2004

Ensuring the comparability of comparison groups: is randomization enough?

Vance Berger
Sherri Rose

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Controlled Clinical Trials 25 (2004) 515–524

Controlled Clinical Trials

www.elsevier.com/locate/conclintrial

Discussion

Ensuring the comparability of comparison groups: is randomization enough?

Vance W. Berger[a,b,*], Sherri Weinstein[a,c]

[a]*National Cancer Institute, EPN, Suite 3131, 6130 Executive Boulevard, MSC-7354, Bethesda, MD 20892-7354, USA*
[b]*Department of Mathematics and Statistics, University of Maryland at Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA*
[c]*The George Washington University, 2121 Eye St, Washington, DC 20052, USA*

Received 22 January 2004; accepted 8 April 2004

Abstract

Background: It is widely believed that baseline imbalances in randomized trials must necessarily be random. In fact, there is a type of selection bias that can cause substantial, systematic and reproducible baseline imbalances of prognostic covariates even in properly randomized trials. It is possible, given complete data, to quantify both the susceptibility of a given trial to this type of selection bias and the extent to which selection bias appears to have caused either observable or unobservable baseline imbalances. Yet, in articles reporting on randomized trials, it is uncommon to find either these assessments or the information that would enable a reader to conduct them. Nevertheless, there have been a few published reports that contain descriptions of either this type of selection bias or indicators that it may have occurred.

Objective: To document that the same type of selection bias has been described in numerous randomized trials and therefore that it represents a problem deserving of greater attention.

Study selection: Computerized searches were not useful in locating trials with one or more elements that contribute to or are indicative of selection bias in randomized trials. We limit our treatment to trials that were previously questioned for susceptibility to selection bias or for large baseline imbalances.

Results: We found 14 randomized trials that appear to be suspicious for selection bias. This may represent only the tip of the iceberg, because the status of other trials is inconclusive.

Conclusions: Authors of clinical trial reports should be required to disclose sufficient details to allow for an assessment of both allocation concealment and selection bias. The extent to which a randomized study was

* Corresponding author. National Cancer Institute, EPN, Suite 3131, 6130 Executive Boulevard, MSC-7354, Bethesda, MD 20892-7354, USA. Tel.: +1 301 435 5303; fax: +1 301 402 0816.
E-mail address: vb78c@nih.gov (V.W. Berger).

0197-2456/$ - see front matter
doi:10.1016/j.cct.2004.04.001

susceptible to selection bias should be considered in determining the relative contribution it makes to any subsequent meta-analysis, policy or decision.

Keywords: Allocation concealment; Broken experiment; Confounding; Selection bias

1. Introduction

Because nonrandomized studies are most susceptible to confounding [1], randomized clinical trials (RCTs) are the gold standard for comparative medical studies, especially when they are free of missing data and noncompliance [2]. Yet, some of the benefits ascribed to randomization, for example that it eliminates all selection bias (e.g., Ref. [3]), can better be described as fantasy than reality. Indeed, even with allocation concealment, proper randomization, no missing data and full compliance (so the experiment is not "broken" according to the definition implicit in [2]), selection bias can still occur [1], and cause substantial, systematic and reproducible baseline imbalances [4], inflate the type I rate [5] and offer a plausible alternative explanation for observed post-treatment between-group differences [6]. As such, definitive attribution of observed treatment effects to the treatments themselves is impossible without ruling out the confounding that can occur as a result of this type of selection bias.

