Special Article

Randomization in Controlled Clinical Trials: Why the Flip of a Coin Is So Important

Jeffrey Blume, Ph.D., and Jeffrey F. Peipert, M.D., M.P.H.

Abstract ***(J Am Assoc Gynecol Laparosc 2004, 11(3):320–325)***

The randomized controlled trial (RCT) is considered the highest level of medical evidence. In this brief overview, we discuss several key principles of the RCT. First, balance is paramount. Comparison groups must have similar proportions of participants with "important" prognostic and confounding factors. Randomization may or may not achieve this balance; if it does not, statistical adjustments should be used. Second, a statistical analysis should emphasize comparability and not mask dissimilarity. If the trial was indeed randomized, certain analysis techniques, such as an intention to treat analysis, should always be presented. Third, additional bias-reducing techniques, such as concealing treatment assignments from treating physicians and participants (i.e., masking) and using clearly defined exclusion and inclusion criteria, should be used wherever possible.

The randomized controlled trial (RCT) is considered the highest level of medical evidence (level I).[1] This is, of course, a stereotype. A trial should not confer the highest level of medical evidence simply because it was randomized; rather the trial should have been well conducted, preferably large enough to differentiate signal from noise, protected in design from selection biases, and, most importantly, *balanced* on prognostic and confounding factors so that comparisons among treatment groups are fair. The process of randomization helps to guard against selection bias and imbalance. However, randomization is neither necessary nor sufficient for valid scientific inference; other quality criteria are clearly important for scientific validity.

The Basics of Randomization

The idea that randomly assigning experimental units to study groups could promote balance on known and unknown confounding factors was noted and popularized in the 1920s by the famous statistician and geneticist Sir Ronald A. Fisher.[2] However, randomization would not be widely accepted in medical studies until the 1940s when it found influential advocates in Bradford Hill and Richard Doll.[3–5] Since then, randomization has become a fundamental tool of medical science and, in particular, in clinical trials.

Randomization is the act of assigning participants or experimental units (e.g., hospitals or workplaces) to study groups by means of a chance process in which the probability of assignment is known or can be determined (but is not necessarily equal for all assignments).[6] A haphazard assignment procedure does not substitute for a random one, because the probability of assignment is unknown and cannot be controlled or determined (e.g., when treatment assignment depends on the participant's birth date or social security number, or when treatment assignment alternates depending on the day of the week).[7] This is precisely when selection and confounding biases seep into clinical trials.[8,9] Details on the practical implementation of randomization in clinical trials can be found elsewhere.[10,11]

Randomization has known statistical properties, while a haphazard assignment procedure does not. For example, randomization imparts a mathematical property that allows p values to be calculated without making any (distributional) assumptions about the data. Although there is some statistical benefit in this sense, the tangible benefit of randomization is the scientific value in producing balanced comparison groups.

Any outcome from a random assignment procedure is a randomized outcome. This means that even unbalanced outcomes are, by definition, random. As an illustration, consider the familiar random process of flipping a fair coin 10 times. More often than not, we would observe that the number of heads is roughly equal to the number of tails—the number of heads and tails are approximately balanced. However it is possible, albeit unlikely, that we will observe 10 heads in a row (an outcome which is 252 times less likely than observing exactly 5 heads out of 10 tosses). Or consider a clinical trial with two treatment groups. It is possible, although highly improbable, to randomly assign only older women to the placebo group and younger women to the

From the Center for Statistical Sciences, Brown University, Division of Research (Dr. Blume); and the George Anderson Outcomes Measurement Unit, Department of Obstetrics and Gynecology, Women & Infants Hospital, Providence, Rhode Island (both authors).

Corresponding author Jeffrey F. Peipert, M.D., M.P.H., Division of Research, Women and Infants Hospital, 101 Dudley Street, Providence, RI 02905.

Supported in part by National Institutes of Health grant K24 HD01298-03, Midcareer Investigator Award in Women's Health Research from the National Institutes of Child Health and Human Development.

Submitted February 15, 2004. Accepted May 1, 2004.

treatment group. This outcome is clearly unbalanced, but it is a randomized outcome.

Randomization and Balance

"Randomized" and "balanced" are perhaps the two most readily confused properties of a clinical trial. Many use these two terms interchangeably, when in fact they have distinct scientific and statistical meanings. Not all randomized trials are balanced, and not all balanced trials are randomized. This distinction is important because valid scientific inference from data requires that potentially influential factors and/or confounding variables, both known and unknown, be balanced among comparison groups. That is, we want our comparison groups to be comparable. Randomization is one way to promote such balance, and it is, perhaps, the most scientifically sound way when not all influential factors are known.

Why is balance so essential in clinical research? Without comparable groups, the causal link between the intervention or treatment under investigation and the outcome is broken. Many sophisticated statistical tools, like regression analysis, can force this balance after the fact so as to provide valid scientific comparison. But prevention is the best policy, and this is the role of randomization. Randomization promotes balance on known *and unknown* factors, and it does this exceptionally well *on average*. But therein lies the caveat: randomization does not always yield balanced comparison groups. Thus, one cannot simply ignore the issue of balance; it must be carefully investigated in the analysis stage.

Investigating the Degree of Imbalance

How large of an imbalance is problematic? It is difficult to provide a general answer to this question because the potential for bias due to imbalance is proportional to the degree of imbalance and to the disparity in the outcome measurement.

A simple illustration may be helpful: an RCT was performed to evaluate SprayGel (Confluent Surgical, Inc., Waltham, MA) as a barrier to adhesion formation.[12] The investigators randomized 45 women undergoing open or laparoscopic myomectomy to receive either SprayGel or routine care. The treated group contained five patients (20.8%) with prior myomectomy; the control group had one patient (4.8%) with prior myomectomy. If prior myomectomy is associated with increased (or decreased) adhesion formation, then this is a potential confounder. Is prior myomectomy an important confounding variable? Does this distribution from randomization appear balanced? We leave this judgment to the reader.

Consider dividing a group of mothers into two groups at random (groups A and B) and determining the average time in labor. We would expect these averages to be almost identical, all things being equal. But, if for some reason, one group had a higher proportion of parous women, then we might expect the average labor time to be different (because we know labor is more rapid for parous women compared to nulliparous women). Suppose that the average labor time for nulliparous women is 14 hours, while the labor duration is 7 hours for parous participants. Suppose further that the proportion of parous women in group A is 45%, compared with 70% in group B. Then the average labor time in group A is 10.85 ($.45 \times 7 + .55 \times 14$) hours, while the average labor time in group B is 9.1 ($.7 \times 7 + .3 \times 14$) hours. This yields a difference of 1.75 ($[0.7 - 0.45] \times [14-7]$) hours, all due to the imbalance of 25%. Whether or not this is a statistically significant difference would depend on the degree of sampling variability.

It is always important to investigate the degree of imbalance, in a trial, even if the trial was randomized. It is not wise to randomize and ignore observed imbalances or factors that are known to be effect modifiers and/or confounders. This has been mockingly referred to as the closeurization principle: randomize and "close your eyes!"[13] Many investigators and journals already avoid this by providing comparisons among the treatment groups on demographic and baseline characteristics. Here the magnitude of the observed imbalance is what is important and not the statistical significance. This is because, as we saw earlier, the degree of bias is directly proportional to the magnitude of imbalance. And while a noticeable imbalance on one or two factors may be problematic inferentially, it is *not*, by itself, evidence that the randomization process has broken down. It simply means we were unlucky with respect to the specific unbalanced covariates, and we will have to account for them in the analysis. Some baseline imbalance in covariates is to be expected, especially if the number of covariates is large.

Ways to Randomize

Randomization does not imply that the probability of assignment to each study group is equal. For example, biased-coin designs can be used to assign more participants to one treatment arm than another by simply varying the assignment probability.[14–16] Several other randomization schemes, such as urn randomization[17] or play the winner rules,[18,19] vary the selection probabilities during the trial depending on the observed outcomes. These schemes are known as adaptive randomization schemes, and they have two important limitations: (1) the outcome of one participant must be measured before the next is randomized (ruling out studies with survival endpoints for example); and (2) standard p values are incorrect and must be adjusted according to the randomization scheme because they depend on the selection probabilities, which now contain information about the outcome. Unfortunately, it is not always clear how to obtain the correct p value, and this can lead to drastic action, such as having to repeat the trial when statisticians can't agree on the adjustment (for one such example consider the ECMO trial).[20–22] In practice, situations ripe for adaptive randomization are rare.

A more commonly employed technique is block randomization. The idea is this: treatment assignments are bunched together in groups or blocks of a given size, and these groupings are special, because they always have the

same number of treatment and placebo assignments. However the order of assignment varies from block to block. For example, consider treatment arms A and B. A block of size eight might be ABBABBAA or BBAABBAA. Multiple blocks are appended together to create an assignment schedule. Thus, after every eight assignments, we would have equal numbers of participants in each arm. Varying the block size will further disguise the assignment schedule. Blocking is often used in studies with small samples or in multicenter studies to ensure that equal numbers of participants are assigned to each arm within each site. Block randomization with varying block design was used in the STOP-DUB multicenter clinical trial.[23]

Allocation Concealment: Keeping the Randomization Process Pure

After randomization, the allocation must be preserved with adequate concealment. Concealment keeps participants and clinicians unaware of the treatment group assignments, thus limiting unconscious (or conscious) bias that may lead to preferential treatment or evaluation. Several empirical investigations have demonstrated that trials using inadequate or unclear allocation concealment yielded up to 40% larger estimates of the effect, compared with trials that used adequate concealment.[9,24,25] Trials deemed of lower quality tended to exaggerate the treatment effects. The results of trials with the worst methods of concealment fluctuated extensively above and below the estimates from better studies.[9]

The minimum accepted standard for allocation concealment includes: sequentially numbered, opaque, sealed envelopes; sequentially numbered containers; pharmacy-controlled; and central randomization (e.g., by telephone to a central office).[26] These minimum criteria are met by only about a quarter of published trials.[27] Given that many individuals involved in trials will be tempted to undermine randomization assignments if afforded the opportunity, it is essential to implement allocation concealment mechanisms so that assignments cannot be deciphered. Enrollment and randomization via the Internet is an excellent mechanism for randomly assigning participants, because it is hard to tamper with, hard to misplace, and readily available.

Why Randomize?

The key to valid scientific comparisons is that the comparison be a fair one. That is, the comparison groups should be balanced on important factors that may affect the outcome. Randomization often imparts the desired balance on known and unknown factors alike. In large studies, the chance that randomization will yield approximately balanced groups is, in general, quite high.

Randomization seldom yields a drastically unbalanced outcome for known or unknown influential factors. This is why randomization is routinely employed in medical studies even though it will seldom yield a perfectly balanced sample. In this way, researchers avoid the impractical task of having to force balance among comparison groups themselves (e.g., by choosing a sample that is balanced on all the important factors, assuming, of course, that all are known). In large studies, the risk that the study will be unbalanced when all is said and done is small. But even this risk can be minimized with modified randomization schemes.[15]

Practically, randomization shifts the burden of proof from the investigator to the critical reader. If the investigator has failed to identify an important risk factor when designing the study, chances are good that the study groups are balanced on this factor. Thus, it is the reader who must identify a reason as to why the comparison groups would not be balanced on important predictor variables. In contrast, the burden to show comparability in an observational study rests with the investigators.

The Intention to Treat Principle

The intention to treat (ITT) principle is a statistical principle designed to guide the analysis of a randomized study. It states that participants should be analyzed according to the study group that they were randomized to, regardless if they received that treatment or another. This applies to participants who withdraw at any point after randomization, participants who cross over to another study arm, or participants who do not fully comply with the treatment intervention. Statisticians sometimes refer to this as "preserving the randomization," because it retains the original treatment assignments and protection against selection bias.

The opposite of an ITT analysis is the "as-treated" analysis, which analyzes participants according to the treatment they actually received. While this approach appears to be more relevant scientifically, it has the fatal drawback of allowing bias to creep into the analysis. A common mistake is the following: in the analysis of clinical trial, the investigators excluded all participants who withdrew or did not complete the entire treatment regimen (arguing that they would dilute the results because they did not receive the full therapy). However, this approach actually introduces bias because the participants who dropout or fail to comply with the assigned therapy might be sicker or have more treatment-related side effects. Thus, the analysis will tend to have healthier participants, who may be more likely to respond to the treatment, and the results may be exaggerated in favor of the new treatment. This problem is compounded only if the dropout rate varies significantly among comparison groups.

Remember that randomization helps to eliminate bias in the assignment of participants to study groups. Hence, if we allow subjects to be re-assigned by some other mechanism (e.g., desire, motivation, adverse events, etc.), then we also allow selection bias to creep back in. If there is a high amount of dropout in a study due to side effects, the estimated treatment effect may be biased downward (i.e., toward the null hypothesis of no effect) from a scientific or causal perspective, but it tends to be unbiased for observed population effect because many in the general population will also stop taking the intervention. So the ITT analysis assures that selection bias does not affect the result, but it

can lack scientific meaning if there is a large proportion of noncompliance or dropouts. For this reason, the ITT analysis is often referred to as the policy analysis, because it generalizes to the population at large. Experience shows that the ITT analysis is a good predictor of policy implementations in the population at large (see Piantadosi 1997, p 278 for some examples).[11]

Of course, if the percentage of participants dropping out and crossing over is minimal, the ITT and as-treated analytic approaches will yield similar results (hence the importance of retention!). As a general rule, the ITT analysis will be more conservative than the as-treated analysis. The only problem is that when the dropout rate is high, the ITT analysis will be so conservative as to be uninformative. Unfortunately, situations like these are often packed with bias, and they provide little reliable information no matter what analysis is done. A good general rule is to always present an ITT analysis and then supplement it with other analyses as appropriate.

Should We Exclude Ineligible Participants?

An important question related to the ITT analysis is how to deal with ineligible participants (i.e., subjects found to be ineligible for the trial by specific inclusion and exclusion criteria). Should they be excluded from the analysis, or do they need to be included to preserve the benefits of randomization? The answer depends on the type of eligibility criteria that is violated.

Assume for the moment that the violated eligibility criteria cannot change and the participant's enrollment was due to a clerical or transcription error that was discovered after the participant had been randomized. In this case, the participant can be excluded from the analysis without introducing bias.[28,29] The reasoning is that such a participant was ineligible to begin with, and we know that his or her inclusion into the study did not adversely affect any other participants' treatment assignment because everybody was independently randomized. Hence, excluding the ineligible participant does not introduce bias and keeps the study population representative of the target population.

However, if the violated eligibility criteria can vary over time (and we strongly dissuade this; all eligibility criteria should be defined at a single point in time), then the participant should not be excluded from the analysis. This is because the criteria might vary differentially with the assigned treatment and then excluding the participant would result in confounding with the treatment. Here it is better to preserve the randomization and deal with the "contaminated" nature of the study population in another way.

Consider a hypothetical randomized trial of a new endoscopic device used to evaluate tubal patency at the time of laparoscopy. The trial recruited women age 18 to 35. However, we later discover that a participant lied about her age and was actually 45. It is reasonable to exclude the participant even after randomization. The key is that the criteria with which we exclude subjects is defined at baseline and, therefore, cannot be confounded by treatment. So excluding the participant does not in any way void the randomization, as long as subjects are excluded based on baseline characteristics or fixed covariates that cannot vary with treatment.

Now if we assume that randomization was completed before the procedure, an example of a problematic exclusion would be excluding all participants who never actually underwent laparoscopy. This is because said participants might very well be sicker (having had the procedure done elsewhere or the procedure put off for other medical reasons). This is an example of a poor eligibility criterion because participants cannot meet this criterion at the time of randomization. This can be fixed in two ways: (1) change the eligibility criteria to participants intending to undergo laparoscopy; or (2) randomize the participants only after they are on the operating table.

There are two possible exceptions to the above discussion. The first is with adaptive randomization, where participants can *never* be excluded because the randomization probabilities depend on the outcomes of those participants (excluding them causing all sorts of analysis problems). The second possible exception is when large numbers of participants purposefully lie about eligibility criteria in an attempt to get the experimental treatment. These participants often drop out of the control arm if they are randomized to it, but remain undiscovered if assigned to the experimental arm. Unfortunately, excluding these participants does tend to introduce selection bias because it is the sicker participants who are more likely to try this tactic and excluding them leaves their undiscovered counterparts in the experimental arm. In this sense, participants who are ineligible may represent a select subset, and systematically removing them from any single arm of a study without removing them from all other arms has the potential to introduce bias.

Forcing Balance in the Analysis

Even though randomization tends to promote balance, influential factors should still be accounted or adjusted for in the ensuing analysis. The most common example of this is adjusting for site-to-site imbalances in baseline and demographic factors by including site as a fixed or random effect in a regression. Regression analysis is probably the most common statistical method used to adjust for important factors. Regression is essentially a sophisticated stratification technique that combines all of the data under the assumption of a common variance, but is otherwise the same as stratification. Regression can be a great tool in this respect, because it allows one to efficiently use all the data at the cost of a relatively minor assumption: that a common variance can be used to estimate the variability of everyone.

An important, but often overlooked point is that the statistical analysis may actually create imbalance when none existed if important confounding factors are ignored. This is common in studies of survival (because of the way survival curves are estimated) and with logistic regression (because collapsing two by two tables can actually create associations when none exists), where the failure to adjust for important factors (even if perfectly balanced at baseline)

will lead to biased results.[30,31] We realize that it is not possible to include every possible factor in a model, nor do we think this is necessary. Our point here is simply that known influential factors cannot be ignored if we expect to get scientifically meaning results.

Our recommendation is that every primary comparison in a study (i.e., straight unadjusted comparisons between two study groups by two means, proportions, or hazard rates), should be augmented with an analysis that adjusts for observed imbalances and important factors.

Subgroups Benefit from Randomization

An interesting and useful property of randomization is that the promotion of balance among comparison groups applies to all subgroups defined by baseline or fixed covariates. For example, if we randomly assign subjects to a treatment and placebo arm, subgroup comparisons within a baseline-defined subgroup also benefit from the randomization. These comparisons will be less efficient because they have a smaller sample, but they are not any less valid and will not be any more biased. These comparisons also tend to be balanced, on average, with respect to known and unknown factors.

Keep in mind, however, that subgroups are more likely to be unbalanced just because they have a smaller sample. However, if the subgroups themselves are moderately large, the decrease in "protection," as compared with the primary comparison, is negligible. Subgroups are also afforded protection by randomizing as long as they are not defined by covariates in the causal pathway of the treatment.

Randomization as a Quality Measure

In the era of evidence-based medicine, randomized studies are routinely hailed as the highest form of empirical evidence simply because they are randomized.[1,32] Nonrandomized studies, no matter how pristine, are not as highly regarded because their design does nothing to address the potential influence of unknown (and unmeasured) confounding variables. Whether or not these factors exist and whether or not they are balanced (or unbalanced for that matter), is left open to debate. In contrast, it is an inherent design property of randomized trials that unknown influential factors are more likely to be balanced.

Moreover, randomized trials are often conducted "better" (i.e., higher quality). For example, in randomized trials eligibility criteria are often selected and enforced to a higher degree of scrutiny. Masking (or blinding) is an additional quality measure. Double masking is seldom employed in observational studies, leaving open the possibility of detection bias. These are two quality criteria that often make randomized trials more scientifically sound than observational studies, and these two characteristics have nothing to do with the randomization process. This is an important point, because a well-conducted and controlled observational study with clear inclusion/exclusion criteria, masking of evaluators, etc., can provide better evidence than a poorly conducted randomized trial.

While RCTs have traditionally been considered the "gold standard" and the highest level of medical evidence, recent reports have demonstrated the value and validity of well-done observational research. One study evaluated five clinical topics and compared the results of observational studies and RCTs. The authors concluded that the results of well-designed observational studies (e.g., cohort or case-control design) do not systematically overestimate the magnitude of the effects of treatment.[33] Thus, high-quality observational studies clearly have an important place in health care research.

Conclusions

Randomization changes the fundamental nature of a study, affecting everything from the budget, physician conduct, and participant care to interim reporting, statistical analyses, and the interpretation of the study results. Randomization is an important tool and should be employed whenever possible. However it does not substitute for something that is even more important: balance. The ITT principle helps retain the protection from selection bias and provides a conservative estimate of the treatment effect if there is noncompliance, dropout, or crossover. While it is true that randomized trials will often provide a higher quality of empirical evidence, it is also true that large well-conducted, balanced, nonrandomized studies can provide strong empirical evidence.[33]

References

1. Harris RP, Helfand M, Woolf SH, et al: Current methods of the US Preventive Services Task Force: A review of the process. Am J Prev Med 2001, 20:21–35.
2. Fisher RA: The Design of Experiments. London: Oliver and Boyd, 1935.
3. Hill AB: The clinical trial. N Engl J Med 1952, 247:133–49.
4. Armitage P: Fisher, Bradford Hill, and randomization. Int J Epidemiol 2003, 32:925–8; discussion 945–8.
5. Chalmers I: Fisher and Bradford Hill: Theory and pragmatism? Int J Epidemiol 2003, 32:922–4; discussion 945–8.
6. Meinert CL: Clinical Trials Dictionary: Terminology and Usage Recommendations. Baltimore, Johns Hopkins University Press, 1996.
7. Grimes DA: Randomized controlled trials: "It ain't necessarily so." Obstet Gynecol 1991, 78:703–4.
8. Altman DG: Randomisation. BMJ 1991, 302:1481–2.
9. Schulz KF, Chalmers I, Hayes RJ, et al: Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995, 273:408–12.
10. Meinert CL, Tonascia S: Clinical Trials: Design, Conduct, and Analysis. New York, Oxford University Press, 1986.
11. Piantadosi S: A Methodologic Perspective. New York, John Wiley & Sons, Inc., 1997.

12. Mettler L, Audebert A, Lehmann-Willenbrock E, et al: Prospective clinical trial of SprayGel as a barrier to adhesion formation: An interim analysis. J Am Assoc Gynecol Laparosc 2003, 10:339–44.

13. Royall RM: Current advances in sampling theory: Implications for human observational studies. Am J Epidemiol 1976, 104:463–74.

14. Hoel DG, Sobel M, Weiss GH: A survey of adaptive sampling for clinical trials. In Perspectives in Biometrics. Edited by RM Elashoff. New York, Academic Press, 1975, pp 29–61.

15. Efron B: Forcing a sequential experiment to be balanced. Biometrika 1971, 58:403–17.

16. Pocock SJ, Simon R: Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics 1975, 31:103–15.

17. Wei LJ, Lachin JM: Properties of the urn randomization in clinical trials. Control Clin Trials 1988, 9:345–64.

18. Wei LJ, Durham S: The randomized play-the-winner rule in medical trials. JASA 1978, 73:840–3.

19. Zelen M: Play the winner rule and the controlled clinical trial. J Am Stat Assoc 1969, 64:131–46.

20. Royall RM, Bartlett RH, Cornell RG, et al: Ethics and statistics in randomized clinical trials. Stat Sci 1991, 6:52–88.

21. Ware JH, Epstein MF: Extracorporeal circulation in neonatal respiratory failure: A prospective randomized study. Pediatrics 1985, 76:849–51.

22. O'Rourke PP: ECMO: Where have we been? Where are we going? Respir Care 1991, 36:683–92; discussion 692–4.

23. Dickersin K, