Empirical Evaluation of Very Large Treatment Effects of Medical Interventions

Tiago V. Pereira, PhD
Ralph I. Horwitz, MD
John P. A. Ioannidis, MD, DSc

Context Most medical interventions have modest effects, but occasionally some clinical trials may find very large effects for benefits or harms.

Objective To evaluate the frequency and features of very large effects in medicine.

Data Sources Cochrane Database of Systematic Reviews (CDSR, 2010, issue 7).

Study Selection We separated all binary-outcome CDSR forest plots with comparisons of interventions according to whether the first published trial, a subsequent trial (not the first), or no trial had a nominally statistically significant ($P<.05$) very large effect (odds ratio [OR], ≥5). We also sampled randomly 250 topics from each group for further in-depth evaluation.

Data Extraction We assessed the types of treatments and outcomes in trials with very large effects, examined how often large-effect trials were followed up by other trials on the same topic, and how these effects compared against the effects of the respective meta-analyses.

Results Among 85 002 forest plots (from 3082 reviews), 8239 (9.7%) had a significant very large effect in the first published trial, 5158 (6.1%) only after the first published trial, and 71 605 (84.2%) had no trials with significant very large effects. Nominally significant very large effects typically appeared in small trials with median number of events: 18 in first trials and 15 in subsequent trials. Topics with very large effects were less likely than other topics to address mortality (3.6% in first trials, 3.2% in subsequent trials, and 11.6% in no trials with significant very large effects) and were more likely to address laboratory-defined efficacy (10% in first trials,10.8% in subsequent, and 3.2% in no trials with significant very large effects). First trials with very large effects were as likely as trials with no very large effects to have subsequent published trials. Ninety percent and 98% of the very large effects observed in first and subsequently published trials, respectively, became smaller in meta-analyses that included other trials; the median odds ratio decreased from 11.88 to 4.20 for first trials, and from 10.02 to 2.60 for subsequent trials. For 46 of the 500 selected topics (9.2%; first and subsequent trials) with a very large-effect trial, the meta-analysis maintained very large effects with $P<.001$ when additional trials were included, but none pertained to mortality-related outcomes. Across the whole CDSR, there was only 1 intervention with large beneficial effects on mortality, $P<.001$, and no major concerns about the quality of the evidence (for a trial on extracorporeal oxygenation for severe respiratory failure in newborns).

Conclusions Most large treatment effects emerge from small studies, and when additional trials are performed, the effect sizes become typically much smaller. Well-validated large effects are uncommon and pertain to nonfatal outcomes.

JAMA. 2012;308(16):1676-1684 www.jama.com

MOST EFFECTIVE INTERVENTIONS in health care confer modest, incremental benefits.[1,2] Randomized trials, the gold standard to evaluate medical interventions, are ideally conducted under the principle of equipoise[3]: the compared groups are not perceived to have a clear advantage; thus, very large treatment effects are usually not anticipated. However, very large treatment effects are observed occasionally in some trials. These effects may include both anticipated and unexpected treatment benefits, or they may involve harms.

Large effects are important to document reliably because in a relative scale they represent potentially the cases in which interventions can have the most impressive effect on health outcomes and because they are more likely to be adopted rapidly and with less evidence. Consequently, it is important to know whether, when observed, very large effects are reliable and in what sort of experimental outcomes they are commonly observed. The importance of very large effects has drawn attention mostly in observational studies[4,5] but has not been well studied in randomized evidence. It is unknown how often very large effects are replicated in subsequent trials of the same comparison, disease and outcome. If data observed in 1 experiment happen to be at the extreme of a distribution, subse-

For editorial comment see p 1691.

Author Affiliations: Health Technology Assessment Unit, Institute of Education and Sciences, German Hospital Oswaldo Cruz, Sao Paulo, Brazil (Dr Pereira); GlaxoSmithKline, King of Prussia, Pennsylvania, and Yale University School of Medicine, New Haven, Connecticut (Dr Horwitz); and Stanford Prevention Research Center, Departments of Medicine and Health and Research, and Policy, Stanford University School of Medicine, and Department of Statistics, School of Humanities and Sciences, Stanford University, Stanford, California (Dr Ioannidis).
Corresponding Author: John P. A. Ioannidis, MD, DSc, Stanford Prevention Research Center, Medical School Office Bldg, 1265 Welch Rd, Room X306, Stanford, CA 94305 (jioannid@stanford.edu).

quent observations will be smaller and closer to the mean value because of regression-to-the-mean effect.[6] Some large treatment effects may represent entirely spurious observations. It is unknown how often studies with seemingly very large effects are repeated.

To our knowledge, there has been no previous empirical evaluation of how often very large effects occur in randomized evidence, what outcomes they pertain to, whether they remain constant or decrease over time, and whether any of them pertain to the most serious of outcomes, death. Answering these questions is important for appreciating whether observed very large effects are common, real, clinically important, or not. Thus, the objective of this study is to describe the features and investigate the evolution and validation of nominally statistically significant ($P < .05$) very large treatment effects of medical interventions when such effects are first recorded in a clinical trial.

METHODS

Empirical Data

We used the Cochrane Database of Systematic Reviews (CDSR), 2010, issue 7. The CDSR organizes quantitative data on treatment comparisons and outcomes in forest plots. We considered forest plots regardless of the number of included studies (eg, some forest plots encompass only a single trial) and regardless of whether the Cochrane authors had performed quantitative synthesis (meta-analysis) of the data.[7]

Definition of a Very Large Effect

We focused for consistency on binary outcomes only and used the odds ratio (OR) metric. There is no widely accepted definition for very large treatment effects in randomized evidence. The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) uses a scale for relative risks for nonrandomized data, separating relative risks of 2 to 5 as large and those greater than 5 as very large and preferring the risk ratio over the OR because the OR may be larger when outcomes are common.[8] We used the OR metric in the main analyses, and we also report some key RR data for comparison. Moreover, point estimates of effects alone may not offer sufficient information, if the confidence intervals are wide and the effect is not even nominally significant. Therefore, our a priori operational definition of very large effect trials was defined to include those with an OR of 5.0 or more (or an OR $\leq$0.20) that had a nominally statistically significant effect based on a Fisher exact test ($P < .05$). When necessary, a continuity correction (ie, addition of 0.5 to all cells) was used in studies with sparse data for an OR estimation.

