Why so many significant Phase 3 results in clinical trials?*

Jérôme Adda† Christian Decker‡ Marco Ottaviani§

June 10, 2019

Abstract

Planning and execution of clinical research and publication of results should conform to the highest ethical standards, 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 the data. Analyzing p-values reported to the *ClinicalTrials.gov* registry with two different methodologies, we find suspicious patterns only for results from trials conducted by smaller industry sponsors, with presumably less reputation at stake. First, a density discontinuity test reveals an upward jump at the classical threshold for statistical significance for Phase 3 results by small industry sponsors, suggesting some selective reporting. Second, we find an excess mass of significant results in Phase 3 compared to Phase 2. However, we can explain almost completely this excess mass for large industry sponsors, once we account for the incentives to selectively continue research through a novel technique linking trials across phases. In contrast, for trials sponsored by small pharmaceutical companies, selective continuation of trials economizing on research costs only explains one third of the increase in the number of significant results from Phase 2 to Phase 3. The different pattern depending on the type of sponsor suggests that reputational concerns can mitigate the short-run economic incentives that undermine integrity in reporting results of clinical research.

Keywords: Clinical trials; Drug development; Selective reporting; p-Hacking; Economic incentives in research

*Funding by the European Research Council through grant 295835 (EVALIDEA) is gratefully acknowledged. We thank Marco Bonetti, Tarani Chandola, Sylvain Chassang, Francesco Decarolis, Edina Hot, John Ioannidis, Tony Tse, and Deborah Zarin for helpful comments. All authors have contributed equally. The authors declare no competing interests. A complete replication package is available upon request from the authors. This paper draws on Christian Decker's master thesis "P-Hacking in Clinical Trials?", supervised by Marco Ottaviani and Jérôme Adda, and defended on April 20, 2017 at Bocconi University.

†Department of Economics and IGIER, Bocconi University, Via Roberto Sarfatti 25, 20136 Milan, Italy. Phone: +39–02–5836–5572. E-mail: jerome.adda@unibocconi.it.

‡Department of Economics, University of Zurich, Schönberggasse 1, 8001 Zurich, Switzerland. Phone: +41–44–634–61–26. E-mail: christian.decker@econ.uzh.ch.

§Department of Economics and IGIER, Bocconi University, Via Roberto Sarfatti 25, 20136 Milan, Italy. Phone: +39–02–5836–3385. E-mail: marco.ottaviani@unibocconi.it.

The evidence produced in clinical trials is susceptible to many kinds of biases [1–3]. While some of these biases could occur accidentally, even unbeknownst to the investigators who carry out the studies, other more egregious 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 *U.S. Food and Drug Administration* (FDA). Given the sizeable research and development costs involved [4] and the lure of large potential profits, investigators end up suffering from conflicts of interest [5–8] and pressure to withhold or "beautify" unfavorable results [9, 10] or even fabricate or falsify data [11].

In the 1990s and 2000s many medical scholars started 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. This is the first such analysis, building on the literature that investigates "p-hacking", publication bias, and the "file-drawer problem"[22, 23] for academic journal publications in a number of fields, including psychology [24, 25], political science [26, 27], and economics [28–30].

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 [31–33] and reporting bias. While the previous literature focused mostly on scientific publications in academic journals for which pre-publication 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.

Our focus is on pre-approval interventional superiority studies on drugs carried out as Phase 2 and Phase 3 trials. Trials in Phase 2 investigate drug safety and efficacy, typically with a small sample of experimental subjects. Phase 3 trials investigate 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 Food and Drug Administration (FDA). To facilitate the analysis, we transform the p-values back to 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 allows us to investigate both the overall shape of the distribution and the region around the thresholds for statistical significance more easily (see SI Appendix for further information on the data and the p-z transformation).

The Distribution of z-Scores: Discontinuity Tests

Figure 1 displays density estimates of the constructed z-statistics for tests performed for primary outcomes of Phase 2 and Phase 3 trials. We present results for all trials in panel A and then give the break down by affiliation of the lead sponsor: non-industry in panel B (NIH, US federal agencies, universities, etc.), top ten industry in panel C (the ten pharmaceutical companies in the sample with largest revenues in 2016 [35]: Pfizer, Novartis, Roche, Merck & Co., Sanofi, Johnson & Johnson, Gilead Sciences, GlaxoSmithKline, AbbVie, and Amgen), and small industry in panel D (the remaining smaller pharmaceutical companies).

Next, we diagnose 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: (1) a spike in the density function right above 1.96, (2) a discontinuity of the density for values over 1.96, and (3) an excess mass of significant results in Phase 3 compared to Phase 2. Our findings are as follows:

1. We detect no spikes in the densities (or discontinuities in the distribution functions) right 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 [36] as well as economics [30] and business studies [37]. Thus, the distribution of z-scores from *ClinicalTrials.gov* appears to be more natural and credible compared to results reported for publications in scientific journals. This lack of egregious evidence of manipulation of results is a reassuring first piece of good news about the integrity of clinical trials.

2. We investigate the presence of a discontinuity in the density of z-statistics with a test that relies on a simple local polynomial density estimator [34]. The densities for Phase 2 trials are smooth and do not show a significant upward shift at the 1.96 threshold in all cases. In contrast, the densities of z-statistics for industry-sponsored (both small and top ten) Phase 3 trials display a break at 1.96. The break is statistically significant only for Phase 3 trials undertaken by small pharmaceutical companies (panel D), with a persistent upward shift to the right of the threshold, indicating an abnormal amount of significant results. See SI Appendix for details. This pattern is suggestive of *selective reporting*, i.e., strategic concealment of some non-significant results.

3. Figure 1 indicates an excess mass of favorable results over the 1.96 threshold in Phase 3 compared to Phase 2. More generally the Phase 3 distribution of z-statistics stochastically

