Comparison of Evidence of Treatment Effects in Randomized and Nonrandomized Studies

John P. A. Ioannidis, MD
Anna-Bettina Haidich, MSc
Maroudia Pappa, MSc
Nikos Pantazis, MSc
Styliani I. Kokori, MD
Maria G. Tektonidou, MD
Despina G. Contopoulos-Ioannidis, MD
Joseph Lau, MD

Context There is substantial debate about whether the results of nonrandomized studies are consistent with the results of randomized controlled trials on the same topic.

Objectives To compare results of randomized and nonrandomized studies that evaluated medical interventions and to examine characteristics that may explain discrepancies between randomized and nonrandomized studies.

Data Sources MEDLINE (1966–March 2000), the Cochrane Library (Issue 3, 2000), and major journals were searched.

Study Selection Forty-five diverse topics were identified for which both randomized trials (n=240) and nonrandomized studies (n=168) had been performed and had been considered in meta-analyses of binary outcomes.

Data Extraction Data on events per patient in each study arm and design and characteristics of each study considered in each meta-analysis were extracted and synthesized separately for randomized and nonrandomized studies.

Data Synthesis Very good correlation was observed between the summary odds ratios of randomized and nonrandomized studies (r=0.75; P<.001); however, nonrandomized studies tended to show larger treatment effects (28 vs 11; P=.009). Between-study heterogeneity was frequent among randomized trials alone (23%) and very frequent among nonrandomized studies alone (41%). The summary results of the 2 types of designs differed beyond chance in 7 cases (16%). Discrepancies beyond chance were less common when only prospective studies were considered (8%). Occasional differences in sample size and timing of publication were also noted between discrepant randomized and nonrandomized studies. In 28 cases (62%), the natural logarithm of the odds ratio differed by at least 50%, and in 15 cases (33%), the odds ratio varied at least 2-fold between nonrandomized studies and randomized trials.

Conclusions Despite good correlation between randomized trials and nonrandomized studies—in particular, prospective studies—discrepancies beyond chance do occur and differences in estimated magnitude of treatment effect are very common.

JAMA. 2001;286:821-830 www.jama.com

Randomized controlled trials have often been considered as the reference standard for evaluating the efficacy of therapeutic and preventive interventions.[1] However, for many medical questions of interest, a large amount of evidence is often accumulated through nonrandomized studies. There has been substantial controversy about whether the results of nonrandomized studies agree with the results of randomized trials. Earlier evaluations suggested that nonrandomized studies may spuriously overestimate treatment benefits yielding misleading conclusions.[2-5]

Recently, the debate has been renewed.[6-8] Much of the debate has been conducted on theoretical grounds about the biases that may affect each type of study design with an emphasis on the fact that nonrandomized studies may be more susceptible to unaccounted confounding. However, empirical evidence has also been accumulating. On the one hand, specific examples have arisen in the recent literature in which randomized studies have found different results compared with the epidemiologic literature that preceded them. Such examples included hormone replacement therapy and the risk of coronary artery disease; beta carotene and alpha tocopherol and their impact on coronary mortality; and the relationship between dietary fiber and colon cancer.[9-12] On the other hand, recent evaluations have suggested that for selected medical topics, both randomized and nonrandomized studies may yield very similar results.[7,8,13]

There is a need to address these issues using empirical data from a large number of diverse medical topics. Using such data, one would like to answer the following questions: How do the results of randomized trials and nonrandomized studies compare when both are performed for the same question? Do

Author Affiliations: Clinical Trials and Evidence-Based Medicine Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina (Drs Ioannidis and Contopoulos-Ioannidis, and Ms Haidich), Department of Hygiene and Epidemiology, University of Athens School of Medicine (Ms Pappa and Mr Pantazis) and Laikon General Hospital (Drs Kokori and Tektonidou), Athens, Greece; Department of Pediatrics, George Washington University School of Medicine, Washington, DC (Dr Contopoulos-Ioannidis); and Division of Clinical Care Research, Department of Medicine, Tufts University School of Medicine, Boston, Mass (Drs Ioannidis and Lau).

Corresponding Author: Joseph Lau, MD, Division of Clinical Care Research, New England Medical Center, Box 63, 750 Washington St, Boston, MA 02111 (e-mail: JLau1@lifespan.org).

nonrandomized studies tend to give more favorable results than randomized trials? Finally, are there design or other characteristics that may explain the discrepancies between randomized trials and nonrandomized studies? To address these issues, we performed a systematic evaluation using data from a large number of medical questions about which the efficacy of therapeutic or preventive interventions had been assessed with both randomized trials and nonrandomized studies.

METHODS

Search for Meta-analyses and Selection of Topics and Outcomes

We identified meta-analyses that had considered both randomized and nonrandomized evidence. The pertinent subjects and meta-analyses were identified using 5 different complementary approaches to maximize the yield of topics and to ensure that a wide variety of topics was retrieved. First, we reviewed the previous literature on comparisons of randomized and nonrandomized studies until mid-1998,[2-6] and we screened all the examples of such comparisons that the articles cited. Second, we perused our personal database of meta-analyses published between 1991 and 1997 in *JAMA*, *Lancet*, *BMJ*, *Annals of Internal Medicine*, and *Archives of Internal Medicine*. Third, we searched MEDLINE (last search updated on March 2000) for articles categorized as meta-analyses (type of publication) that contained a combination of at least 1 Medical Subject Heading suggestive of randomized clinical trials (such as *randomized controlled trials*, *randomized clinical trials*) and 1 Medical Subject Heading suggestive of a nonrandomized design (such as *prospective cohorts*, *retrospective cohorts*, *case-control studies*, etc). Fourth, we screened all the completed systematic reviews of the Cochrane Library (last screen on issue 3, 2000, containing 859 reviews). Fifth, we used meta-analyses that had been performed by investigators in our group with both randomized and nonrandomized comparisons included.

From all these sources, we selected the meta-analyses in which both randomized and nonrandomized studies were cited with at least 1 primary outcome being in binary form. Data on the binary outcome had to be presented in the meta-analysis. For the meta-analyses identified by perusing our personal database of meta-analyses and MEDLINE, we also considered meta-analyses in which some of the binary data might be unreported but might still be retrievable by reviewing the primary articles of each study cited by the meta-analysis. An effort was made to identify all the primary study articles whenever either the primary binary outcome information or important study characteristics were not reported in the meta-analysis. A few primary studies that could not be retrieved (primarily abstracts from conferences and very old studies) had to be excluded whenever the binary outcome data were not available in the published meta-analysis. For final inclusion of a topic in our evaluation, binary data for the same outcome had to be available on at least 1 randomized trial and at least 1 nonrandomized study.