Despite this, articles reporting on RCTs rarely offer an adequate description even of allocation concealment, let alone of susceptibility to or evidence of selection bias [7–9]. This lack of attention is both a cause and an effect of what appears to be the subconscious relegation of selection bias (at least in RCTs) to the ranks of the "fourth decimal point" problems. Selection bias contributes to its own concealment by inflating the magnitude of apparent treatment effects, so that artificially low p-values may mistakenly be attributed to the treatments themselves, thereby producing the illusion that robustness was proved beyond what could be explained by selection bias. It is for this reason that both convincing results and (not or) solid methodology are required [10]. Because the information required to detect selection bias and to quantify its effect is rarely available from publications, there is no way to estimate how prevalent or extreme selection bias may be in practice. In this article, we synthesize, but do not attempt to authenticate or refute, published claims that specific RCTs may have been tainted by selection bias. The overwhelming majority of trials avoid the issue of selection bias altogether; in fact, we are aware of only three [11–13] that have tested for it and not found it. The rest are inconclusive, so these examples may represent only the tip of the iceberg. We hope to encourage medical journals to require RCT reports to disclose sufficient details to allow for an assessment of selection bias, which could be used in determining one aspect of trial quality and rigor.

2. Mechanism of selection bias in RCTs

There are many ways that the methods by which units are selected can create a bias, or limit the extent to which the sample is representative of the population to which it is to apply. Hence, selection bias has been defined in various ways in medical studies. We study a particular type of selection bias related to

the methods of allocation of the intervention. Allocation sequences tend to be prepared in advance, before patients are screened for enrollment, so patient characteristics and preferences cannot influence treatment decisions [14]. This leads to a perception that within the RCT framework selection bias can interfere with external validity, but not with internal validity. But consider that the investigator who screens patients for enrollment observes both measured and unmeasured patient characteristics, and is therefore in a position to estimate the potential outcomes [15] each patient might experience under each treatment [1]. Note that the subjective health perceived by a patient can predict clinical outcomes and even mortality, even after adjusting for other observed predictors [16]. If the upcoming treatment assignments are either known with certainty or predictable (even imperfectly), then this advance knowledge allows for strategic patient selection. That is, better responders may be enrolled when one treatment group is due (or expected) to be assigned, and worse responders may be enrolled when the other treatment group is due to be assigned [17]. Randomization refers only to the use of randomization in the generation of the allocation sequence (although some trials labeled "randomized" do not even do that [14]); it says nothing about prediction of future allocations [1].

Allocation concealment is defined as the inability to observe any allocation prior to its execution [4,17–22]. This is a valid objective, but a good faith attempt to conceal the allocation sequence may not ensure the ignorance for which one would hope. The patterns created by restricted randomization [23] allow future assignments to be predicted from past ones. In unmasked studies, the potential for prediction is then clear. Even in masked trials, masking may be defined as either the *process* (researchers not revealing the treatment codes until the database is locked) or the *effect* (attainment of complete ignorance of all parties involved in the trial as to which patients received which treatments) [1]. Like encryption or setting a speed limit, the *process* is always possible; the *effect* may never be [14,24–26]. The *process* of masking is the only legitimate claim that can be made; it may help to ensure the ignorance of some parties, but it does not prevent tell-tale adverse events from revealing treatment assignments, and so it is unlikely to ensure the desired state of complete ignorance [1].

The unmasking of treatment allocations already made may lead to observer bias, which can be a serious problem. However, as the patients have already been selected at the time of the unmasking, selection bias would appear not to be an issue. Recall, however, that randomization is generally restricted, meaning that future allocations can be predicted from past ones. This opens the door to the type of selection bias with which we are concerned. We see, then, that as is the case with masking, so is the *process* of allocation concealment distinct from the *effect* of allocation concealment, and the process is not sufficient for an unbiased comparison [1,4,27]. As such, specific design features should accompany a claim of allocation concealment.

Selection bias can be sufficient to drastically inflate the chances that the active treatment is found superior to the control treatment, even in the absence of a real benefit [5]. In fact, the "bias caused by sub-optimum randomization methods can be larger than the treatment effects that might be detected" [28]. Selection bias can affect both the estimate of the magnitude of the treatment effect and the between-group p-value of each endpoint. Concordance of such estimates of treatment effects and p-values across endpoints might be considered to offer robust evidence, but again, it is not inconsistent with a complete lack of any real treatment effect if there is selection bias. While this may seem a bit far-fetched to be considered anything other than a hypothetical concern, "practitioners involved in conducting a trial that does not have proper procedures for sequence generation and allocation concealment may find the challenge of deciphering the allocation scheme irresistible" [19]. Indirect evidence of selection bias is provided by the finding that RCTs without

adequate allocation concealment produce larger estimates of treatment effects than those with adequate allocation concealment [29–33]. Some direct examples of trials with evidence of selection bias have been described but not identified [4,34]. The claims that follow provide additional evidence of selection bias.