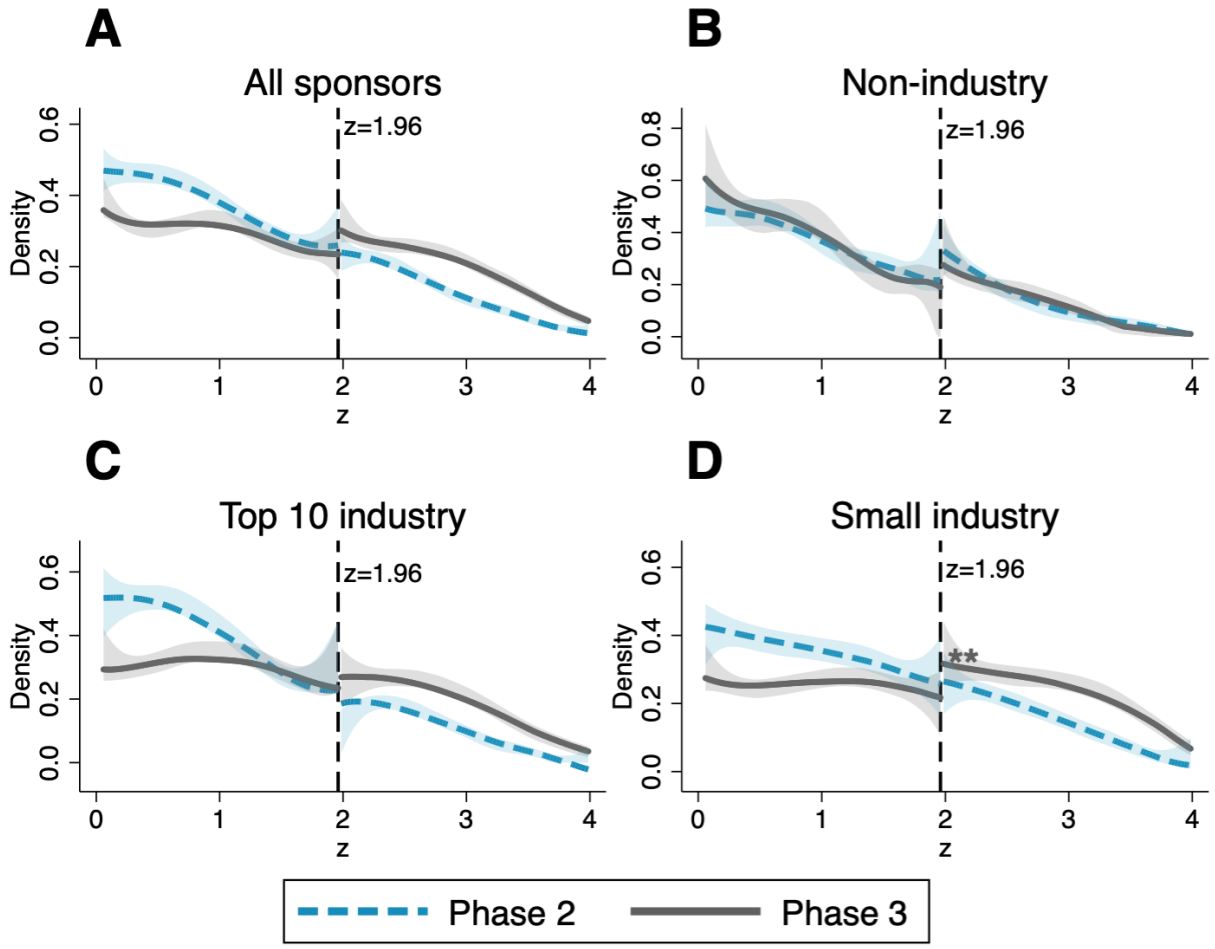

Figure 1: Comparison of Phase 2 and Phase 3 densities of the z-score and tests for discontinuity at $z = 1.96$, depending on affiliation of lead sponsor. Density estimates of the constructed z-statistics for primary outcomes of Phase 2 (dashed blue lines) and Phase 3 (solid grey lines) trials. The shaded areas are 95%-confidence bands and the vertical lines at 1.96 correspond to the threshold for statistical significance at 0.05 level. Sample sizes: A: Phase 2 $n = 3,277$, Phase 3 $n = 3,137$; B: Phase 2 $n = 984$, Phase 3 $n = 598$; C: Phase 2 $n = 1,122$, Phase 3 $n = 1,122$; D: Phase 2 $n = 1,171$, Phase 3 $n = 1,417$. Significance levels for discontinuity tests [34]: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

dominates the Phase 2 distribution; thus more favorable results are more likely to be observed in Phase 3 than in Phase 2. Dominance is particularly strong for industry-sponsored trials (Panels C and D). Even though not as alarming as a spike at the significance threshold, this pattern appears to be suspicious. Whereas only 36.0% of Phase 2 trial results by non-industry sponsors fall above 1.96 (and 35.1% respectively for Phase 3, a difference that is not statistically significant), for industry-sponsored trials the fraction of significant results rises to 46.7% in Phase 2 and even to 71.0% in Phase 3.

Recall that the analysis above considers only p-values associated to primary outcomes of trials. These results constitute the main measure for success of the treatment being trialled, for both the investigators themselves and the evaluating authorities. We provide a similar analysis for p-values from lower-stake secondary outcomes (Figure S1 and Table S4). The densities for all groups of sponsors and both phases do not display any discontinuities at the significance threshold. Moreover, the excess mass of significant results from industry-sponsored trials in Phase 3 relative to Phase 2 is much smaller compared to the distribution for primary outcomes. We find 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 3 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 3 trials could be interpreted as evidence of tampering (*p-hacking*) or non-disclosure of negative results (*selective reporting*). However, this conclusion would be premature before taking a careful look at the dynamic incentives underlying clinical research, as we set out to do.

An alternative explanation for the excess mass of significant results in Phase 3 relative to Phase 2 is *selective continuation* of drug testing to the next phase only when initial results are sufficiently encouraging. *Selective continuation* saves on costly clinical research and is thus socially desirable. Also, from an ethical point of view, no further trials with volunteer patients should be conducted when a drug is highly unlikely to have a positive impact. We outline such a model of the sponsor's continuation decision in Materials and Methods.

To identify the impact of *selective continuation*, we develop a procedure to link Phase 2 and Phase 3 trials in our dataset based on the main intervention (i.e., the tested drug or combination of drugs), the condition to be treated, and the timing. This procedure is illustrated in Figure 2 and explained in detail in the SI Appendix. A given Phase 2 trial may fall into two different categories: it may have no corresponding Phase 3 trial with the same intervention and same condition, or it may have one or multiple matches in Phase 3. The resulting linked data, which we make available to the research community, is a key input in the methodology we develop to estimate a selection

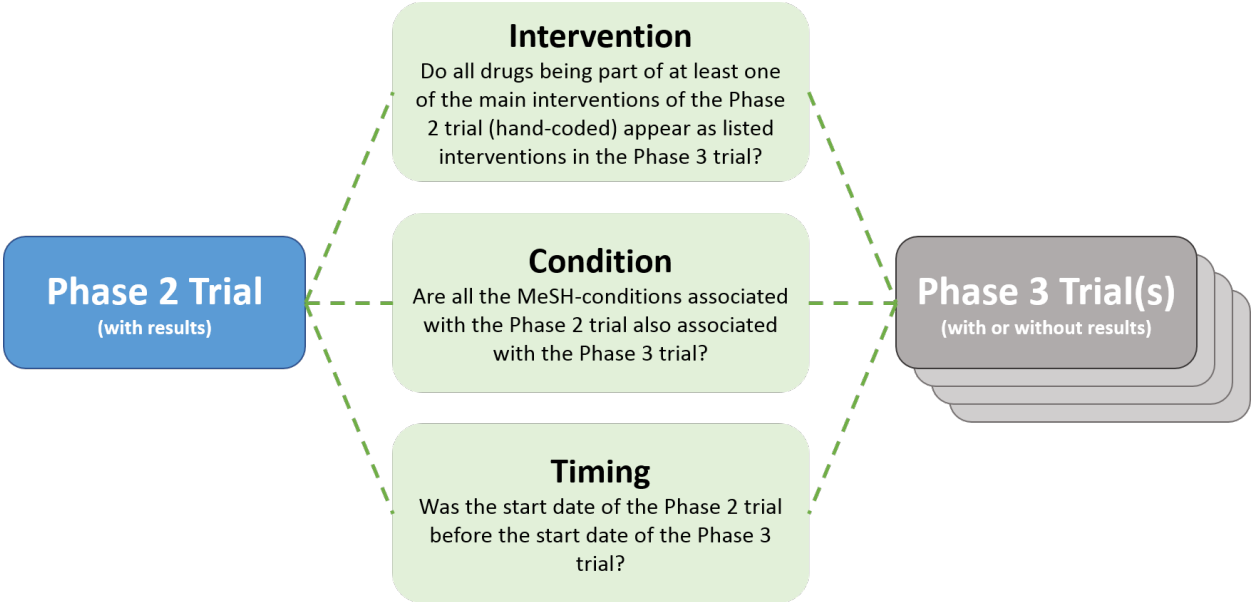

Figure 2: Linking Phase 2 and Phase 3 trials. We consider a Phase 2 trial as continued if we found at least one Phase 3 trial registered in the database (regardless of whether associated results are reported or not) fulfilling all three criteria. See SI Appendix for a more detailed description of the linking procedure.

function capturing *selective continuation* for industry-sponsored trials.

Following our model of the firm's continuation decision, we estimate the selection function with a logistic regression of a dummy variable indicating if there is at least one match among the Phase 3 trials in the database (regardless of whether Phase 3 results are reported or not) on the Phase 2 z-score. We control for adjustment for multiple hypothesis testing, a time trend, and other covariates that might influence the perceived persuasiveness of Phase 2 results (square root of overall enrollment to each trial a proxy for power of the statistical tests, active comparator vs. placebo) or the economic incentives to undertake research (market size 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 of a drug moving to Phase 3 conditional on the information available at the end of Phase 2: the Phase 2 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 estimate 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.

Column 1 in Table 1 shows the estimated logit coefficients for all industry sponsored trials. A higher z-score in Phase 2 significantly increases the probability of continuation to Phase 3. Column 2 and 3 show the results when considering only trials sponsored by small or respectively the ten

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 2 z-score	0.326***	0.269***	0.341**
	(0.0793)	(0.0900)	(0.137)
Dummy for Phase 2 z-score reported as "$z > 3.29$"	0.979***	0.765**	1.437***
	(0.287)	(0.381)	(0.436)
Dummy for Phase 2 z-score reported as "$z > 3.89$"	1.219***	0.732	1.788***
	(0.354)	(0.448)	(0.499)
Constant	-1.285***	-0.573	-2.745***
	(0.422)	(0.484)	(0.613)
Controls	yes	yes	yes
Observations	3,229	1,640	1,580
No. of trials	940	567	372

Notes: Included controls: square root of the overall enrollment, dummy for placebo comparator, logarithm of market size, dummy for multiple hypothesis testing adjustment, and post 2011 dummy. Standard errors in parentheses are clustered at the trial level; significance levels (based on a two-sided t-test): * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

largest industry sponsors. The estimates exhibit two patterns that suggest that larger companies are more selective in which research projects to continue. First, the overall share of matched trials is lower for large industry sponsors, captured by the smaller constant. In the context of our model of the firm's continuation decision, the absolute value of the estimated constant can be interpreted as the opportunity cost of continuing a specific project. This cost is greater for large sponsors with many alternative projects. Second, the coefficients for the Phase 2 z-score and the two dummies are larger. Hence, an increase in the Phase 2 z-score leads to a larger increase in the estimated continuation probability for large sponsors compared to small sponsors.