Whenever a meta-analysis used different binary outcomes/end points and several of them had available data both for randomized and nonrandomized studies, we selected a priori the primary outcome, as stated by the meta-analysis. Whenever it was not clear which was the primary outcome, we selected a priori the outcome that was most important clinically, using consensus among the data extractors. Generally, mortality had a priority over other hard clinical outcomes, soft clinical outcomes, and laboratory outcomes, provided that there were at least some events for the most severe clinical outcome so that calculations of effects would be meaningful.

In some meta-analyses, comparisons of different interventions against each other or against no intervention or placebo had been considered. In this case, each eligible comparison qualified as a separate topic.

Data Extraction for Primary Studies

For each primary study in a meta-analysis we extracted the following information: type of design, year of publication, events per patients in each arm for the outcome of interest, age of the population (adult, children, or mixed), and duration of follow-up (in months, when available [or at least whether it was more than or up to 1 year]). We did not update systematically the eligible meta-analyses to include additional studies published after the meta-analysis. However, we tried to ensure that the comparison of the summary treatment effects between randomized and nonrandomized studies would not be totally offset from missing or recent information. Thus, we screened all the identified topics for missing information (eg, abstracts without binary data); adjusted odds ratio estimates in individual patient data meta-analysis that could not be accounted in our analyses; and major widely known recent trials that might offset the comparison of the magnitude of effects.

Nonrandomized Study Designs

Nonrandomized designs were categorized into prospective nonrandomized studies (all subjects were recruited and evaluated prospectively, but the control arm had not been created through randomization); retrospective cohort studies (subjects were evaluated retrospectively and the study arms were concurrent [without matching]); case-control studies (studies in which the compared groups were defined on the basis of the outcome and/or matching was used); historical control studies (studies with retrospective, nonconcurrent controls); and other or not-specified design. In each topic, we limited the analyses to the study designs that were included or systematically cited through the original meta-analysis.

Statistical Analysis

For each topic we combined the data from randomized and nonrandomized studies separately. We used the odds ratio (OR) as the metric of choice since case-control studies would also be included; moreover, the OR has statistical advantages.[14] We used both random-effects (DerSimonian and Laird[15]) and fixed-effects (Mantel-Haenszel)[16] calculations. Random effects models are reported, unless stated otherwise, because they incorporate an estimate of the between-study variance in the calculations and they tend to give wider (more conservative) confidence intervals than fixed effects.[17] Fixed-effects calculations are also given when substantially different. Heterogeneity between the studies of each type of design was assessed using the Q statistic and was considered significant for $P<.10$.[17]

To evaluate the concordance between the results of randomized and nonrandomized studies, we performed the following analyses: (1) We evaluated the Spearman correlation coefficient for the summary OR estimates between randomized and nonrandomized studies; (2) We assessed in how many cases the summary OR of the nonrandomized studies suggested a larger treatment effect for the experimental intervention than the summary OR of the randomized trials; (3) We evaluated whether the difference in the ORs of randomized and nonrandomized studies for the same topic was larger than what would be anticipated by chance alone. To do this, we estimated the z score, as follows:

$$z = \ln(\mathrm{OR}_{\mathrm{RCT}}) - \ln(\mathrm{OR}_{\mathrm{NRS}}) / \{\mathrm{var}[\ln(\mathrm{OR}_{\mathrm{RCT}})] + \mathrm{var}[\ln(\mathrm{OR}_{\mathrm{NRS}})]\}^{1/2}$$

where $\ln(\mathrm{OR}_{\mathrm{RCT}})$ is the natural logarithm of the OR of randomized trials, $\ln(\mathrm{OR}_{\mathrm{NRS}})$ is the natural logarithm of the OR of nonrandomized studies, and *var* stands for variance. A z score above 1.96 or less than −1.96 suggests that the difference between the randomized trials and nonrandomized studies is beyond chance (.05 level of statistical significance).[18] We also used alternative rules to define discrepancies based on differences in the relative magnitude of the treatment effect: (1) the OR of nonrandomized studies being at least double or less than half of the OR of randomized trials, and (2) the natural logarithm of the OR of nonrandomized studies being at least 50% larger or smaller than the natural logarithm of the OR of randomized trials. The magnitude of the treatment effect is important because it shows how much a treatment works.

Discrepancy rates were estimated for comparisons of randomized trials against all nonrandomized studies; all studies, excluding historical controls; prospective studies; retrospective studies with concurrent controls; and historical control studies only. We also performed analyses limited to studies published in 1986 or later. Furthermore, we evaluated whether the odds of a discrepancy beyond chance depended on the average year of publication in the studies included in each meta-analysis. Finally, study and topic characteristics were scrutinized to see whether there is an explanation for the statistically significant discrepancies. In this regard, we evaluated whether randomized trials and nonrandomized studies differed in years of publication, length of follow-up (less or more than 1 year), age of population (children, adults, elderly people), sample size, or other protocol characteristics.

Analyses were conducted in SPSS 10.0 (SPSS Inc, Chicago, Ill) and in Meta-Analyst (J. L., Boston, Mass). All P values are 2-tailed.

RESULTS

Characteristics of Medical Topics

A total of 45 topics were identified in which both randomized and nonrandomized studies had been performed on the same topic (TABLE 1).[2,3,19-52] Among 408 primary studies with available binary data, there were 240 randomized trials and 168 nonrandomized studies. The latter group included 71 prospective nonrandomized studies, 40 retrospective cohort studies, 25 case-control studies, 29 studies with historical controls, 1 cohort study with individual patient data assembled from several centers (unclear if prospective or retrospective), and 2 studies without clear design (presumably retrospective). The topics covered a wide range of medical specialties. In 29 topics there were more randomized trials than nonrandomized studies. In 26 topics there were more patients in randomized trials than in nonrandomized studies.

Estimates of Treatment Effects and Between-Study Heterogeneity

FIGURE 1 shows side by side the summary ORs for randomized trials and nonrandomized studies in each topic. In all, statistically significant heterogeneity was seen between randomized trials in 9 of 39 topics for which at least 2 randomized trials (23%) had been included. Statistically significant heterogeneity was seen between nonrandomized studies in 13 of 32 topics for which at least 2 nonrandomized studies (41%) had been included. The respective figure was 6 (40%) of 15 topics, when limited to prospective nonrandomized studies. The between-study variance was smaller among randomized trials than among nonrandomized studies in 18 topics while the opposite occurred in 6 cases, and it was the same in both designs in 4 cases (exact $P=.07$ by Wilcoxon test). The between-study variance was smaller among randomized trials than among prospective nonrandomized studies in 10 topics while the opposite occurred in 1 case, and it was the same in both designs in 3 cases (exact $P=.03$ by Wilcoxon test).

