Evolution of treatment effects over time: Empirical insight from recursive cumulative metaanalyses

John P. A. Ioannidis*[†] and Joseph Lau*[‡]

*Division of Clinical Care Research, New England Medical Center, Tufts University School of Medicine, Boston, MA 02111; and [†]Clinical Trials and Evidence-Based Medicine Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece

Communicated by Frederick Mosteller, Harvard University, Cambridge, MA, November 6, 2000 (received for review April 27, 2000)

Evidence on how much medical interventions work may change over time. It is important to determine what fluctuations in the treatment effect reported by randomized trials and their metaanalyses may be expected and whether extreme fluctuations signal future major changes. We applied recursive cumulative metaanalysis of randomized controlled trials to evaluate the relative change in the pooled treatment effect (odds ratio) over time for 60 interventions in two medical fields (pregnancy/perinatal medicine, $n = 45$ interventions; myocardial infarction, $n = 15$ interventions). We evaluated the scatter of relative changes for different numbers of total patients in previous trials. Outlier cases were noted with changes greater than 2.5 standard deviations of the expected. With 500 accumulated patients, the pooled odds ratio may change by 0.6- to 1.7-fold in the immediate future. When 2000 patients have already been randomized, the respective figures are between 0.74- and 1.35-fold for pregnancy/perinatal medicine and between 0.83- and 1.21-fold for myocardial infarction studies. Extreme early fluctuations in the treatment effect were observed in three interventions (magnesium in myocardial infarction, calcium and antiplatelet agents for prevention of preeclampsia), where recent mega-trials have contradicted prior metaanalyses, as well as in four other examples where early large treatment effects were dissipated when more data appeared. Past experience may help quantify the uncertainty surrounding the treatment effects reported in early clinical trials and their metaanalyses. Early wide oscillations in the evolution of the treatment effect for specific interventions may sometimes signal further major changes in the future.

mega-trials | heterogeneity | randomized trials

Randomized trials and metaanalyses are often considered primary means for assessing the efficacy of medical interventions (1–4). However, clinical evidence evolves over time: new trials continue to be performed, replacing, updating, and supplementing the knowledge obtained from earlier ones. Heterogeneity (i.e., variability) among trial results is unavoidable. New trials may strengthen our prior beliefs about the magnitude of a treatment effect (i.e., how much a treatment works), or, in some cases, they may alter these beliefs or invalidate them. Several recent examples (5–8) have been encountered where large randomized trials reached entirely different conclusions when compared with metaanalyses of earlier trials of small or even large sample size on the same question. For example, recent large trials seemed to invalidate our prior beliefs about the efficacy of treatments such as magnesium salts and nitrates for reducing overall mortality in acute myocardial infarction (9) or aspirin (10) and calcium supplementation (11) for the prevention of preeclampsia during pregnancy. Prior expectations of 30–60% reductions in mortality and preeclampsia, respectively, based on early trials, were not confirmed.

Important questions arise: Could we have anticipated these discrepancies? Also, how uncertain should we be about the treatment effects reported by metaanalyses? How much may treatment effects change as data from trials accumulate? In other words, how much uncertainty should there be on how much treatments work?

Cumulative metaanalysis (12) provides a framework for updating the summary results from all trials in a given question as evidence accumulates. An extension of the method, recursive cumulative metaanalysis (13), shows the relative change in the magnitude of the treatment effect as each piece of evidence is obtained. With the advent of evidence-based medicine, many metaanalyses have been performed. These metaanalyses offer empirical evidence on how much the treatment effect has changed over time for several interventions in various medical fields. This empirical evidence may be used to estimate our uncertainty about a given reported treatment effect, based on what has ensued in previous similar circumstances. In the present report, we used two large databases of 60 metaanalyses to obtain empirical evidence of the expected range of change in the treatment effect over time in two medical fields. Furthermore, we attempted to determine whether the evolution of the changes in the treatment effect for certain interventions over time could predict major changes altering our belief in the efficacy of these interventions in the future.[§]

Methods

Databases. We used 60 metaanalyses of randomized trials of therapeutic and preventive interventions in pregnancy and perinatal medicine ($n = 45$) and management of myocardial infarction ($n = 15$). These were derived from a previous database of empirical work on metaanalysis, details on which are described elsewhere (14). A metaanalysis was included if it contained more than five trials that had been published in more than three different calendar years. This rule was applied to target interventions where there was already some meaningful history on the evolution of the treatment effect over time.

The pregnancy/perinatal database was derived by screening all of the metaanalyses of the *Cochrane Pregnancy and Childbirth Database* (1994 edition) (15). Pregnancy and perinatal medicine is a field in which metaanalyses have been performed extensively. It thus offers a unique opportunity for examining the evolution of treatment effects over time in a given medical field, avoiding strong selection biases. The metaanalyses of the management of myocardial infarction (acute therapy and secondary prevention) were derived from a comprehensive screening of six journals likely to publish high-quality metaanalyses in this field (*Lancet, JAMA, New England Journal of Medicine, Annals of Internal Medicine, Archives of Internal Medicine*, and *Circulation*) for the years 1988 to 1995. This field was chosen because there were already several large-scale metaanalyses on the management of myocardial infarction. These

Abbreviation: AMI, acute myocardial infarction.

