American Journal of Epidemiology
Vol. 166, No. 10
DOI: 10.1093/aje/kwm189
Advance Access publication August 21, 2007

Practice of Epidemiology

Should Meta-Analyses of Interventions Include Observational Studies in Addition to Randomized Controlled Trials? A Critical Examination of Underlying Principles

Ian Shrier[1], Jean-François Boivin[1,2], Russell J. Steele[1,3], Robert W. Platt[2,4], Andrea Furlan[5], Ritsuko Kakuma[2], James Brophy[2,6], and Michel Rossignol[1,2]

[1] Centre for Clinical Epidemiology and Community Studies, Lady Davis Institute for Medical Research, SMBD-Jewish General Hospital, McGill University, Montreal, Quebec, Canada.
[2] Department of Epidemiology and Biostatistics, McGill University, Montreal, Quebec, Canada.
[3] Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada.
[4] Department of Pediatrics, McGill University, Montreal, Quebec, Canada.
[5] Institute for Work and Health, Toronto, Ontario, Canada.
[6] Department of Medicine, McGill University Health Centre, McGill University, Montreal, Quebec, Canada.

Received for publication December 13, 2006; accepted for publication May 25, 2007.

Some authors argue that systematic reviews and meta-analyses of intervention studies should include only randomized controlled trials because the randomized controlled trial is a more valid study design for causal inference compared with the observational study design. However, a review of the principal elements underlying this claim (randomization removes the chance of confounding, and the double-blind process minimizes biases caused by the placebo effect) suggests that both classes of study designs have strengths and weaknesses, and including information from observational studies may improve the inference based on only randomized controlled trials. Furthermore, a review of empirical studies suggests that meta-analyses based on observational studies generally produce estimates of effect similar to those from meta-analyses based on randomized controlled trials. The authors found that the advantages of including both observational studies and randomized studies in a meta-analysis could outweigh the disadvantages in many situations and that observational studies should not be excluded a priori.

intervention studies; meta-analysis; observation; randomized controlled trials

Abbreviation: RCT, randomized controlled trial.

When randomized controlled trials (RCTs) are unethical (e.g., randomization to a probable harmful exposure, randomization to a drug that has proven benefits but uncertain side effects), observational studies are essential. However, the merit of the observational study compared with the merit of the RCT when a question can be addressed through either methodology has remained an ongoing debate over the last 10 years (1–6). An extension of this argument pertains to systematic reviews and meta-analyses. In 2001, Oxman suggested that ''coherent and transparent decision rules are needed for deciding when only to include RCTs, when to include non-randomized controlled trials and when to include other types of evidence. So far as possible, there should be an empirical basis for these decision rules, as well as logical arguments'' (7, p. 468). The issue is important because meta-analyses are frequently conducted on a limited number of RCTs. Shrier (8) reviewed a random 1 percent sample of meta-analyses published by the Cochrane Collaboration in 2003 and found that six of 16 reviews included two studies or fewer. Furthermore, 158 of 183 analyses conducted in seven additional studies were limited to two or fewer studies. In meta-analyses such as these, adding more

Correspondence to Dr. Ian Shrier, Centre for Clinical Epidemiology and Community Studies, SMBD-Jewish General Hospital, 3755 Cote Ste Catherine Road, Montreal, Quebec H3T 1E2, Canada (e-mail: ian.shrier@mcgill.ca).

information from observational studies may aid in clinical reasoning and establish a more solid foundation for causal inferences.

