsors. For small sponsors (Fig. 4*B*), the share of the explained difference

Table 1. Estimates of logit selection function for selective continuation, based on primary outcomes

Sponsor	(1) All industry	(2) Small industry	(3) Top-10 industry
Phase II z score	0.331*** (0.0793)	0.266*** (0.100)	0.404*** (0.130)
Dummy for phase II z score reported as "$z > 3.29$"	1.063*** (0.226)	0.756** (0.329)	1.750*** (0.373)
Dummy for phase II z score reported as "$z > 3.89$"	1.232*** (0.255)	0.787*** (0.285)	1.643*** (0.446)
Mean dependent variable	0.296	0.344	0.246
P value Wald test (2) = (3)		0.00480	0.00480
Controls	Yes	Yes	Yes
MeSH condition fixed effects	Yes	Yes	Yes
Completion year fixed effects	Yes	Yes	Yes
Observations	3,925	2,017	1,908
No. of trials	1,167	674	493

Unit of observation: trial-outcome; included controls: square root of the overall enrollment, dummy for placebo comparator, and dummy for multiple hypothesis testing adjustment. See *Materials and Methods* for the exact specification. Categories for condition fixed effects are based on Medical Subject Headings (MeSH) terms associated to the trials (21); for more details, see *SI Appendix*. "*P* value Wald test (2) = (3)" reports the *P* value of a Wald test of the null hypothesis of joint equality of the coefficients in the first three rows and the constant between columns 2 and 3. SEs in parentheses are clustered at the MeSH condition level; significance levels (based on a two-sided *t*-test) are indicated. **$P < 0.05$; ***$P < 0.01$.

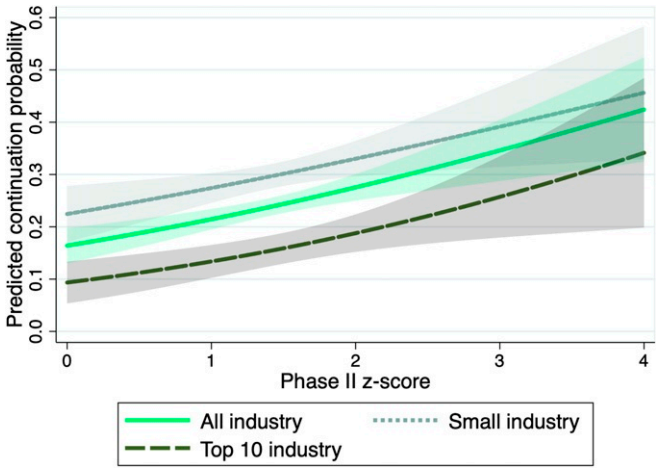

Fig. 3. Predicted continuation probability as function on the phase II z score, depending on affiliation of lead sponsor. Predictions are based on the estimated logit selection functions for selective continuation; see Table 1 for the estimated coefficients. All control variables are fixed at their mean values. The shaded areas are 95% confidence bands.

ranges between 19% and 44% with the majority of results being very close to the estimate in our main specification (29%). Also, for different definitions of large sponsors (Fig. 4*C*), the estimates are quite close to the result from our main specification (85%), ranging between 57% and 101%.

These findings are consistent with our earlier observation that small industry is the only group of sponsors for which the phase III z density exhibits a statistically significant discontinuity at the 1.96 threshold. Along the same lines, a recent evaluation of compliance with FDA requirements for reporting of trial results to ClinicalTrials.gov finds that compliance improves with sponsor size (41).

Discussion and Conclusion

Overall, the distribution of z scores from ClinicalTrials.gov does not indicate widespread manipulation of results reported to the registry. Given the increasing adoption of randomized control trials across life and social sciences, our findings speak in favor of setting up repositories similar to ClinicalTrials.gov in these other domains to monitor results and improve the credibility of research.

