Large Trials vs Meta-analysis of Smaller Trials

How Do Their Results Compare?

Joseph C. Cappelleri, PhD, MPH; John P. A. Ioannidis, MD; Christopher H. Schmid, PhD;
Sarah D. de Ferranti, MD, MPH; Michael Aubert; Thomas C. Chalmers, MD†; Joseph Lau, MD

Objective.—To evaluate the results of large clinical trials vs the pooled results of smaller trials.

Data Identification.—Meta-analyses with at least 1 "large" study were identified from the Cochrane Pregnancy and Childbirth Database and from MEDLINE (1966-1995).

Study Selection.—We used a sample size approach to select 79 meta-analyses with at least 1 large study of 1000 or more patients. We used a statistical power approach to select 61 meta-analyses with at least 1 large study based on statistical power considerations.

Data Extraction.—The outcome of interest for each meta-analysis was the primary one stated in the original publication or, when not clearly specified, was decided on clinically.

Data Synthesis.—By random effects calculations, we found agreement between large and smaller trials in 90% of the meta-analyses selected by the sample size approach and in 82% of the meta-analyses selected by the statistical power approach. Twice as many disagreements appeared when the variability among large studies and among smaller studies was not considered (ie, fixed effects calculations). Of the 15 disagreements between results of large and smaller trials using the random effects model, plausible explanations were identified in 10 meta-analyses: 5 with differences in the control rate of events between large and smaller trials, 4 with specific protocol or study differences, and 1 with potential publication bias. Two other disagreements were not clinically important, and tentative reasons could be identified for 2 of the remaining 3 disagreements.

Conclusions.—Results of smaller studies are usually compatible with the results of large studies, but discrepancies do occur even when the diversity among both large studies and smaller studies is considered. Clinically important differences without a potential explanation are extremely uncommon. Future research should further examine sources of heterogeneity between the results of large and smaller trials.

JAMA. 1996;276:1332-1338

META-ANALYSES of randomized controlled trials and individual large-scale trials have been increasingly used to obtain evidence for treatment decisions. Often the results of meta-analyses of earlier, smaller trials are subsequently confirmed in larger trials.[1] Much recent attention and debate, however, have focused on some disagreements between the results of megatrials (usually trials with more than 10 000 patients) and previous meta-analyses of smaller trials, as exemplified by the use of nitrates and magnesium in the treatment of acute myocardial infarction.[2-8] The combined results of smaller trials showed evidence that nitrates and magnesium reduced mortality by about 32% and 48%, respectively, but the respective megatrials showed no evidence of any real effect (3% risk reduction from nitrates, 5% risk increase from magnesium). Such disagreements have raised concerns about the reliability, interpretation, and adequacy of both large trials and meta-analyses of smaller trials. Several important questions must be addressed. How well do large and smaller studies agree in their results? How frequent are the significant disagreements? Why do these disagreements occur? Are the disagreements clinically important?

Little work has been conducted on these important issues. Two previous evaluations have explored how well the results of a meta-analysis of smaller randomized trials predicted the results of a large "gold standard" trial.[9,10] One of them defined the large trial as the largest cooperative (multicenter) study[9] and the other de-

From the Division of Clinical Care Research (Drs Cappelleri, Schmid, de Ferranti, Chalmers, and Lau, and Mr Aubert), the Division of Geographic Medicine and Infectious Diseases (Dr Ioannidis), the Biostatistics Research Center (Dr Schmid), Department of Medicine, New England Medical Center, Boston, Mass. Dr Cappelleri is now with Pfizer Central Research, Groton, Conn; Dr Ioannidis is now with the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Md; and Dr de Ferranti is now with Children's Hospital, Boston.

†Deceased.

Reprints: Joseph Lau, MD, New England Medical Center, PO Box 63, 750 Washington St, Boston, MA 02111.

fined the large trial as the study with the largest sample size (having at least 1000 patients).[10] To assess how well the results of the largest trial agreed with the combined results of the smaller trials, both investigations assessed whether the direction of the 2 estimates of treatment effect agreed and whether these estimates agreed in terms of statistical significance, as well as whether the pair of confidence intervals of treatment effect overlapped.

These assessments raise several issues.[11-13] When should a study be declared "large"? Defining a study as large because it has more than 1000 patients is convenient, but even studies of more moderate sample sizes may be large enough, if they have sufficient power to detect a postulated treatment effect. Moreover, previous investigations[9,10] assumed that small and large trials, including megatrials, share a common treatment effect. In fact, diversity between trials may be unavoidable as large multicenter studies are designed, conducted, and analyzed with a different level of complexity than small studies. Thus, we believe that an approach that recognizes the existing diversity between trials should also be assessed. Finally, none of the previous evaluations addressed the reasons for which discrepancies occur and their clinical significance; both of these issues are of major importance to clinicians and trialists alike.

Our study addresses these concerns by implementing novel strategies to define large studies and to analyze the concordance of the results of large and smaller trials across a wide variety of clinical areas. Large studies are defined in terms of both their sample size and their statistical power, and the impact of different analytical approaches is investigated. Furthermore, we compile the plausible reasons given by other authors for the genuine disagreements between large and smaller trials, and assess the frequency of unexplained clinically important discrepancies.