Correlation and Comparison of Treatment Effects

The correlation coefficient between the treatment effect in randomized trials and in nonrandomized studies was 0.75 ($P<.001$). This became 0.83 ($P<.001$) when historical control studies were excluded (FIGURE 2).

In 25 of the 45 cases, the nonrandomized studies showed a larger treatment effect for the experimental treatment than the randomized studies. The opposite occurred in 14 cases, but it was probably due to data artifacts in 3 of these: in 1 case, aspirin had shown a larger preventive effect for pregnancy-induced hypertension in randomized trials than in nonrandomized studies in an early meta-

Table 1. Topics of Meta-analyses Considering Both Randomized Trials and Nonrandomized Studies*

ID No.	Topic of Meta-analysis	Study, y	Outcome	No. of Randomized Trials (No. of Patients)	No. of Nonrandomized Studies (No. of Patients)
1	Anticoagulants in acute myocardial infarction	Chalmers et al,[2] 1977	Mortality	6 (3854)	12 (7497)
2	Antiarrhythmic treatment for chronic atrial fibrillation	Reimold et al,[19] 1992	Mortality	6 (808)	5 (604)
3	Diethylstilbestrol for habitual abortion	Sacks et al,[3] 1982	Infant mortality	3 (2187)	2 (508)
4	TENS vs sham in acute postoperative pain	Carroll et al,[20] 1996	Pain relief	7 (256)	2 (94)
5	TENS vs control in postoperative pain	Carroll et al,[20] 1996	Pain relief	2 (68)	4 (308)
6	Coronary artery surgery vs medical treatment for CAD	Sacks et al,[3] 1982	Mortality	7 (1556)	5 (989)
7	Esophageal varices: portacaval anastomosis	Sacks et al,[3] 1982	Mortality	8 (698)	2 (448)
8	Allogenic leukocyte immunotherapy for recurrent abortion	Recurrent Miscarriage Immunotherapy Trialists Group,[21] 1994	No live birth	8 (430)	1 (1133)
9	BCG immunotherapy for malignant melanoma	Sacks et al,[3] 1982	Mortality	1 (42)	3 (267)
10	5-FU adjuvant therapy for colon cancer	Sacks et al,[3] 1982	Mortality	7 (1385)	1 (4359)
11	Hormonal treatment (hCG) for cryptorchism	Pyorala et al,[22] 1995	Descended testis	1 (310)	1 (57)
12	Local vs general anesthesia for carotid endarterectomy	Tangkanakul et al,[23] 2000	Stroke or death, 30 d	3 (154)	14 (5186)
13	Low-level laser therapy for osteoarthritis	Brosseau et al,[24] 2000	No improvement	3 (139)	1 (8)
14	Oil- vs water-soluble media in hysterosalpingography	Vandekerckhove et al,[25] 2000	Pregnancy	5 (829)	6 (1806)
15	Oil-soluble hysterosalpingography vs no treatment	Vandekerckhove et al,[25] 2000	Pregnancy	1 (190)	1 (460)
16	Naltrexone for alcohol dependence	Srisurapanont and Jarusuraisin,[26] 2000	Discontinuation	6 (399)	1 (865)
17	Vaccines for serogroup A meningococcal meningitis	Patel and Lee,[27] 2000	Meningitis in 1 y	5 (291 147)	2 (56 806)
18	Microsurgical vs macrosurgical salpingostomy in subfertility	Watson et al,[28] 2000	Pregnancy	1 (18)	2 (178)
19	Fecal occult blood screening for colorectal cancer	Towler et al,[29] 1998	Cause-specific deaths	4 (329 642)	1 (21 756)
20	Screening mammography	Kerlikowske et al,[30] 1995	Cause-specific deaths	8 (417 742)	3 (1422)
21	Intrathecal therapy for tetanus	Abrutyn and Berlin,[31] 1991	Mortality	8 (640)	1 (134)
22	Vitamin A supplementation	Glasziou and Mackerras,[32] 1993	Mortality	12 (115 848)	3 (16 077)
23	Interferon for hepatitis C	Camma et al,[33] 1996	Normal transaminase	5 (213)	4 (133)
24	Trial of labor vs no trial of labor in breech delivery	Gifford et al,[34] 1995	Apgar score <7 at 5 min	2 (310)	6 (2595)
25	Anterior colporrhaphy vs needle suspension	Black and Downs,[35] 1996	Cure of incontinence	2 (266)	5 (831)
26	Needle suspension vs colposuspension	Black and Downs,[35] 1996	Cure of incontinence	3 (321)	9 (875)
27	Anterior colporrhaphy vs colposuspension	Black and Downs,[35] 1996	Cure of incontinence	4 (379)	11 (1478)
28	BCG vaccine for prevention of tuberculosis	Colditz et al,[36] 1994	Tuberculosis	12 (180 565)	10 (6511)
29	Nonoxynol-9 spermicides for sexually transmitted diseases	Cook and Rosenberg,[37] 1998	Gonorrhea	3 (1473)	1 (241)
30	Safety of early postpartum discharge	Grullon and Grimes,[38] 1997	Maternal problems	1 (131)	3 (839)
31	High-dose diuretics as first-line treatment for hypertension	Psaty et al,[39] 1997	Coronary heart disease	12 (30 783)	2 (2379)

(continued)

analysis,[40] but a major recent trial has shown no effect at all.[53] In another case, BCG immunotherapy for melanoma, the 1 published randomized trial had more favorable data than the nonrandomized studies, but several other randomized studies with less favorable results were not included (only available in abstract form without binary data).[3] A meta-analysis of allogeneic leukocyte immunotherapy showed more favorable results in randomized studies, but this was not true in the main original analysis, which was based on individual patient data with adjustment for significant predictors.[21] Finally, in 6 topics, it was not possible to identify clearly which study design produced more favorable results: in 2 cases (high-dose diuretics for hypertension and antiarrhythmic therapy for chronic atrial fibrillation) different conclusions were reached with fixed- and random-effects calculations; in another topic (hormonal therapy of cryptorchism), the control groups showed 0 efficacy in both types of studies; and in 3 topics the compared treatments (surgical interventions for urinary incontinence) were equally experimental, and thus there was no notion of a more favorable result.

