PNAS
www.pnas.org

Supplementary Information for

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

Jérôme Adda, Christian Decker, and Marco Ottaviani

Correspondence to Marco Ottaviani.
E-mail: marco.ottaviani@unibocconi.it

This PDF file includes:

Supplementary text
Figs. S1 to S4
Tables S1 to S8
References for SI reference citations

Supporting Information Text

p-z Transformation. Our analysis focuses on the reported p-values for the statistical evaluation of trial results. However, the p-density is not particularly well suited to perform discontinuity tests at the significance threshold because it is highly nonlinear in the relevant range. Neither is the p-density well suited for graphical representation, given that it is not possible to display both the region around the significant threshold and the overall distribution conveniently in the same graph.

To overcome these problems, 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 the $[0, 1]$ interval to the whole positive real axis with smaller p-values being stretched more. Hence, the region close to the significance threshold becomes more prominent without losing the other parts of the distribution. Moreover, in the range around the significance threshold the z-density is close to linear, making it easier to identify discontinuities (1, 2). A similar transformation has been applied in the literature on experimental biases across life sciences (3).

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.

To alleviate concerns that the discontinuity we find in the z-density for phase III trials by small industry sponsors (panel D of Figure 1 in the main text) may be driven by the specific transformation we choose, we provide density discontinuity tests for industry sponsored trials with p-values transformed to one-sided instead of two-sided test statistics. That is, $z_{1-sided} = -\Phi^{-1}(p)$. The results, displayed in Figure S2 and Table S4, resemble closely those relying on the transformation to two-sided z-scores. We still find a sizable and statistically significant upward shift at the classical significance threshold for phase III trials by small sponsors. Also, the densities for phase III top ten and phase II (both types of sponsors) are smooth.

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, 30.8% 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, Figures S1–S3, and Tables S2–S5) and plots of densities (Figure S4), we consider only p-values which are reported precisely (i.e., not as inequality).

The Definition of Large vs. Small Industry Sponsors. As our analysis relies on the estimation of densities, comparing trials by different groups of sponsors requires a discrete split of the sample. We focus on the impact of the size of the sponsoring corporations on their incentives. Therefore, we need a definition of "large vs. small" sponsors. In our main analysis, we compare the top ten sponsors in terms of 2018 revenues to the remaining smaller sponsors. These top ten are the ten companies in italics in the first column of Table S1. This particular definition is not only salient but also splits the sample of p-values roughly in half, maximizing statistical power in both subsamples. This is of particular importance for the density discontinuity tests, which require large sample sizes to be reliable.

To check robustness, we repeat our analysis for 56 alternative definitions of "large" and show that our main results hold across this wide range of alternatives for splitting the sample. As displayed in Table S1, we rank sponsors not only by their 2018 revenues (column 1), but also by the volume of prescription drug sales in 2018 (column 2), R&D spending in 2018, and the number of trials reported to the registry (column 4). It is not surprising that these four rankings are correlated. For each of the four criteria, we create fourteen different definitions of "large vs. small": top seven vs. remainder, top eight vs. remainder, and so on up to top twenty vs. remainder. Hence, overall we have 14*4=56 different definitions, one of which is the top ten revenues definition we use for our main analysis.

Figure S1 shows histograms of the p-values of density discontinuity tests across these 56 different definitions. In panels A and B we can see that the phase II and phase III z-densities for large industry sponsor never exhibit significant breaks at the 1.96 threshold, no matter which definition we use. As shown in panel C, for a number of definitions we find a significant discontinuity for phase II small industry, but at the same time in many cases we have p-values far above 0.05. Phase III small

industry (panel D) is the only subgroup for which we find a significant break in our main specification. For the great majority of alternative definition, this finding is confirmed and the p-value never exceeds 0.146.

We also repeat the counterfactual exercise of predicting the share of significant phase III results based on *selective continuation* for each of the 56 different definitions. As discussed in the main text, the different patterns between large and small industry sponsors are robust across this wide range of alternative ways to define "large" sponsors (Figure 4, panels B and C).

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 (1, 2). 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 II. For phase III, the p-values are lower, but when splitting the sample 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 industry sponsored phase III trials, for which we found a break at 1.96.

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 have still a positive sign, the latter being even statistically significant. These findings confirm that there is a persistent 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 just above 1.96.

Similar discontinuity tests for the z-density from secondary outcomes do not display any noteworthy break at the significance threshold (Figure S3 and Table S5). Moreover, the excess mass of significant results from industry-sponsored trials in phase III relative to phase II is much smaller compared to the distribution for primary outcomes.