A strict rule for excluding all non-RCT evidence would lead to the conclusion that, prior to 2005, we should not have indicated to the public that smoking increases mortality (an RCT published in 2005 showed reduced mortality following a smoking cessation program (9), but, even so, there was no statistically significant effect in subjects over age 45 years or in those smoking fewer than 39 cigarettes per day). Although some proponents might be willing to allow non-RCT evidence where no RCTs exist, does one RCT provide sufficient evidence to ignore all the non-RCT evidence? What if the quality of the one (or two or three) RCT(s) is poor? In this situation, one has to question whether recommendations based on all the evidence, experimental and nonexperimental, might not have been more appropriate. As a concrete example of the potential usefulness of this approach, the prevailing opinion prior to 1999 was that stretching immediately before exercise was of benefit, a recommendation mostly based on four small RCTs. Shrier (10) published a systematic review that included these RCTs, but also the observational and basic science evidence, and concluded that stretching immediately before exercise would not reduce injury. A subsequent large RCT that directly addressed the question supported Shrier's hypothesis (11), as have other systematic reviews and meta-analyses (12, 13).

In this article, we address Oxman's (7) request for logical arguments. To date, the logical arguments against including observational studies reflect the view that observational studies are more prone to bias because 1) randomization removes the chance of confounding (because of the infinite population assumption), and 2) the double-blind process minimizes biases caused by the placebo effect.

We review evidence that questions whether these arguments are justified and propose alternative perspectives on this topic. We also critically examine the question of confounding in observational studies and raise questions as to whether one might be able to assess the likelihood of bias due to confounding. Next, we explore the empirical evidence comparing meta-analyses based on RCTs with meta-analyses based on observational studies on the same topic.

We conclude that the totality of the evidence suggests that, under some conditions, including observational studies would increase precision appropriately and may produce equally or more relevant and valid results for the question being asked. Doing so could reduce false inferences based solely on RCTs (such as occurred in the example of stretching noted above (10)) or the inability to make any inferences because of the paucity of RCTs on the subject. This article focuses on health care interventions. We believe that our results are applicable to the development of patient care guidelines or to create institutional policy, regardless of whether it is important to establish causality.

MAIN DIFFERENCES BETWEEN RCTS AND OBSERVATIONAL STUDIES

Many advantages routinely ascribed to the RCT design can be achieved through careful observational designs, such as risk of detection bias, ensuring that the people assessing the outcome are blinded to exposure status, and problems with data quality. In this section, we discuss the two main limitations ascribed to observational studies: confounding and potential for the placebo effect.

Confounding

In this paper, confounding is defined with respect to its effect on causal inference for an individual study, that is, whether the intervention is responsible for the observed differences in outcome. Therefore, confounding exists if the two groups "differ in their probability distribution for the outcome . . . for reasons other than effects of exposure" (14, p. 40), regardless of whether the difference in probability distribution occurred by chance (as may occur in an RCT) or for other reasons. That said, if a disease is well understood, a well-conducted observational study may yield an unconfounded estimate by adjusting for the correct mix of potential confounders at the design stage by matching or restriction, or at the analysis stage with a multiple regression analysis, propensity scores, and so forth (14, 15) (as long as the underlying statistical assumptions are valid; e.g., is transformation of variables required, directed acyclic graphs suggest that structural selection bias would not be introduced (16)). The problem of confounding occurs when potential confounders have not been incorporated in the model because one is unaware of them, they have not been measured, or they have not been measured well.

In the presence of unknown, unmeasured, or poorly measured confounders, treatment allocation by randomization improves one's ability to make causal inferences about the treatment because there is an expectation that potential confounders will be equally distributed in each group. Although randomization is intended to create equal distributions of the potential confounders (and therefore remove any association between exposure and the potential confounder), this result is obtained on average and confounding can still occur in individual studies (17). For example, let us assume that a randomized trial results in an unequal distribution of a variable (e.g., diabetes) that affects the outcome (e.g., mortality), such that 15 percent of the treatment group and 25 percent of the comparison group are diabetic. Concluding that the treatment had a causal effect on mortality without adjusting for the proportion of diabetics in each group would be incorrect in this particular study (i.e., the estimate from this study is confounded) even though the trial was randomized because the difference in mortality might be solely or partly due to the distribution of diabetic patients (17).