Overall, these data suggested that larger treatment effects were somewhat more frequent to occur with nonrandomized studies than randomized trials (25 vs 14, exact $P=.11$; 28 vs 11 [correcting the 3 artifacts], exact $P=.009$ by Wilcoxon test). In 5 topics for which randomized trials suggested more favorable results than nonrandomized studies, there had been only 1 randomized trial performed (n=18, n=42, n=59, n=131, and n=190).

Discrepancies Between Randomized Trials and Nonrandomized Studies

In 7 (16%) of the 45 topics, the difference between the randomized trials and nonrandomized studies based on random effects calculations was beyond what would be expected by chance alone (TABLE 2). By fixed-effects calculations, this occurred in another 5 of the 45 topics (total 27%). The rates of discrepancies were substantially higher when their definition was based on the relative magnitude of the treatment effects in the compared designs. The natural logarithms of the ORs differed by at least 50% in 28 (62%) of the 45 topics and in 15 cases (33%), the OR varied at least 2-fold between nonrandomized studies and randomized trials (Table 2).

There were trends for higher rates of discrepancies in comparisons involving historical control studies (Table 2) and the rates of discrepancies beyond chance tended to decrease when only prospective studies were considered (8% by random effects, 15% by fixed effects) or when simply historical control studies were excluded (11% by random effects, 21% by fixed effects). How-

Table 1. Topics of Meta-analyses Considering Both Randomized Trials and Nonrandomized Studies (cont)*

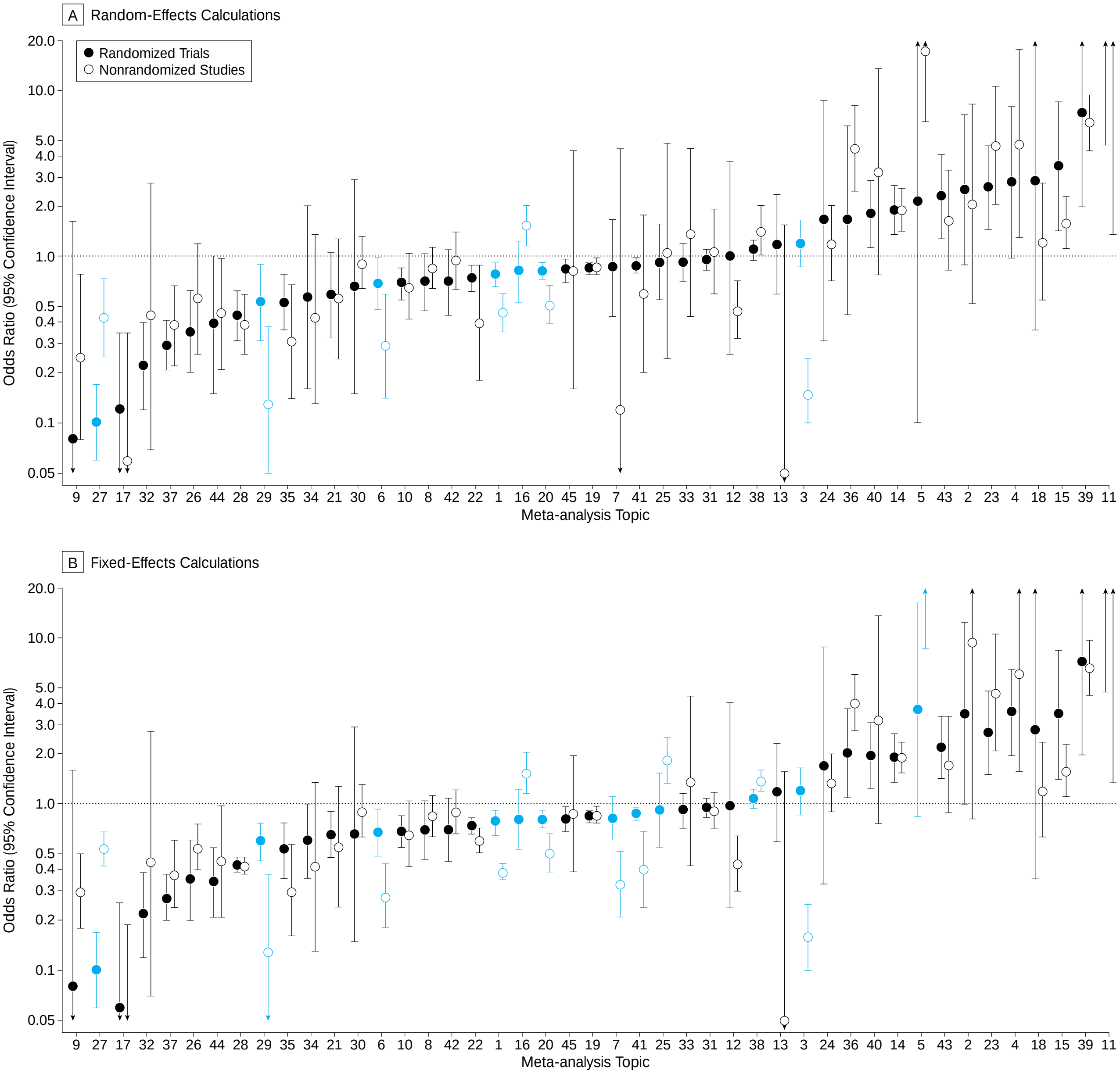
ID No.	Topic of Meta-analysis	Study, y	Outcome	No. of Randomized Trials (No. of Patients)	No. of Nonrandomized Studies (No. of Patients)
32	Aspirin for prevention of hypertension in pregnancy	Imperiale and Petrulis,[40] 1991	Hypertension	5 (337)	1 (48)
33	Allogenic blood transfusion in cancer	McAlister et al,[41] 1998	Death	5 (1923)	1 (273)
34	Sc vs IV heparin in deep venous thrombosis	Hommes et al,[42] 1992	Extension or recurrence	6 (770)	1 (48)
35	Low-protein diets in chronic renal insufficiency	Fouque et al,[43] 1992	Renal death	6 (890)	3 (248)
36	Corticosteroids for idiopathic facial nerve paralysis	Ramsey et al,[44] 2000	Improved function	3 (230)	6 (661)
37	Graduated compression stockings	Wells et al,[45] 1994	Deep vein thrombosis	12 (1844)	3 (532)
38	Aspirin for primary prevention of stroke	Hart et al,[46] 2000	Stroke	5 (52 251)	3 (108 706)
39	Fixed nail plates vs sliding hip for femoral fractures	Chinoy and Parker,[47] 1999	Total complications	1 (59)	6 (936)
40	Corticosteroids for stable COPD	Callahan et al,[48] 1991	Response to therapy	10 (598)	1 (62)
41	Treatment of hypertension in elderly people	Insua et al,[49] 1994	Mortality	7 (14 977)	3 (356)
42	Selective decontamination of the digestive tract	Vandenbroucke-Grauls and Vandenbroucke,[50] 1991	Mortality	6 (491)	6 (998)
43	Cyclosporine withdrawal vs no withdrawal	Kasiske et al,[51] 1993	Acute graft rejection	8 (650)	5 (316)
44	Cyclosporine withdrawal vs no cyclosporine	Kasiske et al,[51] 1993	Acute graft rejection	3 (551)	2 (139)
45	Prehospital thrombolysis	Ioannidis et al,[52] 2001	Mortality	7 (6549)	3 (1321)