In Table S5 we report estimates of the same logistic model when considering the Phase 2 z-scores associated to secondary outcomes instead of primary outcomes. Notwithstanding the much larger sample size, none of the coefficients for the z-score is statistically significant. 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 2 and Phase 3

Under the assumption that, conditional on our control variables, the expected z-statistic in Phase 3 equals the z of a similar Phase 2 trial, we can construct a hypothetical Phase 3 distribution for primary outcomes accounting for *selective continuation*. To do so, we estimate the kernel density of Phase 2 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 2 trial. The resulting counterfactual density can be compared to the actual Phase 2 and Phase 3 densities which we estimate using a standard unweighted kernel estimator.

Since the selection function is increasing in the Phase 2 z-score, the counterfactual z-density rotates counter-clockwise, increasing the share of significant results (see Figure S2). To calculate the overall share of significant results under the hypothetical regime, we combine 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 decompose the difference in the share of significant results in Phase 2 and Phase 3 into two parts: *selective continuation* and an unexplained residual. As illustrated in Figure 3 and Table S6, when we consider all industry sponsored trials, *selective continuation*, i.e., economizing on the cost of trials that are not promising enough, accounts for more than half of the difference, leaving unexplained 44.2% of the difference.

Next, we repeat the estimation procedure separately for trials sponsored by large and small industry. The difference is the share of significant results between Phase 2 and Phase 3 is similar for both groups (22.7 percentage points for top 10 industry vs. 22.4 percentage points for small indus-

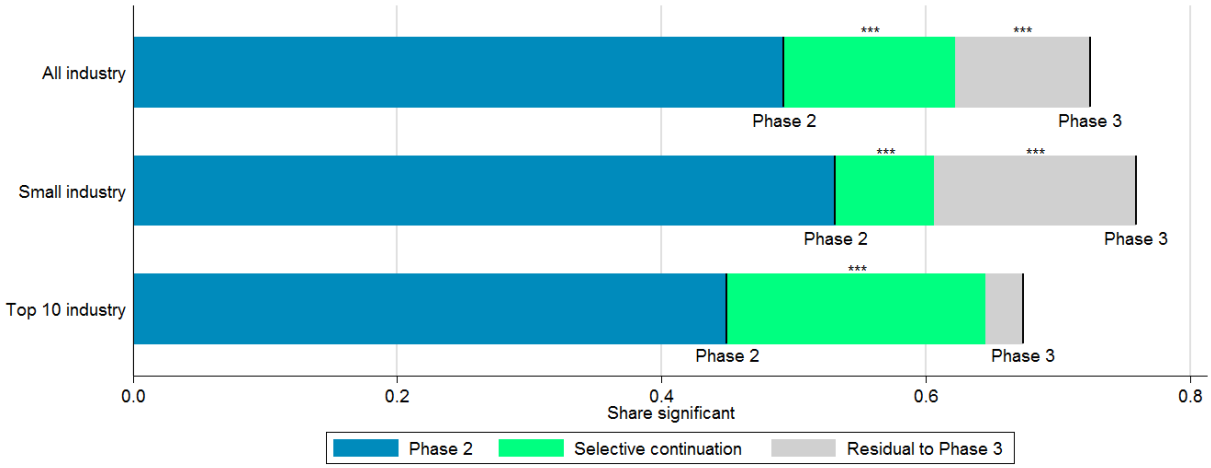

Figure 3: Selection-based decomposition of the difference in significant results from primary outcomes between Phase 2 and Phase 3, depending on affiliation of lead sponsor. Phase 2 and 3 lines represent the shares of trials with a p-value below 5 percent (or equivalently a z-score above 1.96). The green segments represent the parts of the differences explained by *selective continuation*, based on counterfactuals constructed from the Phase 2 distribution. For precise numbers and sample sizes see Table S6. Significance levels for the differences (based on a two-sided t-test): * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

try). However, trials sponsored by the ten largest companies report a smaller share of significant results in both Phase 2 (45.0% vs. 53.2%) and Phase 3 (67.4% vs. 75.9%). The difference between the prediction controlled for *selective continuation* and actual Phase 3 shrinks to 2.8 percentage points and is not statistically significant anymore. Thus, for top 10 industry sponsors *selective continuation* can explain almost the entire difference and our methodology suggests no indication of *selective reporting* or potential tampering.

To the contrary, as already suggested by the estimates of the selection function, for small sponsors *selective continuation* is less pronounced and can account for only 32.5% of the difference between Phase 2 and Phase 3. For small sponsors, we are left with a statistically significant unexplained residual of 15.3 percentage points. Recall that this is also the only group of sponsors for which the Phase 3 z-density exhibits a statistically significant discontinuity at the 1.96 threshold.

Discussion

Overall, the distribution of z-scores from *ClinicalTrials.gov* does not reveal egregious evidence for manipulation of results reported to the registry. However, the residual of excessive significant Phase 3 results unexplained by our methodology might be attributed to some *selective reporting* (withholding unfavorable studies entirely), or tampering with the data, or p-hacking (withholding

single unfavorable results). The fact that this suspicious pattern, like the discontinuity in the Phase 3 density at 1.96, appears especially for smaller sponsors is consistent with economic incentives. Similar patterns have been found for manipulation of online reviews for hotels, depending on whether they are affiliated to large chains [38].

Larger companies should naturally be more concerned about their reputation, as they have more to lose in case they get caught with shady practices. Consequently, their incentives to report results completely and truthfully are stronger than for the fringe, which is potentially less concerned about reputational losses associated with allegations of *selective reporting* or tampering. These incentives are also reflected in the internal disclosure policies that most large companies have established in recent years.

Our results suggest that reputational concerns can mitigate short-run economic incentives that potentially undermine integrity in reporting the results of clinical research. These findings contribute to the understanding of the efficient organization of science and the research process more generally [39–43]. Reputational concerns can be sufficient to discipline larger organizations, whereas smaller organizations might need additional regulatory supervision. To conclude, disclosure regulations of clinical research should pay particular attention to smaller industry sponsors for whom the discipline of reputational concerns has less bite.

Materials and Methods

Database for Aggregate Analysis of ClinicalTrials.gov (AACT)

The *Database for Aggregate Analysis of ClinicalTrials.gov* (AACT) was launched in September 2010 to extend the ability of researchers to use the *ClinicalTrials.gov* data [19–21]. The database aggregates all the data of the registry and makes it available for free bulk download. The project is administered by the *Clinical Trials Transformation Initiative* (CTTI), 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 43 sub-tables with information on timing, conditions, interventions, facilities, locations, sponsors, investigators, responsible authorities, eligible participants, outcome measures, adverse events, results, and descriptions of the 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 analyze the largest possible portion of the data on an aggregate level. Given the aggregate level of the analysis, rather than using the estimated coefficients, we focus on p-values, the only measures reported uniformly and comparably for many trials, independent of their characteristics and the statistical method used for the analysis.

For the study, we use the AACT database as of August 14, 2018. We concentrate on Phase 2 and Phase 3 interventional (as opposed to observational) superiority (as opposed to non-inferiority) 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 drop the trials of the sponsor *Colgate Palmolive* which reported p-values exactly equal to 0.05 for 137 out of its 150 results. We attribute 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 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.

Altogether, we obtain a sample of 10,678 p-values from tests performed on primary outcomes of 4,237 trials. These single p-values constitute the units of observation for our analysis. The average number of p-values for primary outcomes per trial is 2.52 (standard deviation 4.62). As a consequence of the *FDA Amendments Act* (FDAAA), the largest part of our results data pertains to trials conducted after 2007.

p-z Transformation

We transform 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. More details can be found in the SI Appendix.

Density Discontinuity Tests

We implement 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, Jansson, and Ma [34]. This test builds on a local polynomial density estimation technique that avoids pre-binning of the data. More details on the testing procedure and supplementary results can be found in the SI Appendix.

Linking Phase 2 and Phase 3 Trials

To analyze *selective continuation* from Phase 2 to Phase 3, we link the Phase 2 trials in our dataset for which we have results to Phase 3 trials.