As we show, to correctly interpret the distribution of research results, it is important to understand the sequential nature of research and its interplay with economic incentives. Although phase III trials appear to deliver too many positive results, we can explain a large part of this excess of favorable results by linking them to phase II outcomes and accounting for selective continuation.

However, we find that selective continuation cannot explain fully the high number of significant results in phase III trials sponsored by smaller firms. For the same group of trials, we also identified a discontinuity in the density at the classical significance threshold. These patterns suggest that enforcers of registration should pay particular attention to smaller industry sponsors, for which reputational concerns may be less consequential—a channel that should be investigated more thoroughly by future work.

In conclusion, our exploratory findings indicate that current levels of regulation and enforcement are not sufficient to fully

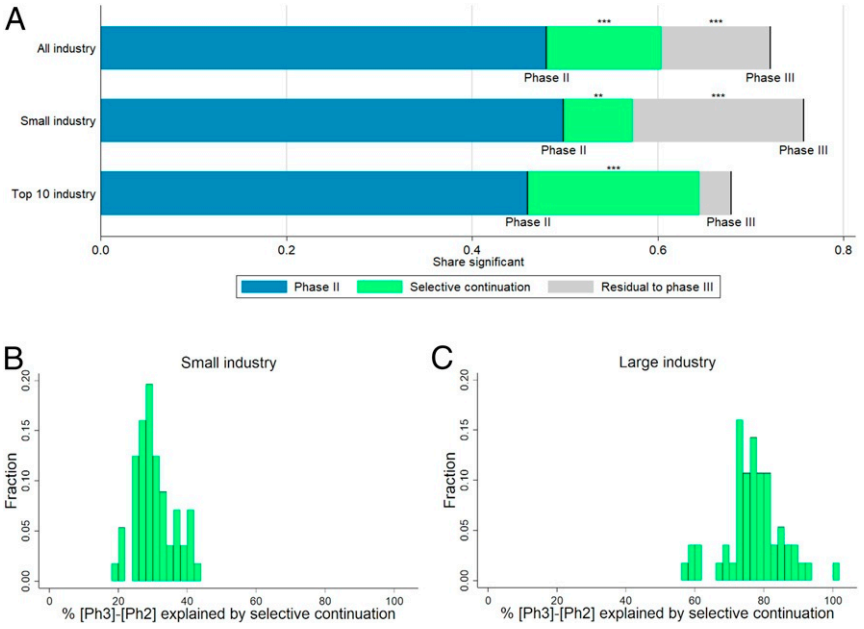

Fig. 4. (*A*) Selection-based decomposition of the difference in significant results from primary outcomes between phase II and phase III, depending on affiliation of lead sponsor (top-10 revenues criterion). Phase II and III lines represent the shares of trials with a *P* value below 5% (or, equivalently, a z score above 1.96). The green segments represent the parts of the differences explained by selective continuation, based on counterfactuals constructed from the phase II distribution. For precise numbers and sample sizes, see *SI Appendix*, Table S7. Significance levels for the differences (based on a two-sided *t*-test) are indicated. $^{**}P < 0.05$; $^{***}P < 0.01$. (*B* and *C*) Histograms of the percentage share of the difference in the share of significant results between phase III and phase II explained by selective continuation across different definitions for large vs. small industry sponsors. The shares correspond to the green area in *A* divided by the sum of the green and the gray areas. The sample of industry-sponsored trials is split according to 56 different definitions of large sponsors. These definitions are obtained by ranking sponsors by their 2018 revenues, volume of prescription drug sales in 2018, research and development spending in 2018, and the number of trials reported to the registry. For each of these four criteria, 14 different definitions of "large vs. small" were created: top seven vs. remainder, top eight vs. remainder, and so on, up to top 20 vs. remainder. Further details are provided in *SI Appendix*.

discipline reporting. To evaluate opportunities for reform, policymakers might want to weigh the ex post information benefits of mandatory registration against the reduced incentives of investigators to undertake clinical trials (42–46). An empirical quantification of this chilling effect could serve as an important input for a social cost–benefit analysis for tightening current rules.