3. Claims of selection bias in randomized studies

We now discuss RCTs for which a claim has been made of the specific type of selection bias we consider, or at least baseline imbalances have been questioned. First, the Coronary Artery Surgery Study compared coronary bypass surgery to medical therapy. Assignments were made via telephone communication from the Coordinating Center. Of the 2099 randomizable patients, 780 agreed to participate. The randomized and randomizable groups differed in the degree of baseline coronary artery disease [35]. Within the medical therapy group, there was *more* extensive baseline coronary artery disease among randomized than randomizable patients, whereas within the surgery group, there was *less* extensive coronary artery disease among randomized than randomizable patients [17].

In the Western Washington Intracoronary Streptokinase Trial, comparing intracoronary streptokinase to standard medical attention for the treatment of acute myocardial infarction, 250 patients were randomized, 134 to streptokinase and 116 to control, but one patient was mistakenly assigned to streptokinase [36]. Given the planned 1:1 allocation ratio, the probability of observing a difference of at least 18 in the group sizes is 0.004 [36]. The probability of observing even the corrected difference of 16 or larger is still very low (0.009). The statisticians involved in the study were "particularly concerned in verifying that the randomization process had been carried out as planned", as these values suggest the possibility of basing treatment assignments on patient characteristics, a form of selection bias [36].

Groothuis et al. [37] did not fully describe the method by which treatments were assigned in the study of RSV Immune Globulin in Infants and Young Children with Respiratory Syncytial Virus, but it is clear that randomization was unmasked, developed and performed separately by each center [38]. Any restrictions on the randomization would allow for prediction of future assignments. The possibility that this information was used to selectively enroll healthier or sicker patients across groups was raised [38], as "the person maintaining the list [at each center] would have been aware of the upcoming treatment assignments". The methods they used for randomization "did not protect against such influences, conscious or unconscious" [38].

A trial to assess episiotomy was "affected by physician noncompliance with the randomly assigned therapy" [19]. That is, it appeared that some physicians assigned episiotomy to certain patients even when it was not the treatment indicated by the randomization. While assigning a treatment deterministically does not require advance knowledge of upcoming allocations, it is a form of treatment allocation based on patient characteristics and introduces selection bias.

In a surgical trial conducted at 23 centers, some centers used the sealed envelope system and others used the centralized telephone system [34]. The median age of patients randomized to the experimental treatment was considerably lower than those in the conventional treatment group (59 vs. 63 years, $p<0.01$) when envelopes were used. For three clinicians there were even larger age imbalances (57 vs. 72 years, $p<0.01$; 33 vs. 69 years, $p<0.001$; 47 vs. 72 years, $p=0.03$). These imbalances were not observed when using the telephone system [34], so the implication is that these

imbalances resulted from the use of the sealed envelope system itself. The mechanism would then be through the ability to observe upcoming allocations to be made. Likewise, "the process of randomization by sealed numbered envelopes was frequently violated", as envelopes at some treatment centers may have been unsealed, thus influencing assignments [39], in the Captopril Prevention Project [40]. There were highly significant between-group baseline differences in height, weight and both systolic and diastolic blood pressure, with p-values of 10^{-4}, 10^{-3}, 10^{-8} and 10^{-18}, respectively. The Hypertension Detection and Follow-up Program also used sealed envelopes for the randomization, which, as noted by Psaty et al. [40], is subject to manipulation. In fact, the randomization "was tampered with at one clinic and as a result, 446 participants from that clinic were excluded". In another RCT using sealed envelopes, 125 patients with myocardial infarction were randomized to either heparin or a control group, but the envelopes containing the treatment codes were not numbered consecutively, and there may have been "prejudice in the selection of therapy by alteration of the sequence of the envelopes. . .envelopes were transilluminated for this purpose" [41].