[‡]To whom reprint requests should be addressed at: Division of Clinical Care Research, New England Medical Center Hospitals, Box 63, 750 Washington Street, Boston, MA 02111. E-mail: jlau1@lifespan.org.

[§]Supplemental information for this article is available at http://www.nemc.org/dccr/Projects/RCMA/supplements.htm.

Article published online before print: *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.021529998. Article and publication date are at www.pnas.org/cgi/doi/10.1073/pnas.021529998

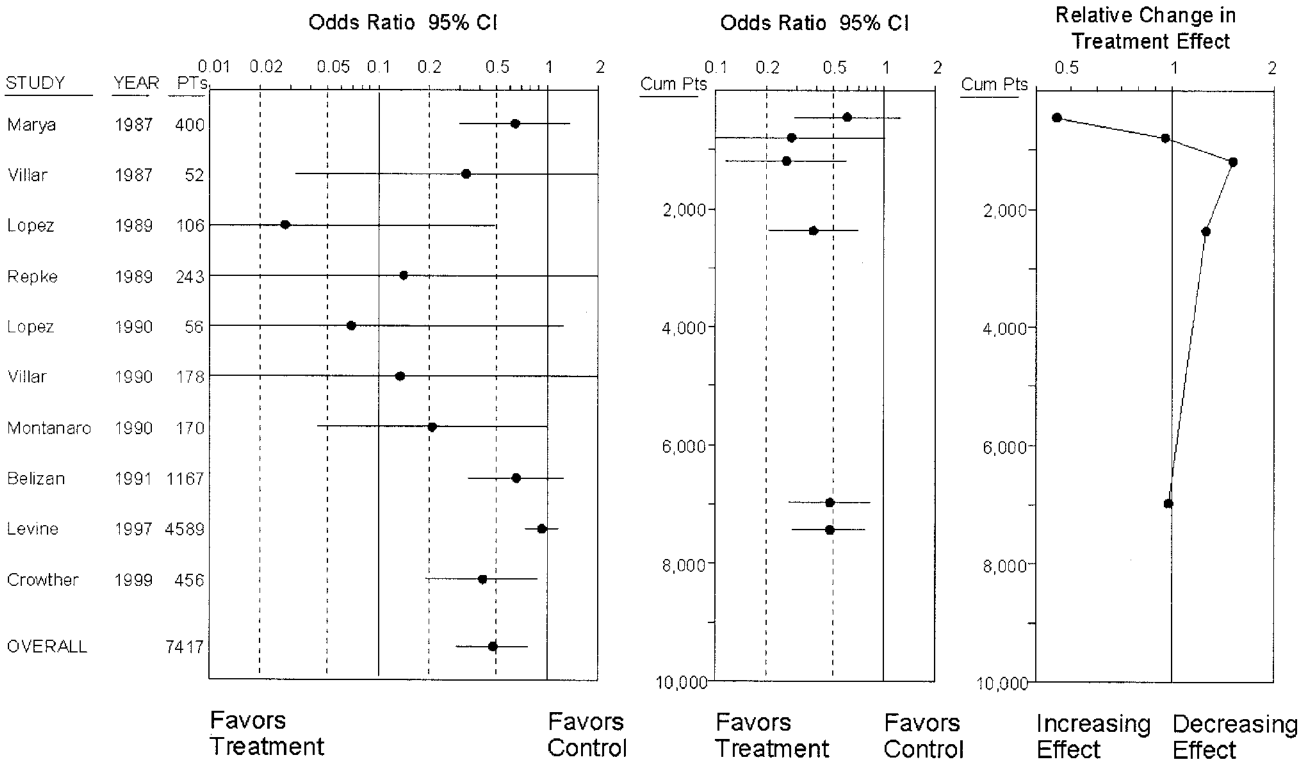

Fig. 1. Illustrative example of a standard metaanalysis on calcium supplementation for prevention of preeclampsia (*Left*), the respective cumulative metaanalysis (*Center*), and the respective recursive cumulative metaanalysis (*Right*). Cumulative metaanalysis is performed at the end of each year when new trials have been published in this year. The odds ratio of the cumulative metaanalysis at the end of each year is displayed according to the cumulative number of patients. The relative change in the odds ratio is also displayed according to the same scale. Calculations are based on random effects. PTs, patients; Cum Pts, cumulative patients; CI, confidence interval.

also include some of the most hotly debated discrepancies between metaanalyses and large randomized trials (16–19).

Trials published until early 1994 (Cochrane database) or early 1995 (myocardial infarction database) were considered in the main analysis. The inclusion of these trials allowed us to evaluate whether major discrepant results in trials published more recently could have been predicted on the basis of the prior evolution of the treatment effect for these interventions. To this end, metaanalyses were updated (to June 1999) by consulting the most recent Cochrane database, conducting MEDLINE searches, and perusing bibliographies of recent retrieved articles.

Metaanalysis, Cumulative Metaanalysis, and Recursive Cumulative Metaanalysis. For each intervention, trials were chronologically ordered per publication year, and cumulative metaanalysis (12) was performed to obtain pooled odds ratios at the end of each calendar year. We also noted the total number of patients randomized in published clinical trials (cumulative sample size) at the end of each calendar year. Each calendar year was considered as an information step, in which evidence was updated by trials published in the interim.