We read one by one the protocols for all the Phase 2 trials in the dataset for which p-values are reported and which were completed before 2018. With this restriction on the completion date, there could potentially be a follow-up registered before August 2018. This method allowed us to determine for 1,461 trials the main experimental intervention(s), i.e., the main drug or combination of drugs whose efficacy and safety is to be established. As conditions for the matching we use the MeSH terms determined by the curators for the purpose of making the *ClinicalTrials.gov* webpage searchable [21], disregarding overly generic categories such as simply "Disease". We consider a Phase 2 trial as matched if we found at least one Phase 3 trial registered in the database (regardless of whether associated results are reported or not) fulfilling all of the following criteria:

1. All drugs being part of at least one of the determined main interventions of the Phase 2 trial appear as listed interventions in the Phase 3 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. All the MeSH-conditions associated with the Phase 2 trial are also associated with the Phase 3 trial.

3. The start date of the Phase 2 trial was before the start date of the Phase 3 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 2. 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 firm decides to stop the development of the drug it gets a payoff of $\underline{V}(I_2)$ and an unobserved payoff $\underline{\eta}$. In case of continuation into Phase 3, the firm pays a development cost $c+\eta$, where η is only observable to the firm. The future payoff is denoted V^{Ph3} and is increasing in the Phase 3 z-score, which is uncertain at the time of the decision to set up a Phase 3 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 unobserved cost shocks $\underline{\eta}$ and η are both iid and extreme value distributed, the probability of undertaking a Phase 3 trial is a logistic function [44]

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

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 2 z-scores, restricted as explained in the section above, we estimate the logistic model

$$continuation_i = logistic\left[\alpha+\beta_0(1-D1_i-D2_i)z_i^{Ph2}+\beta_1 D1_i+\beta_2 D2_i+\mathbf{x}_i'\gamma+\varepsilon_i\right],$$

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 3 trial matched to Phase 2 trial to which z-score i belongs (regardless of whether results are reported), and z_i^{Ph2} is the Phase 2 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 2 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), the logarithm of the market size for different classes of conditions (determined from the total Medicare D expenditure), 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.

As registration of trials and reporting of results occurs often with a substantial time lag, we also control for a time trend by including a dummy variable indicating whether the trial was completed after 2011.

Summing up, z^{Ph2}, D_1, D_2, and $\mathbf{x}$ correspond to I_2, the information relevant for the continuation decision at the end of Phase 2, in the model above. The predicted values $\widehat{continuation}_i$ can be interpreted as the probability of a drug moving to Phase 3 conditional on the Phase 2 z-score (and other informative covariates observable at the end of Phase 2).

Kernel Density Estimation

Let $Z_1, Z_2, \ldots, Z_n$ be the sample of z-score 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 the bandwidth which we choose with the Sheather-Jones plug-in estimator [45]. To estimate the actual Phase 2 and Phase 3 densities, we set all weights w_i equal to one. To construct the hypothetical densities controlled for *selective continuation*, we estimate the kernel density of the Phase 2 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 Figure S2.

This procedure is similar in spirit to the weight function approach used to test for publication bias in meta-analyses [46], 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 [47].

References

1. Ioannidis, J. P. A. Why most published research findings are false. *Plos Med* **2** (2005).
2. Garattini, S. *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. Brown, A. W., Kaiser, K. A. & Allison, D. B. Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proc Natl Acad Sci USA* **115,** 2563–2570 (2018).
4. DiMasi, J. A., Hansen, R. W. & Grabowski, H. G. The price of innovation: new estimates of drug development costs. *J Health Econ* **22,** 151–185 (2003).
5. Relman, A. S. Economic incentives in clinical investigation. *N Engl J Med* **320,** 933–934 (1989).
6. Angell, M. Is academic medicine for sale? *N Engl J Med* **342,** 1516–1518 (2000).
7. Lexchin, J., Bero, L. A., Djulbegovic, B. & Clark, O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *BMJ* **326,** 1167–1170 (2003).
8. Budish, E., Roin, B. N. & Williams, H. Do firms underinvest in long-term research? evidence from cancer clinical trials. *Am Econ Rev* **105,** 2044–2085 (2015).
9. Boutron, I. & Ravaud, P. Misrepresentation and distortion of research in biomedical literature. *Proc Natl Acad Sci USA* **115,** 2613–2619 (2018).
10. Li, G. *et al.* Enhancing primary reports of randomized controlled trials: three most common challenges and suggested solutions. *Proc Natl Acad Sci USA* **115,** 2595–2599 (2018).
11. Fanelli, D. How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *Plos One* **4,** 1–11 (2009).
12. Young, N. S., Ioannidis, J. P. A. & Al-Ubaydli, O. Why current publication practices may distort science. *Plos Med* **5,** 1–5 (2008).
13. Simes, R. J. Publication bias: the case for an international registry of clinical trials. *J Clin Oncol* **4,** 1529–1541 (1986).
14. Easterbrook, P., Gopalan, R., Berlin, J. & Matthews, D. Publication bias in clinical research. *Lancet* **337,** 867–872 (1991).
15. Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A. & Rosenthal, R. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med* **358,** 252–260 (2008).
16. Rosen, C. J. The Rosiglitazone story – lessons from an FDA advisory committee meeting. *N Engl J Med* **357,** 844–846 (2007).
17. Harris, G. Drug maker hid test data, files indicate. *New York Times (July 13, 2010),* A1 (2010).

18. Zarin, D. A. & Tse, T. Moving toward transparency of clinical trials. *Science* **319,** 1340–1342 (2008).

19. Zarin, D. A., Tse, T., Williams, R. J. & Rajakannan, T. Update on trial registration 11 years after the ICMJE policy was established. *N Engl J Med* **376,** 383–391 (2017).

20. Zarin, D. A., Tse, T., Williams, R. J., Califf, R. M. & Ide, N. C. The ClinicalTrials.gov results database – update and key issues. *N Engl J Med* **364,** 852–860 (2011).

21. Tasneem, A. *et al.* The database for aggregate analysis of ClinicalTrials.gov (AACT) and subsequent regrouping by clinical specialty. *Plos One* **7,** 1–12 (2012).

22. Rosenthal, R. The file drawer problem and tolerance for null results. *Psychol Bull* **86,** 638–641 (1979).

23. Franco, A., Malhotra, N. & Simonovits, G. Publication bias in the social sciences: unlocking the file drawer. *Science* **345,** 1502–1505 (2014).

24. Simonsohn, U., Nelson, L. D. & Simmons, J. P. P-curve: a key to the file-drawer. *J Exp Psychol Gen* **143,** 534–547 (2014).

25. Hartgerink, C. H., van Aert, R. C., Nuijten, M. B., Wicherts, J. M. & van Assen, M. A. Distributions of *p*-values smaller than .05 in psychology: what is going on? *PeerJ* **4,** e1935 (2016).

26. Gerber, A. & Malhotra, N. Do statistical reporting standards affect what is published? Publication bias in two leading political science journals. *Q J Polit Sci* **3,** 313–326 (2008).

27. Gerber, A. S., Malhotra, N., Dowling, C. M. & Doherty, D. Publication bias in two political behavior literatures. *Amer Polit Res* **38,** 591–613 (2010).

28. De Long, J. B. & Lang, K. Are all economic hypotheses false? *J Polit Econ* **100,** 1257–1272 (1992).

29. Stanley, T. D. Beyond publication bias. *J Econ Surv* **19,** 309–345 (2005).

30. Brodeur, A., Lé, M., Sangnier, M. & Zylberberg, Y. Star wars: the empirics strike back. *Am Econ J Appl Econ* **8,** 1–32 (2016).

31. Guedj, I. & Scharfstein, D. *Organizational scope and investment: evidence from the drug development strategies and performance of biopharmaceutical firms* NBER Working Paper 10933 (2004).

32. Krieger, J. L. *Trials and terminations: learning from competitors' R&D failures* Harvard Business School Working Paper 18-043 (2017).

33. Cunningham, C., Ederer, F. & Ma, S. *Killer acquisitions* SSRN Working Paper 3241707 (2019).

34. Cattaneo, M. D., Jansson, M. & Ma, X. Manipulation testing based on density discontinuity. *Stata J* **18,** 234–261(28) (2018).

35. EvaluatePharma. World preview 2017. *https://info.evaluategroup.com/rs/607-YGS-364/ images/WP17.pdf* (2017).

36. Holman, L., Head, M. L., Lanfear, R. & Jennions, M. D. Evidence of experimental bias in the life sciences: why we need blind data recording. *Plos Biol* **13,** 1–12 (2015).