In a randomized study of a culturally sensitive AIDS education program [42], Marcus [43] hypothesized that "subjects with lower baseline knowledge scores. . .may have been channeled into the treatment group", because the treatment group had significantly lower baseline AIDS knowledge scores (39.89 vs. 36.72 on a 52 question test, p=0.005).

In a study of etanercept for children with juvenile rheumatoid arthritis [44], patients in the etanercept group were younger (p=0.0026), less likely to be Caucasian (p=0.022) and of lower weight (p=0.027) than patients in the placebo group. The publication [44] makes no attempt to explain these baseline differences, but the FDA statistical review [45] reveals several issues. For one thing, there was a three-month open-label run-in on etanercept. Because it is easier to discern similarity to, or difference from, that which has already been experienced than it is to unmask an assignment when neither treatment has been previously experienced [46], this run-in increases the likelihood of unmasking treatment allocations. Also, the randomization used blocks of size two, the worst situation for selection bias [5]. Worse still, corresponding blocks in the two strata were mirror images of each other. Four patients were randomized from the wrong stratum; in three of these cases, it was foreseeable that the treatment received would likely be affected; in two cases, the treatment received was actually reversed. Neither this, nor the fact that some patients were randomized out of order, were mentioned in the publication [44].

Oxytocin was compared to amniotomy for induction of labor in 223 women [47]. While the trial used even or uneven dates at birth as the method of "randomization", and hence does not qualify as truly randomized [14], we include it nevertheless because the trial was labeled as randomized, and a casual reading would not reveal that it in fact was not. A re-analysis "hypothesized that clinicians knowing in advance the method of induction of labor to be used for each woman would be influenced in their decision to use induction at all (enroll the woman in the trial). It was demonstrated that obstetricians were very reluctant to induce labor with amniotomy in a woman born on an uneven date when she had an unfavorable cervix (low Bishop score). Thus, randomization by date of birth was an unsatisfactory method in this case, because it produced selection bias at trial entry" [48].

Eight mammography RCTs have been identified by location as Malmo, Canada, Goteborg, Stockholm, Kopperberg, Ostergotland, New York and Edinburgh [28]. Due to substantial baseline imbalances, only the Canadian and Malmo trials were rated, by some, to have even medium quality [49]. That is, using baseline imbalances as a measure of the inadequacy of a trial [50], six of the

eight trials were said to have had their randomization fail. In the Swedish trials, the mean age of the screened and control groups were 55.05 and 54.54 years ($p=3\times10^{-27}$). Also, 28% of the women in the study group vs. 51% of the women in the control group ($p=8\times10^{-42}$) had undergone a prior mammogram [50]. This led to the allegation that the women were not randomized properly and that these results were not due to chance [28]. The Canadian National Breast Screening Study, found to be of sufficient quality by some [28], has been criticized by others. In 14 of 15 screening centers, women were randomized only *after* the clinical examination [51], contrary to what is stated in the protocol. There were baseline imbalances in prior health claims for breast cancer in women aged 40–59 ($p=0.05$), alterations of names in allocation books in women aged 40–49 ($p=0.01$), and advanced breast cancer detected at baseline by physical examination in women aged 40–49 ($p=0.003$) [52]. The lack of allocation concealment [52] allowed for the types of selection bias we consider, and the description of allocation concealment was lacking in many of the other trials as well [53].