Linking Phase II and Phase III Trials. To analyze *selective continuation* from phase II to phase III, we link phase II and phase III trials in our dataset, based on the main intervention, the medical condition to be treated, and the timing. This is not such a straightforward exercise to implement for two reasons:

- The AACT dataset is a mere digitization of the reported trial protocols. Hence, most variables are not well codified and have non-generic entries. 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 differently by different reporting parties. 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 II trial followed by one phase III trial and then a request for FDA approval. In most cases, there is a number of phase II 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 II trials are typically followed by an even larger number of phase III 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 II trials in the dataset for which at least one p-value is reported and which were completed before end of December 2018. With this restriction on the completion date, there could potentially be a follow-up phase III trial registered before August 2019. From the protocols, we determine the main experimental intervention(s), i.e., the main drug or combination of drugs whose efficacy and safety is to be established, for 1,773 phase II trials. As indication of the medical condition the trials address, we use the Medical Subject Headings (MeSH) terms determined by the curators for the purpose of making the *ClinicalTrials.gov* webpage searchable (4), disregarding overly generic categories such as simply "Disease".

We consider 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 are 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 the MeSH-conditions 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.

This linking is not perfect, for instance because it disregards whether all the drugs in the phase III trial were part of one combination in one arm. Moreover, we do not take into consideration other 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 33.3% of the industry-sponsored phase II trials in our restricted dataset to at least one phase III trial. These numbers are in line with the ones reported in previous studies (5) and on the FDA webpage (6). For non-industry sponsored trials, however, reporting in phase III is very meager and we can find phase III matches for only 18.0% of the phase II trials. Given this low number and the fact that there are no significant differences between the phase II and phase III 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 II trial to be prior to the start date of the phase III 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 III trials start before the corresponding phase II trials are fully completed. For instance, some phase II results on long-run impacts might still be pending but the collected evidence is already strong enough for the investigators to start a phase III 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.

MeSH Condition Fixed Effects and Market Size Data. To account for potential systematic differences across drugs for the treatment of different kinds of conditions, we include condition fixed effects in the estimation of the selection functions for *selective continuation*. For this purpose, we assign each trial in one of the 15 largest categories of conditions (in terms of frequency in our data), based on the MeSH terms determined by the curators of the database (4). These categories are displayed in Table S8. Some strongly overlapping categories have been merged. Trials that could not be assigned to a specific group or belong to one of the smaller groups constitute the omitted category. In case a trial is associated with more than one category, we assign it to the one with the largest expected market size.

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. Overall, these drugs make up for 90% of the expenditure on the drugs in the dataset. Table S8 shows also the total spending by category.

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) (7). Even though in January 2017 those rules have been redefined more precisely (8), 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. According to https://clinicaltrials.gov/ct2/manage-recs/fdaaa (accessed Jun 23, 2017), 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 I 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.

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. Since September 2008, these submitted results are publicly accessible in the *ClinicalTrials.gov* results database so as to reach an even higher level of transparency. However, there are some loopholes in the legislation (9); for instance, the required level of details of the results is not clearly defined and phase I trials and trials of not-approved products are exempt. 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 (10–13). 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 (10) 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 (10) found that their methodology based on assumptions about the approval status of the drug tended to underestimate reporting rates. Later studies document for a sample of 329 industry-sponsored phase II-IV US trials completed or terminated 2007-2009 a result reporting rate to *ClinicalTrials.gov* of 58% by December 2014 (14) and an increase of the overall reporting rate for applicable trials from 58% to 72% in the two years before September 2017, driven not by fear of sanctions but by public pressure on the responsible parties (11).

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 (8). However, the FDA's efforts to police compliance are still very limited (11–13). 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 (10, 11, 13).

The most recent and complete evaluation of compliance with the FDAAA Final Rule finds 64.5% of industry-sponsored trials to report any results and 50.3% to be fully compliant with the rules; that is, to report results within one year of the primary completion date (12).

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.

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 (15), consistency of reported primary outcomes (16, 17), comparisons to results published in academic journals (14, 18), and the provision of Individual Participant Data (IPD) (19). All these studies, as well as overall assessments by the curators of the database (20, 21), find ambiguous results and see scope for improvement (22).

The biggest challenge when working with the AACT 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, reporting parties do not always interpret the different cells in the reporting form in the same way. For instance, when reporting 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, reporting parties indicate differently which drug constitutes the experimental treatment and which one is the control. Often, one can find a clarification in other parts of the protocol. Similar issues arise with many of the self-reported variables. Even though for most trials the reported content is complete and the whole study protocol embedded in the context gives a clear picture, different parties often report the same information in different cells. This non-uniformity prevents the mechanical evaluation of large parts of the data, even with natural language processing algorithms.

Consequently, we are forced to either codify the data by hand (like the main intervention of phase II 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 are numerical entries or entries that allow only for a finite, prespecified number of answers (e.g., binary variables).