37. Meyer, K. E., van Witteloostuijn, A. & Beugelsdijk, S. What's in a p? Reassessing best practices for conducting and reporting hypothesis-testing research. *J Int Bus Stud* **48,** 535–551 (2017).

38. Mayzlin, D., Dover, Y. & Chevalier, J. Promotional reviews: an empirical investigation of online review manipulation. *Am Econ Rev* **104,** 2421–2455 (2014).

39. Wuchty, S., Jones, B. F. & Uzzi, B. The increasing dominance of teams in production of knowledge. *Science* **316,** 1036–1039 (2007).

40. Azoulay, P., Graff Zivin, J. S. & Manso, G. Incentives and creativity: evidence from the academic life sciences. *Rand J Econ* **42,** 527–554 (2011).

41. Park, I.-U., Peacey, M. W. & Munafo, M. R. Modelling the effects of subjective and objective decision making in scientific peer review. *Nature* **506,** 93 (2014).

42. Li, D. & Agha, L. Big names or big ideas: do peer-review panels select the best science proposals? *Science* **348,** 434–438 (2015).

43. Wu, L., Wang, D. & Evans, J. A. Large teams develop and small teams disrupt science and technology. *Nature* **566,** 378–382 (2019).

44. McFadden, D. Modeling the choice of residential location. *Transp Res Rec* (1978).

45. Sheather, S. J. & Jones, M. C. A reliable data-based bandwidth selection method for kernel density estimation. *J R Stat Soc Series B Stat Methodol* **53,** 683–690 (1991).

46. Hedges, L. V. Modeling publication selection effects in meta-analysis. *Statist Sci* **7,** 246–255 (1992).

47. DiNardo, J., Fortin, N. M. & Lemieux, T. Labor market institutions and the distribution of wages, 1973-1992: a semiparametric approach. *Econometrica* **64,** 1001–1044 (1996).

Supporting Information (SI)

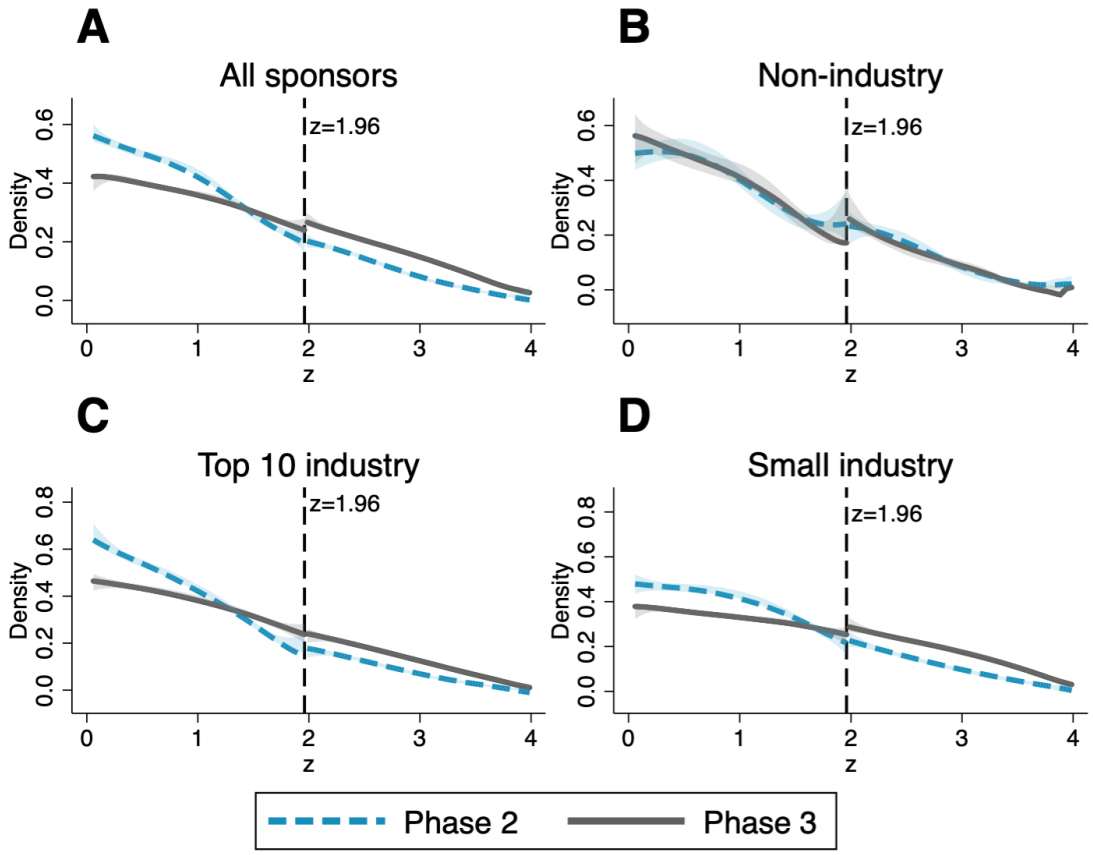

Figure S1: Comparison of Phase 2 and Phase 3 z-score distributions and test for a discontinuity at $z = 1.96$ for *secondary outcomes*, depending on affiliation of lead sponsor. Density estimates of the constructed z-statistics for tests on secondary outcomes of Phase 2 (dashed blue lines) and Phase 3 (solid grey lines) trials. The shaded areas are 95%-confidence bands and the vertical lines at 1.96 correspond to the threshold for statistical significance at 0.05 level. Sample sizes: A: Phase 2 $n = 15,219$, Phase 3 $n = 20,832$; B: Phase 2 $n = 2,118$, Phase 3 $n = 1,652$; C: Phase 2 $n = 7,071$, Phase 3 $n = 8,587$; D: Phase 2 $n = 6,030$, Phase 3 $n = 10,593$. Significance levels for discontinuity tests [34]: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

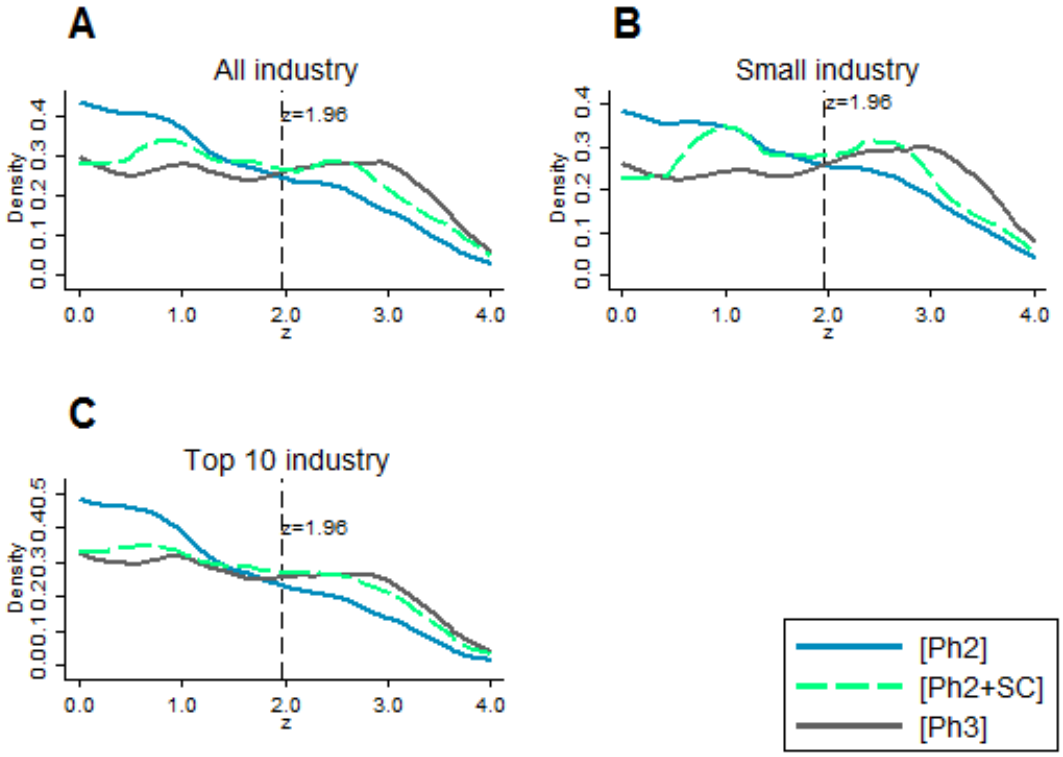

Figure S2: Kernel density estimates for Phase 2 and Phase 3 z-scores and constructed counterfactuals accounting for *selective continuation* , depending on affiliation of lead sponsor. Estimated densities based only on p-values which are reported precisely (i.e. not as inequality). Shorthand notation: Ph2=Phase 2, Ph3=Phase 3, and SC=*selective continuation*. Sample sizes: A: Ph2 $n = 2,293$, Ph3 $n = 2,539$; B: Ph2 $n = 1,171$, Ph3 $n = 1,417$; C: Ph2 $n = 1,122$, Ph3 $n = 1,122$.