*ID indicates identification; TENS, transcutaneous electrical nerve stimulation; CAD, coronary artery disease; 5-FU, 5-fluorouracil; hCG, human chorionic gonadotropin; Sc, subcutaneous; IV, intravenous; and COPD, chronic obstructive pulmonary disease.

ever, the magnitude of the treatment effect often differed substantially between randomized trials and nonrandomized studies regardless of which study designs were included in the latter group. Even when only prospective studies were considered, the natural logarithm of the ORs still differed by at least 50% in 16 (62%) of the 26 topics (Table 2).

When limited to studies published in 1986 or later, there were 5 discrepancies beyond chance by random effects among 23 topics that had at least 1 randomized trial and at least 1 nonrandomized study. The odds of having a discrepancy beyond chance tended to decrease when the average year of publication of the considered studies was more recent, but the change was not formally significant (OR, 0.93; *P*=.12).

Figure 1. Comparison of the Summary Odds Ratio and 95% Confidence Interval in Randomized Trials vs Nonrandomized Studies for the 45 Topics



The topic numbers correspond to the identification numbers in Table 1. Calculations have been performed with random effects in the panel A and, for comparison, by fixed effects in the panel B. The topics have been ordered according to increasing odds ratio estimates in randomized trials using random-effects calculations. Data shown in blue indicate the topics in which the difference between randomized trials and nonrandomized studies was beyond what would be expected by chance alone. For 1 topic (No. 11), both of the summary estimates lie outside the depicted range.

In 6 of the 7 disagreements beyond chance by random effects (TABLE 3), the estimated treatment benefit was larger in nonrandomized studies than in randomized trials while in 1 case both treatments were equally experimental. Overall, more favorable results were significantly more common with nonrandomized studies vs randomized trials (exact *P*=.03 by Wilcoxon test; *P*=.02 when fixed effects disagreements were included).

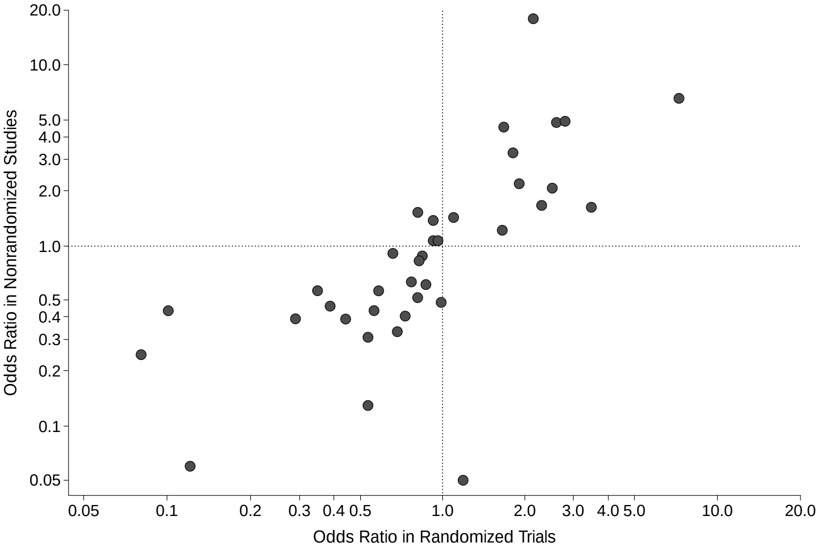
The age of the study populations was largely similar in nonrandomized studies and randomized trials on topics with discrepancies (data not shown). There was also no clear difference in the mean follow-up, perhaps with the exception of 1 disagreement (anterior colporrhaphy vs needle suspension) for which nonrandomized studies tended to have longer follow-up. In 2 cases (screening mammography, hypertension in elderly people), the randomized trials were of much larger sample size than the nonrandomized studies. In 4 cases, randomized trials had been published on average 5 or more years later than the nonrandomized studies (Table 3). Typically, the included randomized and nonrandomized studies on the same topic administered treatment in the same way and outcome measures were similarly defined. Selection criteria could have differed between studies, but differences could occur even within randomized trials or within nonrandomized studies and not necessarily only between randomized trials and nonrandomized studies.

COMMENT

Our empirical evaluation of 45 medical topics has found that randomized trials and nonrandomized studies show a high correlation in their estimates of efficacy of medical interventions. However, a high correlation does not necessarily also mean a similar magnitude of effect. Randomized trials and nonrandomized studies often disagree substantially on how much a treatment works. In fact, we observed that it was somewhat more frequent to find larger treatment effects in nonrandomized studies vs randomized trials than for the opposite to occur. However, it is precarious to claim that a study design arriving at a more favorable effect is necessarily spurious while a study design showing a smaller benefit is always more reliable. For example, sometimes a flawed study may fail to identify an existing benefit, because of the "noise" caused by its errors.

Discrepancies beyond what could be explained by chance were not uncommon between the 2 types of designs. When we allowed also for the between-study variability for each type of design by using random effects calculations, discrepancies beyond chance still occurred in 7 of 45 topics. Recently, using different methods for identification of topics, Concato et al[8] claimed no major disagreement for 5 randomized vs nonrandomized study comparisons while Benson and Hartz[7] found that in 2 of the 19 comparisons the point estimate of the nonrandomized

Figure 2. Comparison of the Summary Odds Ratio in Randomized Trials vs Nonrandomized Studies



Historically controlled studies are excluded from the calculations. Calculations are performed with random effects. Odds ratios are shown in a natural logarithmic scale. Not shown is 1 topic with very large summary odds ratios (>25) for both types of designs.