We classify trials and link them across phases based on the MeSH terms associated to the treated conditions. The MeSH thesaurus is a controlled list of vocabulary produced by the *National Library of Medicine* and used for indexing, cataloging, and searching biomedical and health-related information. The MeSH classification is provided by *ClinicalTrials.gov* administrators based on natural language processing algorithms.

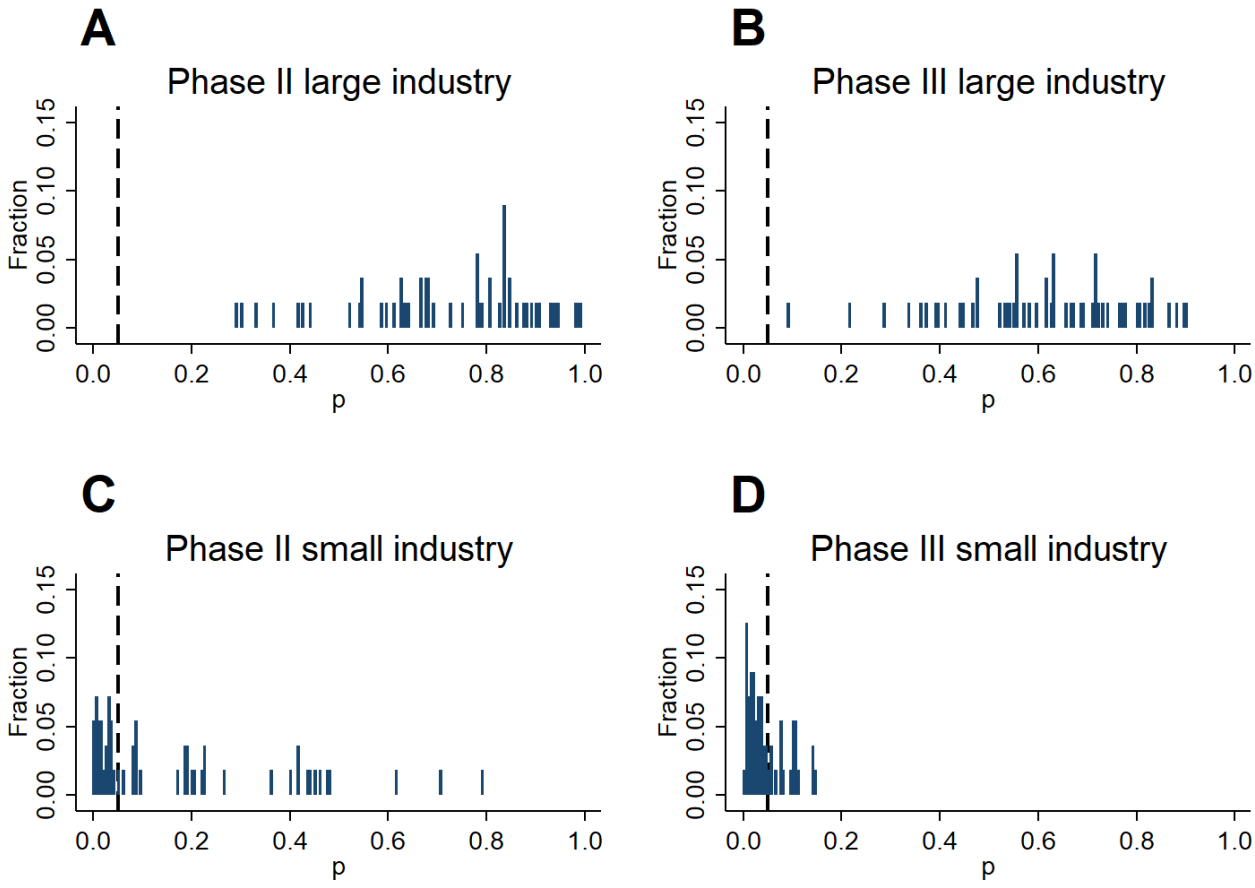

Fig. S1. Robustness check: histograms of p-values from density discontinuity tests at $z = 1.96$ across 56 different definitions for large vs. small industry sponsors. The p-values result from discontinuity tests (1) at $z = 1.96$ in the densities of constructed z-statistics for primary outcomes. The dashed vertical lines indicate $p = 0.05$. 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 revenue, volume of prescription drug sales in 2018, R&D 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" are created: top seven vs. remainder, top eight vs. remainder, and so on up to top twenty vs. remainder. Further details are provided in the *supplementary text*.