Categorization of Forest Plots and Definition of Index Trial

Forest plots were categorized into 3 groups: those for which a nominally significant very large effect was found in the first published trial (group A); those for which a nominally significant very large effect was observed in a subsequent trial, but not the first one (group B); and those for which no trial had a nominally significant very large effect (group C).

When 2 or more trials had the earliest publication year and it was impossible to identify which was published first, we randomly selected 1 trial as the first published study.

The index trial is the first published trial with a nominally significant very large effect for group A; the earliest such trial for group B; and the first published trial for group C.

Eligible Topics

We accepted very large effects regardless of either the treatment comparison or choice of intervention. We excluded forest plots using outcomes measured on continuous scales and those not including the year of publication of each trial (thus it would be impossible to identify the first published trial). We also excluded reviews with problems in their structure (ie, information that could not be parsed or with inconsistent data hierarchy), protocols, and methodological reviews.

Data Extraction

We used an automated data extraction approach to gather information of all eligible forest plots from the CDSR. Briefly, raw data from each of the 3545 available reviews (stored under a hierarchical structure; CDSR, 2010, issue 7) were systematically and automatically extracted. Specifically, computer scripts written in Python, a general-purpose dynamic programming language, were developed to load raw structured extensible markup language files. This format encodes all of the quantitative information from a review into a single file. Next, loaded files for each review were parsed considering the fixed structure of the CDSR. For the latter step, data for each forest plot were extracted using a combination of a Linux C shell-processing language (AWK), a stream editor (SED), the general-purpose programming language from the statistical package Stata 11.0 as well as Python. Validation of the automated approach was performed with the manual extraction of 200 randomly chosen forest plots (100% of agreement).

For each eligible forest plot, we extracted the title, comparison, outcome, total number of trials, year of publication of each study, years elapsed between the index trial and the next trial, total number of participants and events across all trials, OR, relative risk (RR), risk difference (RD), and *P* value of the index trial and we extracted the OR, RR, RD, and *P* value of the meta-analysis using a random-effects model[9] whenever additional trials were available beyond the index trial.

Categorization of Types of Outcomes and Interventions

Detailed categorization of outcomes and types of interventions for 10s of thousands forest plots was impractical. Thus, we randomly selected (using a random number generator) for further in-depth evaluation 250 forest plots from each group. We estimated that in-

depth evaluation of 750 forest plots could be done with approximately 1 person-year of effort. For this sample, we further extracted detailed information on type of outcome, involvement of drug(s) in the comparison (yes/no), or type of study (randomized or nonrandomized). Outcomes were classified[10] as death, composites including death, clinically defined benefit, laboratory-defined benefit, pain response, withdrawals, and harms. Withdrawals due to specific reasons (eg, lack of clinical benefit or harms) were counted under withdrawals. Drugs included biologics, but excluded supplements, minerals, vitamins, vaccines, and immunoglobulins.

Statistical Analysis

Descriptive statistics are expressed as median with interquartile range (IQR) or absolute counts and percentages. Comparisons between independent groups were performed with Fisher exact, Mann-Whitney *U*, and Kruskal-Wallis tests, as appropriate. Kaplan-Meier plots with the log-rank test were used to compare groups in terms of the probability and time to publication of the next trial after the index trial. Follow-up was censored in 2010. Index trials were compared with the corresponding meta-analyses on the same topic for effect sizes and *P* values using the Wilcoxon signed-rank test. For meta-analyses, we present the results from a random-effects model (DerSimonian-Laird method) to account for any potential between-study heterogeneity across trials.[9] For all analyses, OR estimates from index trials are presented as 1.0 or higher (ie, when the OR was <1.0, we took the opposite comparison [A vs B was switched to B vs A] and then the meta-analysis estimate on the same topic was computed using the respective comparison). We also performed sensitivity analyses limited to forest plots with index trials published after the year 2000 to see whether regression-to-the-mean continued to be an issue in more recent trials.

We calculated how many of the 500 selected topics of groups A and B maintained very large effects with a 2-tailed $P<.001$ including all the available evidence.[11] With a $P<.001$ (2-tailed), the treatment effect estimate is more than 2.58-times larger than the standard deviation of the estimate and some nonnull effect is likely to be present (>99.5% likely to be present unless the total sample size is still very small or the prior probability of an effect being present was <10%).[11,12] These topics can be considered to have effects that have been very well-validated.[11] We then recorded what these topics were and what kind of outcomes and interventions they pertained to overall. We also did the same focusing on those topics for which 2 or more trials had been performed.

Finally, we also surveyed all the Cochrane forest plots to identify whether there are any with $P<.001$ and with a very large effect pertaining strictly to death as an outcome (not a mortality-related composite) that the systematic reviewers also considered to represent reliable evidence for which they had no major concerns about biases that may decrease the credibility of the very large effect.

Given that some mortality estimates may have been presented as log-relative risks (eg, log-hazard ratio) and standard error, we also performed automatic searches across the CDSR using specifically terms *mortal**, *fatal*, *survival*, and *death* to see whether any such forest plots may have been missed by focusing on forest plots with binary outcome data, but none were found.

All data analyses were performed using Stata (version 11.0, Stata Corp). All *P* values are 2-tailed with nominal statistical significance claimed for $P<.05$.