Table 2. Frequency of Discrepancies Among Randomized Trials and Nonrandomized Studies for Various Definitions and Types of Studies*

	No. (%) of Discrepancies Among Randomized Trials and Nonrandomized Studies or Eligible Topics				
Discrepancy Definition	All Studies (n = 45)	All Studies Except Historical Control (n = 38)	Prospective Studies (n = 26)	Retrospective Studies With Current Controls (n = 17)	Historical Control (n = 10)
Statistical (absolute *z* score >1.96)	7 (16)	4 (11)	2 (8)	2 (12)	2 (20)
Magnitude of effect					
≥2-Fold difference in odds ratio	15 (33)	11 (29)	7 (27)	5 (29)	5 (50)
≥50% Difference in odds ratio	28 (62)	23 (61)	16 (62)	9 (53)	7 (70)

*All data are based on random-effects calculations. See "Methods" section for definitions of discrepancies.

studies lay outside the 95% confidence intervals of the effect found by randomized trials. These 2 studies suggested a relatively higher concordance between the randomized and nonrandomized studies.[54] Several of the topics covered by these 2 surveys were also included in our evaluation, but 13 topics were not. If these 2 evaluations and our own are merged, statistically significant discrepancies between randomized and nonrandomized studies occur in 7 of 58 topics by random-effects calculations and in 13 topics by fixed-effects calculations.

Of interest, significant between-study variability was seen as frequently among the randomized trials as between the randomized and nonrandomized studies. Furthermore, significant variability was seen more than 40% of the time among the nonrandomized studies on the same topic. Thus variability seems to be very common both in randomized and nonrandomized studies and perhaps more frequent in the latter. Variability may be due to bias, but it may also reflect differences in the true treatment effect under different study settings and in different populations.[55]

Part of the variability could have been due to the fact that we considered a wide spectrum of nonrandomized designs. Several of the discrepancies beyond chance occurred in cases where nonrandomized studies were represented by historical control studies or other retrospective designs that may be more susceptible to bias than prospective designs. In fact, there were relatively few discrepancies beyond chance when randomized trials were compared with prospective nonrandomized studies. Still, perfect agreement was not seen even for these comparisons, and it was very common to see major differences in the estimates of the treatment effect.

Perfect agreement is perhaps impossible to expect between different types of study design or even within the same study design. Even the best designed studies may differ in several parameters and may form a continuum in the spectrum of medical evidence that we can obtain from them. We observed discrepancies, such as the cases of screening mammography or the treatment of hypertension in elderly people, for which randomized studies had a very different sample size than their nonrandomized counterparts. It is conceivable that larger studies in which large-scale effectiveness is probed may yield more conservative results than smaller studies in which efficacy is assessed, regardless of study design. The same may hold true when the timing of each study is considered. We encountered examples, in which nonrandomized studies had been published earlier than the randomized trials. Early studies in selected populations may yield promising results that may lead to subsequent trials with the aim of validating the benefits in larger populations. Publication bias and a time lag for negative studies may also be operating, regardless of study design.[56,57] Quality is also important to consider, and in this regard, sometimes nonrandomized or randomized studies,[58] or both may have important quality defects. For example, a recent meta-analysis suggests that even within randomized trials, the ones with greater methodological rigor show no benefit while the ones with potential flaws may be spuriously overestimating the benefit.[59]

It is not known whether meta-analyses that examine both random-

Table 3. Discrepancies Beyond Chance Between Randomized Trials and Nonrandomized Studies*

		Odds Ratio (95% Confidence Interval) Estimate by Random Effects		No. of Studies			
					Nonrandomized Studies		
ID No.	Topics of Meta-analysis	Randomized Trials	Nonrandomized Studies	RCT	Prospective	Retrospective	Mean, y RCT/NR
		Discrepancy by Fixed and Random Effects					
1	Anticoagulants in acute myocardial infarction	0.77 (0.65-0.92)	0.46 (0.35-0.59)†	6	3	9	1969/1958
3	Diethylstilbestrol for habitual abortion	1.20 (0.87-1.67)	0.15 (0.10-0.24)	3	0	2	1954/1949
6	Coronary artery surgery vs medical treatment for CAD	0.68 (0.48-0.97)	0.29 (0.14-0.59)†	7	0	5	1978/1976
16	Naltrexone for alcohol dependence	0.81 (0.53-1.23)	1.54 (1.16-2.04)	6	1	0	1996/1997
20	Screening mammography	0.81 (0.72-0.92)	0.51 (0.39-0.67)	8	0	3	1988/1987
27	Anterior colporrhaphy vs colposuspension	0.10 (0.06-0.17)	0.43 (0.25-0.73)†	4	11	0	1990/1987
29	Nononxynol-9 spermicides for sexually transmitted diseases	0.53 (0.31-0.90)†	0.13 (0.05-0.38)	3	0	1	1990/1982
		Discrepancy by Fixed Effects Only					
5	TENS vs control in acute postoperative pain	2.13 (0.10-47.9)†	17.5 (6.51-46.9)	2	0	4	1986/1984
7	Esophageal varices: portacaval anastomosis	0.85 (0.43-1.67)†	0.12 (0.00-4.48)†	8	0	2	1971/1975
25	Anterior colporrhaphy vs needle suspension	0.92 (0.55-1.55)	1.06 (0.24-4.77)†	2	5	0	1989/1989
38	Aspirin for primary prevention of stroke	1.09 (0.95-1.25)	1.43 (1.02-2.01)†	5	3	0	1993/1993
41	Treatment of hypertension in elderly people	0.87 (0.78-0.98)	0.60 (0.20-1.76)†	7	0	3	1987/1976

*ID indicates identification; RCT, randomized controlled trial; NR, nonrandomized studies; CAD, coronary artery disease; and TENS, transcutaneous electrical nerve stimulation.
†Statistically significant heterogeneity among studies ($P < .10$ by the Q statistic).

ized and nonrandomized evidence may do so because the results of the 2 types of designs are fairly concordant. If this is true, then meta-analyses with both types of data may be a biased sample and this could explain in part the relatively good overall correlation that we observed. Avoidance of this potential bias makes a strong case for examining information from all types of studies in meta-analysis. Nevertheless, even with this selection approach, the frequency of discrepancies was quite high when based on the comparison of the magnitude of the observed treatment effects. On the other hand, the selection of topics from published meta-analyses also leaves the possibility of publication bias affecting the results of specific meta-analyses. However, it is not known whether such bias would affect nonrandomized studies more than randomized trials and whether there would be an overall net bias affecting our comparisons.