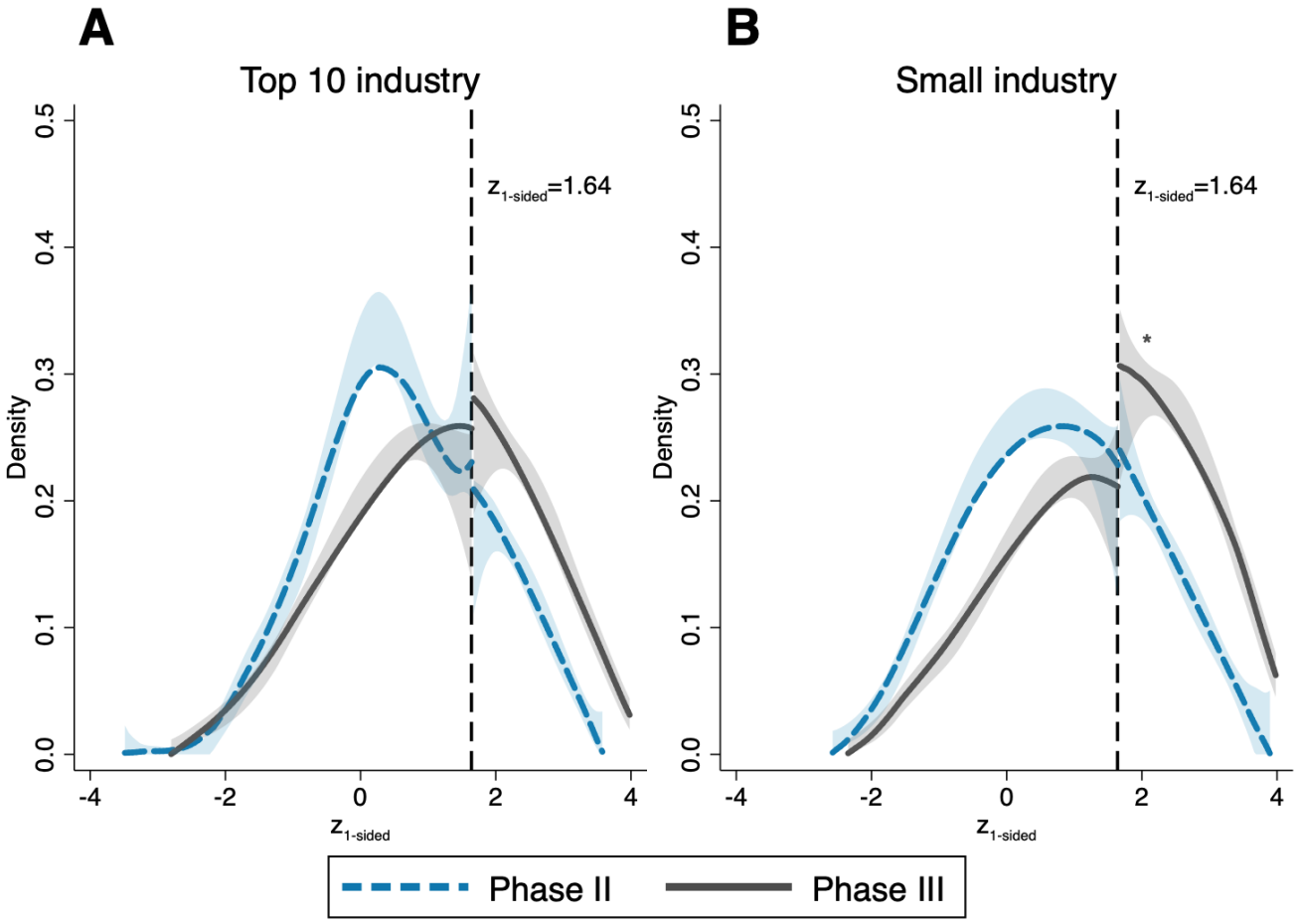

Fig. S2. Robustness check: density discontinuity tests for large and small industry sponsored trials with transformation to *one-sided* test scores. Density estimates of constructed one-sided z-statistics for primary outcomes of phase II (dashed blue lines) and phase III (solid grey lines) trials. The shaded areas are 95%-confidence bands and the vertical lines at 1.64 correspond to the threshold for statistical significance at 0.05 level. Sample sizes: A: $n = 1,332$ (phase II), $n = 1,424$ (phase III); B: $n = 1,450$ (phase II), $n = 1,520$ (phase III). Significance levels for discontinuity tests (1): * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; exact p-values reported in Table S4.