METHODS

Selection of Meta-analyses

We screened meta-analyses of randomized controlled trials involving human subjects with binary outcomes for the presence of large trials. Only meta-analyses with at least 1 large study and 2 smaller trials were included. To collect meta-analyses, we conducted a MEDLINE search (1966-1995) of the English literature with the Medical Subject Heading "meta-analysis" and at least 1 of the following: the Medical Subject Heading "models, statistical or placebos"; the EXPLODE command on "clinical trials"; the EXPLODE command on "research design"; or the text words "random" or "placebo." We also screened the complete Cochrane Pregnancy and Childbirth Database,[14] a comprehensive database of systematic reviews of controlled trials on perinatal topics. Meta-analyses on nitrates and magnesium for patients with acute myocardial infarction[1] were updated to reflect the controversial disagreement in mortality results between recent mega-trials and earlier smaller trials.[3-8]

Definition of a Large Study

In order to define a large study, we used 2 distinct, but complementary, approaches: 1 based on sample size and 1 based on statistical power. The sample size definition identified a large study as having at least 1000 patients in total, as previously proposed.[10] Originating from an idea in experimental design,[15] the statistical power definition qualified a study as large regardless of its sample size, provided that it had sufficient power to detect the postulated treatment effect obtained by pooling the remaining smaller trials. This approach was applied to all meta-analyses identified as having large studies by the sample size approach and to all meta-analyses in the Cochrane Pregnancy and Childbirth Database.

For the statistical power definition, we used the following 4-step algorithm:

1. Obtain pooled estimates of the relative risk reduction and the weighted control rate (the proportion of patients in the control arm who had an event of interest, weighted by the sample size in that group) from all studies except the study with the largest sample size (N).
2. Use these estimates to calculate the minimum total sample size (n) needed for 80% statistical power and a 5% level of significance.[16]
3. If N is less than n, stop the procedure and, if this is the first study considered, reject the meta-analysis as not containing any large study. If N is greater than or equal to n, select the meta-analysis and declare that study large.
4. If the study is deemed large, repeat the first 3 steps assessing the study with the next largest sample size based on the pooled estimates of the remaining smaller studies.

Data Extraction

The most important primary outcome, as stated in the original publication, was chosen for each meta-analysis and data for this outcome were extracted from each trial. When the most primary outcome was not clearly specified, we were blinded to all results before identifying by consensus the most clinically important outcome.

Statistical Methods

The combined results of the large studies (or the results of the only large study) were compared with those of smaller studies. In each meta-analysis, the Mantel-Haenszel (fixed effects)[17] and the DerSimonian and Laird (random effects)[18] methods were used to pool risk ratios (RRs) from multiple large studies and multiple smaller studies. The fixed effects model assumes that all trials are similar in that they share the same underlying treatment effect. Thus, the observed differences in their results are considered to be due to chance alone (sampling error within each study). The random effects model in addition incorporates the potential heterogeneity of the treatment effect among different studies by assuming that each study estimates a unique treatment effect that, even given a large amount of data, might still differ from the effect in another study. Compared with the fixed effects model, the random effects model therefore weights smaller trials more heavily in its pooled estimate of treatment effect. The fixed effects model is equivalent to the random effects model when there is no heterogeneity of the treatment effect among different studies. Since 1 of the aims of our study was to compare the 2 approaches, both models were used but with priority given to the random effects model, because it also incorporates the potential diversity among different trials.

A standardized z statistic was used to assess whether sufficient evidence exists for agreement ($-1.96<z<1.96$) or disagreement ($z\geq1.96$ or $z\leq-1.96$) between large and smaller studies in each meta-analysis. This statistic uses the variances of treatment effects of large studies and smaller studies to define the SE of the difference in their treatment effects. The statistic takes the following form: $z=[\ln(RR_{\text{large}})-\ln(RR_{\text{smaller}})] \div [\text{variance of } \ln(RR_{\text{large}})+\text{variance of } \ln(RR_{\text{smaller}})]^{1/2}$ where $\ln(RR_{\text{large}})$ denotes the natural logarithm of the pooled RR of the large studies or of the RR of the only large study, and $\ln(RR_{\text{smaller}})$ denotes the natural logarithm of the pooled RR ratio of the smaller studies. Used to quantify discordance, this z statistic is mathematically equivalent to the χ^2 test statistic (with 1 *df*) to test whether the natural logarithm of each RR in the 2 strata is the same.[17] The computations of the 4 terms in z are given in Rothman[17] for the fixed effects model and in Ioannidis et al[18] for the random effects model.

We probed disagreements between large and smaller trials found by the random effects model in several ways. First, we examined the association between the treatment effect and the control rate (the proportion of patients in the control arm who had an event of interest). The control rate has been gaining acceptance as an available summary measure of patient or study differences, such as the baseline risk of patients and length of study follow-up. These differences may account for heterogeneity of treatment effects across studies.[19-25] We performed regression analyses of the natural logarithm of the RR on the control rate in each study and used a hierarchical model fit by the EM algorithm[26] to account for random measurement error.[27] We then tested whether a statistical difference existed ($P \leq .05$) between the control rate of the largest study and the pooled control rate of all studies, as well as between the pooled control rate of the large studies and that of the smaller studies. Control rates were pooled with the logistic transformation,[28] and each control rate was weighted by the inverse of its variance and the among-study variance of control rates (ie, random effects pooling).

Second, we assessed the possibility that the results may have been affected by publication bias. Publication bias is the phenomenon where studies with "negative" (nonsignificant) results may be less likely to be published either because their investigators do not feel it is important to report negative results or because peer reviewers and journal editors find such negative studies unappealing for publication. Theoretical and empirical evidence[29] has demonstrated that publication bias may mostly affect negative studies of small sample size. These are typically studies that have low precision (high variance). Therefore, if the treatment effects of the studies included in a meta-analysis are found to be related to the sample size or to the variance of the treatment effects, this association may suggest publication bias. A formal approach to detect this association uses the rank correlation test based on the Kendall rank correlation coefficient (τ) to evaluate whether the (standardized) natural logarithm of the RR is correlated with its variance.[29] Because this test has relatively low statistical power,[29] we based evidence of publication bias on $P \leq .10$.