Several diabetes treatments were compared in The University Group Diabetes Program RCT, but severe baseline imbalances disadvantaged the tolbutamide group relative to the placebo group [55]. In fact, "After proper randomization one does not expect to find absolutely similar percentages in both groups for every characteristic. However, one does expect to find certain characteristics, which bias the study against tolbutamide to be balanced by other characteristics, which bias in favor of tolbutamide. This simply did not happen in this study... It would appear to any reasonable statistician that for some reason or other the randomization procedure broke down in these three clinics" [54]. Miettinen and Cook [55] added "...by chance or otherwise, the tolbutamide series might be of higher risk in terms of its distribution by familiar coronary heart disease risk indicators than the group given placebo".

The Lifestyle Heart Trial was conducted to assess the progression of coronary atherosclerosis achieved through drastic lifestyle changes without lipid lowering drugs [56]. Of 193 patients, 93 remained eligible after a quantitative coronary angiography. Of these, 53 were randomized (randomization details are unclear) into the experimental group and 40 to the control group. However, only 28 patients randomized to the experimental group and 20 patients randomized to the control group agreed to participate in the study. Patients that did participate were more likely to have a history of angina (87% vs. 65%, $p=0.02$), a larger number of lesions (4.5 vs. 3.5, $p=0.04$) and more severely stenosed lesions (2.3 vs. 2.0 on their 3.0 scale, $p=0.05$). Not all patients that agreed to participate reported their data, and therefore the analyses were based on 20 patients in the experimental group and 15 patients in the control group. When comparing these two groups, there were significant baseline imbalances in gender, which is the likely cause of the large weight and height differences. The experimental group had all men and also had lower high density lipoprotein cholesterol levels ($p=0.04$).

A randomized comparison of talc to mustine for control of pleural effusions, with 23 patients in each group [57], had severe baseline imbalances in age (50.3 vs. 55.3 years), stage (52% 1 or 2 vs. 74%), mean interval between breast cancer diagnosis and effusion diagnosis (33.1 vs. 60.4), and proportion post-menopausal (43% vs. 74%) [58]. These imbalances were quite large, yet did not reach statistical significance because of the small sample size. The imbalances did reach significance in a randomized trial of tonsillectomy for recurrent throat infection in children [59], with $p=0.0309$ and $p=0.0076$, respectively, by the two-sided Smirnov test [60], for history of episodes of throat infection and parents' socioeconomic status. This was an unmasked trial and used blocks of fixed

size four. Altman [58] described large imbalances in other RCTs as well, but did not mention selection bias per se.

4. Discussion

RCTs provide the most reliable between-group evidence, as treatment comparisons tend to be most reliable when the treatment groups being compared are as similar as possible in every way other than the experimental intervention. In fact, the need to create similar treatment groups at baseline is why something as distasteful (to some) as randomization [4,19] is used so often. Yet, unquestioning belief that randomization eliminates all selection bias may be the cause of the common perception that if a between-group p-value is sufficiently low, then it is beyond the reach of selection bias, or any other such minor methodological concern. That is, the p-value is often used as "the arbiter of validity" [61] to prove that the effect is real and not due to selection bias. However, even proper randomization (the proper generation of the allocation sequence, but not its subsequent application to the patients enrolled) guarantees neither the comparability of the comparison groups at baseline nor that any observed imbalances are random. In fact, it has been recognized [2] that unreliable results are not inconsistent with proper randomization, and that even randomized experiments can be "broken" by missing data and/or noncompliance. Our contention is that these are not the only ways a properly randomized experiment may be broken.

Baseline imbalances may be due to chance alone, just as post-treatment comparisons may be, but a disproportionate number of baseline imbalances seem to be dismissed as random and unimportant. Baseline imbalances may also indicate selection bias, especially when unmasked designs conspire with highly restricted randomization to allow for deciphering allocation sequences, which appears not to be an uncommon event [4]. Greenhouse [62] even contends that it "is obvious that if there are serious imbalances in observable baseline variables, it can only be because clinicians were manipulating patient assignment to a treatment. This by definition should give rise to a selection bias". Contrary to popular opinion [63,64], then, baseline covariate testing *is* logical (to test for selection bias), although better tests of selection bias exist, especially considering the multiplicity of covariates and the fact that some are not observed.