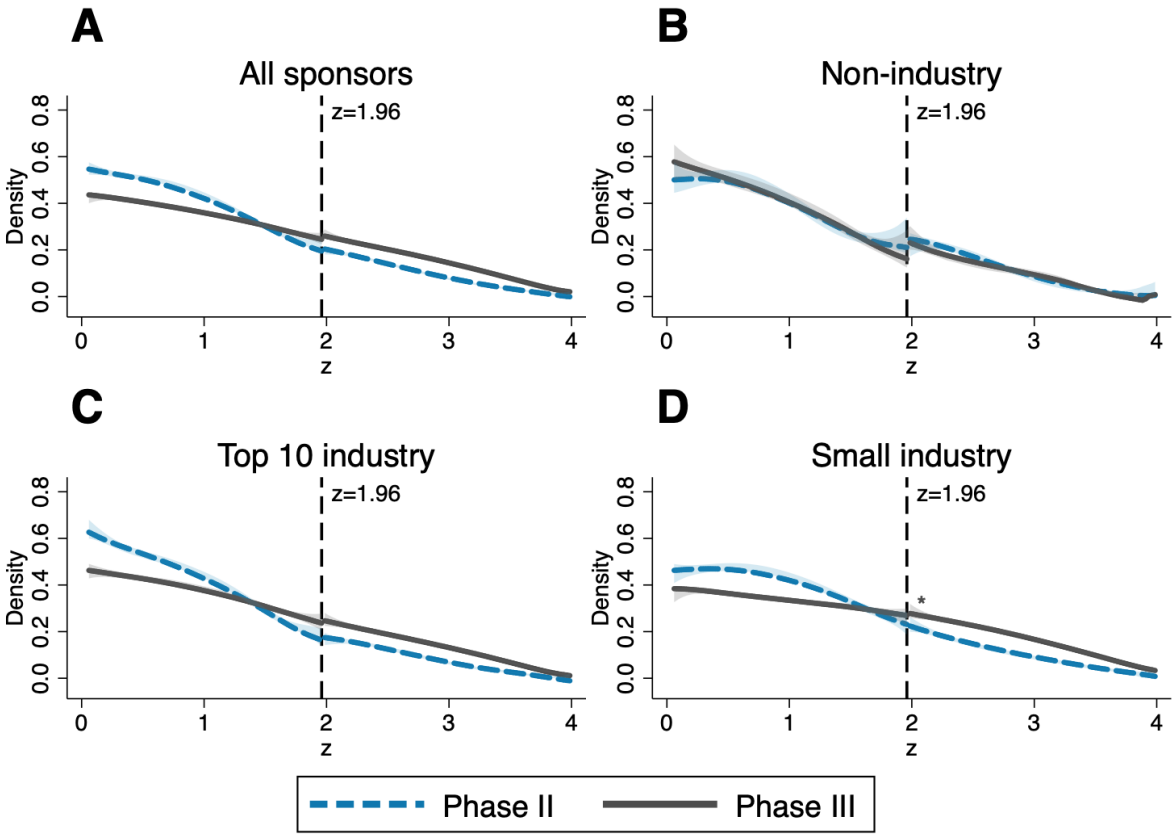

Fig. S3. Comparison of phase II and phase III z-score distributions and test for a discontinuity at $z = 1.96$ for *secondary outcomes*, depending on affiliation of lead sponsor. Density estimates of the constructed z-statistics for tests on secondary outcomes of phase II (dashed blue lines) and phase III (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: $n = 17,840$ (phase II), $n = 25,050$ (phase III); B: $n = 2,553$ (phase II), $n = 2,102$ (phase III); C: $n = 8,579$ (phase II), $n = 11,480$ (phase III); D: $n = 6,672$ (phase II), $n = 11,486$ (phase III). Significance levels for discontinuity tests (1): * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; exact p-values reported in Table S5.