How likely is such a situation? The expectation of a perfectly equal distribution of covariates with randomization is based on the assumption of an infinitely large sample (18). In practice, all studies are conducted on finite samples. For example, if 20 percent of subjects in a moderate-sized trial of 400 people are expected to have diabetes (a potential confounder for the outcome mortality), there is a 95 percent probability that the proportion of diabetes in one group will be 15.6–24.4 percent (18) (representing the 95 percent probability for the sampling interval) and a 5 percent probability that it will lie outside this interval (refer to the Appendix for

calculations). Furthermore, diabetes may not be the only potential confounder (e.g., hypertension, smoking, hypercholesterolemia, physical inactivity), and we must therefore consider the probability of an unequal distribution of any one of the confounders. If there are five important confounders, each with a 20 percent prevalence and independent of each other, there is a 23 percent probability that the distribution of at least one covariate will lie outside the 15.6–24.4 percent interval. In randomized trials of 50 subjects per group, the corresponding sampling interval is 11.2–28.8 percent. These conditions affect the ability to make appropriate causal inferences, which could be referred to as ''confounding by chance.'' Even though confounding by chance might be rare in very large RCTs, it would be expected to occur more frequently within subgroup analyses because these comparisons are based on much smaller sample sizes.

Confounding by chance will occur with the same probability in observational studies because it is, by definition, due to chance. In both RCTs and observational studies, the *p* value automatically incorporates the uncertainty due to confounding by chance. Furthermore, in the context of a meta-analysis, confounding by chance in one direction in one study is expected to be matched by confounding by chance in the other direction in another study. The main problem with observational studies is the additional potential limitation of ''confounding by indication.'' Confounding by indication occurs when a treatment is specifically provided to a subject because of his or her probability of experiencing the outcome (19).

Confounding by indication

We have already mentioned that when a disease is well understood and prognostic factors are measured appropriately, the appropriate statistical model can yield an unconfounded estimate (14, 15). In this situation, the difference between the RCT and observational study concerns potential unknown or imperfectly measured confounders. Although confounding by indication is always possible in an observational study because the assumption of no unknown confounders is unverifiable, important nuances can help qualitatively determine the plausibility of this situation occurring; it is accepted practice in science to accept small probabilities that the data/interpretation will not be reproducible (e.g., type I error rates of 0.05 or confounding by chance).

First, we must distinguish between potential confounders and potential confounding because it is not necessary to adjust for every potential confounder (17, 20). Even when a variable is associated with the exposure and the outcome, an unconfounded estimate of the effect would still be obtained if the statistical model included a variable that lay along a causal pathway (i.e., an intermediary variable) between either the ''confounder'' and the exposure or the confounder and the outcome (it would not be appropriate to include a covariate that lay along the causal pathway between the main exposure of interest and the outcome) (14, 15, 17, 20). That said, the unknown confounder could still create a problem if it partially acted through a mechanism that does not include any of the measured variables.

Second, including all variables associated with both exposure and outcome (i.e., ''potential confounders'') in a statistical model will sometimes introduce bias rather than remove it. This situation occurs because including a covariate caused by two otherwise independent variables (i.e., a covariate that is a common effect) creates a conditional association between the otherwise independent variables (a full discussion of this topic is beyond the scope of this paper (14, 15, 21)). Identification of the appropriate subset of covariates required to obtain an unbiased estimate requires an underlying theory of the causal pathways (21, 22). Therefore, even if an existing unknown potential confounder became known later, it is not correct to assume that it should have necessarily been included in the statistical model.

Within the context described above (i.e., a well-understood disease for which known prognostic factors are appropriately measured and included in the statistical model by investigators), a systematic approach using basic epidemiologic principles allows for a qualitative assessment of the plausibility of confounding by indication in observational studies. We now discuss the following three situations: 1) the covariate is not considered important, 2) treatment is allocated by someone unaware of the patient's probability of experiencing the outcome, and 3) treatment is allocated by someone knowledgeable about the patient's probability of experiencing the outcome.