Finally, we collected alternative specific reasons that the authors of the meta-analyses or other authors provided to account for disagreements. If none were identified, we examined whether the disagreement was clinically important.

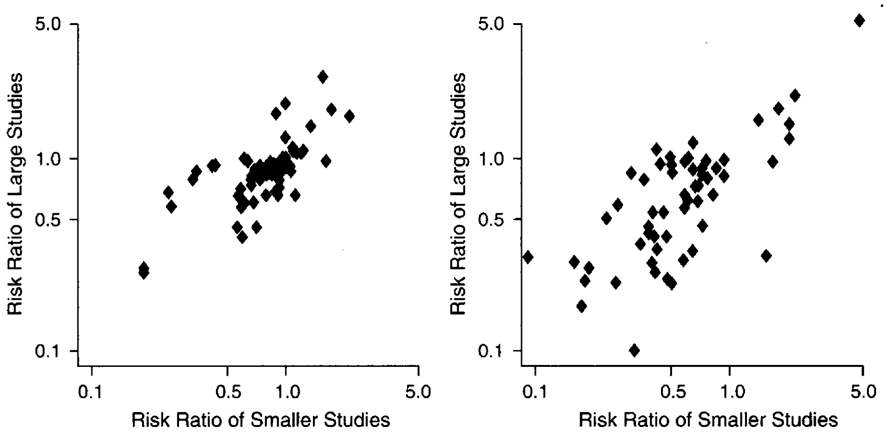
Calculations were performed in Epi Info,[16] Meta-Analyst,[30] and Mathcad.[31] Control-rate meta-regressions were performed in S-PLUS,[32] and rank correlations were calculated in SAS.[33]

Table 1.—Meta-analyses of Randomized Controlled Trials Included in the Comparison of Large and Smaller Studies

	Large Study Definition	
	Large Sample Size (≥1000 Patients)	Sufficient Power (≥80%)
No. of meta-analyses identified	**79**	**61**
Cochrane[14]	33	43
Other	46*	18†
No. of large trials per meta-analysis, median (range)	1 (1-10)	2 (2-21)
No. of patients per single large trial or per meta-analysis of large trials, median (range)	2829 (1009-356 025)	1741 (75-358 021)
No. of smaller trials per meta-analysis, median (range)	8 (2-56)	4 (2-57)
No. of patients per meta-analysis of smaller trials, median (range)	2062 (87-13 195)	982 (52-37 351)
No. (%) of disagreements between large and smaller trials		
Fixed effects	14 (17.7)	21 (34.6)
Random effects	8 (10.1)	11 (18.0)

*Includes 1 additional meta-analysis on perinatal medicine,[34] as well as meta-analyses on acute myocardial infarction[1,35-39] (n=12 meta-analyses), secondary prevention of vascular events[1,21,23,39-42] (n=16), breast cancer[43,44] (n=6), surgery[45-47] (n=3), and miscellaneous areas[18,48-54] (n=8).

†Addresses treatments for acute myocardial infarction[1,37-39] (n=6 meta-analyses), secondary prevention of vascular events[21,39-40,42] (n=5), surgery[46,47] (n=2), and other areas[18,49,52-54] (n=5).



Plots of the pooled risk ratio of large trials (or the risk ratio of the single large trial, when only 1 large trial was available) against the pooled risk ratio of the smaller trials according to the sample size approach (left) and the statistical power approach (right). Risk ratios are shown on a logarithmic scale and are pooled according to the DerSimonian and Laird (random effects) model.[18] Not shown are 1 outlier for the sample size approach and 1 outlier for the statistical power approach, each having very low risk ratios (<0.1) for both large and smaller studies.

RESULTS

Concordance of Large Studies With Smaller Studies

About 2100 MEDLINE citations and 500 Cochrane systematic reviews were initially screened. As shown in Table 1, a total of 79 meta-analyses included at least 1 study of large sample size. A screening of these 79 meta-analyses and of the entire Cochrane database revealed 61 meta-analyses that included at least 1 large study having at least 80% statistical power. Twenty-nine meta-analyses met both definitions. When the sample size criterion was used, most of the identified meta-analyses had only 1 study with 1000 or more patients. When the power criteria were used, typically more than 1 study was deemed large.

The Figure displays the scatterplot of the random effects pooled RRs of large studies (or the RR of the only large study) and of smaller studies. It can be observed that the 2 sets of RRs mostly clustered around an RR of 1 when the sample size approach was applied (the left panel), while the 2 sets of RRs were spread over a wider range when the statistical power approach was applied (the right panel). For both approaches, the natural logarithms of the RRs between large and smaller studies were highly positively correlated ($r=0.75$).