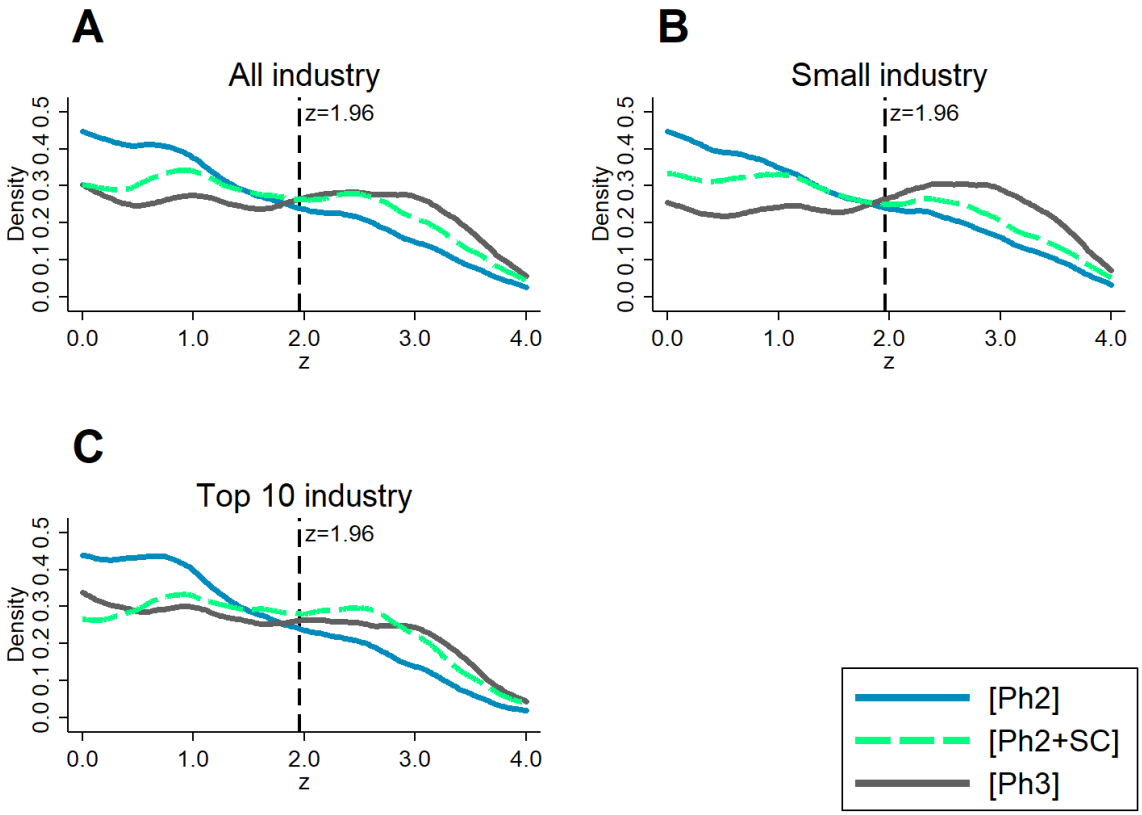

Fig. S4. Kernel density estimates for phase II and phase III z-scores and constructed counterfactuals accounting for *selective continuation* , depending on affiliation of lead sponsor. Estimated densities based only on p-values which are reported precisely (i.e. not as inequality). Shorthand notation: Ph2=phase II, Ph3=phase III, and SC=*selective continuation*. Sample sizes: A: $n = 4,135$ (phase II), $n = 5,957$ (phase III); B: $n = 2,181$ (phase II), $n = 3,209$ (phase III); C: $n = 1,954$ (phase II), $n = 2,748$ (phase III).

Table S1. Ranking of Industry Sponsors by Different Criteria.

Rank	Revenues 2018	Rx Sales 2018	R&D Spending 2018	No. Trials Reported
1	*Johnson & Johnson*	Pfizer	Roche	GlaxoSmithKline
2	*Roche*	Roche	Johnson & Johnson	Pfizer
3	*AbbVie/Abbott Laboratories*	Novartis	Novartis	Merck Sharp & Dohme Corp.
4	*Pfizer*	Johnson & Johnson	Pfizer	Eli Lilly & Co
5	*Novartis*	Merck Sharp & Dohme Corp.	Merck Sharp & Dohme Corp.	Boehringer Ingelheim
6	*Bayer*	AbbVie/Abbott Laboratories	Sanofi	AstraZeneca
7	*GlaxoSmithKline*	Sanofi	AbbVie/Abbott Laboratories	Roche
8	*Merck Sharp & Dohme Corp.*	GlaxoSmithKline	AstraZeneca	Novartis
9	*Sanofi*	Amgen	Bristol-Myers Squibb	Takeda Pharmaceutical
10	*Eli Lilly & Co*	Gilead Sciences	Eli Lilly & Co	Shire
11	Amgen	Bristol-Myers Squibb	GlaxoSmithKline	Amgen
12	Bristol-Myers Squibb	AstraZeneca	Celegne	Bayer
13	Gilead Sciences	Eli Lilly & Co	Gilead Sciences	Sanofi
14	AstraZeneca	Bayer	Amgen	Johnson & Johnson
15	Danaher Corporation	Novo Nordisk	Bayer	Gilead Sciences
16	Boehringer Ingelheim	Takeda Pharmaceutical	Boehringer Ingelheim	Bristol-Myers Squibb
17	Takeda Pharmaceutical	Celegne	Takeda Pharmaceutical	Otsuka Holdings
18	Teva Pharmaceutical Industries	Shire	Biogen	AbbVie/Abbott Laboratories
19	Novo Nordisk	Boehringer Ingelheim	Novo Nordisk	Novo Nordisk
20	Merck KGaA	Allergan	Regeneron Pharmaceuticals	Merck KGaA