We then estimated the relative change in the treatment effect in each information step (13). The relative change was defined as the pooled odds ratio at the next information step divided by the pooled odds ratio at the current information step. Therefore, it provided a measure of how much the treatment effect changes as evidence accumulates. For example, if two trials were published in 1987 and their pooled odds ratio was 0.80 and then another two trials were published in 1990 and the pooled odds ratio of all four trials was 0.96, the relative change at the 1987 information step was estimated as $0.96/0.80 = 1.20$.

In a typical recursive cumulative metaanalysis graph (13), the relative change in treatment effect is plotted as a function of the information steps. In this case, we used the cumulative sample size for each information step. Plots of the relative change as a function of the cumulative sample size show the evolution of the relative changes in the treatment effect for a specific intervention at different numbers of accumulated randomized patients. For visual comparison, Fig. 1 shows side by side a typical metaanalysis, a cumulative metaanalysis, and a recursive cumulative metaanalysis.

Scatter of Relative Change of Treatment Effect for Various Cumulative Sample Sizes. The scatter of the relative change values is expected to be substantially wider in information steps where the cumulative sample size is small and should shrink as cumulative sample size increases. Pooled treatment effects may experience greater change, when based on fewer patients. We therefore obtained a measure of the scatter of the relative change for different values of cumulative sample size across all metaanalyses in each of the two medical fields. We searched with an iterative algorithm for the power, g, that maximizes the log-likelihood function in a linear regression of the form $\log_{10}$(relative change in the pooled odds ratio) $=$ $b{\cdot}\log_{10}$(cumulative sample size) $+ a$, weighted by $w =$ (cumulative sample size)^g. From this weighted regression, 95% prediction intervals were obtained for the range of the relative change given various cumulative sample sizes.

Fixed and Random Effects Calculations. For all analyses, separate calculations were performed with fixed-effects (Mantel–Haenszel) (20) and random-effects (DerSimonian–Laird) (21) pooling methods. The fixed-effects approach assumes no significant heterogeneity between the results of the individual studies being pooled, whereas random-effects calculations allow for

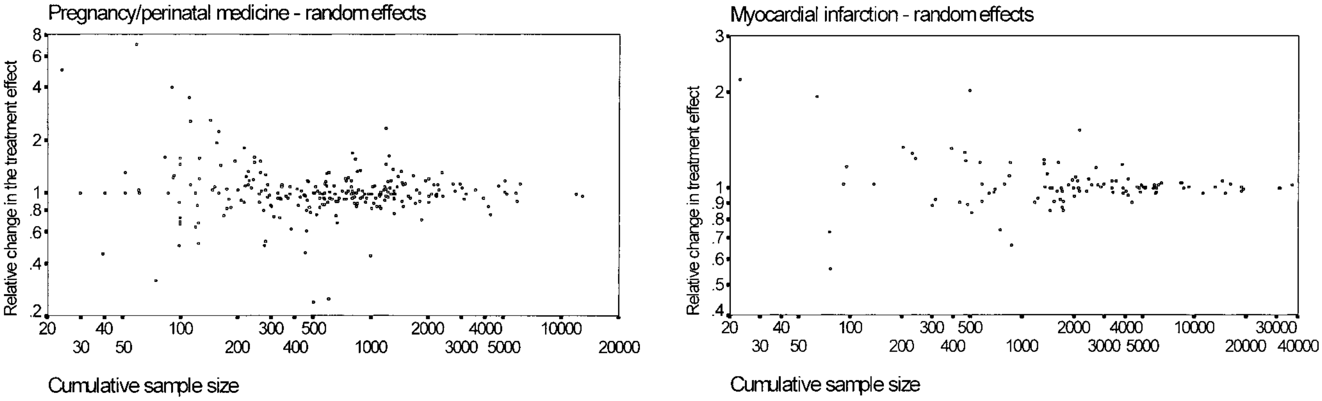

Fig. 2. Scatter plots of the relative change in the pooled odds ratio as a function of the cumulative number of patients in previously published randomized trials. An outlier relative change of 0.05 for sample size = 50 is not shown in the myocardial infarction graph. A relative change greater than 1 means that the odds ratio tends to increase with the accumulation of more data. Fixed effects graphs were similar.

such heterogeneity, and they add an empirical estimate of the between-study variance τ^2 to the within-study variance (22). Theoretically, if the sample size and event rates of the trial(s) at information step $t + 1$ are small, the odds ratio cannot move that far with the fixed-effects approach, but it may still move substantially with random-effects calculations.

Determination of Extreme Cases/Outliers—Prediction of Major Changes in the Future. Finally, we attempted to determine whether extreme/outlier points with extreme width of treatment effect fluctuations in a recursive cumulative metaanalysis for a specific intervention also predicted large relative changes of the treatment effect in the future, as suggested by large discrepancies in the results of recent randomized trials as compared with prior evidence. For each of the two medical fields we estimated for each point the studentized residual from the weighted regression, i.e., how many standard deviations away from the predicted value of relative change each point was. Points that were 2.5 or more deviations away from the value predicted by either fixed-effects or random-effects calculations and 1.8 or more deviations away from the predicted value by the other pooling method were considered as extreme cases/outliers. The rule was set *a priori* to avoid determining outliers largely dependent on the statistical modeling.