Materials and Methods

Database for Aggregate Analysis of ClinicalTrials.gov. The Database for Aggregate Analysis of ClinicalTrials.gov (AACT) was launched in September 2010 to allow for free bulk download of all of the data contained in the ClinicalTrials.gov registry (19–21). The project is administered by the Clinical Trials Transformation Initiative, a partnership of the FDA and Duke University with the aim of improving quality and efficiency of clinical trials. The database, which is updated daily and directly accessible in the cloud, contains over 40 subtables with information on timing, conditions, interventions, facilities, locations, sponsors, investigators, responsible authorities, eligible participants, outcome measures, adverse events, results, and descriptions of trials.

The trials in the database cover a wide range of different diseases, interventions, and study designs. Hence, also, the reported results are very diverse in nature. In contrast to a meta-analysis on a specific disease or treatment, which typically uses only a narrowly defined subgroup of the dataset, we analyzed the largest possible portion of the overall data. Given the aggregate level of our analysis, rather than using the estimated coefficients, we focused on *P* values, the only measure reported uniformly and comparably for many trials, independent of their characteristics and the statistical method used for the analysis.

This study is based on the AACT data available on August 15, 2019. Over the last 2 y, we obtained similar results in earlier drafts of this paper based on less data. We concentrated on phase II and phase III interventional (as opposed to observational) superiority (as opposed to noninferiority) studies on drugs (as opposed to medical devices and others) which report at least one proper *P* value for a statistical test on a primary outcome of the trial.

We dropped the trials of the sponsor Colgate Palmolive, which reported *P* values exactly equal to 0.05 for 137 out of its 150 results. We attributed these exact *P* values of 0.05 to a reporting mistake; clearly, these were intended to be reported as significant results with *P* value lower than or equal to 0.05. Leaving Colgate Palmolive's results in the sample would lead to a substantial spike at $z = 1.96$, which could be wrongly interpreted as evidence for p-hacking. Moreover, we dropped the trial with the identifier NCT02799472, as it reported 211 *P* values for primary outcomes and would therefore have much more impact than all other trials (average number of *P* values for primary outcomes per trial: 2.5; median: 1).

Altogether, we obtained a sample of 12,621 *P* values from tests performed on primary outcomes of 4,977 trials. These single *P* values constituted the units of observation for our analysis. As a consequence of the FDA Amendments Act, the largest part of our results data pertains to trials conducted after 2007.

***P*–z Transformation.** We transformed the *P* values taken from the AACT database to corresponding z statistics by supposing that all *P* values would originate from a two-sided Z test of a null hypothesis that the drug has the same effect as the comparison. Given that under the null hypothesis, this statistic is normally distributed, we have the one-to-one correspondence $z = -\Phi^{-1}(\frac{p}{2})$, where z is the absolute value of the test statistic, and Φ^{-1} is the inverse of the standard normal cumulative distribution function. This transformation facilitates both the graphical analysis and the identification of discontinuities at the significance threshold, given that the z density is close to linear around the significance threshold, whereas the corresponding p density is highly nonlinear in this range.

Density Discontinuity Tests. We implemented tests of discontinuity in the z-score density at the $z = 1.96$ significance threshold based on the state-of-the-art procedure developed by Cattaneo et al. (38). This test builds on a local polynomial density-estimation technique that avoids prebinning of the data. More details on the testing procedure and supplementary results can be found in *SI Appendix*.

Linking Phase II and Phase III Trials. To analyze selective continuation from phase II to phase III, we linked phase II and phase III trials in our dataset, based on the main intervention, the medical condition to be treated, and the timing.