Although we included a substantial number of comparisons, larger than in any previous evaluation in this field, this is still a small sample compared with the number of medical questions that are being probed with randomized and nonrandomized studies. It is conceivable that for many questions of interest, randomized trials may never be performed, if early nonrandomized studies show either clear harm or a large benefit. The ethical barrier may become insurmountable in such cases. Conversely, nonrandomized studies may be considered unworthy of consideration if randomized evidence is available on a topic. Although we perused several hundreds of meta-analyses, the vast majority regarded the randomized design as a prerequisite for eligibility and most of them did not even cite the nonrandomized studies. This is unfair for epidemiological research that may often offer some complementary insights to those provided by randomized trials. We propose that future systematic reviews and meta-analyses should pay more attention to the available nonrandomized data. It would be wrong to reduce the efforts to promote randomized trials so as to obtain easy answers from nonrandomized designs.[13] However, nonrandomized evidence may also be useful and may be helpful in the interpretation of the randomized results. Whenever discrepancies occur, such discrepancies should be carefully scrutinized since they may yield valuable information for designing future research.

Author Contributions: *Study concept and design:* Ioannidis, Lau.
Acquisition of data: Haidich, Pappa, Pantazis, Kokori, Tektonidou, Contopoulos-Ioannidis.
Analysis and interpretation of data: Ioannidis, Haidich.
Drafting of the manuscript: Ioannidis.
Critical revision of the manuscript for important intellectual content: Ioannidis, Haidich, Pappa, Pantazis, Kokori, Tektonidou, Contopoulos-Ioannidis, Lau.
Statistical expertise: Haidich, Pappa, Pantazis, Kokori, Tektonidou, Contopoulos-Ioannidis.
Obtained funding: Ioannidis, Lau.
Study supervision: Ioannidis, Lau.
Other Contributors: Giota Touloumi, PhD, and Kyriaki Boki, MD, contributed to the organization of the protocol and assisted in securing funding for the project; Panagiotis Anastassopoulos, MD, MSc, PhD, contributed to methodologic aspects of the protocol; Konstantinos Soukis, MSc, contributed expertise in computer searches; and Maria Katsorida, MD, contributed to the data extraction and analyses.
Funding/Support: The work was supported by grant PENED ED27 (974) from the Program for Support of Research Potential, Greek Secretariat for Research and Technology, funded through the European Union. Also supported in part through the New England Medical Center Research Fund.
Acknowledgment: We thank Dimitrios Trichopoulos, MD, for his support and encouragement and Christopher H. Schmid, PhD, Norma Terrin, PhD, Ingram Olkin, PhD, and Les Irwig, PhD, for interesting discussions. Georgia Kolokouri offered valuable secretarial assistance. We also thank Marian Perez and Priscilla Chew, MPH, for helping in the retrieval of primary articles.

REFERENCES

1. Pocock SJ. *Clinical Trials: A Practical Approach*. Chichester, England: John Wiley & Sons; 1983.
2. Chalmers TC, Matta RJ, Smith H Jr, Kunzler AM. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *N Engl J Med*. 1977;297:1091-1096.
3. Sacks H, Chalmers TC, Smith H Jr. Randomized versus historical controls for clinical trials. *Am J Med*. 1982;72:233-240.
4. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy, I: medical. *Stat Med*. 1989;8:441-454.
5. Miller JN, Colditz GA, Mosteller F. How study design affects outcomes in comparisons of therapy, II: surgical. *Stat Med*. 1989;8:455-466.
6. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ*. 1998;317:1185-1190.
7. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med*. 2000;342:1878-1886.
8. Concato J, Shah N, Horwitz RI. Randomized controlled trials, observational studies and the hierarchy of research designs. *N Engl J Med*. 2000;342:1887-1892.
9. Hulley S, Grady D, Bush T, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary artery disease in postmenopausal women. *JAMA*. 1998;280:605-613.
10. The Alpha Tocopherol, Beta Carotene Cancer Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *N Engl J Med*. 1994;330:1029-1035.
11. Yusuf S, Dagenais G, Pogue J, Bosch J, Sleight P for the Heart Outcomes Prevention Study Investigators. Vitamin E supplementation and cardiovascular events in high-risk patients. *N Engl J Med*. 2000;342:154-160.
12. Schatzkin A, Lanza E, Corle D, et al for the Polyp Prevention Trial Study Group. Lack of effect of a low-fat, high-fiber diet on the recurrence of colorectal adenomas. *N Engl J Med*. 2000;342:1149-1155.
13. Pocock SJ, Elbourne DR. Randomized trials or observational tribulations? *N Engl J Med*. 2000;342:1907-1909.
14. Rothman KJ, Greenland S. *Modern Epidemiology*. 2nd ed. Philadelphia, Pa: Lippincott-Raven; 1998.
15. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7:177-188.
16. Mantel N, Haenszel WH. Statistical aspects of the analysis of data from retrospective studies of diseases. *J Natl Cancer Inst*. 1959;22:719-748.
17. Lau J, Ioannidis JPA, Schmid CH. Quantitative synthesis for systematic reviews. *Ann Intern Med*. 1997;127:820-826.
18. Ioannidis JP, Cappelleri JC, Lau J. Issues in comparisons of meta-analyses and large trials. *JAMA*. 1998;279:1089-1093.
19. Reimold SC, Chalmers TC, Berlin JA, Antman EM. Assessment of the efficacy and safety of antiarrhythmic therapy for chronic atrial fibrillation: observations on the role of trial design and implications of drug-related mortality. *Am Heart J*. 1992;124:924-932.
20. Carroll D, Tramer M, McQuay H, Nye B, Moore A. Randomization is important in studies with pain outcomes: systematic review of transcutaneous electrical nerve stimulation in acute postoperative pain. *Br J Anaesth*. 1996;77:798-803.
21. Recurrent Miscarriage Immunotherapy Trialists Group. Worldwide collaborative observational study and meta-analysis of allogenic leukocyte immunotherapy for recurrent spontaneous abortion. *Am J Reprod Immunol*. 1994;32:55-72.
22. Pyorala S, Huttunen NP, Uhari M. A review and meta-analysis of hormonal treatment of cryptorchidism. *J Clin Endocrinol Metab*. 1995;80:2795-2799.
23. Tangkanakul C, Counsell C, Warlow C. Local vs general anaesthesia for carotid endarterectomy [Cochrane Review on CD-ROM]. Oxford, England: Cochrane Library, Update Software; 2000:Issue 3.
24. Brosseau L, Welch V, Wells G, et al. Low-level laser therapy (classes I, II and III) for the treatment of osteoarthritis [Cochrane Review on CD-ROM]. Oxford, England: Cochrane Library, Update Software; 2000:Issue 3.
25. Vandekerckhove P, Watson A, Lilford R, Harada T, Hughes E. Oil-soluble vs water-soluble media for assessing tubal patency with hysterosalpingography or laparoscopy in subfertile women [Cochrane Review on CD-ROM]. Oxford, England: Cochrane Library, Update Software; 2000:Issue 3.
26. Srisurapanont M, Jarusuraisin N. Opioid antagonists for alcohol dependence [Cochrane Review on CD-ROM]. Oxford, England: Cochrane Library, Update Software; 2000:Issue 3.
27. Patel MK, Lee CK. Polysaccharide vaccines for preventing serogroup A meningococcal meningitis [Cochrane Review on CD-ROM]. Oxford, England: Cochrane Library, Update Software; 2000:Issue 3.
28. Watson A, Vandekerckhove P, Lilford R. Techniques for pelvic surgery in subfertilility [Cochrane Review on CD-ROM]. Oxford, England: Cochrane Library, Update Software; 2000:Issue 3.