The change in the cumulative odds ratio is based on the ratio of two measures that are mutually dependent to some extent; i.e., the information at step $t + 1$ also contains the information at step t. Therefore, in the evaluation of extreme fluctuations, we also considered an approach comparing independent odds ratios. In this approach, the pooled odds ratio at information step t is compared with the odds ratio of the new trials of information step $t + 1$. In this way, the two sets are independent. The extent of heterogeneity between the natural logarithms of these two independent odds ratios is calculated by the standardized z score:

$$z \text{ score} = (\ln \mathrm{OR}_{t,\text{pooled}} - \ln \mathrm{OR}_{t+1})/$$
$$[\mathrm{se}(\ln \mathrm{OR}_{t,\text{pooled}})^2 + \mathrm{se}(\ln \mathrm{OR}_{t+1})^2]^{0.5}$$

This latter approach answers the question "Do the results of the recent trials differ from the results of the previous ones on the same topic?" The disadvantage of this approach is that it does not consider the evidence from other topics on the same medical field. We evaluated the frequency of disagreements between independent sets of trials and whether disagreements could predict major discrepancies as compared with the recursive cumulative metaanalysis approach. Furthermore, we examined the evolution of the DerSimonian–Laird estimate of τ^2 at sequential information steps to see whether early inflations of τ^2 might predict late large fluctuations in the treatment effect.

Results

Scatter and Predicted Range of Relative Changes in the Treatment Effect. The graphs in Fig. 2 show the scatter of the relative changes in the treatment effect for different numbers of accumulated patients in the two chosen medical fields. As expected, when evidence is based on only few patients, there is substantial uncertainty about how much the pooled treatment effect will change in the future. The scatter of the points was somewhat less influenced by sample size in the case of pregnancy/perinatal medicine than for myocardial infarction. The values of all of the regression coefficients were very close to 0 and did not differ significantly from 0 ($P > 0.6$ for all), suggesting that, at all cumulative sample sizes, it would be equally likely for the pooled odds ratio to increase or to decrease in the future. Table 1 shows the 95% prediction intervals for the relative change in the treatment effect for different numbers of patients in the two medical fields. For both fields, when only 100 patients have been randomized, tripling or making one-third of the pooled odds ratio in the immediate future should not be surprising. Even when 500 patients have been randomized, the 95% intervals for the relative change in the odds ratio are between 0.6 and 1.7 approximately. When 2000 patients have already been randomized, it is expected that 95% of the time the relative change in the odds ratio may be approximately between 0.74 and 1.35 for pregnancy/perinatal medicine and between 0.83 and 1.21 for myocardial infarction studies (fixed-effects calculations).

Extreme/Outlier Cases. Table 2 shows the identified extreme value/outlier cases where the relative change in the treatment effect was substantially more pronounced than what would have been anticipated on the basis of the previously accumulated sample size. The Fig. 3 panels show recursive cumulative metaanalysis graphs for four examples of major discrepancies between prior evidence and recent mega-trials mentioned above, along with several other examples of interventions without extreme/outlier oscillations.

Pregnancy/Perinatal Medicine. Both cases where recent large trials altered our prior beliefs in efficacy had shown prominent outliers during the early accumulation of randomized evidence many years before the hotly debated controversial data appeared. In the case of calcium supplementation for the prevention of preeclampsia, the random-effects odds ratio changed from 0.61 to 0.28 between 1987 (when there were 452 randomized patients)

Table 1. 95% prediction intervals for the relative change in the treatment effect (odds ratio) for different numbers of accumulated patients (cumulative sample size, *N*)

	Pregnancy/perinatal		Myocardial infarction	
Patients, *N*	Fixed effects	Random effects	Fixed effects	Random effects
100	0.37–2.78	0.32–3.13	0.18–5.51	0.23–4.43
500	0.59–1.71	0.56–1.71	0.60–1.67	0.63–1.58
1,000	0.67–1.49	0.65–1.53	0.74–1.35	0.76–1.32
2,000	0.74–1.35	0.73–1.37	0.83–1.21	0.84–1.20
15,000	0.85–1.14	0.86–1.15	0.96–1.05	0.96–1.05

For example, if the observed treatment effect for drug A in the treatment of myocardial infarction is provided by a fixed-effects odds ratio of 0.60 based on $N = 2000$ patients randomized to date in published trials until now, then we can be 95% certain that the odds ratio when a next trial(s) appear(s) would be between 0.60·0.83 = 0.50 and 0.60·1.21 = 0.73.

The following regressions were the best-fit ones to the discipline-wide scatter plots presented in the paper: Pregnancy/perinatal medicine–fixed-effects calculations: $\log_{10}$(relative change in odds ratio) = 0.0208−0.0068 log(N). Regression weight: $N^{0.80}$; P values for coefficients: $P = 0.60$ and $P = 0.60$.

Pregnancy/perinatal medicine–random effects calculations: $\log_{10}$(relative change in odds ratio) = 0.0048−0.0017 log(N). Regression weight: $N^{0.85}$; P values for coefficients: $P = 0.91$ and $P = 0.90$.