If the covariate is not thought to be important by the person allocating the treatment, it is highly unlikely to be causally related to exposure to treatment, and any unequal distribution must have occurred solely by chance. For example, if physicians are unaware that a drug affects the liver, there is an expectation that subjects with preexisting liver disease would be randomly distributed between the two groups; there would be no confounding by indication, and the risk of confounding due to this factor would be the same as in an RCT. Therefore, it is important to know who allocated the treatment and why.

If treatment allocation in an observational study is determined by a method/person uninformed about the probability of the outcome for the individuals (e.g., a hospital changes the type of drug used to treat a condition because of cost), then the situation is similar to when the factor is considered unimportant: the likelihood of equal distributions of disease severity in the treatment and comparison groups is the same as in an RCT. Therefore, under these conditions, the risk of confounding in an observational study should be minimal, and including observational studies in a meta-analysis would lead to increased precision of the estimate.

If the probability of the outcome influences treatment allocation, confounding by indication remains a problem, but it is still possible to adjust for the confounding if one appropriately measures the covariate. For example, if a physician prescribes one medication over another because a patient appears fatigued, including the physician's assessment of ''appearance of fatigue'' in the model would remove the confounding by indication. It is important that one measures the covariate accurately (e.g., the physician's assessment of patient fatigue and not patient fatigue itself). In addition, this method is obviously not possible when the relevant variable has not been recorded, which may occur in database studies.

What if physicians' prescribing practices are determined on non-evidence-based criteria that are difficult to measure (as opposed to easy-to-measure variables that could be included in the model)? In this case, one must remember that to be a confounder, the covariate must cause the outcome (or be associated with a cause of the outcome) (15). Therefore, for this to be used as an argument against including observational studies in a meta-analysis, one would have to argue that these non-evidence-based criteria are often direct or indirect causes of the outcome and act through mechanisms (i.e., covariates) not already included in the model. In addition, the criteria for prescription must remain an important confounder even after correcting for other known potential confounders. Thus, the probability of bias under the conditions mentioned above may be low enough to be acceptable and should be evaluated in each individual study.

A more difficult situation arises when subjects influence the treatment allocation. For example, patients may specifically choose one treatment over another because they are aware of some of their own personal factors that might affect prognosis. In this case, the probability of confounding by indication in observational studies becomes much greater, and interpretations need to be more cautious. It is in this context that randomized trial methods have a significant advantage over observational studies. That being said, if the prognostic factor is known, appropriately measured by the investigators, and included in the model, any bias would be minimized. Furthermore, observational studies are still likely to yield equal distributions of disease severity when one investigates ''unintended effects'' or side effects of a treatment, that is, effects that are not the reason that the subject is exposed to the treatment (23). In postmarketing observational studies, patients taking minoxidil for hypertension were observed to have increased hair growth. Because this effect was unintended, it is likely that there was an equal distribution of male-pattern baldness at baseline in the groups of hypertensive individuals who decided to take or not to take this blood pressure medication.

Bias due to the placebo effect

One of the fundamental assumptions underlying clinical trials is that the double-blind, placebo-controlled method (or any intervention used as the comparison group) yields an estimate of the true effect of the treatment in a specific population by comparing subjects with similar expectations (neither group knows whether they receive the active exposure). This method is generally possible with RCT methodology only, and it is an important strength when interpreted correctly. However, the results of some recent studies have suggested that the methodological advantages may be overstated (24, 25). Perhaps more importantly, some authors challenge its status as a ''gold standard'' based both on theory and on the results of studies that compare different methods of eliminating the placebo effect (26, 27). The following studies represent some empirical examples illustrating why the causal parameter estimated by a double-blind RCT may not be the parameter we are most interested in.

The double-blind RCT is not the only method that addresses the bias due to the placebo effect. In one study, researchers created similar expectations for pain relief in patients with mild cancer pain by telling them they would receive naproxen and then giving 50 percent of the subjects the active drug and 50 percent placebo (deception method); the results differed from those obtained when subjects were asked to participate in a traditional randomized double-blind trial (28). Similar differences between double-blind and deception methods were obtained for smoking cessation (29), caffeine (30), and insomnia (31), although the intervention effect was sometimes larger with the deception technique and sometimes larger with the double-blind technique.