Notes: The companies in italics are defined as the top ten industry sponsors in our main analaysis. Known large-scle subsidiaries are grouped with their mother corporation (e.g. Janssen Research & Development as part of Johnson & Johnson). Small companies that may have alliances with (or are later acquired by) larger companies are coded as separate sponsors. Shire was acquired by Takeda Pharmaceutical in early 2019 but we treat the two companies separately, as this acquisitaion happend at the very end of our sample period. Sources: Revenues 2018: https://en.wikipedia.org/wiki/List_of_largest_biomedical_companies_by_revenue (revenues data collected from financial statements on company websites, accessed Oct 23, 2019); Rx (i.e. prescription drugs) Sales 2018 and R&D Spending 2018: (23) (based on data from EvaluatePharma®); No. Trials Reported: own calculations based on *ClinicalTrials.gov* data.

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

Sponsor	(1) Phase II	(2) Phase III
All	0.09*	0.00***
	(3,953)	(3,664)
Non-industry	0.23	0.35
	(1,171)	(720)
All industry	0.30	0.52
	(2,782)	(2,944)
Small industry	0.20	0.032**
	(1,450)	(1,520)
Top 10 industry	0.91	0.67
	(1,332)	(1,424)

Notes: P-values result from the density discontinuity test (1), described in detail in the *supplementary text*, 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 III trials.

Sponsor \Cutoff value	(1) z=1.96	(2) z=2.01	(3) z=2.46
All industry	0.031	0.087**	0.11*
Small industry	0.166**	0.056	0.177**
Top 10 industry	0.029	0.075	0.015

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 (1), described in detail in the *supplementary text*, for primary outcomes. Sample sizes: All industry $n = 2,944$, Small industry $n = 1,520$, Top 10 industry $n = 1,424$. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table S4. Robustness check – transformation to *one-sided* test scores: P-values for tests of density discontinuity at the $z_{1-sided} = 1.64$ threshold.

Sponsor	(1) Phase II	(2) Phase III
Small industry	0.20	0.076*
	(1,450)	(1,520)
Top 10 industry	0.31	0.74
	(1,332)	(1,424)

Notes: P-values result from the density discontinuity test (1), described in detail in the *supplementary text*, for primary outcomes; significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Sample sizes in parentheses.

Table S5. P-values for tests of density discontinuity at the z=1.96 threshold – *secondary outcomes*.

Sponsor	(1) Phase II	(2) Phase III
All	0.54	0.21
	(17,804)	(25,050)
Non-industry	0.34	0.35
	(2,553)	(2,102)
All industry	0.34	0.07*
	(15,251)	(22,948)
Small industry	0.87	0.06*
	(6,672)	(11,468)
Top 10 industry	0.44	0.36
	(8,579)	(11,480)

textitNotes: P-values result from the density discontinuity test (1), described in detail in the *supplementary text*, for secondary outcomes; significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Sample sizes in parentheses.

Table S6. Estimates of logit selection function for *selective continuation*, based on *secondary outcomes*.

Sponsor	(1) All industry	(2) Small industry	(3) Top 10 industry
Phase II z-score	0.109*	0.197**	-0.0612
	(0.0557)	(0.0839)	(0.0674)
Dummy for phase II z-score reported as "$z > 3.29$"	0.465	0.600	0.0737
	(0.416)	(0.628)	(0.461)
Dummy for phase II z-score reported as "$z > 3.89$"	0.512	0.279	0.779**
	(0.353)	(0.395)	(0.351)
Mean dependent variable	0.353	0.360	0.347
Controls	yes	yes	yes
MeSH condition fixed effects	yes	yes	yes
Completion year fixed effects	yes	yes	yes
Observations	17,724	7,502	10,222
No. of trials	720	402	318

Notes: Unit of observation: trial-outcome; included controls: square root of the overall enrollment and dummy for placebo comparator. Categories for condition fixed effects are based on Medical Subject Headings (MeSH) terms associated to the trials (4); for more details, see *supplementary text*. Standard errors in parentheses are clustered at the MeSH condition level; significance levels (based on a two-sided t-test): * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table S7. Selection-based decomposition of the difference in significant results from primary outcomes between phase II and phase III, depending on affiliation of lead sponsor.