Berger and Exner [17] proposed testing for selection bias by examining, within each treatment group, the ability of the investigator's expected likelihood of a patient to receive the active treatment to predict the response variable. This gold standard for detecting selection bias has been applied to very few trials [11–13], and until this test is applied to more trials, there is no way to estimate how prevalent or extreme selection bias may be in practice. Certainly, the serendipitous methods by which we encountered trials that we could evaluate for selection bias are not amenable to offering any credible indicator of the true extent to which selection bias occurs in randomized trials, so our examples may represent only the tip of the iceberg. Certainly, many RCTs with severe baseline imbalances have been published, and in fact many more baseline imbalances may go unreported, as evidenced by the fact that a survey of baseline comparisons found only 2% unbalanced at the 5% level (5% would be expected) [64]. We focused attention here on the 14 we found that have already been criticized as suspicious for selection bias of the type we consider. It may be argued that 14 is not a large number of trials, relative to all the trials that have been conducted. Yet, we found only 17 trials (the 14 we reported and the three [11–13] that reported testing for selection bias) for which

there is any indication whatsoever that there was or was not selection bias. Using this appallingly small number as a denominator, 14 no longer seems so small a numerator, as 14/17=82%.

One may argue that it is the baseline imbalances themselves, and not their cause, that matters most, and so the classification of baseline imbalances in RCTs as random or systematic (selection bias) is useless. We disagree, because if there is selection bias, then methods other than simple covariate adjustment would be required to correct for it [65]. We would recommend, therefore, that the test proposed by Berger and Exner [17] for detecting selection bias be conducted routinely, because there is no other credible indicator that a given trial was free of selection bias. That is, given the very real possibility for selection bias and the ease with which its presence can be assessed, the safe approach might be to check for selection bias even in randomized, masked trials with allocation concealment, and even if there are no obvious baseline imbalances. Unless selection bias is falsified as an explanation for observed treatment effects, these effects cannot legitimately be attributed to the treatments themselves [66].

Acknowledgement

The review team offered helpful comments that improved the clarity.

References

[1] Berger VW, Christophi CA. Randomization technique, allocation concealment, masking, and susceptibility of trials to selection bias. J Mod Appl Stat Methods 2003;2(1):80-6.
[2] Barnard J, Du J, Hill JL, Rubin DB. A broader template for analyzing broken randomized experiments. Sociol Methods Res 1998;27(2):285–317.
[3] Strauss GM. Measuring effectiveness of lung cancer screening: from consensus to controversy and back. Chest 1997; 112:216S–28S.
[4] Schulz KF. Subverting randomization in controlled trials. JAMA 1995;274:1456–8.
[5] Proschan M. Influence of selection bias on type I error rate under random permuted block designs. Stat Sin 1994;4: 219–31.
[6] Fine PEM. Implications of different study designs for the evaluation of acellulara pertussis vaccines. Dev Biol Stand 1997;89:123–33.
[7] Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. JAMA 1994;272:125–8.
[8] Smith PJ, Moffatt MEK, Gelskey SC, Hudson S, Kaita K. Are community health interventions evaluated appropriately? A review of six journals. J Clin Epidemiol 1997;50(2):137–46.
[9] 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;273:408–12.
[10] Moye LA. Endpoint interpretation in clinical trials: the case for discipline. Control Clin Trials 1999;20:40–9.
[11] Fiellin DA, O'Connor PG, Chawarski M, Pakes JP, Pantalon MV, Schottenfeld RS. Methadone maintenance in primary care—a randomized controlled trial. JAMA 2001;286(14):1724–31.
[12] Kroenke K, West SL, Swindle R, Gilsenan A, Eckert GJ, Dolor R, et al. Similar effectiveness of paroxetine, fluoxetine, and sertraline in primary care—a randomized trial. JAMA 2001;286(23):2947–55.
[13] Van Dijk D, Jansen EWL, Hijman R, Nierich AP, Diephuis JC, Moons KGM, et al. Cognitive outcome after off-pump and on-pump coronary artery bypass graft surgery—a randomized trial. JAMA 2002;287(11):1405–12.
[14] Berger VW, Bears JD. When can a clinical trial be called "randomized"? Vaccine 2003;21:468–72.
[15] Frangakis CE, Rubin DB. Principal stratification in causal inference. Biometrics 2002;58:21–9.