We read one by one the protocols for all of the phase II trials in the dataset for which at least one *P* value was reported. We considered only phase II trials that were completed before the end of December 2018 to allow for enough time such that a follow-up phase III trial could have been registered by August 2019. From the protocols, we determined the main experimental intervention(s), i.e., the main drug or combination of drugs whose efficacy and safety was to be established, for 1,773 phase II trials.

We considered a phase II trial as continued if we could link it to at least one phase III trial; that is, if we found at least one phase III trial registered in the database (regardless of whether associated results were reported or not) fulfilling all of the following criteria:

1) Intervention: All drugs being part of at least one of the determined main interventions of the phase II trial appear as listed interventions in the phase III trial. This is either with exactly the same name or with a synonym which the reporting party states to refer to the same drug.
2) Condition: All of the MeSH conditions (21) associated to the phase II trial are also associated to the phase III trial.
3) Timing: The start date of the phase II trial was before the start date of the phase III trial.

For more details on the linking procedure, see *SI Appendix*.

Selection Function. Denote by I_2 a vector collecting the relevant information pertaining to the clinical trial at the end of phase II. It contains the z score, z^{Ph2}, and other variables describing the circumstances of the trial (such as sample size to proxy for statistical power). If the sponsor firm decides to stop the development of the drug, it obtains a payoff of $\underline{V}(I_2) + \underline{\eta}$. In case of continuation into phase III, the firm pays a development cost $c + \eta$. The idiosyncratic payoff and cost shocks $\underline{\eta}$ and η are only observable to the firm, but not to the econometrician. The future payoff is denoted V^{Ph3} and is increasing in the phase III z score, which is uncertain at the time of the decision to set up a phase III trial. The firm has an expectation on the distribution of the z score, based on the information available in I_2. The decision of the firm is thus,

$$V^{Ph2}(I_2) = \max\left[\underline{V}(I_2) + \underline{\eta}; -c - \eta + \delta E_{z_3|I_2} V^{Ph3}(z_3)\right],$$

where δ is the discount factor. Assuming that the idiosyncratic shocks $\underline{\eta}$ and η are both independent and identically extreme value distributed, the probability of undertaking a phase III trial is a logistic function (47).

$$\begin{aligned} Prob(continuation) &= \frac{\exp(-c + \delta E_{z_3|I_2} V^{Ph3}(z_3))}{exp(\underline{V}(I_2)) + \exp(-c + \delta E_{z_3|I_2} V^{Ph3}(z_3))} \\ &= logistic(I_2). \end{aligned}$$

Following this model, we use a logistic regression to estimate a selection function that captures selective continuation for industry-sponsored trials. In the sample of phase II z scores, restricted as explained in the section above, we estimate the logistic model

$$\begin{aligned} continuation_i = logistic\,[\,&\alpha + \beta_0(1 - D1_i - D2_i)z_i^{Ph2} + \beta_1 D1_i \\ &+ \beta_2 D2_i + \mathbf{x}_i'\boldsymbol{\gamma} + \phi_{ci} + \tau_{ti} + \varepsilon_i\,], \end{aligned}$$

where $continuation_i$ is a dummy variable which results from our linking of trials across phases and equals one if there is at least one phase III trial matched to a phase II trial to which z-score i belongs (regardless of whether results are reported), and z_i^{Ph2} is the phase II z score associated to a primary outcome. $D1_i$ and $D2_i$ are dummy variables for a statistic to be reported as $z > 3.29$ or $z > 3.89$, respectively. As explained above, those cases are so frequent that we treat them separately.

Moreover, the vector $\mathbf{x}_i$ gathers further control variables which might influence the perceived persuasiveness of phase II results or the economic incentives to carry on with the research on top of the z score. These include the square root of the overall enrollment to each trial (as proxy for the power of the tests), a dummy indicating whether there was a placebo involved in the trial (as opposed to an active comparator), and a dummy indicating whether the *P* value is explicitly declared as adjusted for multiple hypothesis testing. For the last variable, the baseline corresponds to no adjustment of the critical value of the testing procedure or no information provided. We codified this variable manually from the *P*-value descriptions; only 2.9% of the relevant observations are explicitly adjusted.