Table S1: Market size for treated conditions from total Medicare D spending.

MeSH code	Category	Total Medicare D Spending in 2011 in bn US$
C14	Cardiovascular Diseases	13.215
F03	Mental Disorders	12.336
C18	Nutritional and Metabolic Diseases	8.957
C19	Endocrine System Diseases	8.45
C10	Nervous System Diseases	5.956
C08/C09	Respiratory Tract Diseases/Otorhinolaryngologic Disease	5.945
C06	Digestive System Diseases	4.377
C02	Virus Diseases	3.195
C05	Musculoskeletal Diseases	2.888
C04	Neoplasms	2.64
C12/C13	Male Urogenital Diseases/ Female Urogenital Diseases and Pregnancy Complications	2.262
C20	Immune System Diseases	1.355
C01	Bacterial Infections and Mycoses	1.216
C11	Eye Diseases	1.163
C23	Pathological Conditions, Signs and Symptoms	0.812
C17	Skin and Connective Tissue Diseases	0.683
C15	Hemic and Lymphatic Diseases	0.343
C25	Chemically-Induced Disorders	0.17
C16	Congenital, Hereditary, and Neonatal Diseases and Abnormalities	0.101
C26	Wounds and Injuries	0.049
C03	Parasitic Diseases	0.032

Notes: Details of the calculations provided in the SI Appendix.

Table S2: P-values for tests of density discontinuity at the $z = 1.96$ threshold.

Sponsor	(1) Phase 2	(2) Phase 3
All	0.43	0.14
	(3,277)	(3,137)
Non-industry	0.50	0.37
	(984)	(598)
All industry	0.66	0.34
	(2,293)	(2,539)
Small industry	0.82	0.021**
	(1,171)	(1,417)
Top 10 industry	0.83	0.24
	(1,122)	(1,122)

Notes: P-values result from the density discontinuity test [34], described in detail in the SI Appendix, for primary outcomes; significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Sample sizes in parentheses.

Table S3: Size of discontinuities in the z-density at the significance threshold, as well as at $z = 2.01$ and $z = 2.46$, for industry-sponsored Phase 3 trials.

Sponsor \Cutoff value	(1) z=1.96	(2) z=2.01	(3) z=2.46
All industry	0.055	0.044	0.085
Small industry	0.195**	0.122	0.109
Top 10 industry	-0.102	-0.017	0.016

Notes: Differences of the bias-corrected density estimates to the right and to the left of the respective cutoff, resulting from the density discontinuity test [34], described in detail in the SI Appendix, for primary outcomes. Sample sizes: All industry $n = 2,539$, Small industry $n = 1,417$, Top 10 industry $n = 1,122$. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table S4: P-values for tests of density discontinuity at the z=1.96 threshold – *secondary outcomes*.

Sponsor	(1) Phase 2	(2) Phase 3
All	0.29	0.31
	(15,219)	(20,832)
Non-industry	0.48	0.52
	(2,118)	(1,652)
All industry	0.24	0.18
	(13,101)	(19,180)
Small industry	0.58	0.48
	(6,030)	(10,593)
Top 10 industry	0.45	0.71
	(7,071)	(8,587)

Notes: P-values result from the density discontinuity test [34], described in detail in the SI Appendix, for secondary outcomes; significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Sample sizes in parentheses.

Table S5: Estimates of logit selection function for *selective continuation*, based on *secondary outcomes*.

Sponsor	**(1) All industry**	**(2) Small industry**	**(3) Top 10 industry**
Phase 2 z-score	0.106	0.168	-0.0475
	(0.0874)	(0.107)	(0.145)
Dummy for Phase 2 z-score reported as "$z > 3.29$"	0.550	0.757	0.159
	(0.481)	(0.515)	(0.633)
Dummy for Phase 2 z-score reported as "$z > 3.89$"	0.676	0.276	0.553
	(0.446)	(0.573)	(0.714)
Constant	-0.959	-0.662	-1.865**
	(0.682)	(0.811)	(0.907)
Controls	yes	yes	yes
Observations	14,704	6,810	7,894
No. of trials	587	349	238

Notes: Included controls: square root of the overall enrollment, dummy for placebo comparator, logarithm of market size, dummy for multiple hypothesis testing adjustment, and post 2011 dummy. Standard errors in parentheses are clustered at the trial level; significance levels (based on a two-sided t-test): * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table S6: Selection-based decomposition of the difference in significant results from primary outcomes between Phase 2 and Phase 3, depending on affiliation of lead sponsor.

Share of significant results			
Sponsor	**(1) All industry**	**(2) Small industry**	**(3) Top 10 industry**
[Ph2]	0.493	0.532	0.450
	(0.00777)	(0.0114)	(0.0133)
[Ph3]	0.724	0.759	0.674
	(0.00591)	(0.00680)	(0.00969)
[Ph2+SC]	0.622	0.606	0.645
	(0.0188)	(0.0275)	(0.0527)

Differences			
Sponsor	**(4) All industry**	**(5) Small industry**	**(6) Top 10 industry**
[Ph3]-[Ph2]	0.231***	0.227***	0.224***
	(0.00966)	(0.0133)	(0.0167)
[Ph3]-[Ph2+SC]	0.102***	0.153***	0.0284
	(0.0201)	(0.0283)	(0.0536)
[Ph2+SC]-[Ph2]	0.129***	0.0738***	0.195***
	(0.0175)	(0.0255)	(0.0511)
Observations	8,700	4,917	3,783
Observations Ph2	3,477	1,838	1,639
Observations Ph3	5,223	3,079	2,144
No. of trials Ph2	1,027	632	395
No. of trials Ph3	2,276	1,376	900

Notes: Columns 1-3 display the share of significant results based on kernel density estimates and adjustment for selection, with shorthand notation Ph2=Phase 2, Ph3=Phase 3, and SC=*selective continuation*. Columns 4-6 display the differences in these shares. The standard errors in parentheses are obtained by bootstrapping the whole estimation procedure (500 repetitions); significance levels (based on a two-sided t-test): * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Market Size Data

To obtain a proxy for the expected market size for a newly developed drug, we evaluate the Medicare D spending for existing drugs in 2011 according to information from the *Centers for Medicare & Medicaid Services* publicly available at https://www.cms.gov/Research-Statistics-Data-and-Systems/Research-Statistics-Data-and-Systems.html. *Part D Prescription Drug Event* (PDE) data is provided for a subset ($\sim 70\%$) of Medicare beneficiaries.

We classify manually 1,056 marketed drugs, among which the 420 with the highest Medicare D spending, into the MeSH categories for the treated conditions. Some strongly overlapping categories have been merged. Overall, these drugs make up for 90% of the expenditure on the drugs in the dataset. Table S1 shows the total spending for those categories.

The MeSH categorization allows us to connect these numbers to the trials in the *ClinicalTrials.gov* registry. In case a trial is associated with more than one category, we assign it to the one with the highest spending. The trials for which the MeSH categorization is missing (207 trials) are assigned the mean market size of the remaining sample.

p-z Transformation

Our analysis focuses on the reported p-values for the statistical evaluation of the results of the trials. However, the p-distribution is not ideal for graphical representation, given that the most interesting part of the distribution is the region around the classical significance threshold of 0.05. A proper analysis of this part requires zooming in on the distribution for small p-values (say < 0.1). It is also important to compare the overall p-distributions for different groups of trials. For instance, a clockwise rotation of the density represents a higher share of more significant results. To overcome this problem, we transform the p-values 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 "stretches" the distribution from support $[0,1]$ to the whole positive real axis and it does so the more the smaller the original p-value. Hence, the part around the significant threshold becomes more prominent without losing the other parts of the distribution. A similar transformation has been applied in previous literature on experimental biases across life sciences [36].