[16] Fayers PM, Sprangers MA. Understanding self-rated health. Lancet 2002;359:187–8.
[17] Berger VW, Exner DV. Detecting selection bias in randomized clinical trials. Control Clin Trials 1999;20:319–27.
[18] Blackwell D, Hodges Jr JL. Design for the control of selection bias. Ann Math Stat 1957;28:449–60.
[19] Schulz KF. Unbiased research and the human spirit: the challenges of randomized controlled trials. Can Med Assoc J 1995;153:783–6.
[20] Mark DH. Interpreting the term selection bias in medical research. Fam Med 1997;29(2):132–6.
[21] Matts JP, McHugh RB. Conditional Markov chain design for accrual clinical trials. Biom J 1983;25:563–77.
[22] Vamvakas EC. Evaluation of clinical studies of the efficacy of therapeutic apheresis. J Clin Apher 2000;15:6–17.
[23] Schulz KF, Grimes DA. Generation of allocation sequences in randomized trials: chance, not choice. Lancet 2002; 359:515–9.
[24] Oxtoby A, Jones A, Robinson M. Is your 'double-blind' design truly double blind? Br J Psychiatry 1989;155:700–1.
[25] Day S. Blinding or masking. Encycl Biostat 1998;1:410–7.
[26] Cook DJ, Hébert PC, Heyland DK, Guyatt GH, Brun-Buisson C, Marshall JC, et al. How to use an article on therapy or prevention: pneumonia prevention using subglottic secretion drainage. Crit Care Med 1997;25:1502–13.
[27] Grimes DA, Schulz KF. Randomized controlled trials in *contraception*: the need for 'CONSORT' guidelines. Contraception 2001;64:139–42.
[28] Gotzsche PC, Olsen O. Is screening for breast cancer with mammography justifiable? Lancet 2000;355:129–34.
[29] Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomized trials affect estimates of intervention efficacy reported in meta-analysis? Lancet 1998;352(9128):609–13.
[30] Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomized and nonrandomized clinical trials. BMJ 1998;317:1185–90.
[31] 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(5):408–12.
[32] Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999; 282:1054–60.
[33] Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. BMJ 2001;323:42–6.
[34] Kennedy A, Grant A. Subversion of allocation in a randomized controlled trial. Control Clin Trials 1997;18(3S):77S–8S.
[35] CASS Investigators A. Coronary Artery Surgery Study (CASS): a randomized trial of coronary bypass surgery. Comparability of entry characteristics and survival in randomized patients and nonrandomized patients meeting randomized criteria. J Am Coll Cardiol 1984;3:114–28.
[36] Hallstrom A, Davis K. Imbalance in treatment assignments in stratified blocked randomization. Control Clin Trials 1988; 9:375–82.
[37] Groothuis JR, Simoes EAF, Levin MJ, Hall CB, Long CE, Rodriguez WJ, et al. Prophylactic administration of respiratory syncytial virus immune globulin to high-risk infants and young children. NEJM 1993;329(21):1524–30.
[38] Ellenberg SS, Epstein JS, Fratantoni JC, Scott D, Zoon KC. A trial of RSV immune globulin in infants and young children: the FDA view. N Engl J Med 1994;331(3):203–4.
[39] Peto R. Failure of randomisation by 'sealed' envelope. Lancet 1999;354:73.
[40] Psaty BM, Furberg CD, Pahor M, Alderman M, Kuller LH. National guidelines, clinical trials, and quality of evidence. Arch Intern Med 2000;160(17):2577–80.
[41] Carleton RA, Sanders CA, Burack WR. Heparin administration after acute myocardial infarction. NEJM 1960;263: 1002–1005.
[42] Stevenson HC, Davis G. Impact of culturally sensitive AIDS video education on the AIDS risk knowledge of African American adolescents. AIDS Educ Prev 1994;6:40–52.
[43] Marcus SM. A sensitivity analysis for subverting randomization in controlled trials. Stat Med 2001;20:545–55.
[44] Lovell DJ, Giannini EH, Reiff A, Cawkwell GD, Silverman ED, Nocton JJ, et al. Etanercept in children with polyarticular juvenile rheumatoid arthritis. N Engl J Med 2000;342(11):763–9.
[45] Berger V. FDA Product Approval Information-Licensing Action: Statistical Review. http://www.fda.gov/cber/products/etanimm052799.htm 1999, accessed 3/7/02.
[46] Leber PD, Davis CS. Threats to the validity of clinical trials employing enrichment strategies for sample selection. Control Clin Trials 1998;19:178–87.