To account for potential systematic differences across drugs for the treatment of different kinds of conditions, we included condition fixed effects

ϕ_c. For this purpose, we assigned each trial in one of the 15 largest categories of conditions, based on the MeSH terms determined by the curators of the database (21). For more details, see *SI Appendix*.

As registration of trials and reporting of results occurs often with a substantial time lag, we also controlled for a flexible time trend by including completion year fixed effects τ_t.

Summing up, z^{Ph2}, D_1, D_2, $\mathbf{x}$, and ϕ_c correspond to I_2, the information relevant for the continuation decision at the end of phase II, in the model above. The predicted values $\widehat{continuation}_i$ can be interpreted as the probability of a drug moving to phase III conditional on the phase II z score (and other informative covariates observable at the end of phase II).

Kernel Density Estimation. Let $Z_1, Z_2, \ldots, Z_n$ be the sample of z scores in a given group of trials. To estimate the density, we use the standard weighted kernel estimator

$$\hat{f}(z) = \frac{1}{W}\sum_{i=1}^{n}\frac{w_i}{h}K\left(\frac{z-Z_i}{h}\right),$$

where $W = \sum_{i=1}^{n} w_i$, $K(\cdot)$ is the Epanechnikov kernel function, and h is the bandwidth which we choose with the Sheather–Jones plug-in estimator (48). To estimate the actual phase II and phase III densities, we set all weights w_i equal to one. To construct the hypothetical densities controlled for selective continuation, we estimated the kernel density of the phase II statistics, using the predicted probabilities from our selection function as weights, i.e., $w_i = \widehat{continuation}_i$. The resulting densities for precisely reported (i.e., not as inequality) test statistics by different groups of sponsors are plotted in *SI Appendix*, Fig. S4.

This procedure is similar in spirit to the weight-function approach used to test for publication bias in meta-analyses (49, 50), but it allows the weights to depend on more than one variable. The construction of counterfactual distributions by weighted kernel-density estimation has also been used in other strands of the economics literature, e.g., for the decomposition of the effects of institutional and labor-market factors on the distribution of wages (51).

Data Availability. A complete replication package of the econometric analysis presented in the paper, including all data files and our constructed linking of phase II and phase III trials, is deposited at the Harvard Dataverse at https://doi.org/10.7910/DVN/NBLYSW. The clinical trials data mainly analyzed in the paper are freely available for download at http://aact.ctti-clinicaltrials.org/.

ACKNOWLEDGMENTS. This work was supported by European Research Council Grant 295835 EVALIDEA (Designing Institutions to Evaluate Ideas). We thank Marco Bonetti, Tarani Chandola, Sylvain Chassang, Francesco Decarolis, Edina Hot, John Ioannidis, Melissa Newham, Nicolas Serrano-Velarde, Tony Tse, and Deborah Zarin for helpful comments. This paper draws on C.D.'s master's thesis, "P-Hacking in Clinical Trials?", supervised by M.O. and J.A. and defended on April 20, 2017, at Bocconi University.