Myocardial infarction–fixed-effects calculations: $\log_{10}$(relative change in odds ratio) = 0.0033−0.0013 log(N). Regression weight: $N^{1.45}$; P values for coefficients: $P = 0.87$ and $P = 0.78$.

Myocardial infarction–random-effects calculations: $\log_{10}$(relative change in odds ratio) = 0.0002−0.0002 log(N). Regression weight: $N^{1.40}$; P values for coefficients: $P = 0.91$ and $P = 0.996$.

and 1989. It changed again from 0.26 to 0.38 between 1990 (when there were 1205 patients) and 1991. Fixed-effects changes were similar. All of this change preceded the appearance of a large trial ($n = 4589$) (11) in 1997 showing no treatment benefit at all. Of interest, still another trial published in 1999 ($n = 456$) showed a large benefit suggesting a continuation of fluctuations and controversy regarding the treatment effect. In the case of antiplatelet agents for the prevention of preeclampsia, the fixed-effects odds ratio changed dramatically from 0.52 to 0.95 between 1991 (when there were 825 randomized patients) and 1993. This change was before the CLASP trial, with almost 10,000 subjects, also showed no efficacy in 1994 (12).

Outliers were also observed in four other interventions in the pregnancy/perinatal field. In the case of active management of preterm rupture of membranes, the pooled random-effects odds ratio for chorioamnionitis changed from 0.29 (1978) to 1.01 (1984) and continued to perform substantial oscillations [1.52 in 1986, 0.93 in 1992 (1585 patients), 0.86 in 1995 (3286 patients), 0.75 in 1996 (7068 patients)]. Actually, the most recent large trial (23) showed a large, statistically significant effect that was not apparent in any of the previous trials. In the case of elective induction of labor at term, the fixed-effect odds ratio for cesarean section changed from 0.40 to 0.84 between 1989 and 1992. Subsequent trials, including a large one (23), have consistently given results showing a small benefit regarding this outcome. It is conceivable that the large treatment effect reported in the early trials may have been due to publication bias or publication lag (24). Publication lag may have also operated in the case of prophylactic syntometrine vs. ergot derivative, where the pooled random-effects odds ratio for postpartum hemorrhage dramatically changed from 4.33 in 1961 to 1.08 in 1963, and then further decreased to 0.80 by 1965; and in the case of beta-mimetic tocolytics in preterm labor, where the pooled random-effects odds ratio changed from 0.12 in 1979 to 0.85 in 1980 and 0.92 in 1992 and practically remained at this level for the next 7 years. Beta-mimetic agents are very effective in postponing delivery for short periods of time, but their effect on serious complications is much more limited, probably because such complications may be prevented by several other interventions. For the last two examples, mega-trials have not been performed.

Myocardial Infarction. Two of the three outliers occurred in the case of magnesium salts for the treatment of acute myocardial infarction (AMI). Well before the appearance of ISIS-4 data in 1995 (9), which changed dramatically our appreciation of the efficacy of magnesium, in two other calendar times (1987 and 1988) the pooled efficacy of magnesium had changed dramatically. By fixed effects, the pooled odds ratio changed from 0.38 in 1987 to 0.75 in 1988 and 0.51 in 1989. By random effects the respective figures were 0.39, 0.79, and 0.52. Notably, the treatment effect also changed substantially again in 1992, when with the publication of the results of LIMIT-2 [a trial of 2316 patients (25) added to the accumulated evidence of 1621 previously randomized patients] the pooled odds ratio changed from 0.39 to 0.60 by fixed effects (3.92 deviations) and from 0.40 to 0.48 by random effects (1.72 deviations).

The only other outlier case in the myocardial infarction field was one in which an odds ratio of 20.5 had been reported by a small trial on oral anticoagulation for secondary prevention of myocardial infarction, where it would be obvious that the estimated effect

Table 2. Extreme changes/outliers in the relative change of treatment effect

		Relative change in OR (s. res.)	
Topic	Patients, *N*	Fixed effects	Random effects
Pregnancy/perinatal medicine field			
Prophylactic syntometrine vs. ergot derivative	600	0.37 (−4.02)	0.25 (−5.17)
Beta-mimetic tocolytics in preterm labor	59	7.00 (3.07)	7.08 (2.71)
Antiplatelet agents for prevention of preeclampsia	825	1.83 (2.75)	1.56 (1.90)
Elective induction of labor at term	1248	2.10 (4.01)	1.63 (2.49)
Active management of prelabor ROM	112	3.72 (2.67)	3.48 (2.27)
Calcium for prevention of preeclampsia	452	0.57 (−2.04)	0.46 (−2.57)
Myocardial infarction field			
Magnesium salts in AMI*	500	1.97 (2.63)	2.03 (3.01)
Magnesium salts in AMI*	868	0.68 (−2.24)	0.66 (−2.66)
Oral anticoagulants for secondary prevention	50	0.04 (−2.35)	0.05 (−2.56)

*For magnesium salts in AMI, the formal criteria for an outlier were met at two early time-points. OR, odds ratio; s. res., studentized residual (deviations away from predicted); AMI, acute myocardial infarction; ROM, rupture of membranes.