RESULTS

Prevalence of Very Large Effects

Among 3545 available reviews, 3082 contributed usable information on 85 002 forest plots (eFigure 1, available at http://www.jama.com). These 85 002 forest plots included a total of 228 220 trial entries. Of those, 20 573 (9%) had a very large effect. The median number of comparisons per review was 2 with an (IQR, 1-4). A total of 103 666 of 228 220 trial entries (45.4%) were identified as having an OR estimate of 2 or higher or 0.5 or less. Among 85 002 eligible forest plots, 52 088 (61.3%) had at least 1 trial with an OR estimate 2 or higher or 0.5 or less.

Overall, 8239 forest plots (9.7%) had a nominally statistically significant very large effect in the first published trial, group A; 5158 (6.1%) had a nominally statistically significant very large effect found only after the first published trial, group B; and 71 605 (84.2%) had no trials with significant very large effects, group C.

Nominally significant very large effects arose mostly from small trials with few events (TABLE 1). For the index trials, the median number of events was only 18 in group A and 15 in the group B. Nevertheless, the median number of events was even smaller with a median of 14 in the group C index trials. Index trials tended to have more extreme effect sizes and *P* values in group A than in group B, but the differences were not substantial in absolute magnitude.

FIGURE 1 shows the scatterplot of ORs against risk differences for index trials in groups A and B. Extreme very large effects were more common in group A. Risk differences with an absolute value of more than 0.5 appeared in 21.6% of group A trials vs 12.4% of group B trials ($P<.001$); ORs that were greater than 40 appeared in 12.5% of group A trials vs 5.8% of group B trials ($P<.001$).

Probability of and Time-to-Next-Published Trial

Forest plots with a very large effect had a higher likelihood of having subsequent trials published than forest plots that did not show a very large effect (eFigure 2, available at http://www.jama.com). Forest plots with first trials with large effects had a statistically significant, but very small, higher risk of having subsequent trials published than forest plots of trials not showing a large effect. Significant differences for the

Table 1. Amount of Evidence and Effect Sizes Observed in Index Trials[a]

	Median (Interquartile Range)			
	Group A (n = 8239)	Group B (n = 5158)	Group C (n = 71 605)[b]	Group C (n = 65 492)[c]
Participants	78 (43-161)	93 (48-183)	102 (50-253)	110 (54-270)
Events	18 (10-36)	15 (9-30)	14 (3-41)	16 (5-45)
Odds ratio	12.03 (7.363-21.48)	10.02 (6.285-17.12)	1.562 (1.075-2.804)	1.688 (1.173-2.983)
Relative risk	5.737 (3.25-11.37)	5.656 (3.400-11.00)	1.400 (1.092-2.223)	1.489 (1.148-2.406)
$-\text{Log}_{10}$ (*P* value)	2.720 (1.881-4.878)	2.412 (1.748-3.888)	0.252 (0.000-0.676)	0.305 (0.000-0.736)
Absolute risk difference	0.355 (0.200-0.479)	0.279 (0.146-0.410)	0.035 (0.005-0.098)	0.042 (0.011-0.104)

[a] All values are medians (interquartile range). For consistency, all odds ratio and relative risk values are presented as 1.0 or higher, ie, values less than 1.0 have been inverted (eg, 0.1 has become 10). By Kruskal-Wallis test for the comparison across groups followed by a Bonferroni-corrected Mann-Whitney *U* test for the comparison between groups. Except for relative risk estimates between group A and group B ($P = .591$), all multiple comparisons yielded nominally significant results with a Bonferroni-corrected $P < .01$. *P* values of 1, .10, .05, and .001, for example, equal 0, 1, 1.30, and 3 on the $-\log_{10}$ scale, respectively.
[b] For group C, 6113 forest plots did not contribute with effect sizes (ie, indefinite estimates) and a *P* value because the index trials had either 0 events in both groups or all participants presented with the event. Estimates based on all forest plots replaced indefinite values of OR and *P* values with 1 and indefinite risk difference values with 0.
[c] Sensitivity analyses excluding all forest plots with indefinite estimates.

time-to-next-published trial existed for group A vs B (log-rank test, $P < .001$), group A vs C (log-rank test, $P = .0084$) and group B vs C (log-rank test, $P < .001$). The 5-year cumulative probability of not having a new trial published was 64.8% in group A, 35.9% in group B, and 65.6% in Group C. Overall, at least 1 trial was published subsequent to the index trial in 45.1% (3718 of 8239) for group A, 70.5% (3 638 of 5158) for group B, and 42.7% (30 579 of 71 of 605) for group 3.

Comparison of Index Trials vs Respective Meta-analyses

When additional trials were published in the group A topics, the cumulative evidence encompassed a median of 412 (IQR, 196-948) participants and 91 (IQR, 40-224) events. As shown in FIGURE 2 ("Group A" panel), the summary effects of the resulting meta-analyses were more conservative than index trials in 3341 of 3718 topics (90%). The median random-effects OR for the cumulative evidence was 4.20 (IQR, 2.56-7.81), significantly smaller than the first reported estimates (OR, 11.88; IQR, 7.24-21.15). Of the 3718 topics, the summary effect still exceeded 5 in 1570 (42.2%), it was between 2 and 5 in another 1591 (42.8%). A second trial claiming a very large effect was seen in 1806 topics (56.8%). For group A, nominal statistical significance was lost at the meta-analysis level in 1277 of 3718 cases (34.3%).