Note that the p-values in the dataset originate from diverse statistical procedures (e.g., ANCOVA, ANOVA, Chi-squared-test, mixed models analysis, linear regression, logistic regression, 1-sided t-test, 2-sided t-test, etc.), with test statistics that follow different distributions, some continuous, some discrete. Even though the sample size of the trials is sufficiently large that, according to the Central Limit Theorem, many of the resulting statistics are approximately normally

distributed, in general the actual test-statistic of the trial and our calculated z do not coincide. Nevertheless, the p-z transformation allows us to conveniently compare the results of all trials.

Beyond the graphical aspect, another advantage of looking at z-statistics instead of p-values is that in the usual process of statistical inference the test statistic precedes the p-value. Hence, one can expect that the test statistic is a more immediate object for selection or potential tampering by the researchers. For this reason, we use the z-transformation also for our estimation of the selection functions and the share of significant results.

The Missing Tail of the z-Distribution

Not all p-values in the registry are reported precisely, but some are only stated in comparison to a certain threshold, e.g. $p < 0.05$ or $p > 0.1$. Whereas for most parts of the distribution this is a minor issue and affects only a small number of observations, relative reporting becomes the rule for very low p-values, corresponding to high z-statistics. In particular, 39.9% of the p-values in our sample of tests for primary outcomes are reported as $p < 0.001$ (corresponding to $z > 3.29$) or $p < 0.0001$ (corresponding to $z > 3.89$). There are barely any p-values reported with equality below these thresholds. For the z-distribution, this implies that we know the size of the right tail (i.e., the mass above a certain threshold) but we do not have any information about the exact shape.

For our analysis of the share of significant results, we deal with this issue as following. As indicated in the regression equation, we include the dummies $D1$ for "$z > 3.29$" and $D2$ for "$z > 3.89$" into the estimation of the selection function, so that the probability of continuation is estimated separately for those two cases. Moreover, we include p-values which are reported as exactly zero (as a result of rounding) and hence cannot be transformed into a z-score in the group $D2$. For the few cases in which a z-score is reported as inequality with respect to a level $\bar{z}$ other than 3.29 and 3.89, we replace the respective z with the mean of the precisely reported z-statistics conditional on being above or respectively below $\bar{z}$.

For the discontinuity tests (Figure 1, Figure S1, Table S2, Table S3 and Table S4) and plots of densities (Figure S2), we consider only p-values which are reported precisely (i.e., not as inequality).

Testing for Discontinuities of Distributions and Densities of z-scores

We provide a formal test of discontinuity in the z-score density at the $z = 1.96$ significance threshold. We implement manipulation tests based on a state-of-the-art procedure developed by Cattaneo, Jansson, and Ma [34]. This test builds on a local polynomial density estimation technique that avoids pre-binning of the data . Table S2 shows the p-values of the tests performed on the densities from primary outcomes, depending on the affiliation of the lead sponsors of the trials, as described

in the main text. We do not find any evidence of manipulation for trials in Phase 2. For Phase 3, the p-values are lower, but only significant for trials sponsored by small industry.

Figure 1 in the main text suggests that the breaks we find are not due to a spike, i.e., a concentration of mass right above 1.96 (leading to a discontinuity in both the density and the cumulative distribution function), but due to a persistent upward shift in the density with an increased frequency of results also further to the right of 1.96 (leading to a discontinuity only in the density but not in the cumulative distribution function). To reinforce this claim and distinguish the two cases, we perform further density discontinuity tests with cutoffs 0.05 and 0.5 above the significance threshold, corresponding to $z = 2.01$ and $z = 2.46$, for the group of which we found to display a break at 1.96, i.e., for the industry-sponsored Phase 3 trials.

With this method we can implicitly test for a discontinuity in the cumulative distribution function. If the discontinuity in the density was due to a spike at 1.96, we would expect our test to find a downward jump in the density at some point above. If there was manipulation and all inflated results were concentrated exactly at 1.96 (sharp discontinuity in the cumulative distribution function at 1.96), we should have a sharp downward discontinuity in the density right above the threshold (captured by the test at 2.01). Assuming more realistically that investigators want to push their results above the significance threshold but cannot perfectly target a p-value of 0.05, we would expect an excess mass above 1.96 that slowly vanishes (captured by the test at 2.46). Even in the absence of a sharp discontinuity, also in this case we would expect a downward tendency in the density.

The differences of the bias-corrected density estimates to the right and to the left of the respective cutoffs tabulated in Table S3 do not display such a downward tendency. To the contrary, for small industry sponsors, the differences at 2.01 and 2.46 get smaller and are not statistically significant anymore, but have still a positive sign. These findings confirm that there is a persisting upward shift in the density around the significance threshold, but there is no break in the cumulative distribution function with an excess mass concentrated only right above 1.96.

Similar discontinuity tests for the z-density from secondary outcomes do not display any break at the significance threshold (Figure S1 and Table S4). Moreover, the excess mass of significant results from industry-sponsored trials in Phase 3 relative to Phase 2 is much smaller compared to the distribution for primary outcomes.

Linking Phase 2 and Phase 3 Trials

To analyze *selective continuation* from Phase 2 to Phase 3, we link the Phase 2 trials in our dataset for which we have results to Phase 3 trials. We match Phase 2 and Phase 3 trials based on the specific drug or combination of drugs (in medical terms *intervention*) the trial investigates for the treatment of a specific disease (in medical terms *condition*). This is not such a straightforward exercise to implement for two reasons:

- The dataset is a mere digitization of the reported trial protocols. Hence, most variables are not well codified and have non-generic entries. Moreover, even though the information on interventions and conditions of the trials for which results are reported is rather complete, the cells in the reporting forms are interpreted in varying ways by different investigators. For instance, in the specification of a trial's intervention, in many cases all the drugs involved in the trial are inserted in one cell, without specifying whether the drugs are given as a combination or separately to different arms of the trial. Often, it is not specified which drug constitutes the experimental treatment rather than the control. Hence, it is not possible to mechanically identify a trial's main experimental intervention. As an additional complication, many drugs appear in the data with different names; some times the drugs are referred by the chemical composition, while other times by their commercial name.

- The process of drug development is not linear in the sense that we usually do not have one Phase 2 trial followed by one Phase 3 trial and then a request for FDA approval. In most cases, there is a number of Phase 2 trials looking at similar but potentially slightly different interventions/conditions, such as different drug dosages, different characteristics of eligible patients, or different control interventions. These Phase 2 trials are followed by an even larger number of Phase 3 trials with similar interventions/conditions but slightly varying specifications.

We address these hurdles in the following way. We read one by one the protocols for all the Phase 2 trials in the dataset for which p-values are reported and which were completed before 2018. With this restriction on the completion date, there could potentially be a follow-up registered before August 2018. This method allowed us to determine for 1,461 trials the main experimental intervention(s), i.e., the main drug or combination of drugs whose efficacy and safety is to be established. As conditions for the matching we use the MeSH terms determined by the curators for the purpose of making the *ClinicalTrials.gov* webpage searchable [21], disregarding overly generic categories such as simply "Disease". We consider a Phase 2 trial as matched if we found at least one Phase 3 trial registered in the database (regardless of whether associated results are reported or not) fulfilling all of the following criteria:

1. All drugs being part of at least one of the determined main interventions of the Phase 2 trial appear as listed interventions in the Phase 3 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. All the MeSH-conditions associated with the Phase 2 trial are also associated with the Phase 3 trial.

3. The start date of the Phase 2 trial was before the start date of the Phase 3 trial.

This linking is not perfect, for instance because it disregards whether all the drugs in the Phase 3 trial were part of one combination in one arm. Moreover, we do not take into consideration other potentially important details of the trials like the exact population of eligible patients. However, given the limitations of the data, this procedure appears reasonably accurate. We manage to link 34.9% of the industry-sponsored Phase 2 trials in our restricted dataset to at least one Phase 3 trial. These numbers are roughly in line with the ones reported in previous studies [4] and on the FDA webpage [48]. For non-industry sponsored trials, however, reporting in Phase 3 is very meager and we can find Phase 3 matches for only 17.1% of the Phase 2 trials. Given this low number and the fact that there are no significant differences between the Phase 2 and Phase 3 distribution for non-industry sponsors to begin with, we investigate selection only for industry-sponsored trials.

Note that criterion 3 considers only the start dates of the trials. It might appear to be more intuitive to require the completion date of the Phase 2 trial to be prior to the start date of the Phase 3 trial. Indeed, most of our linked trials fulfill also this stronger condition. However, in some cases this condition is too strong. That is, some Phase 3 trials start before the corresponding Phase 2 trials are fully completed. For instance, some Phase 2 results on long-run impacts might still be pending but the collected evidence is already strong enough for the investigators to start a Phase 3 trial. Moreover, we consider the reported start dates to be more reliable. The responsible parties might have incentives to report a later completion date than the actual, in order to meet the requirements for timely reporting of results.