[47] Bakos O, Backstrom T. Induction of labor: a prospective, randomized study into amniotomy and oxytocin as induction methods in a total unselected population. Acta Obstet Gynecol Scand 1987;66:537–41.
[48] Villar J, Carroli G. Methodological issues of randomized clinical trials for the evaluation of reproductive health interventions. Prev Med 1996;25:365–75.
[49] Olsen O, Gotzsche PC. Cochrane review on screening for breast cancer with mammography. Lancet 2001;358:1340–2.
[50] Gotzsche PC, Olsen O. Screening mammography re-evaluated, Reply. Lancet 2000;355:752.
[51] Bailar JC, MacMahon B. Randomization in the Canadian National Breast Screening Study: a review for evidence of subversion. Can Med Assoc J 1997;156:193–9.
[52] Boyd NF. The review of randomization in the Canadian National Breast Screening Study. Can Med Assoc J 1997;156(2): 207–209.
[53] Humphrey LL, Helfand M, Chan BKS, Woolf SH. Breast cancer screening: a summary of the evidence for the US Preventive Services Task Force. Ann Intern Med 2002;137(5, Part 1):347–60.
[54] Schor S. The University Group Diabetes Program: a statistician looks at the mortality results. JAMA 1971;217(12): 1671–1675.
[55] Miettinen OS, Cook EF. Confounding: essence and detection. Am J Epidemiol 1981;114(4):593–603.
[56] Ornish D, Scherwitz LW, Billings JH, Gould KL, Merritt TA, Sparler S, et al. Intensive lifestyle changes for reversal of coronary heart disease. JAMA 1998;280(23):2001–7.
[57] Fentiman IS, Rubens RD, Hayward JL. Control of pleural effusions in patients with breast cancer. Cancer 1983;52:737–9.
[58] Altman DG. Comparability of randomized groups. Statistician 1985;34:125–36.
[59] Paradise JL, Bluestone CD, Bachman RZ, Colborn DK, Bernard BS, Taylor FH, et al. Efficacy of tonsillectomy for recurrent throat infection in severely affected children results of parallel randomized and nonrandomized clinical trials. NEJM 1984;310(11):674–83.
[60] Berger VW, Permutt T, Ivanova A. The convex hull test for ordered categorical data. Biometrics 1998;54(4):1541–50.
[61] Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet 2002;359:248–52.
[62] Greenhouse SW. The growth and future of biostatistics. Stat Med 2003;22:3323–35.
[63] Senn S. Testing for baseline balance in clinical trials. Stat Med 1994;13:1715–26.
[64] Altman DG. In: Armitage P, Colton T, editors. Covariate imbalance, adjustment for, encyclopedia of biostatistics. Chichester: John Wiley & Sons, 1998. p. 1000–5.
[65] Barrier R, Ivanova A, Berger VW. Adjusting for selection bias in randomized trials. Stat Med 2005 [in press].
[66] Berger VW. Selection bias and baseline imbalances in randomized trials. Drug Inf J 2004;38:1–2.