These results suggest that each specific method used to address the bias due to the placebo effect results in a different parameter of interest being measured and that it is not possible to ''eliminate'' the placebo bias. Although the differences in results between the study designs may or may not be minimized with more objective outcome measures (this issue remains to be studied), the results from deception studies would remain relevant to the many standardized, but subjective outcomes currently receiving wide attention (e.g., pain scales, quality-of-life scales).

Is the estimate of the effect from the double-blind method or deception method more relevant to the clinician? When prescribing treatment to a patient, physicians most often tell the patient the treatment will work. Observational studies reproduce what happens in real life, where patients know what they are receiving and they comply with the intervention if they believe it helps them. If in reality there is no biologic effect, the context for the patient is that of a deception study (i.e., the patient is told a treatment is effective when it is not). However, there are obvious ethical problems with performing deception studies when exposures may cause harm. Still, the evidence suggests that although the double-blind technique may be the most acceptable method to remove the bias associated with the placebo effect for many interventions, it does not always lead us to the unbiased estimate of effect we are interested in. At this point, we do not have answers but simply raise the question as to whether meta-analyses based on only RCTs yield the effect estimate we are most interested in.

EMPIRICAL EVIDENCE FOR SIMILAR RESULTS

Although the findings of RCTs sometimes contradict the findings of highly publicized observational studies, they also sometimes contradict the findings of highly publicized randomized trials (32). Given the methodological heterogeneity of individual studies, Oxman's request for empirically based decision rules (7) is better addressed by examining the conclusions based on meta-analyses using only RCTs versus meta-analyses using only observational studies. We searched the literature for articles using the search strategy (meta-analysis or meta-analyses or ''systematic review'') AND (observational or non-randomised or non-randomized) AND (randomised or randomized) and examined all articles in which the purpose was specifically to compare the results of meta-analyses based on the two study designs over a broad range of conditions. We also hand-searched the bibliographies and then conducted a citation search on the most

comprehensive report published (MacLehose et al. (33)). The following paragraphs summarize the results and the pros and cons of each authors' approach.

Sacks et al. (34) examined the inclusion of studies with historical controls versus RCTs and found that historical control studies produce effect estimates of much larger magnitude. The underlying assumption when using historical controls is that there has not been a change in the management of the disease and that outcome incidence (e.g., mortality rate) is constant over time. These assumptions were not evaluated for each study, and this problem alone might explain much of the large discrepancies observed in the review.

Because epidemiology has made recent advances in study design and the ability to adjust for potential confounders, one must be careful to compare studies of similar quality. Several papers have attempted to do so, and, even though they all have limitations (6, 35), the findings are still informative. In brief, Concato et al. (36) found similar estimates of effect for meta-analyses based on RCTs versus high-quality cohort studies, MacLehose et al. (33) concluded that discrepancies for high-quality studies were small but that discrepancies for low-quality studies were large, Benson and Hartz (37) found similar results between meta-analyses based on RCTs and on cohort studies performed after 1984 (a proxy for better quality studies), and Ionnadis et al. (5) found discrepancies in only 8 percent of topics covered by prospective studies. At the Fourth Canadian Cochrane Symposium in Montreal, Canada, in 2005, Furlan et al. (38–40) showed that the discrepant results between cohort and RCT studies regarding low back pain were almost all attributable to the quality of the studies and to the homogeneity (i.e., similarities in settings, population, interventions, control group, and outcomes) of the pair cohort RCT.