As shown in Table 1, regardless of the definition of a large study, we found agreement between large and smaller studies

Table 2.—Specific Disagreements Between Large and Smaller Studies: Comparison of Risk Ratios*

		Large Study Defined by Size				Large Study Defined by Power			
		Smaller Studies		Large Studies		Smaller Studies		Large Studies	
Meta-analysis	Outcome	No.	Risk Ratio (95% CI)	No.	Risk Ratio (95% CI)	No.	Risk Ratio (95% CI)	No.	Risk Ratio (95% CI)
Magnesium in acute myocardial infarction[1]	Mortality	9	0.44 (0.27-0.72)	2	0.92 (0.67-1.26)	9	0.44 (0.27-0.72)	2	0.92 (0.67-1.26)
Nitrates in acute myocardial infarction[1]	Mortality	13	0.68 (0.49-0.95)	2	0.97 (0.92-1.02)	12	0.59 (0.44-0.80)	3	0.97 (0.92-1.02)
Antiplatelet agents in pregnancy (4000†)	Preeclampsia	14	0.35 (0.23-0.53)	3	0.86 (0.76-0.98)	13	0.37 (0.23-0.58)	4	0.78 (0.59-1.04)
BCG vaccination[53]	Tuberculosis	3	0.26 (0.14-0.48)	10	0.58 (0.42-0.82)	3	0.26 (0.14-0.48)	10	0.58 (0.42-0.82)
Calcium supplementation in pregnancy (5938†)	Hypertension	5	0.25 (0.14-0.42)	1	0.67 (0.49-0.92)	NA	NA	NA	NA
Diethylstilbestrol in pregnancy (2891†)	Miscarriage	5	1.03 (0.82-1.29)	1	1.92 (1.08-3.41)	NA	NA	NA	NA
Warfarin vs aspirin in nonvalvular atrial fibrillation[40]	Stroke	3	0.43 (0.30-0.61)	1	0.92 (0.61-1.38)	NA	NA	NA	NA
Anti–Rh-D prophylaxis post partum (3314†)	Immunization after 6 mo	8	0.09 (0.04-0.20)	2	0.01 (0.00-0.05)	NA	NA	NA	NA
Free bleeding from placental end of umbilical cord (4004†)	Placentomaternal infusion	NA	NA	NA	NA	2	0.61 (0.49-0.75)	1	1.01 (0.66-1.52)
Advice as a strategy for reducing smoking in pregnancy (3394†)	Continued smoking	NA	NA	NA	NA	4	0.94 (0.90-0.98)	1	0.99 (0.98-1.01)
Continuous vs interrupted sutures for perineal repair (3252†)	Dyspareunia	NA	NA	NA	NA	2	0.65 (0.50-0.84)	1	1.19 (0.94-1.51)
Transabdominal vs transcervical sampling (6005†)	Adequate sample	NA	NA	NA	NA	5	0.51 (0.33-0.78)	1	0.85 (0.66-1.12)
Amnioinfusion for meconium-stained liquor (5379†)	Meconium below vocal cords	NA	NA	NA	NA	4	0.32 (0.17-0.59)	2	0.10 (0.04-0.22)
Prophylactic oxytocin in third stage of labor (2974†)	Postpartum hemorrhage	NA	NA	NA	NA	5	0.64 (0.37-1.12)	3	0.33 (0.25-0.45)
Ergonovine maleate vs oxytocin in third stage of labor (2999†)	Blood loss >500 mL	NA	NA	NA	NA	3	0.51 (0.37-0.69)	1	0.90 (0.77-1.05)

*Risk ratios for multiple trials were pooled with the random effects model[18]; No. indicates number of trials; CI, 95% confidence interval; and NA, not applicable.
†Review number from the Cochrane database.[14]

in most (82%-90%) of the meta-analyses with the random effects model, whereas about twice as many disagreements were observed with the fixed effects model. For most of the statistical disagreements between large and smaller studies, a smaller treatment benefit was estimated in large studies than in smaller studies (Table 2). As shown in Table 2, of the 15 discrepancies identified with the power approach, in 9 cases large trials failed to confirm the benefit shown by the meta-analysis of smaller trials; in 2 cases large trials showed a significant treatment effect that was not evident in the meta-analyses of smaller trials; and in 4 cases both large trials and smaller trials indicated a statistically significant benefit.

Accounting for Disagreements

Table 2 contains the results of 15 specific meta-analyses in which large and smaller studies disagreed by random effects calculations. Four disagreed by both ways of defining a large study, and each of those 4 involved megatrials. What follows are possible explanations for the set of 15 disagreements (Table 3).

Control Rate

In 5 of the 15 disagreements, the RRs were significantly related to the control rate of events across studies (Table 3), and in 1 additional case a suggestion of a relationship was found ($P=.09$). Therefore, when a higher proportion of patients in the control arm suffered adverse events, magnesium and nitrates in acute myocardial infarction became more efficacious in lowering rates of mortality; antiplatelet agents became more efficacious in lowering rates of preeclampsia; free bleeding from the placental end of the umbilical cord became more efficacious in lowering rates of placentomaternal infusion; continuous sutures for perineal repair became more efficacious than interrupted sutures in lowering rates of dyspareunia; and oxytocin became more efficacious than ergonovine maleate (Syntometrine) in the third stage of labor in lowering rates of blood loss in excess of 500 mL.

For 4 of these 6 meta-analyses, both the control rate in the largest study and the pooled control rate of multiple large studies (when more than 1 large study was available) were significantly different from the corresponding pooled control rate of smaller studies (Table 4). In the meta-analysis on magnesium in acute myocardial infarction, only the control rate of the largest study was significantly different from the pooled control rate of the smaller studies (Table 4). The pooled control rate of the smaller studies was significantly higher than the pooled control rate of multiple large studies and the control rate of the largest study for those meta-analyses with significant differences, with the exception of the meta-analysis on ergonovine maleate vs oxytocin in the third stage of labor. In the meta-analysis on perineal repair, the control rate of the large study was less than half the control rate of the 2 smaller studies, but the difference did not reach formal statistical significance ($P=.26$).