For the 5158 forest plots included in group B, the cumulative evidence encompassed a median 732 (IQR, 313-1767) participants with a median of 123.5 (IQR, 48-318) events. The summary effects of the meta-analyses had a median random-effects OR of 2.60 (IQR, 1.72-4.11), significantly smaller than the index trials' effects (OR, 10.02; IQR, 6.28-17.12; Figure 2, "Group B" panel). A reduction in effect size was observed in 5050 (98%) forest plots, in which nominal statistical significance was lost in 2159 cases (41.8%). For group B, the summary effect still exceeded 5 in 909 forest plots (17.6%), it was between 2 and 5 in another 2503 forest plots (48.5%). A second trial claiming a very large effect was only seen in 1386 topics (26.9%).

Sensitivity Analysis

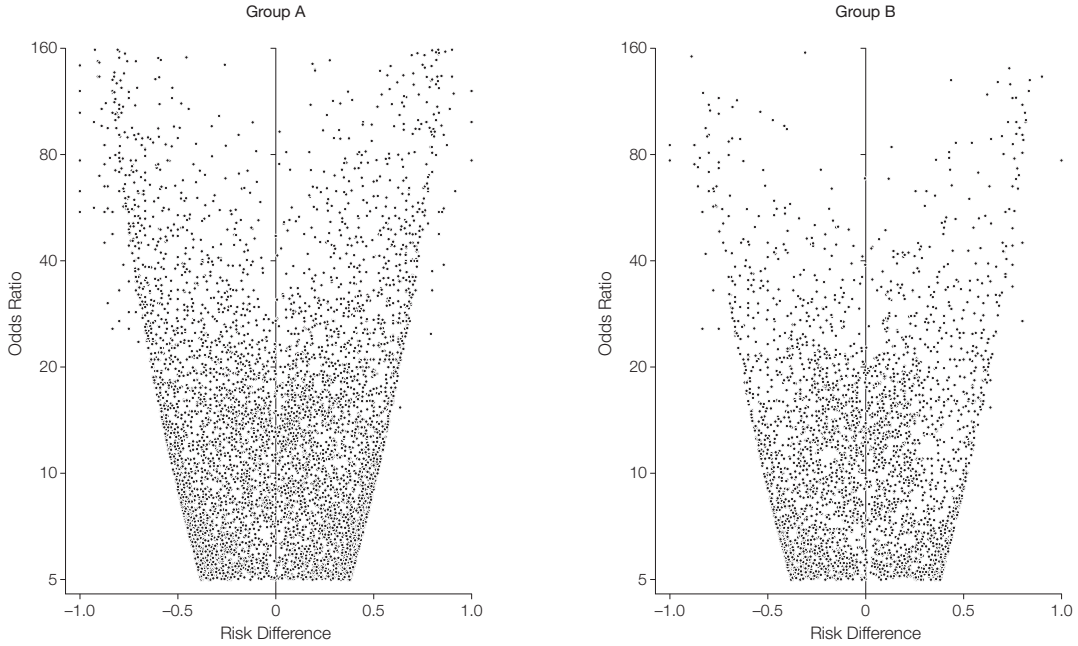
We also performed sensitivity analyses limited to index trials published more recently (after 2000). For group A, when limited to 1973 index trials that were published after 2000, the median effect was 11.70 (IQR, 7.33-19.55). For 441 of them, at least 1 subsequent trial was performed. The summary effect of their corresponding meta-analyses yielded a median OR of 5.2 (IQR, 3.04-8.37), remaining 5 or higher in 229 trials (52%) and 2 or higher in 389 trials (88.2%). For that group of 441 forest plots with 2 or more trials, the summary effect was smaller in the meta-analysis than in the index trial in 374 (85%) instances. For group B, when limited to 1434 index trials that were published after 2000, the median effect was 9.00 (IQR, 6.17-15.35). For 746 of them, at least 1 subsequent trial was performed. The summary effect of their corresponding meta-analyses was found to have a median of 2.40 (IQR, 1.57-3.54), it remained greater than 5 in 101 trials (13.5%) and greater than 2 in 442 trials (59.2%). For that group of 746 forest plots with 2 or more trials, the summary effect was smaller in the meta-analysis than in the index trial in 737 (98.8%) instances.

Outcomes and Type of Intervention

As shown in TABLE 2, based on a randomly chosen sample of 250 forest plots from each group, the 3 groups differed significantly in the type of outcomes, but not in the involvement of drugs in the comparison or type of study (randomized vs nonrandomized). Overall, forest plots including nonrandomized data represented less than 1% of the forest plots evaluated. This random sample of 750 forest plots corresponded to 673 unique comparisons from 536 independent reviews. Sixty-six comparisons were represented by at least 2 more forest plots in different outcomes.

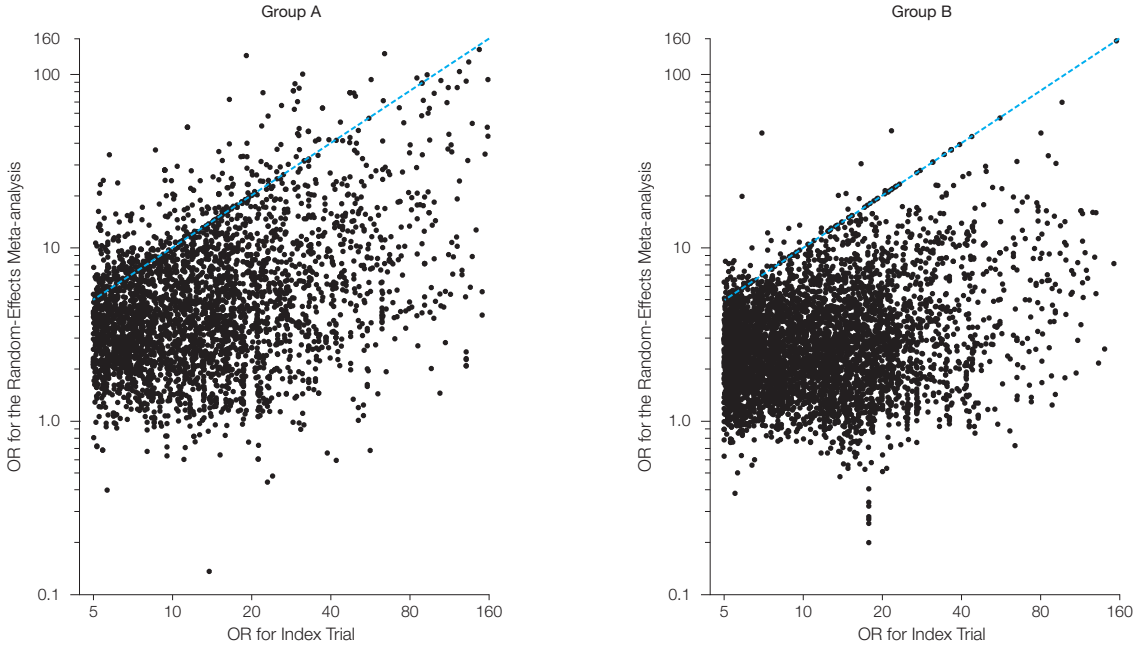
Death as an outcome was less common in the groups of very large effects (3.6% in group A and 3.2% in group B)