29. Towler B, Irwig L, Glasziou P, Kewenter J, Weller D, Silagy C. A systematic review of the effects of screening for colorectal cancer using the faecal occult blood test, hemoccult. *BMJ*. 1998;317:559-565.
30. Kerlikowske K, Grady D, Rubin SM, Sandrock C, Ernster VL. Efficacy of screening mammography: a meta-analysis. *JAMA*. 1995;273:149-154.
31. Abrutyn E, Berlin JA. Intrathecal therapy for tetanus: a meta-analysis. *JAMA*. 1991;266:2262-2267.
32. Glasziou PP, Mackerras DE. Vitamin A supplementation in infectious diseases: a meta-analysis. *BMJ*. 1993;306:366-370.
33. Camma C, Almasio P, Craxi A. Interferon as treatment for acute hepatitis C: a meta-analysis. *Dig Dis Sci*. 1996;41:1248-1255.
34. Gifford DS, Morton SC, Fiske M, Kahn K. A meta-analysis of infant outcomes after breech delivery. *Obstet Gynecol*. 1995;85:1047-1054.
35. Black NA, Downs SH. The effectiveness of surgery for stress incontinence in women: a systematic review. *Br J Urol*. 1996;78:497-510.
36. Colditz GA, Brewer TF, Berkey CS, et al. Efficacy of BCG vaccine in the prevention of tuberculosis: meta-analysis of the published literature. *JAMA*. 1994;271:698-702.
37. Cook RL, Rosenberg MJ. Do spermicides containing nonoxynol-9 prevent sexually transmitted infections? a meta-analysis. *Sex Transm Dis*. 1998;25:144-150.
38. Grullon KE, Grimes DA. The safety of early postpartum discharge: a review and critique. *Obstet Gynecol*. 1997;90:860-865.
39. Psaty BM, Smith NL, Siscovick DS, et al. Health outcomes associated with antihypertensive therapies used as first-line agents: a systematic review and meta-analysis. *JAMA*. 1997;277:739-745.
40. Imperiale TF, Petrulis AS. A meta-analysis of low-dose aspirin for the prevention of pregnancy-induced hypertensive disease. *JAMA*. 1991;266:260-264.
41. McAlister FA, Clark HD, Wells PS, Laupacis A. Perioperative allogeneic blood transfusion does not cause adverse sequelae in patients with cancer: a meta-analysis of unconfounded studies. *Br J Surg*. 1998;85:171-178.
42. Hommes DW, Bura A, Mazzolai L, Buller HR, ten Cate JW. Subcutaneous heparin compared with continuous intravenous heparin administration in the initial treatment of deep vein thrombosis. *Ann Intern Med*. 1992;116:279-284.
43. Fouque D, Laville M, Boissel JP, Chifflet R, Labeeuw M, Zech PY. Controlled low protein diets in chronic renal insufficiency: meta-analysis. *BMJ*. 1992;304:216-220.
44. Ramsey MJ, DerSimonian R, Holtel MR, Burgess LP. Corticosteroid treatment for idiopathic facial nerve paralysis: a meta-analysis. *Laryngoscope*. 2000;110:335-341.
45. Wells PS, Lensing AW, Hirsh J. Graduated compression stockings in the prevention of postoperative venous thromboembolism. *Arch Intern Med*. 1994;154:67-72.
46. Hart RG, Halperin JL, McBride R, Benavente O, Man-Son-Hing M, Kronmal RA. Aspirin for the primary prevention of stroke and other major vascular events. *Arch Neurol*. 2000;57:326-332.
47. Chinoy MA, Parker MJ. Fixed nail plates vs sliding hip systems for the treatment of trochanteric femoral fractures: a meta-analysis of 14 studies. *Injury*. 1999;30:157-163.
48. Callahan CM, Dittus RS, Katz BP. Oral corticosteroid therapy for patients with stable chronic obstructive pulmonary disease. *Ann Intern Med*. 1991;114:216-223.
49. 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.
50. Vandenbroucke-Grauls CM, Vandenbroucke JP. Effect of selective decontamination of the digestive tract on respiratory tract infections and mortality in the intensive care unit. *Lancet*. 1991;338:859-862.
51. Kasiske BL, Heim-Duthoy K, Ma JZ. Elective cyclosporine withdrawal after renal transplantation: a meta-analysis. *JAMA*. 1993;269:395-400.
52. Ioannidis JPA, Salem D, Lau J. Accuracy and clinical effect of out-of-hospital electrocardiography in the diagnosis of acute cardiac ischemia: a meta-analysis. *Ann Emerg Med*. 2001;37:461-470.
53. CLASP. A randomized trial of low-dose aspirin for the prevention and treatment of pre-eclampsia among 9,364 pregnant women. *Lancet*. 1994;343:619-629.
54. Ioannidis JPA, Haidich A-B, Lau J. Any casualties in the clash of randomized and observational evidence? *BMJ*. 2001;322:879-880.
55. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet*. 1998;351:123-127.
56. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet*. 1991;337:867-872.
57. Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA*. 1998;279:281-286.
58. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA*. 1995;273:408-412.
59. Gotzsche PC, Olsen O. Is screening for breast cancer with mammography justified? *Lancet*. 2000;355:129-134.

I am convinced that it is of primordial importance to learn more every year than the year before. After all, what is education but a process by which a person begins to learn how to learn?
—Peter Ustinov (1921-)