Publication Bias

Evidence of a relationship between the treatment effect and its variance was found in the meta-analysis on advice as a strategy for reducing smoking in pregnancy (Table 3). This raises the concern of potential publication bias against small studies with negative results. Two other meta-analyses were also suggestive of such a relationship, although the finding did not reach formal statistical significance ($P<.10$ level).

Protocol or Study Differences

Authors of published meta-analyses for 4 treatments have mentioned specific protocol or study differences that are likely to explain, at least in part, the statistically significant difference in RRs between

Table 3.—Potential Reasons for Explaining Discrepancies Between Large and Smaller Studies*

Meta-analysis	Definition of a Large Study	Change in Risk Ratio (%) With Control Rate†	Publication Bias Rank Correlation‡	Specific Protocol Reason?‡	Other Reason?‡
Magnesium in acute myocardial infarction[1]	Size, power	−10.0 (−14.0, −6.0)§	0.16	No	No
Nitrates in acute myocardial infarction[1]	Size, power	−2.8 (−5.2, −0.3)§	−0.06	No	No
Antiplatelet agents in pregnancy (4000‖)	Size, power	−9.0 (−18.4, 1.3)¶	0	No	No
BCG vaccination[53]	Size, power	−2.5 (−6.8, 1.9)	−0.02	Yes	No
Calcium supplementation in pregnancy (5938‖)	Size	−2.0 (−4.8, 6.7)	−0.06	No	Yes
Diethylstilbestrol in pregnancy (2891‖)	Size	−0.8 (−2.0, 0.4)	0.06	Yes	No
Warfarin vs aspirin in nonvalvular atrial fibrillation[40]	Size	−2.8 (−7.6, 2.3)	−0.33	No	Yes
Anti–Rh-D prophylaxis post partum (3314‖)	Size	−6.4 (−23.0, 14.4)	0.22	No	Yes
Free bleeding from placental end of umbilical cord (4004‖)	Power	−1.1 (−1.9, −0.3)§	1#	No	No
Advice as a strategy for reducing smoking in pregnancy (3394‖)	Power	3.2 (−3.3, 10.1)	−0.73¶	No	No
Continuous vs interrupted sutures for perineal repair (3252‖)	Power	−1.2 (−3.0, −0.01)§	−1#	Yes	No
Transabdominal vs transcervical sampling (6005‖)	Power	13.4 (−9.8, 42.5)	0.33	Yes	No
Amnioinfusion for meconium-stained liquor (5379‖)	Power	−1.9 (−13.6, 11.2)	0.2	No	Yes
Prophylactic oxytocin in third stage of labor (2974‖)	Power	−1.6 (−9.4, 6.8)	0	No	Yes
Ergonovine maleate vs oxytocin in third stage of labor (2999‖)	Power	6.0 (1.2, 11.1)§	0	No	No

*Outcomes listed in Table 1.
†Represents the percentage of change in the risk ratio for every 1% absolute increase in the control rate. Values in parentheses represent 95% confidence limits.
‡See text for details.
§*P*<.05.
‖Review number from the Cochrane database.[14]
¶.05<*P*<.10.
#*P*=.12.

Table 4.—Comparison of Control Rates for Selected Meta-analyses*

Meta-analysis	Definition of a Large Study	Pooled Control Rate of Smaller Studies (95% CI)	Pooled Control Rate of Multiple Large Studies (95% CI)	Pooled Control Rate of All Studies Except the Largest (95% CI)	Control Rate of Largest Study (95% CI)
Magnesium in acute myocardial infarction[1]	Size, power	0.097 (0.062-0.149)	0.085 (0.061-0.118)	0.101 (0.073-0.139)	0.072 (0.069-0.075)‡
Nitrates in acute myocardial infarction[1]	Size	0.139 (0.103-0.183)	0.069 (0.061-0.078)†	0.123 (0.084-0.176)	0.073 (0.070-0.076)‡
	Power	0.145 (0.107-0.194)	0.073 (0.064-0.085)†	0.123 (0.084-0.176)	0.073 (0.070-0.076)‡
Antiplatelet agents in pregnancy (4000§)	Size	0.155 (0.106-0.223)	0.052 (0.034-0.078)†	0.124 (0.081-0.185)	0.076 (0.068-0.084)‡
	Power	0.173 (0.126-0.233)	0.053 (0.037-0.075)†	0.124 (0.081-0.185)	0.076 (0.068-0.084)‡
Free bleeding from placental end of umbilical cord (4004§)	Power	0.777 (0.540-0.912)	NA	0.770 (0.540-0.912)	0.309 (0.230-0.401)‡
Continuous vs interrupted sutures for perineal repair (3252§)	Power	0.440 (0.105-0.840)	NA	0.440 (0.105-0.840)	0.212 (0.177-0.253)
Ergonovine maleate vs oxytocin in third stage of labor (2999§)	Power	0.048 (0.017-0.129)	NA	0.048 (0.017-0.129)	0.166 (0.149-0.184)‡

*Outcomes are listed in Table 1. CI indicates 95% confidence interval; and NA, not applicable.
†*P*<.05 compared with pooled control rate of smaller studies.
‡*P*<.05 compared with pooled control rate of all studies except the largest.
§Review number from the Cochrane database.[14]

large and smaller studies (Table 3). These include the meta-analyses on BCG vaccination for prevention of tuberculosis,[53] where distance from the equator (degrees of latitude) corresponded to greater vaccine efficacy (mean latitude of 10 large studies = 29.4°; mean latitude of 3 smaller studies = 47.0°; *P*=.06 for test of difference in means); diethylstilbestrol in pregnancy,[14] where only the largest study appeared to have used methods of allocation likely to preclude foreknowledge of the assigned treatment; continuous vs interrupted sutures for perineal repair,[14] where the increased risk of dyspareunia in the single large study may have resulted from too early resumption of sexual intercourse; and transabdominal vs transcervical villous sampling for perinatal diagnosis,[14] where the operators in a large study had greater prior experience with the transabdominal method.