Figure 1. Scatterplots for Index Trials With Very Large Effect



Group A includes 8012 index trials; group B, 5116. For consistency, all odds ratio (OR) values are presented as 5.0 or higher, ie, values of 0.20 or less have been inverted (eg, 0.1 has become 10). Points are not shown for 227 forests plots (2.7%) from group A and 42 (0.8%) from group B because the index trial ORs were greater than 160.

Figure 2. Treatment Effects in Index Trials vs the Meta-analysis of All Trials on the Same Topic (Under a Random-Effects Model).



The group A panel depicts estimates for forest plots from trials that had at least 1 additional study published (n = 3617). The group B panel shows estimates for 5115 forest plots from group B. For group A, results are not shown for 101 of 3718 index trials (2.7%) because the estimated odds ratios (ORs) exceeded 160. For group B, results are not shown for 43 of 5158 forest plots (0.8%) because the estimated OR either in index trials or in the meta-analysis exceeded 160. The blue dashed lines correspond to the situation for which the OR in the index trial is the same as the OR in the meta-analysis.

than in group C (11.6%). Conversely, outcomes related to laboratory-defined benefits were significantly more common in the groups with very large effects (10% in group A and 10.8% in groups B) than in group C (3.2%).

The magnitude of the effect size of index trials did not vary according to the type of outcome in any of the 3 groups (eTable 1, available at http://www.jama.com). Nevertheless, the summary effect size of the meta-analysis including all evidence did vary according to the type of outcome in both groups A and B. For these groups, mortality-related outcomes yielded the smallest magnitude of effect sizes, whereas clinical-benefit and laboratory-defined benefit rendered the largest effects (eTable 2).

Fifteen of the 19 trials with nominally significant very large effects for mortality-related outcomes had at least 1 subsequent trial published. In the respective meta-analyses, only 6 (40%) maintained nominal statistical significance, only 1 (6.6%) with a very large effect.

Table 2. Characteristics of a Random Sample of 750 Forest Plots by Group From the Cochrane Database of Systematic Reviews, 2010, Issue 7

	No. (%) of Forest Plots[a]			
	Group A (n = 250)	Group B (n = 250)	Group C (n = 250)	*P* Value[c]
Type of outcome				
Clinically defined benefit	114 (45.6)	125 (50)	116 (46.4)	<.001
Harms	84 (33.6)	71 (28.4)	77 (30.8)	
Laboratory-defined benefit	25 (10)	27 (10.8)	8 (3.2)	
Mortality	9 (3.6)	8 (3.2)	29 (11.6)	
Mortality-composite	1 (0.4)	1 (0.4)	2 (0.8)	
Pain response	10 (4)	4 (1.6)	5 (2)	
Withdrawals	7 (2.8)	14 (5.6)	13 (5.2)	
Type of interventions				
Drug[b]	155 (62)	149 (59.6)	140 (56)	.40
No drugs involved	95 (38)	101 (40.4)	110 (44)	
Type of studies included				
Randomized only	247 (99)	248 (99.2)	248 (99.2)	>.99
Nonrandomized included	3 (1)	2 (0.8)	2 (0.8)	
No. of different reviews	219 (87.6)	201 (80.4)	214 (85.6)	.08

[a] Results are shown as absolute numbers (percentage). Detailed information on each forest plot are available in a supplemental material available at http://www.jama.com.
[b] Includes biologics and monoclonal antibodies; does not include vaccines or supplements.
[c] Based on Fisher exact test.

Well-Validated, Statistically Significant, Very Large Effects

In 93 of the 500 topics in groups A and B (18.6%), the overall evidence maintained a very large effect and it was also statistically significant with $P<.001$. Outcomes for these topics were related to clinically defined benefit in 41, laboratory-defined benefit in 8, harms in 33, pain in 6, withdrawals in 2, and mortality in 1 topic. The only mortality topic referred to the 10-fold reduction in the odds of *Haemophilus influenzae* type b–related deaths with the respective vaccine in high-income countries.[13] None of the other 92 outcomes would be considered life-threatening.

In 47 of the 93 topics, the index trial included all the available evidence (ie, a forest plot with a single trial), while in the other 46 topics there was evidence from at least 1 additional trial. Among these 46 topics, the outcomes included clinically defined benefits in 21, laboratory-defined benefits in 6, harms in 12, and pain in 7. No mortality or withdrawals outcomes were included in this set. Some representative examples of well-validated very large effects are shown in the upper part of TABLE 3. These include clinical benefits such as control of nocturnal enuresis with alarms in children, or symptomatic improvement with 5-aminosalicylic acid in ulcerative colitis; mostly mild or modest harms such as burning with capsaicin or local tenderness with the influenza vaccine; laboratory-determined response such as induction of hepatitis B surface antigen with hepatitis B vaccination; and control of acute pain with analgesics such as etoricoxib or diclofenac. As shown, all of these effects corresponded to very large absolute risk differences.