1. J. P. A. Ioannidis, Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
2. S. Garattini *et al.*, Evidence-based clinical practice: Overview of threats to the validity of evidence and how to minimise them. *Eur. J. Intern. Med.* **32**, 13–21 (2016).
3. A. W. Brown, K. A. Kaiser, D. B. Allison, Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2563–2570 (2018).
4. J. A. DiMasi, R. W. Hansen, H. G. Grabowski, The price of innovation: New estimates of drug development costs. *J. Health Econ.* **22**, 151–185 (2003).
5. A. S. Relman, Economic incentives in clinical investigation. *N. Engl. J. Med.* **320**, 933–934 (1989).
6. M. Angell, Is academic medicine for sale? *N. Engl. J. Med.* **342**, 1516–1518 (2000).
7. J. Lexchin, L. A. Bero, B. Djulbegovic, O. Clark, Pharmaceutical industry sponsorship and research outcome and quality: Systematic review. *BMJ* **326**, 1167–1170 (2003).
8. E. Budish, B. N. Roin, H. Williams, Do firms underinvest in long-term research? Evidence from cancer clinical trials. *Am. Econ. Rev.* **105**, 2044–2085 (2015).
9. I. Boutron, P. Ravaud, Misrepresentation and distortion of research in biomedical literature. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2613–2619 (2018).
10. G. Li *et al.*, Enhancing primary reports of randomized controlled trials: Three most common challenges and suggested solutions. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2595–2599 (2018).
11. D. Fanelli, How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One* **4**, e5738 (2009).
12. J. P. A. Ioannidis. N. S. Young, O. Al-Ubaydli, Why current publication practices may distort science. *PLoS Med.* **5**, e201 (2008).
13. R. J. Simes, Publication bias: The case for an international registry of clinical trials. *J. Clin. Oncol.* **4**, 1529–1541 (1986).
14. P. J Easterbrook, R. Gopalan, J. A. Berlin, D. R Matthews, Publication bias in clinical research. *Lancet* **337**, 867–872 (1991).
15. E. H. Turner, A. M. Matthews, E. Linardatos, R. A. Tell, R. Rosenthal, Selective publication of antidepressant trials and its influence on apparent efficacy. *N. Engl. J. Med.* **358**, 252–260 (2008).
16. C. J. Rosen, The Rosiglitazone story—lessons from an FDA advisory committee meeting. *N. Engl. J. Med.* **357**, 844–846 (2007).
17. G. Harris. Drug maker hid test data, files indicate. *NY Times*, 13 July 2010, Section A, p. 1.
18. D. A. Zarin, T. Tse, Moving toward transparency of clinical trials. *Science* **319**, 1340–1342 (2008).
19. D. A. Zarin, T. Tse, R. J. Williams, T. Rajakannan, Update on trial registration 11 years after the ICMJE policy was established. *N. Engl. J. Med.* **376**, 383–391 (2017).
20. D. A. Zarin, T. Tse, R. J. Williams, R. M. Califf, N. C. Ide, The ClinicalTrials.gov results database—update and key issues. *N. Engl. J. Med.* **364**, 852–860 (2011).
21. A. Tasneem *et al*. The database for aggregate analysis of ClinicalTrials.gov (AACT) and subsequent regrouping by clinical specialty. *PLoS One* **7**, e33677 (2012).
22. R. Rosenthal, The file drawer problem and tolerance for null results. *Psychol. Bull.* **86**, 638–641 (1979).
23. A. Franco, N. Malhotra, G. Simonovits, Publication bias in the social sciences: Unlocking the file drawer. *Science* **345**, 1502–1505 (2014).
24. L. Holman, M. L. Head, R. Lanfear, M. D. Jennions, Evidence of experimental bias in the life sciences: Why we need blind data recording. *PLoS Biol.* **13**, e1002190 (2015).
25. U. Simonsohn, L. D. Nelson, J. P. Simmons, P-curve: A key to the file-drawer. *J. Exp. Psychol. Gen.* **143**, 534–547 (2014).
26. C. H. J. Hartgerink, R. C. M. van Aert, M. B. Nuijten, J. M. Wicherts, M. A. L. M. van Assen, Distributions of *p*-values smaller than .05 in psychology: What is going on? *PeerJ* **4**, e1935 (2016).