No Specific Reason

There was no apparent specific reason for the statistical differences between the results of large and smaller studies for the remaining 5 meta-analyses (Table 3). In 2 of them (anti–Rh-D prophylaxis post partum and amnioinfusion for meconium-stained liquor) the differences were not clinically important: both large and smaller studies showed a very large and statistically significant benefit from treatment (Table 2).

The meta-analysis on routine calcium supplementation in pregnancy[14] included studies conducted in very different populations,[55] which might have led to disagreement. The meta-analysis on prophylactic oxytocin in the third stage of labor[14] included studies of variable quality, as stated in the Cochrane report,[14] and no statistical difference existed when the RR of 1 of the largest trials taken to have the least bias was compared with the pooled RR of the 5 smaller studies (*P*=.39). Finally, there was no obvious reason for the statistical discrepancy between large and smaller trials for the meta-analysis of warfarin vs aspirin in nonvalvular atrial fibrillation.[40] However, all 4 studies included in this meta-analysis were of comparable size, even though only 1 had a little more than 1000 patients.

COMMENT

Our investigation shows that the results of smaller trials are usually compatible with the results of larger trials. This is true when studies are deemed large by their sample size or power, although the 2 approaches tend to select substantially different sets of studies. While smaller and large studies are likely to agree in their results, clear-cut discrepancies do occur and their frequency is more substantial when the results are analyzed without considering the variability of treatment effect among different large trials and among different smaller trials (ie, with a fixed effects model). Potential explanations for most of the genuine disagreements may be identified in control rate differences, specific protocol or study differences, and publication bias, as well as methodological factors such as the quality of primary studies.[56] Clinically important disagreements without identifiable explanations are uncommon.

Previous investigations assessed agreements and disagreements descriptively,[9,10] based on the concordance of the direction and statistical significance of the treatment effect (whether or not P<.05). However, the results of large and smaller trials may not be statistically different even if they disagree in terms of their own statistical significance, and we encountered several such examples in our analysis. Conversely, studies may give the same direction and agreement about statistical significance for the treatment effect but significantly differ both clinically and statistically. Our analysis has addressed these concerns: clinically, by evaluating the clinical significance and reasons for the discrepancies; and statistically, by taking into account chance fluctuations in the comparison between large and smaller studies (with the z statistic) and potential heterogeneity or diversity among large studies and among smaller studies (with the random effects model).

At least 3 aspects of our statistical methodology should be discussed. First, we have defined a "large" trial in 2 ways—by size and by power—and have found more disagreements between large and smaller trials when *large* was defined by power. This discrepancy suggests that the definition of a large trial affects the comparison of large trials and smaller trials, and that further research is needed to address this issue. Second, although it may be worthwhile to analyze the whole body of published meta-analyses by the power rule in future research, we have no a priori reason to believe that the results and conclusions would be affected by such a laborious undertaking. Third, our analysis used RRs as the measure of effect because most treatment effects tend to be multiplicative; therefore, this measure is probably most appropriate for physicians.[57] In some cases, though, the absolute magnitude of the treatment effect may be at least as important and so future analyses may also need to address risk differences.

There are obvious and unavoidable differences between large and smaller trials in their design, implementation, and analysis such as the number of sites, different locations, and target populations. The unavoidable diversity between smaller and large trials is reflected in the substantial number of disagreements we found with the fixed effects method. The disagreements between large and smaller studies are reduced by half, however, when this heterogeneity is taken into account using random effects calculations. Depending on the situation, a fixed or a random effects model can be justified,[58,59] but it is clear that assessing and evaluating the diversity among studies is at least as important as obtaining their pooled results.

Our investigation is retrospective and not designed to decide whether a meta-analysis of smaller trials is sufficient or a megatrial is warranted. Nonetheless, some practical recommendations can be made to help interpret future discordance. Clinicians and other decision makers should realize that publication bias, study protocol, and variability in the control rate of events in different trials may underlie discordant results. It would be worthwhile to investigate by meta-regression or analysis of individual patient data whether analysis of these factors may even predict potentially discordant results in populations different from those studied in the trials. Hypotheses that emanate from this information can be formally investigated in future trials.

Both large studies and meta-analyses of smaller studies add value to clinical decision making and should be interpreted as offering a continuum of experience. Each has its advantages and limitations.[60,61] The results of many diverse smaller studies may reflect the natural heterogeneity of treatment effectiveness found in day-to-day clinical practice.[9] Large studies may produce a more precise answer to a particular question when the treatment effect is not large but is clinically important.[62] In the absence of megatrials, a well-conducted meta-analysis may reflect the best synthesis of available evidence. In many circumstances, it is difficult to decide whether a megatrial is needed or a meta-analysis of relatively small trials is sufficient for clinical decision making. In that regard, further research in understanding the sources of heterogeneity between large and smaller trials should increase the clinical relevance of both large and smaller trials.

This work was supported by grant T32 AI07389 from the National Institutes of Health and grants R01 HS07782 and R01 HS08532 from the Agency for Health Care Policy and Research of the US Public Health Service. We are thankful to Martin McIntosh, PhD, for his technical assistance.