Well-Validated Very Large Effects for Mortality

Across all eligible forest plots with sufficient information to reconstruct 2×2 contingent tables (n=85 002), there were 2791 (3.2%) with a very large effect and P value of $<.001$. Upon a closer examination of these 2791 forest plots, we found that only 13 concerned mortality. Furthermore, only 3 of those 13 effects were considered reliable by the systematic reviewers (shown in the lower part of Table 3). Two effects actually pertained to increased risk of death (fatal hemorrhagic stroke with thrombolysis in ischemic stroke[14]; and 90-day mortality with lung reduction surgery in emphysema[15]), whereas the third pertained to decreased risk of death for extracorporeal oxygenation for severe respiratory failure in newborns.[16]

The other 10 large and highly significant effects seen in mortality-related outcomes were considered potentially spurious. For example, the reviews CD004403 and CD005096, which described large survival benefits of antibiotics in chronic obstructive pulmonary disease exacerbations[17] and Chinese herbs in gastric cancer,[18] respectively, were later (CDSR 2011, issue 1 and 2, respectively) withdrawn for lack of supportive data. The *H influenzae* type b vaccine review has also been withdrawn.[19] It had showed a large beneficial effect in mortality seen only for *H influenzae* type b–specific deaths and limited to a single trial in high-income countries.[13] Two other systematic reviews of the influence of the

azathioprine[20] and methotrexate[21] on mortality in primary biliary cirrhosis suggested no significant effect in the main analyses, but the reviewers had also performed sensitivity analyses with extreme scenarios about missing data that reached the range of implausible very large effects.

Finally, the reviewers also questioned the validity of very large survival benefits with β-agonists in threatened miscarriage,[22] different types of brachytherapies for stage I cervical cancer,[23] chaunxiong preparations to prevent stroke,[24] continuous vs intermittent infusion of loop diuretics in congestive heart failure,[25] and vitamin C in children with tetanus,[26] since their corresponding cumulative evidence was based on single small trials that were of very poor quality or (in the case of tetanus) apparently were not even a randomized controlled trial.

COMMENT

Trials with nominally significant very large effects appeared in 16% of the 85 000 forest plots with binary outcomes in the CDSR. Often these trials are the first published or even the only

Table 3. Examples of Well-Validated Very Large Effects in Medical Interventions[a]

CDSR No.	Topic	Outcome	Studies, No. of Patients	Events	Summary, OR (95% CI)	I^2 (95% CI)	Summary, RD (95% CI)
CD004115	Rectal 5-aminosalicylic acid vs placebo for induction of remission in ulcerative colitis	Clinically defined benefit: symptomatic improvement	8 (811)	478	8.9 (5.3 to 14.8)	42 (0 to 74)	0.52 (0.41 to 0.62)
CD002911	Alarm interventions vs control for nocturnal enuresis in children	Clinically defined benefit: not achieving 14 consecutive dry nights	14 (576)	357	0.03 (0.02 to 0.06)	0 (0 to 55)	–0.59 (–0.71 to –0.47)
CD006481	Hepatitis B immunization vs control	Laboratory-defined benefit: HBsAg (best-case scenario)	4 (1230)[b]	364	0.005 (0.004 to 0.06)	73 (8 to 92)	–0.34 (–0.78 to 0.11)[c]
CD003407	Erythropoietin or darbepoetin for patients with cancer	Laboratory-defined benefit: hematologic response (increase in hemoglobin ≥2 g/dL or increase in hematocrit ≥6%)	20 (3562)	1344	7.27 (5.6 to 9.4)	42 (1 to 66)	0.35 (0.28 to 0.42)
CD004309	Single-dose oral etoricoxib 120 mg vs placebo for acute postoperative pain in adults	Pain: at least 50% relief over 6 h	5 (655)	285	17.3 (5.7 to 51.9)	78 (48 to 91)	0.54 (0.36 to 0.72)
CD004768	Single-dose oral diclofenac 50 mg vs placebo for acute pain after dental surgery in adults	Pain: at least 50% relief over 4-6 h	11 (1119)	460	6.4 (4.01 to 9.91)	46 (0 to 73)	0.39 (0.29 to 0.49)
CD001269	Inactivated parenteral influenza vaccine vs placebo/do-nothing in healthy adults	Harm: local tenderness/soreness	14 (6833)	2363	5.1 (3.2 to 8.3)	92 (89 to 95)	0.29 (0.21 to 0.67)
CD007393	Topical 0.075% capsaicin vs placebo for chronic neuropathic pain in adults	Harm: burning, stinging, erythema	5 (557)	245	5.4 (2.7 to 11.2)	60 (0 to 85)	0.42 (0.24 to 0.61)
CD000213	Thrombolysis for acute ischemic stroke	Mortality: fatal intracranial hemorrhage (in 7-10 d)	4 (983)	54	10.65 (4.17 to 27.18)	0 (0 to 85)	0.09 (0.04 to 0.14)
CD001340	Extracorporeal membrane oxygenation for severe respiratory failure in newborns	Mortality: death before discharge home	4 (244)	94	0.19 (0.08 to 0.50)	15 (0 to 87)	–0.40 (–0.59 to –0.21)
CD001001	Lung volume reduction surgery for diffuse emphysema: surgery vs control	Mortality: 90-d mortality	4 (1415)	69	6.45 (3.27 to 12.73)	0 (0 to 85)	0.07 (0.05 to 0.09)

Abbreviations: OR, odds ratio. RD, risk difference. HbsAg, hepatitis B surface antigen.