If observational studies include all the biases of RCTs plus additional problems, how can meta-analyses based on observational studies so often yield the same results as meta-analyses based on RCTs? Well-conducted observational studies will yield similar estimates of effect compared with RCTs when the bias created by the potential limitations exclusive to observational studies is small in magnitude compared with the variability and/or bias created by choice of study population, types of subjects willing to enter a study, quality of data acquired, and other random effects (the analogy in electrical terms is a small signal-to-noise ratio). Therefore, it is important to study the discrepancies that occur between studies (whether due to study design or otherwise) because they provide information that can be used for appropriate clinical reasoning and causal inferences. For example, Galbraith plots help identify outliers that might provide clues to differences that matter (41, 42), and meta-regression can help identify the magnitude of effect for hypothesized effect modifiers. The important point here is that one has a choice to either 1) qualitatively/quantitatively assess the probability of bias due to lack of randomization (i.e., confounding by indication) to determine whether it is appropriate to include the observational data, or 2) decide a priori without assessment that the observational study is not worthwhile. In choosing to include observational studies, we must remain cautious not to over-interpret the data because there may be a greater risk of publication bias (43), and any between-study heterogeneity needs to be appropriately explored.

Although the results of meta-analyses based on observational studies are often similar to those based on RCTs, some authors argue against the inclusion of observational studies because the researcher needs to know the estimate of the effect for a particular topic and not whether the observational studies agree with RCTs "on average" (35, 44). We agree with this general idea but hasten to add that all of science, including evidence-based medicine, is based on probabilities and rational decision making. In economic appraisals and Bayesian decision theoretic approaches, decision making is a function of both probabilities and utility (or loss function) (45, 46). Utility is based on two potential situations: 1) How much harm is the decision maker willing to accept if she or he decides to act when the action is actually harmful? and 2) How much benefit is the decision maker willing to lose if she or he does not act and the action is actually helpful? For example, if there is a 99 percent probability that including the observational data would improve our ability to choose the correct treatment for a condition and a 1 percent chance it would hinder our ability, then observational studies should be included. Because of the lack of research in this area, we are currently limited to personal preferences to decide 1) what range of probabilities we should accept as the cutoff figure for benefit and for harm (estimating the probability is necessarily subjective at this time), 2) what other conditions (e.g., how many RCTs exist on a topic) we should insist on (the current decision to exclude observational studies if there is one RCT or a given number of randomized patients is a subjective decision without supporting evidence itself), and 3) the exact decision rules required to determine this (as requested by Oxman (7)).

As a consequence of the arguments described, we believe that study design should be explored as an explanatory variable when there are discrepancies between studies, and doing so requires that different study designs be included in the systematic review. If the study design is not the reason for the discrepancy, then including observational studies increases the sample size and provides additional evidence for interstudy differences. If the study design is associated with a difference in the estimate of the effect, then the potential reasons for the discrepancies should include a discussion about the relative probabilities, directions, and magnitudes of the biases, and the pros and cons of each study design (e.g., confounding by indication, blinding, restricted sample) because it is the totality of all limitations that determines how we interpret the relative strength of the different studies.

Finally, it is important to note that including observational studies in a systematic review without a meta-analysis presents fewer problems than including them in one with a meta-analysis. How does one weight studies of different design in a meta-analysis? Theoretically, studies should be weighted according to the probability of bias. We have argued that observational studies are not always more prone to bias and therefore applying a weighting scheme based exclusively on study design may lead to misclassification. The

most promising statistical technique to date may be response-surface estimation because it can accommodate factors that lead to bias from any cause (including study design) (47, 48). In brief, this analysis appropriately weights the results of any study by the potential for bias from an ''ideal'' study that has no bias, which is not possible in a meta-regression (47). However, the method requires that quality scores be highly correlated with bias; therefore, there must be agreement on which items create which biases, in which direction and of what magnitude. Because the objective of this paper is to argue that well-conducted observational studies are not automatically more biased than well-conducted RCTs, there is clearly not even agreement at this basic level, and more work is necessary before we can take appropriate advantage of this statistical technique.

In conclusion, the theoretical and empirical evidence presented in this paper suggests that excluding observational studies in systematic reviews a priori is inappropriate and internally inconsistent with an evidence-based approach.

ACKNOWLEDGMENTS

Conflict of interest: none declared.

REFERENCES

1. Day NE. Discussion on ''statistical issues arising in the Women's Health Initiative''. Biometrics 2005;61:912–14.
2. Freedman DA, Petitti DB. Discussion on ''statistical issues arising in the Women's Health Initiative''. Biometrics 2005; 61:918–20.
3. Egger M, Smith GD, O'Rourke K. Rationale, potentials, and promise of systematic reviews. In: Egger M, Smith GD, Altman DG, eds. Systematic reviews in health care. Meta-analysis in context. London, United Kingdom: BMJ Publishing Group, 2001:3–19.
4. Guyatt GH, DiCenso A, Farewell V, et al. Randomized trials versus observational studies in adolescent pregnancy prevention. J Clin Epidemiol 2000;53:167–74.
5. Ioannidis JP, Haidich AB, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 2001;286:821–30.
6. Kunz R, Khan KS, Neumayer HH. Observational studies and randomized trials. N Engl J Med 2000;343:1194–5.
7. Oxman AD. The Cochrane Collaboration in the 21st century: ten challenges and one reason why they must be met. In: Egger M, Smith GD, Altman DG, eds. Systematic reviews in health care. Meta-analysis in context. London, United Kingdom: BMJ Publishing Group, 2001:459–73.
8. Shrier I. Cochrane Reviews: new blocks on the kids. Br J Sports Med 2003;37:473–4.
9. Anthonisen NR, Skeans MA, Wise RA, et al. The effects of a smoking cessation intervention on 14.5-year mortality. Ann Intern Med 2005;142:233–9.
10. Shrier I. Stretching before exercise does not reduce the risk of local muscle injury: a critical review of the clinical and basic science literature. Clin J Sport Med 1999;9:221–7.
11. Pope RP, Herbert RD, Kirwan JD, et al. A randomized trial of preexercise stretching for prevention of lower-limb injury. Med Sci Sports Exerc 2000;32:271–7.
12. Thacker SB, Gilchrist J, Stroup DF, et al. The impact of stretching on sports injury risk: a systematic review of the literature. Med Sci Sports Exerc 2004;36:371–8.
13. Herbert RD, Gabriel M. Effects of stretching before and after exercising on muscle soreness and risk of injury: systematic review. BMJ 2002;325:468.
14. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999;10:37–48.
15. Hernan MA. A definition of causal effect for epidemiological research. J Epidemiol Community Health 2004;58:265–71.
16. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004;15:615–25.
17. Greenland S, Morgenstern H. Confounding in health research. Annu Rev Public Health 2001;22:189–212.
18. Agresti A. Categorical data analysis. New York, NY: John Wiley & Sons, 2002.
19. Begaud B. Dictionary of pharmacoepidemiology. New York, NY: John Wiley & Sons, 2000.
20. Pearl J. The art and science of cause and effect. In: Causality: models, reasoning and inference. Cambridge, United Kingdom: University of Cambridge, 2000:331–58.
21. Pearl J. Causality: models, reasoning and inference. Cambridge, United Kingdom: University of Cambridge, 2000.
22. Hernan MA, Hernandez-Diaz S, Werler MM, et al. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol 2002;155:176–84.
23. Miettinen O. The need for randomization in the study of intended effects. Stat Med 1983;2:267–71.
24. Hrobjartsson A, Gotzsche PC. Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. N Engl J Med 2001;344:1594–602.
25. Balk EM, Bonis PA, Moskowitz H, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002;287:2973–82.
26. Kaptchuk TJ. Powerful placebo: the dark side of the randomised controlled trial. Lancet 1998;351:1722–5.
27. Kaptchuk TJ. The double-blind, randomized, placebo-controlled trial: gold standard or golden calf? J Clin Epidemiol 2001;54:541–9.
28. Bergmann JF, Chassany O, Gandiol J, et al. A randomised clinical trial of the effect of informed consent on the analgesic activity of placebo and naproxen in cancer pain. Clin Trials Metaanal 1994;29:41–7.
29. Hughes JR, Gulliver SB, Amori G, et al. Effect of instructions and nicotine on smoking cessation, withdrawal symptoms and self-administration of nicotine gum. Psychopharmacology (Berl) 1989;99:486–91.
30. Kirsch I, Rosadino MJ. Do double-blind studies with informed consent yield externally valid results? An empirical test. Psychopharmacology (Berl) 1993;110:437–42.
31. Dahan R, Caulin C, Figea L, et al. Does informed consent influence therapeutic outcome? A clinical trial of the hypnotic activity of placebo in patients admitted to hospital. Br Med J (Clin Res Ed) 1986;293:363–4.
32. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. JAMA 2005;294:218–28.
33. MacLehose RR, Reeves BC, Harvey IM, et al. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. Health Technol Assess 2000;4:1–154.
34. Sacks H, Chalmers TC, Smith H Jr. Randomized versus historical controls for clinical trials. Am J Med 1982;72:233–40.
35. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003; 7:iii–173.

36. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000;342:1887–92.
37. Benson K, Hartz AJ. Observational studies and randomized trials. N Engl J Med 2000;342:1878–86.
38. Furlan AD, Tomlinson G, Jadad A, et al. Why randomized trials and non-randomized studies of the same interventions agree or disagree. Presented at the 4th Canadian Cochrane Symposium, Montreal, Quebec, Canada, March 12, 2005.
39. Furlan AD, Tomlinson G, Jadad A, et al. Meta-regression of randomized and non-randomized studies of treatments for low-back pain. Presented at the 4th Canadian Cochrane Symposium, Montreal, Quebec, Canada, March 12, 2005.
40. Furlan AD, Tomlinson G, Jadad A, et al. Agreement between randomized trials and non-randomized studies was influenced by methodological quality and homogeneity. J Clin Ethics (in press).
41. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. Stat Med 1988;7: 889–94.
42. Thompson SG. Why and how sources of heterogeneity should be investigated. In: Egger M, Smith GD, Altman DG, eds. Systematic reviews in health care. Meta-analysis in context. London, United Kingdom: BMJ Publishing Group, 2001: 157–75.
43. Easterbrook PJ, Berlin JA, Gopalan R, et al. Publication bias in clinical research. Lancet 1991;337:867–72.
44. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 1998;317:1185–90.
45. Sackett DL, Haynes RB, Guyatt GH, et al. Deciding on the best therapy. In: Sackett DL, Haynes RB, Guyatt GH, eds. Clinical epidemiology: a basic science for clinical medicine. 2nd ed. Toronto, Canada: Little, Brown and Company, 1991: 187–248.
46. Torrance GW. Measurement of health state utilities for economic appraisal. J Health Econ 1986;5:1–30.
47. Greenland S, O'Rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics 2001;2:463–71.
48. Vanhonacker WR. Meta-analysis and response surface extrapolation: a least squares approach. Am Stat 1996;50:294–9.

APPENDIX

Calculation of 95 Percent Sampling Intervals

The expectation of an equal distribution of covariates with randomization is based on the assumption of an infinitely large sample (18). If randomization is conducted so that 200 subjects are in each group and the prevalence of the covariate is 20 percent, the 95 percent probability for the sampling interval lies between 1.96 $\times$ the standard error of 20 percent, that is, between 15.6 percent and 24.4 percent (18). Note that if one group includes 24 percent diabetics, the other group must have 16 percent diabetics (the difference between groups is 8 percent) if there are an equal number of total subjects in each group.

If there are five covariates and each has a 20 percent prevalence, the probability that the previously stated 95 percent sampling intervals (i.e., 15.6 percent, 24.4 percent) will be true for all five confounders is only 77 percent (i.e., 0.95^5). To know the 95 percent sampling region for five covariates, one can use the 99 percent sampling interval for a single confounder because $0.99^5 = 0.95$. The 95 percent probability for the sampling interval is 14.2–25.8 percent.