Appendixes on the description and data of all meta-analyses selected are available on request from the National Auxiliary Publications Service, document 05349. Order from NAPS, c/o Microfiche Publications, PO Box 3513, Grand Central Station, New York, NY 10163-3513. Remit in advance, in US funds only, $7.75 for photocopies or $5 for microfiche. Outside the United States and Canada, add postage of $4.50. The postage charge for any microfiche order is $1.50.

This article is dedicated to the memory of Thomas C. Chalmers, MD.

References

1. Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N Engl J Med.* 1992;327:248-254.
2. Lau J, Chalmers TC. Should intravenous magnesium be given to patients with acute myocardial infarction? *Clin Res.* 1994;42:290A.
3. Borzak S, Ridker PM. Discordance between meta-analyses and large-scale randomized, control trials. *Ann Intern Med.* 1995;123:873-877.
4. Domanski MJ, Friedman LM. Relative role of meta-analysis and randomized controlled trials in the assessment of medical therapies. *Am J Cardiol.* 1994;74:395-396.
5. Woods KL. Mega-trials and management of acute myocardial infarction. *Lancet.* 1995;346:611-614.
6. Yusuf S, Flather M. Magnesium in acute myocardial infarction. *BMJ.* 1995;310:751-752.
7. Egger M, Davey Smith G. Misleading meta-analysis. *BMJ.* 1995;310:752-754.
8. Morris JL, Cowan Campbell J. Nitrates in myocardial infarction: a current perspective. *Can J Cardiol.* 1995;11:5B-10B.
9. Chalmers TC, Levin H, Sacks HS, Reitman D, Berrier J, Nagalingam R. Meta-analysis of clinical trials as a scientific discipline, I: control of bias and comparison with large co-operative trials. *Stat Med.* 1987;6:315-325.
10. Villar J, Carroli G, Belizan JM. Predictive ability of meta-analyses of randomised controlled trials. *Lancet.* 1995;345:772-776.
11. Flournoy N, Olkin I. Do small trials square with large ones? *Lancet.* 1995;345:741-742.
12. Lumley T, Keech A. Meta-meta-analysis with confidence. *Lancet.* 1995;346:576-577.
13. Gonser M, Vetter K, Noack F. Meta-analyses of interventional trials done in populations with different risks. *Lancet.* 1995;345:1304-1305.
14. Enkin MW, Keirse MJNC, Renfrew MJ, Neilson JP, eds. *Pregnancy and Childbirth Module: Cochrane Database of Systematic Reviews (Cochrane Updates on Disk).* Oxford, England: Update Software; 1994. Disk issue 1.
15. Johnson EG, Tukey JW. Graphical exploratory analysis of variance illustrated on a splitting of the Johnson and Tsao data. In: Mallows CL, ed. *Design, Data, and Analysis: By Some Friends of Cuthbert Daniel.* New York, NY: John Wiley & Sons; 1987:171-244.
16. Dean AG, Dean JA, Burton AH, Dicker RC. *Epi Info, Version 5: A Word Processing, Database, and Statistics Program for Epidemiology on Microcomputers.* Atlanta, Ga: Centers for Disease Control; 1990.
17. Rothman KJ. *Modern Epidemiology.* Boston, Mass: Little Brown & Co; 1986:177-236.
18. Ioannidis JPA, Cappelleri JC, Lau J, et al. Early or deferred zidovudine therapy in HIV-infected patients without an AIDS-defining illness: a meta-analysis. *Ann Intern Med.* 1995;122:856-866.
19. Schmid CH, McIntosh M, Cappelleri JC, Lau J, Chalmers TC. Measuring the impact of the control rate in meta-analysis of clinical trials. *Control Clin Trials.* 1995;16:66S.
20. Antman EM. Randomized trials of magnesium in acute myocardial infarction: big numbers do not tell the whole story. *Am J Cardiol.* 1995;75:391-393.
21. Thompson SG. Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. *Stat Methods Med Res.* 1993;2:173-192.

22. Smith GD, Song F, Sheldon RA. Cholesterol lowering and mortality: the importance of considering initial level of risk. *BMJ.* 1993;306:1367-1373.

23. Boissel JP, Collet JP, Lievre M, Girard P. An effect model for the assessment of drug benefit: example of antiarrhythmic drugs in postmyocardial infarction patients. *J Cardiovasc Pharmacol.* 1993; 22:356-363.

24. Abramson JH. *Making Sense of Data: A Self-instruction Manual on the Interpretation of Epidemiologic Data.* 2nd ed. New York, NY: Oxford University Press; 1994:319-389.

25. Rotwell M. Can overall results of clinical trials be applied to all patients? *Lancet.* 1995;345:1616-1619.

26. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc.* 1977;38:1-22.

27. McIntosh M. The population risk as an explanatory variable in research synthesis of clinical trials. *Stat Med.* 1996;15:1713-1728.

28. Cox DR, Snell EJ. *Analysis of Binary Data.* 2nd ed. New York, NY: Chapman & Hall; 1989:19-20.

29. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics.* 1994;50:1088-1101.

30. Lau J. *Meta-Analyst.* Version 0.988. Boston, Mass: New England Medical Center; 1995.

31. MathSoft, Inc. *Mathcad Plus 6.0.* Cambridge, Mass: MathSoft Inc; 1995.

32. Statistical Sciences Inc. *S-PLUS User's Manual. Version 3.2.* Seattle, Wash: StatSci, MathSoft Inc; 1993.

33. SAS Institute Inc. *The SAS System for Windows. Version 6.10.* Cary, NC: SAS Institute Inc; 1994.

34. Halliday HL. Overview of clinical trials comparing natural and synthetic surfactants. *Biol Neonate.* 1995:67(suppl 1):32-47.

35. ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group. ISIS-4: a randomized factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58 050 patients in suspected acute myocardial infarction. *Lancet.* 1995;345:669-685.

36. Jovell AJ, Lau J, Berkey C, Kupelnick B, Chalmers TC. Early angiography and angioplasty following thrombolytic therapy of acute myocardial infarction: metaanalyses of the randomized control trials. *Online J Curr Clin Trials* [serial online]. June 5, 1993; doc 67.

37. Basinki A, Naylor CD. Aspirin and fibrinolysis in acute myocardial infarction: meta-analytic evidence for synergy. *J Clin Epidemiol.* 1991;44:1085-1096.

38. Naylor CD, Jaglal SB. Impact of intravenous thrombolysis on short-term coronary revascularization rates: a meta-analysis. *JAMA.* 1990;264:697-702.

39. Antiplatelet Trialists' Collaboration. Collaborative overview of randomised trials of antiplatelet therapy, I: prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients. *BMJ.* 1994;308: 81-106.

40. Barnett HJM, Eliasziw M, Meldrum HE. Drugs and surgery in the prevention of ischemic stroke. *N Engl J Med.* 1995;332:238-248.

41. Pocock SJ, Henderson RA, Rickards AF, et al. Meta-analysis of randomised trials comparing coronary angioplasty with bypass surgery. *Lancet.* 1995; 346:1184-1189.

42. Cappelleri JC, Lau J, Kupelnick B, Chalmers TC. Efficacy and safety of different aspirin dosages on vascular diseases: a meta-regression analysis. *Online J Curr Clin Trials* [serial online]. March 14, 1995; doc 174.

43. Early Breast Cancer Trialists' Collaborative Group. Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy: 133 randomised trials involving 31 000 recurrences and 24 000 deaths among 75 000 women. *Lancet.* 1992;339:1-15, 71-85.

44. Early Breast Cancer Trialists' Collaborative Group. Effects of radiotherapy and surgery in early breast cancer: an overview of the randomized trials. *N Engl J Med.* 1995;333:1444-1455.

45. Leizorovicz A, Haugh MC, Chapuis FR, Samama MM, Boissel JP. Low molecular weight heparin in prevention of perioperative thrombosis. *BMJ.* 1992;305:913-920.

46. Antiplatelet Trialists' Collaboration. Collaborative overview of randomised trials of antiplatelet therapy, III: reduction in venous thrombosis and pulmonary embolism by antiplatelet prophylaxis among surgical and medical patients. *BMJ.* 1994; 308:235-246.

47. Antiplatelet Trialists' Collaboration. Collaborative overview of randomised trials of antiplatelet therapy, II: maintenance of vascular graft or arterial patency by antiplatelet therapy. *BMJ.* 1994; 308:159-168.

48. Insua JT, Sacks HS, Lau TS, et al. Drug treatment of hypertension in the elderly: a meta-analysis. *Ann Intern Med.* 1994;121:355-362.

49. Collins R, Peto R, MacMahon S, et al. Blood pressure, stroke and coronary heart disease, II: short-term reductions in blood pressure: overview of randomised drug trials in their epidemiological context. *Lancet.* 1990;335:827-838.

50. Garg R, Yusuf S, for the Collaborative Group on ACE Inhibitor Trials. Overview of randomized trials of angiotensin-converting enzyme inhibitors on mortality and morbidity in patients with heart failure. *JAMA.* 1995;273:1450-1456.

51. Nony P, Boissel JP, Lievre M, et al. Evaluation of the effect of phosphodiesterase inhibitors on mortality in chronic heart failure patients: a meta-analysis. *Eur J Clin Pharmacol.* 1994;46:191-196.

52. Glasziou PP, Mackerras DEM. Vitamin A supplementation in infectious diseases: a meta-analysis. *BMJ.* 1993;306:366-370.

53. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Stat Med.* 1995;14:395-411.

54. Fiore MC, Smith SS, Jorenby DE, Baker TB. The effectiveness of the nicotine patch for smoking cessation: a meta-analysis. *JAMA.* 1994;271:1940-1947.

55. Belizan JM, Villar J, Gonzalez L, Campodonico L, Bergel E. Calcium supplementation to prevent hypertensive disorders of pregnancy. *N Engl J Med.* 1991;325:1399-1405.

56. Khan KS, Daya S, Jadad AR. The importance of quality of primary studies in producing unbiased systematic reviews. *Arch Intern Med.* 1996;156: 661-666.

57. Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *J Clin Epidemiol.* 1994;47:881-889.

58. Hedges LV. Fixed effects models. In: Cooper H, Hedges LV, eds. *The Handbook of Research Synthesis.* New York, NY: Russell Sage Foundation; 1994:285-299.

59. Raudenbush SW. Random effects models. In: Cooper H, Hedges LV, eds. *The Handbook of Research Synthesis.* New York, NY: Russell Sage Foundation; 1994:301-321.

60. Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. In: Warren KS, Mosteller F, eds. *Doing More Good Than Harm: The Evaluation of Health Care Interventions.* New York, NY: New York Academy of Sciences; 1993:314-340.

61. Sackett DL, Cook DJ. Can we learn anything from small trials? In: Warren KS, Mosteller F, eds. *Doing More Good Than Harm: The Evaluation of Health Care Interventions.* New York, NY: New York Academy of Sciences; 1993:25-31.

62. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med.* 1984;3: 409-420.