Share of significant results			
Sponsor	**(1) All industry**	**(2) Small industry**	**(3) Top 10 industry**
[Ph2]	0.481	0.499	0.460
	(0.0205)	(0.0226)	(0.0319)
[Ph3]	0.721	0.757	0.679
	(0.0149)	(0.0136)	(0.0245)
[Ph2+SC]	0.604	0.573	0.645
	(0.0316)	(0.0406)	(0.0458)

Differences			
Sponsor	**(4) All industry**	**(5) Small industry**	**(6) Top 10 industry**
[Ph3]-[Ph2]	0.241***	0.258***	0.219***
	(0.0259)	(0.0256)	(0.0422)
[Ph3]-[Ph2+SC]	0.117***	0.184***	0.0339
	(0.0354)	(0.0426)	(0.0529)
[Ph2+SC]-[Ph2]	0.123***	0.0746**	0.185***
	(0.0260)	(0.0347)	(0.0410)
Observations	10,092	5,390	4,702
Observations Ph2	4,135	2,181	1,954
Observations Ph3	5,957	3,209	2,748
No. of trials Ph2	1,244	732	512
No. of trials Ph3	2,655	1,544	1,111

Notes: Columns 1-3 display the share of significant results based on kernel density estimates and adjustment for selection, with shorthand notation Ph2=phase II, Ph3=phase III, 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, clustered at the trial level); significance levels (based on a two-sided t-test): * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table S8. Categories for MeSH condition fixed effects with market size determined 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
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
C23	Pathological Conditions, Signs and Symptoms	0.812
C17	Skin and Connective Tissue Diseases	0.683
C25	Chemically-Induced Disorders	0.17
C16	Congenital, Hereditary, and Neonatal Diseases and Abnormalities	0.101

Notes: Details of the calculations provided in the *supplementary text*.

References

1. MD Cattaneo, M Jansson, X Ma, Simple local polynomial density estimators. *J. Am. Stat. Assoc.* (2019).
2. MD Cattaneo, M Jansson, X Ma, Manipulation testing based on density discontinuity. *Stata J.* **18**, 234–261(28) (2018).
3. L Holman, ML Head, R Lanfear, MD Jennions, Evidence of experimental bias in the life sciences: Why we need blind data recording. *Plos Biol.* **13**, 1–12 (2015).
4. A Tasneem, et al., The database for aggregate analysis of ClinicalTrials.gov (AACT) and subsequent regrouping by clinical specialty. *Plos One* **7**, 1–12 (2012).
5. JA DiMasi, RW Hansen, HG Grabowski, The price of innovation: New estimates of drug development costs. *J. Heal. Econ.* **22**, 151–185 (2003).
6. U.S. Food and Drug Administration, The drug development process. *https://www.fda.gov/ ForPatients/Approvals/Drugs/ucm405622.htm* (2018).
7. AJ Wood, Progress and deficiencies in the registration of clinical trials. *N. Engl. J. Med.* **360**, 824–830 (2009).
8. DA Zarin, T Tse, RJ Williams, S Carr, Trial reporting in ClinicalTrials.gov–the final rule. *N. Engl. J. Med.* **375**, 1998–2004 (2016).
9. DA Zarin, T Tse, Moving toward transparency of clinical trials. *Science* **319**, 1340–1342 (2008).
10. ML Anderson, et al., Compliance with results reporting at ClinicalTrials.gov. *N. Engl. J. Med.* **372**, 1031–1039 (2015).
11. C Piller, T Bronshtein, 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).
12. NJ 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).
13. C Piller, FDA and NIH let clinical trial sponsors keep results secret and break the law. *Science* (2020).
14. DA Zarin, T Tse, RJ Williams, T Rajakannan, KM Fain, 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).
15. M Sekeres, et al., Poor reporting of scientific leadership information in clinical trial registers. *Plos One* **3**, 1–6 (2008).
16. S Mathieu, I Boutron, D Moher, D Altman, P Ravaud, Comparison of registered and published primary outcomes in randomized controlled trials. *Jama* **302**, 977–984 (2009).
17. S Ramagopalan, et al., Prevalence of primary outcome changes in clinical trials registered on ClinicalTrials.gov: A cross-sectional study. *F1000Research* **3** (2014).
18. JS Ross, GK Mulvey, EM Hines, SE Nissen, HM Krumholz, Trial publication after registration in ClinicalTrials.gov: A cross-sectional analysis. *Plos Med.* **6**, 1–9 (2009).
19. DA Zarin, T Tse, Sharing individual participant data (IPD) within the context of the trial reporting system (TRS). *Plos Med- .* **13**, 1–8 (2016).
20. DA Zarin, T Tse, RJ Williams, T Rajakannan, Update on trial registration 11 years after the ICMJE policy was established. *N. Engl. J. Med.* **376**, 383–391 (2017).
21. DA Zarin, T Tse, RJ Williams, RM Califf, NC Ide, The ClinicalTrials.gov results database–update and key issues. *N. Engl. J. Med.* **364**, 852–860 (2011).
22. K Dickersin, E Mayo-Wilson, Standards for design and measurement would make clinical research reproducible and usable. *Proc. Natl. Acad. Sci. USA* **115**, 2590–2594 (2018).
23. M Christel, 2019 Pharm Exec 50. *Pharm. Exec.* **39**, 12–19 (2019).