27. A. Gerber, N. Malhotra, Do statistical reporting standards affect what is published? Publication bias in two leading political science journals. *Q. J. Polit. Sci.* **3**, 313–326 (2008).
28. A. S. Gerber, N. Malhotra, C. M. Dowling, D. Doherty, Publication bias in two political behavior literatures. *Amer. Polit. Res.* **38**, 591–613 (2010).
29. J. B. De Long, K. Lang, Are all economic hypotheses false? *J. Polit. Econ.* **100**, 1257–1272 (1992).
30. T. D. Stanley, Beyond publication bias. *J. Econ. Surv.* **19**, 309–345 (2005).
31. B. Abel, L. Mathias, M. Sangnier, Y. Zylberberg, Star wars: The empirics strike back. *Am. Econ. J. Appl. Econ.* **8**, 1–32 (2016).
32. I. Guedj, D. Scharfstein. Organizational scope and investment: Evidence from the drug development strategies and performance of biopharmaceutical firms (NBER Working Paper 10933, National Bureau of Economic Research, Cambridge, MA, 2004).
33. J. Lev Krieger. Trials and terminations: Learning from competitors' R&D failures (Harvard Business School Working Paper 18-043, Harvard Business School, Boston, MA, 2017).
34. C. Cunningham, F. Ederer, M. Song, Killer acquisitions. http://doi.org/10.2139/ssrn.3241707 (19 April 2020).
35. G. Z. Jin, P. Leslie, Reputational incentives for restaurant hygiene. *Am. Econ. J. Microecon.* **1**, 237–267 (2009).
36. D. Mayzlin, Y. Dover, J. Chevalier, Promotional reviews: An empirical investigation of online review manipulation. *Am. Econ. Rev.* **104**, 2421–2455 (2014).
37. P. Azoulay, A. Bonatti, J. L. Krieger, The career effects of scandal: Evidence from scientific retractions. *Res. Policy* **46**, 1552–1569 (2017).
38. M. D. Cattaneo, M. Jansson, X. Ma, Simple local polynomial density estimators. *J. Am. Stat. Assoc.*, 10.1080/01621459.2019.1635480 (2019).
39. K. E. Meyer, A. van Witteloostuijn, S. Beugelsdijk, What's in a *p*? Reassessing best practices for conducting and reporting hypothesis-testing research. *J. Int. Bus. Stud.* **48**, 535–551 (2017).
40. J. Adda, C. Decker, M. Ottaviani, Replication data for: P-hacking in clinical trials and how incentives shape the distribution of results across phases. Harvard Dataverse. https://doi.org/10.7910/DVN/NBLYSW. Deposited 15 November 2019.
41. N. J. DeVito, S. Bacon, B. Goldacre, Compliance with legal requirement to report clinical trial results on ClinicalTrials.gov: A cohort study. *Lancet* **395**, 361–369 (2020).
42. S. Matthews, A. Postlewaite, Quality testing and disclosure. *Rand J. Econ.* **16**, 328–340 (1985).
43. M. Dahm, P. González, N. Porteiro, Trials, tricks and transparency: How disclosure rules affect clinical knowledge. *J. Health Econ.* **28**, 1141–1153 (2009).
44. E. Henry, Strategic disclosure of research results: The cost of proving your honesty. *Econ. J.* **119**, 1036–1064 (2009).
45. A. M. Polinsky, S. Shavell, Mandatory versus voluntary disclosure of product risks. *J. Law Econ. Organ.* **28**, 360–379 (2010).
46. E. Henry, M. Ottaviani, Research and the approval process: The organization of persuasion. *Am. Econ. Rev.* **109**, 911–955 (2019).
47. D. McFadden, Modeling the choice of residential location. *Transp. Res. Rec.* **673**, 72–77 (1978).
48. S. J. Sheather, M. C. Jones, A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Stat. Soc. B* **53**, 683–690 (1991).
49. V. L. Hedges, Modeling publication selection effects in meta-analysis. *Statist. Sci.* **7**, 246–255 (1992).
50. I. Andrews, M. Kasy, Identification of and correction for publication bias. *Am. Econ. Rev.* **109**, 2766–2794 (2019).
51. J. DiNardo, N. M. Fortin, T. Lemieux, Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica* **64**, 1001–1044 (1996).