[a] All examples are derived from the 500 randomly selected topics of groups A and B, except for the meta-analyses on death outcomes in which none were included in that random sample. Examples of very large effects on mortality were obtained by automatic searches in the 3082 forest plots with sufficient information to reconstruct 2×2 contingency tables. Summary estimates are based on random-effects calculations. I^2 is given as percentage.

[b] Includes a trial with 0 events in both study groups.

[c] Summary results under random-effects calculations presented extreme statistical heterogeneity (eg, I^2=99.5%). The same data under a fixed-effects model (inverse variance) yielded an RD summary estimate of –0.59 (95% CI, –0.62 to –0.57, $P<10^{-40}$).

ones available on the treatment comparison and outcome of interest. Typically trials with very large effects have limited evidence. First trials with very large effects were not less likely to have subsequent trials than other topics. Very large effects seen in nonfirst trials were more likely than others to be followed-up with additional studies. When additional evidence is obtained, most of the very large treatment effects become much smaller and many lose their nominal significance. Very large effects with strong statistical support (as denoted by $P<.001$) account for approximately 3% of the forest plots in the CDSR. These very large effects pertain practically exclusively to nonfatal outcomes. Across the 85 000 topics of the CDSR, there was only 1 example in which a large beneficial effect on mortality with high levels of statistical significance and no major validity concerns has been documented.

Based on this picture, most large treatment effect estimates should be considered with caution: many are spurious findings, while the vast majority may represent substantial overestimations. The overestimation is commensurate with the winner's curse phenomenon (regression-to-the-mean for inflated treatment effects).[6,10,27] Clinical researchers do not seem reluctant to conduct further studies when a prior trial had identified a very large effect. This is not surprising since typically there are many other outcomes on the same comparison. Moreover, some of the subsequently published trials may have been already launched, if not completed, well before a trial with a very large effect is published.

Trials with very large effects are more likely than other trials to pertain to laboratory-defined efficacy. Laboratory measurements have made randomized trials more efficient and very large effects may be easier to obtain. However, the relevance of laboratory end points as surrogates of hard clinical outcomes has long been contested.[28-31]

On the other hand, trials with very large effects are less likely than other trials to pertain to death. Even in these cases, investigators frequently perform additional trials, and, then these large mortality-related effects mostly shrink or disappear. Well-validated very large effects for mortality or even life-threatening clinical outcomes are exceedingly rare.

Limitations of this study must be acknowledged. First, despite the very large amount of data that it has amassed, the CDSR does not cover the entire randomized evidence for all medical interventions. There is no strong reason to believe that topics for which there have not been any Cochrane reviews yet are likely to be different in terms of their representation of very large effects and types of outcomes, but some very large effects for mortality may not have been captured by the CDSR. Second, it is possible that for some extremely effective interventions, there may be reluctance to perform randomized trials. This may lead to an under-representation of very large effects in CDSR. However, such uncontestable life-saving interventions are probably uncommon. Some examples may include insulin for diabetes, blood transfusion for severe hemorrhagic shock, neostigmine for myasthenia gravis, tracheostomy for tracheal obstruction, suturing for large wounds, and ether for anesthesia.[32] Glasziou and colleagues[32] suggest that randomized trials would be unnecessary if there is certainty that the relative risk exceeds 10, but the exact threshold and certainty required to make such a decision is not totally clear. Empirical evidence suggests that when time-honored standards of care are tested in randomized trials, half of the time they are proven ineffective.[33,34]

Third, we did not try to assess the relevance of each outcome in the context of all other outcomes on the same comparison of interventions for the same disease. This would have been subjective and extremely difficult to perform on a large-scale. Fourth, very large effects do not always guarantee that an intervention is useful. For example, there is wide variability of the cost-effectiveness across otherwise seemingly life-saving interventions.[35] Fifth, we used a quantitative rule with $P<.001$ for well-validated effects, as previously proposed,[11] although for each topic the exact credibility of the effect may depend on a multitude of other factors reflecting the quality of the evidence and diverse biases that need to be scrutinized on a case-by-case basis. If anything, this means that the well-validated effects are likely to be even fewer than what we estimate. Sixth, we used a random-effects model in our analysis and this gives more weight to small trials in meta-analyses. When small trials have inflated effects, the choice of model may actually overestimate the summary treatment effect; thus, it may underestimate the extent to which the index trials had spuriously large effects.

Acknowledging these caveats, this empirical evaluation suggests that very large effect estimates are encountered commonly in single trials. Conversely, genuine very large effects with extensive support from substantial evidence appear to be rare in medicine and large benefits for mortality are almost entirely nonexistent. As additional evidence accumulates, caution may still be needed, especially if there is repetitive testing of accumulating trials.[36] Patients, clinicians, investigators, regulators, and the industry should consider this in evaluating very large treatment effects when the evidence is still early and weak.

Author Contributions: Dr Ioannidis had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Pereira, Horwitz, Ioannidis.
Acquisition of data: Pereira, Ioannidis.
Analysis and interpretation of data: Pereira, Horwitz, Ioannidis.
Drafting of the manuscript: Pereira, Horwitz, Ioannidis.
Critical revision of the manuscript for important intellectual content: Pereira, Horwitz.
Statistical analysis: Pereira, Horwitz, Ioannidis.
Study supervision: Horwitz, Ioannidis.
Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.
Funding/Support: Dr Pereira was supported in part by grants from Fundação de Amparo à Pesquisa do Estado de São Paulo (São Paulo Research Foundation, FAPESP)
Role of the Sponsor: The funding agency had no role in the design and conduct of the study; the collection, analysis, and interpretation of the data; or the preparation, review, or approval of the manuscript.

Online-Only Material: The eTables 1 and 2 and eFigures 1 and 2 are available at http://www.jama.com.