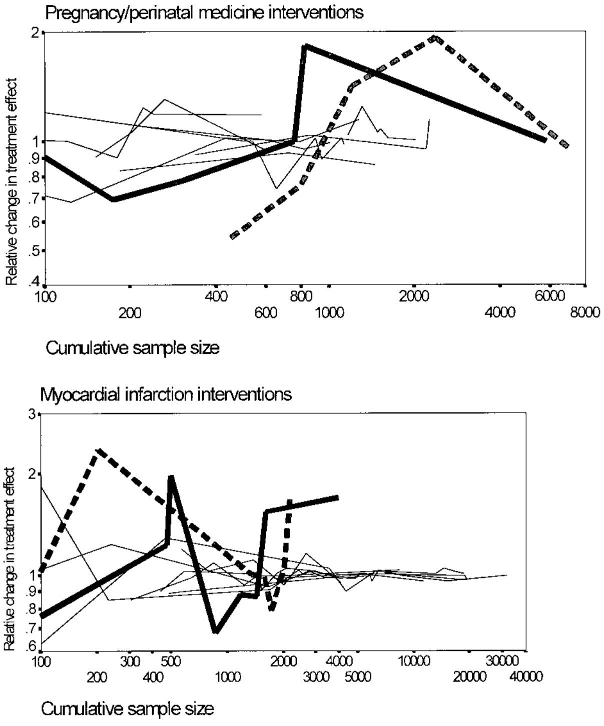

Fig. 3. Recursive cumulative metaanalyses for 10 perinatal topics (*Upper*) and 10 myocardial infarction topics (*Lower*). The controversial cases of antiplatelet agents (bold continuous line) and calcium (bold interrupted line) used to prevent preeclampsia during pregnancy and magnesium salts (bold continuous line) and nitrates (bold interrupted line) to decrease mortality in acute myocardial infarction (AMI) clearly show larger oscillations in the pooled treatment effect over time than the other topics. Recent large trials for these four topics have been included. Other topics depicted in the pregnancy/perinatal interventions panel include balanced protein/energy supplementation in pregnancy, extended spectrum vs. first-generation cephalosporins with cesarean section, prophylactic administration of synthetic surfactant, birthing chair vs. recumbent position for second stage of labor, prophylactic phenobarbital in very-low-birth-weight neonates, prophylactic indomethacin in preterm infants, prophylactic oral betamimetics in pregnancy, and corticosteroids after preterm prelabor rupture of membranes. Other topics depicted in the myocardial infarction interventions panel include beta-blockers in AMI, beta-blockers for secondary prevention, calcium channel blockers for AMI, calcium channel blockers for secondary prevention, oral anticoagulants for AMI, class I antiarrhythmics in AMI, prophylactic lidocaine in AMI, and i.v. streptokinase in AMI. Calculations are based on fixed effects, but random-effects graphs are very similar. A relative change greater than 1 means that the odds ratio tends to increase with the accumulation of more data.

would be unrealistic. In the case of nitrates for myocardial infarction, there were no typical outlier changes before the appearance of the ISIS-4 data, which altered dramatically our appreciation of their efficacy. Nevertheless, the change in the pooled treatment effect between 1984 (when 1731 patients had accumulated) and 1985 was close to the definition of an outlier (2.28 deviations by fixed effects, 1.72 deviations by random effects).

Comparisons of Independent Sets of Odds Ratios. Thirteen metaanalyses (at 17 time points) had absolute z scores greater than 2.50 and 22 metaanalyses had absolute z scores greater than 1.96 (random-effects calculations), when independent odds ratios of past trials and current trials were compared at each year when new trials appeared. These z scores suggest that it is common to see new evidence that disagrees beyond chance with the prior accumulated evidence. However, in the large majority of these cases (10 of the 13, and 17 of the 22), once the new evidence had been incorporated into the past data, the relative change of the treatment effect was not extreme, given the prior sample size and the medical field. More importantly, none of the four metaanalyses where recent mega-trials changed markedly the pooled treatment effect had shown a z score of more than 2.50 in the past, and only one had shown a z score of more than 1.96 at some prior time point. Thus, disagreements between past evidence and new evidence are frequent, but typically the varying new evidence does not change the pooled treatment effect by much, and the presence of such variation does not seem to predict major changes in the future.

Evolution of Between-Study Variance. The τ^2 was greater than 0 in at least one information step in 7/15 myocardial infarction metaanalyses and in 41/45 perinatal metaanalyses. Estimates of τ^2 greater than 0.10 were seen at some point in 4/15 and 35/45 metaanalyses, respectively. These included three of the four metaanalyses where large subsequent fluctuations were observed, whereas for antiplatelet agents to prevent preeclampsia, τ^2 was 0 at all information steps. Inflation of the estimated τ^2 at three successive information steps was seen in 4/15 and 10/45 metaanalyses, respectively. These analyses included magnesium and nitrates for myocardial infarction, but not calcium or antiplatelet agents for preeclampsia.