Background on *ClinicalTrials.gov*

ClinicalTrials.gov is an online registry of clinical research studies in human volunteers. The website is maintained by the *National Library of Medicine* (NLM) at the *National Institutes of Health* (NIH) in collaboration with the *U.S. Food and Drug Administration* (FDA). It was established in February 2000 with the aim to increase transparency in clinical research. Initially, the registry contained only trials to test the efficacy of new experimental drugs for serious or life-threatening diseases or conditions and registration was mainly voluntary. For more information on the history of the registry, related policies, and laws, see https://clinicaltrials.gov/ct2/about-site/history (accessed Jun 23, 2017).

In 2007 the requirements for registration of trials were extended substantially through the *FDA Amendments Act* (FDAAA) [49]. Even though in January 2017 those rules have been redefined more precisely [50], in the following we will refer to the regulation of Section 801 of the FDAAA which was the legislation in force at the time when the great majority of the data in our analysis was generated. The main criteria a trial must meet to be affected by this regulation are the following:

- initiated after September 27, 2007, or initiated on or before that date and still ongoing as of December 26, 2007;

- controlled clinical investigation of drugs, biologics or medical devices other than Phase 1 trials and small feasibility studies;
- the trial has one or more sites in the United States or it involves drugs, biologics or medical devices manufactured in the United States.

For more details see https://clinicaltrials.gov/ct2/manage-recs/fdaaa (accessed Jun 23, 2017).

If these criteria apply, the responsible party (i.e., the sponsor or the principal investigator of the trial) must register the trial and provide the required information no later than 21 days after enrollment of the first participant. In case the investigated drug, biologic, or device is approved, licensed, or cleared by FDA, moreover, the responsible party must submit some basic summary results of the trial no later than twelve months after the completion date. For this purpose, in September 2008, *ClinicalTrials.gov* was extended by a publicly accessible results database. Thereby, an even higher level of transparency is supposed to be reached, even though there are still many loopholes in the legislation [18]. For instance, the required level of details of the results is not clearly defined and also the aforementioned exemptions for Phase 1 studies and not-approved products are criticized. In all the other cases that do not meet the stated criteria, registering and reporting of results is voluntary.

The FDAAA establishes penalties for non-compliance of up to $10,000 per day. However, no enforcement has yet occurred [51, 52]. Assessing compliance rates is not easy because the aforementioned exemptions and imprecisions in the FDAAA legislation complicate identifying which trials are applicable. An early algorithm-based study [51] shows that only 13.4% of applicable clinical trials registered on *ClinicalTrials.gov* between 2008 and 2012 reported results in a timely fashion and only 38.3% reported results at any time at all. However, in a manual review of a subsample of trials the same authors [51] found that their methodology based on assumptions about the approval status of the drug tended to underestimate reporting rates. More recent studies document for a sample of 329 industry-sponsored Phase 2-4 US trials completed or terminated 2007-2009 a result reporting rate to *ClinicalTrials.gov* of 58% by December 2014 [53] and an increase of the overall reporting rate for applicable trials from 58% to 72% in the last two years until September 2017, driven not by fear of sanctions but by public pressure on the responsible parties [52].

Considering the missing enforcement of the FDAAA regulations, lack of reporting does not necessarily mean that the responsible parties intend to hide their results, but rather that they just do not take the time to go through the lengthy reporting process. In this light, notwithstanding the legal requirements, for the purpose of our analysis reporting of results should be seen as mostly voluntary.

Since January 2017 the improved "Final Rule" is in place (hence, it does not affect the great majority of the trials we analyze), addressing many loopholes and broadening the scope of the 2007 legislation [50]. However, the FDA's efforts to police compliance are still very limited [52].

Beyond the disclosure mandate, the FDAAA raised public awareness about the importance of transparency in clinical research and led many large pharmaceutical companies and research institutions to develop internal disclosure policies [51, 52].

Several studies in the medical literature assess the quality of the data reported to the registry and the results database along different dimension, e.g., information about scientific leadership [54], consistency of reported primary outcomes [55, 56], comparisons to results published in academic journals [53, 57], and the provision of Individual Participant Data (IPD) [58]. All these studies, as well as overall assessments by the curators of the database [19, 20], find ambiguous results and see scope for improvement [59].

The *Database for Aggregate Analysis of ClinicalTrials.gov* (AACT) was launched in September 2010 to extend the ability of researchers to use the *ClinicalTrials.gov* data [19–21]. The database aggregates all the data of the registry and makes it available for free bulk download.

The biggest challenge when working with the data is that, as a mere digitization of the trial protocols, most variables have non-generic entries and many of them contain large bodies of text. Moreover, when looking at the data more carefully, it turns out that the reporting parties do not always interpret the different cells in the reporting form in the same way. For instance, when it comes to specifying the intervention of a trial, in many cases all the drugs involved in a trial are inserted in one cell without specifying whether they are given as a combination or separately to different arms of the trial. Furthermore, the way in which reporting parties indicate which drug constitutes the experimental treatment and which one the control is not uniform. Often, one can find a clarification in other parts of the protocol. Similar issues arise with almost all the reported variables. Even though for some trials the reported content is complete and when reading the whole study protocol embedded in the context gives a clear picture, reporting is often made in different cells making it almost impossible to evaluate large parts of the data mechanically at an aggregate level, even with natural language processing algorithms.

Consequently, we are forced to either codify the data by hand (like the main intervention of Phase 2 trials which we use for our linking of trials across phases) or restrict attention to characteristics that are codified uniformly among all the trials in the database. The latter is the case for numerical entries or entries that allow only for a finite, pre-specified number of answers (e.g., binary variables).

To facilitate the browsing of the database, the administrators ran natural language processing algorithms on the descriptions of trials to match interventions and conditions with *Medical Subject Heading* (MeSH) terms [21]. The MeSH thesaurus is a controlled list of vocabulary produced by the *National Library of Medicine* and used for indexing, cataloging, and searching for biomedical and health-related information. It is the standard for classification in medical research. We use the MeSH terms assigned for the conditions to classify trials. We also use this classification to link trials across different phases and for the determination of the market size.

References

48. U.S. Food and Drug Administration. The drug development process. *https://www.fda.gov/ForPatients/Approvals/Drugs/ucm405622.htm* (2018).

49. Wood, A. J. Progress and deficiencies in the registration of clinical trials. *N Engl J Med* **360,** 824–830 (2009).

50. Zarin, D. A., Tse, T., Williams, R. J. & Carr, S. Trial reporting in ClinicalTrials.gov — the final rule. *N Engl J Med* **375,** 1998–2004 (2016).

51. Anderson, M. L. *et al.* Compliance with results reporting at ClinicalTrials.gov. *N Engl J Med* **372,** 1031–1039 (2015).

52. Piller, C. & Bronshtein, T. Faced with public pressure, research institutions step up reporting of clinical trial results. *STAT (January 8, 2018) https://www.statnews.com/2018/01/09/clinical-trials-reporting-nih/* (2018).

53. Zarin, D. A., Tse, T., Williams, R. J., Rajakannan, T. & Fain, K. M. *Evaluation of the ClinicalTrials.gov results database and its relationship to the peer-reviewed literature* in *Eighth International Congress on Peer Review and Scientific Publication, Chicago, IL, September 2017* (2017).

54. Sekeres, M. *et al.* Poor reporting of scientific leadership information in clinical trial registers. *Plos One* **3,** 1–6 (2008).

55. Mathieu, S., Boutron, I., Moher, D., Altman, D. & Ravaud, P. Comparison of registered and published primary outcomes in randomized controlled trials. *Jama* **302,** 977–984 (2009).

56. Ramagopalan, S. *et al.* Prevalence of primary outcome changes in clinical trials registered on ClinicalTrials.gov: a cross-sectional study. *F1000Research* **3** (2014).

57. Ross, J. S., Mulvey, G. K., Hines, E. M., Nissen, S. E. & Krumholz, H. M. Trial publication after registration in ClinicalTrials.gov: a cross-sectional analysis. *Plos Med* **6,** 1–9 (2009).

58. Zarin, D. A. & Tse, T. Sharing individual participant data (IPD) within the context of the trial reporting system (TRS). *Plos Med* **13,** 1–8 (2016).

59. Dickersin, K. & Mayo-Wilson, E. Standards for design and measurement would make clinical research reproducible and usable. *Proc Natl Acad Sci USA* **115,** 2590–2594 (2018).