REFERENCES

1. Guyatt GH, Oxman AD, Vist GE, et al; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924-926.

2. Kent DM, Trikalinos TA. Therapeutic innovations, diminishing returns, and control rate preservation. *JAMA*. 2009;302(20):2254-2256.

3. Djulbegovic B. The paradox of equipoise: the principle that drives and limits therapeutic discoveries in clinical research. *Cancer Control*. 2009;16(4):342-347.

4. Hill AB. The environment and disease: association or causation? *Proc R Soc Med*. 1965;58:295-300.

5. Durrheim DN, Reingold A. Modifying the GRADE framework could benefit public health. *J Epidemiol Community Health*. 2010;64(5):387.

6. Ioannidis JP. Why most discovered true associations are inflated. *Epidemiology*. 2008;19(5):640-648.

7. Ioannidis JP, Patsopoulos NA, Rothstein HR. Reasons or excuses for avoiding meta-analysis in forest plots. *BMJ*. 2008;336(7658):1413-1415.

8. Guyatt GH, Oxman AD, Sultan S, et al; GRADE Working Group. GRADE guidelines: 9. rating up the quality of evidence. *J Clin Epidemiol*. 2011;64(12):1311-1316.

9. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177-188.

10. Pereira TV, Ioannidis JP. Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. *J Clin Epidemiol*. 2011;64(10):1060-1069.

11. Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? *BMJ*. 2001;322(7280):226-231.

12. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.

13. Eskola J, Käyhty H, Takala AK, et al. A randomized, prospective field trial of a conjugate vaccine in the protection of infants and young children against invasive *Haemophilus influenzae* type b disease. *N Engl J Med*. 1990;323(20):1381-1387.

14. Wardlaw JM, Murray V, Berge E, Del Zoppo GJ. Thrombolysis for acute ischaemic stroke. *Cochrane Database Syst Rev*. 2009;(4):CD000213.

15. Tiong LU, Davies R, Gibson PG, et al. Lung volume reduction surgery for diffuse emphysema. *Cochrane Database Syst Rev*. 2006;(4):CD001001.

16. Mugford M, Elbourne D, Field D. Extracorporeal membrane oxygenation for severe respiratory failure in newborn infants. *Cochrane Database Syst Rev*. 2008;(3):CD001340.

17. Ram FS, Rodriguez-Roisin R, Granados-Navarrete A, Garcia-Aymerich J, Barnes NC. Antibiotics for exacerbations of chronic obstructive pulmonary disease. *Cochrane Database Syst Rev*. 2006;(2):CD004403.

18. Gan T, Wu Z, Tian L, Wang Y. Chinese herbal medicines for induction of remission in advanced or late gastric cancer. *Cochrane Database Syst Rev*. 2010;(1):CD005096.

19. Swingler GH, Michaels D, Hussey GG. WITHDRAWN: conjugate vaccines for preventing *Haemophilus influenzae* type B infections. *Cochrane Database Syst Rev*. 2009;(4):CD001729.

20. Gong Y, Christensen E, Gluud C. Azathioprine for primary biliary cirrhosis. *Cochrane Database Syst Rev*. 2007;(3):CD006000.

21. Giljaca V, Poropat G, Stimac D, Gluud C. Methotrexate for primary biliary cirrhosis. *Cochrane Database Syst Rev*. 2010;(5):CD004385.

22. Lede R, Duley L. Uterine muscle relaxant drugs for threatened miscarriage. *Cochrane Database Syst Rev*. 2005;(3):CD002857.

23. Wang X, Liu R, Ma B, et al. High-dose rate versus low-dose rate intracavity brachytherapy for locally advanced uterine cervix cancer. *Cochrane Database Syst Rev*. 2010;(7):CD007563.

24. Yang X, Zeng X, Wu T. Chuanxiong preparations for preventing stroke. *Cochrane Database Syst Rev*. 2010;(1):CD006765.

25. Salvador DR, Rey NR, Ramos GC, Punzalan FE. Continuous infusion vs bolus injection of loop diuretics in congestive heart failure. *Cochrane Database Syst Rev*. 2005;(3):CD003178.

26. Hemilä H, Koivula TT. Vitamin C for preventing and treating tetanus. *Cochrane Database Syst Rev*. 2008;(2):CD006665.

27. Krum H, Tonkin A. Why do phase III trials of promising heart failure drugs often fail? the contribution of "regression to the truth." *J Card Fail*. 2003;9(5):364-367.

28. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med*. 1996;125(7):605-613.

29. Albert JM, Ioannidis JP, Reichelderfer P, et al. Statistical issues for HIV surrogate end points: point/counterpoint: an NIAID workshop. *Stat Med*. 1998;17(21):2435-2462.

30. Buyse M, Sargent DJ, Grothey A, Matheson A, de Gramont A. Biomarkers and surrogate end points—the challenge of statistical validation. *Nat Rev Clin Oncol*. 2010;7(6):309-317.

31. Farmer AJ, Perera R, Ward A, et al. Meta-analysis of individual patient data in randomised trials of self-monitoring of blood glucose in people with non-insulin treated type 2 diabetes. *BMJ*. 2012;344:e486.

32. Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? picking signal from noise. *BMJ*. 2007;334(7589):349-351.

33. Prasad V, Gall V, Cifu A. The frequency of medical reversal. *Arch Intern Med*. 2011;171(18):1675-1676.

34. Prasad V, Cifu A, Ioannidis JP. Reversals of established medical practices: evidence to abandon ship. *JAMA*. 2012;307(1):37-38.

35. Tengs TO, Adams ME, Pliskin JS, et al. Five-hundred life-saving interventions and their cost-effectiveness. *Risk Anal*. 1995;15(3):369-390.

36. Thorlund K, Devereaux PJ, Wetterslev J, et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? *Int J Epidemiol*. 2009;38(1):276-286.