Discussion

Recursive cumulative metaanalysis provides insight into how much evidence-based beliefs about the efficacy of treatments change over time as evidence accumulates. Our empirical analysis from two different medical fields allows us to estimate what uncertainty there is about how much a treatment works when the evidence is based on different numbers of randomized patients. For both disciplines, the uncertainty decreased drastically with increasing cumulative sample size. With 500 randomized patients, interpretations of treatment benefits have to be cautious, because odds ratios in the range of 0.6–1.7 can easily be dissipated by future evidence. At 2000 patients, odds ratios may still change by as much as 0.74- to 1.35-fold, in the case of the pregnancy/perinatal discipline, and somewhat less in the case of myocardial infarction trials. More than 10,000 patients are required to relieve uncertainty about the first decimal point in the odds ratio of a treatment effect reported by a metaanalysis.

One may estimate whether the new effect seen with a new trial is significantly different from the pooled effect that had been estimated based on previous trials. Significance testing is straightforward because the ratio of the odds ratios and its variance can be calculated from the number of observations in the previous and in the new trials and from the observed effects. In our approach, we go one step further in that instead of basing inferences on the data of one specific metaanalysis, inferences are based on data from many metaanalyses. The prediction intervals aim at estimating what uncertainty there is for the change in effect, given an accumulated number of observations and the previous experience from several other metaanalyses in the same field. Within the same field, trials with the same accumulated number of observations may have different effects and/or different event rates and/or may be followed by trials with different effects and event rates; thus variances in the effect change may indeed be different. The prediction intervals generate empirical estimates for the observed fluctuations in the whole medical field rather than for one specific sequential metaanalysis. We have shown in this paper empirically that the extreme fluctuations identified by this approach have predictive ability for future extreme fluctuations, whereas predictive ability is limited when extreme fluctuations are identified by using inferences based on one specific metaanalysis.

We performed calculations with both fixed and random effects, and inferences were generally similar. However, theoretically it is conceivable that if there is very large between-study heterogeneity, then with random-effects calculations the combined estimate may nearly be the unweighted average of the n studies' log(odds ratios) and variability then would depend on n and not so much on the actual number of accumulated observations. Both models should thus be considered and potential discrepancies screened.

The modest difference in uncertainty between pregnancy/perinatal trials and myocardial infarction trials may be real, the result of selection bias or a chance finding due to sampling error. Myocardial infarction metaanalyses were more selected and derived from influential journals. They may therefore deal with more established treatments, and the average quality of trial design and conduct may have been superior.

Typically, clinical trials are reported with P values for the null hypothesis that the treatment does not work at all. However, these are derived from the data of the specific trial. Trials take into account neither the previous evidence on the same question (12, 13) nor the empirical evidence from other interventions studied in the same medical field. What we propose here is an empirical approach that allows researchers to determine how stable a treatment effect is likely to be on the basis of what has typically ensued in similar settings from past experience. For clinical purposes, knowing how much a treatment is likely to work is usually more important than knowing that we are probably correct in rejecting the null hypothesis (26). The strength borrowed from external evidence is the basic advantage of our approach. Nevertheless, it may also be the main source of limitations, if past experience is not generalizable to the current experience. For example, perhaps quality defects are better addressed in current trials as compared with older ones.

The regular updating of randomized evidence by metaanalyses at annual or biannual intervals is becoming common standard practice in initiatives such as the International Cochrane Collaboration (27). There is justified interest in disseminating high-quality, updated information on how much treatments work to keep medical practice up to date and maximally cost-effective (27). It is important to caution that pooled estimates from metaanalyses may sometimes offer a misleading reassurance with their tight confidence intervals. The recursive metaanalysis approach (13) acknowledges and quantifies the uncertainty inherent in pooled estimates. Moreover, the advent of evidence-based medicine has also allowed the accumulation of several metaanalyses in various medical fields. Thus the empirical estimation of uncertainty should become more precise. We should acknowledge that defining which metaanalyses should be included in a medical field is partly subjective, similar to defining criteria for which trials should be included in a single metaanalysis. With more evidence, comparisons of the magnitude of uncertainty in variously defined medical fields may also show whether treatment effect fluctuations differ substantially in various medical fields and settings.

Recursive cumulative metaanalysis also predicts major changes in the treatment effect estimate that may occur in the future. Unexpectedly wide oscillations in the treatment effect early in the course of accumulating clinical evidence are associated with major changes in the treatment effect in the future. Early fluctuations preceded the major surprises of the discordant results of mega-trials evaluating magnesium in acute myocardial infarction and calcium and antiplatelet agents used to prevent preeclampsia in pregnant women. Figuratively, this pattern is similar to that observed when minor earthquakes precede major catastrophic earthquakes. Although our method cannot predict exactly when the "earthquake" will occur, it nevertheless suggests that when cumulative treatment effect estimates change widely over time, results should be interpreted with caution because large changes, in either direction, may sometimes be observed again in the future. We have also identified examples of wide oscillations where large effects reported in early trials have been dissipated in the presence of more data. Early reported treatment benefits have to be interpreted cautiously. In the presence of wide early oscillations, clinicians should also wait for a more complete picture to evolve.

The definition of "extremes" is not absolute, and there is no reason to believe that specific cut-offs of standardized residuals are necessarily better than others; however, to avoid post hoc interpretations we agreed *a priori* to examine prespecified cut-offs. The standardized residuals may also be seen as a continuous variable without cut-offs. The empirical data suggest that all four major future discrepancies were preceded by early fluctuations of at least 1.72 standardized residuals by random effects and 2.04 standardized residuals by fixed effects in their respective medical fields. Obviously, when extremes are defined by less stringent criteria, then the rate of false positives may also increase (a traditional tradeoff between specificity and sensitivity), and "extreme" changes in the odds ratio may be less significant from a clinical viewpoint.

Several mechanisms may be responsible for large fluctuations of the treatment effect. Potential candidates include publication lag and publication bias (24), heterogeneity in the baseline risk of the studied patient populations (14), quality defects in the conduct and design of trials (28), variability in the treatments used over time, and other unknown sources of diversity. Publication bias and publication lag in particular may provide an explanation for the cases where early large treatment effects are subsequently gradually dissipated with the appearance of more evidence. Full registration and publication of all trials is an ethical imperative (29).

The notion that treatment effects reported in randomized trials and their metaanalyses are absolute constants is unrealistic. Many factors could generate diversity in different populations under different circumstances (30). Heterogeneity in the treatment effects is important to detect and, if possible, to predict. In this regard, recursive cumulative metaanalysis offers a means of appreciating the evolution of evidence over time and may offer some insight into what uncertainty there is about how much a treatment works.

1. Mosteller, F. & Colditz, G. A. (1996) *Annu. Rev. Public Health* **17,** 1–23.
2. Chalmers, T. C. & Lau, J. (1996) *Stat. Med.* **15,** 1263–1268.
3. Naylor, C. D. (1997) *Br. Med. J.* **315,** 617–619.
4. Lau, J., Schmid, C. H. & Chalmers, T. C. (1995) *J. Clin. Epidemiol.* **48,** 45–57.
5. Cappelleri, J. C., Ioannidis, J. P. A., Schmid, C. H., de Ferranti, S. D., Aubert, M., Chalmers, T. C. & Lau, J. (1996) *J. Am. Med. Assoc.* **276,** 1332–1338.
6. LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J. & Derderian, F. (1997) *N. Engl. J. Med.* **337,** 536–542.
7. Borzak, S. & Ridker, P. M. (1995) *Ann. Intern. Med.* **123,** 873–877.
8. Ioannidis, J. P. A., Cappelleri, J. C. & Lau, J. (1998) *J. Am. Med. Assoc.* **279,** 1089–1093.
9. ISIS-4 Collaborative Group (1995) *Lancet* **345,** 669–685.
10. CLASP Collaborative Group (1994) *Lancet* **343,** 619–629.
11. Levine, R. J., Hauth, J. C., Curet, L. B., Sibai, B. M., Catalano, P. M., Morris, C. D., DerSimonian, R., Esterlitz, J. R., Raymond, E. G., Bild, D. E., *et al.* (1997) *N. Engl. J. Med.* **337,** 69–76.
12. Lau, J., Antman, E. M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F. & Chalmers, T. C. (1992) *N. Engl. J. Med.* **327,** 248–254.
13. Ioannidis, J. P. A., Contopoulos-Ioannidis, D. G. & Lau, J. (1999) *J. Clin. Epidemiol.* **52,** 281–291.
14. Schmid, C. H., Lau, J., McIntosh, M. & Cappelleri, J. C. (1998) *Stat. Med.* **17,** 1923–1942.
15. Enkin, M. W., Keirse, M. J., Renfrew, M. J. & Neilson, J. P., eds. (1994) *The Cochrane Collaboration Pregnancy and Childbirth Database* (Update Software, Oxford), Disk issue 2.
16. Woods, K. L. (1995) *Lancet* **346,** 611–614.
17. Antman, E. M. (1995) *Am. J. Cardiol.* **75,** 391–393.
18. Ioannidis, J. P. & Lau, J. (1997) *J. Clin. Epidemiol.* **50,** 1089–1098.
19. Bailar, J. C., III (1997) *N. Engl. J. Med.* **337,** 559–561.
20. Mantel, N. & Haenszel, W. (1959) *J. Natl. Cancer Inst.* **22,** 719–748.
21. Der Simonian, R. & Laird, N. M. (1986) *Control Clin. Trials* **7,** 177–188.
22. Lau, J., Ioannidis, J. P. A. & Schmid, C. H. (1997) *Ann. Intern. Med.* **127,** 820–826.
23. Hannah, M. E., Ohlsson, A., Farine, D., Hewson, S. A., Hodnett, E. D., Myhr, T. L., Wang, E. E., Weston, J. A. & Willan, A. R. (1996) *N. Engl. J. Med.* **334,** 1005–1010.
24. Ioannidis, J. P. A. (1998) *J. Am. Med. Assoc.* **279,** 281–286.
25. Woods, K. L., Fletcher, S., Roffe, C. & Haider, Y. (1992) *Lancet* **339,** 1553–1558.
26. Goodman, S. N. (1993) *Am. J. Epidemiol.* **137,** 485–496.
27. Chalmers, I., Dickersin, K. & Chalmers, T. C. (1992) *Br. Med. J.* **305,** 786–788.
28. Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D. Schulz, K. F., Simel, D., *et al.* (1996) *J. Am. Med. Assoc.* **276,** 637–639.
29. Horton, R. & Smith, R. (1999) *Lancet* **354,** 1138–1139.
30. Lau, J., Ioannidis, J. P. A. & Schmid, C. H. (1998) *Lancet* **351,** 123–127.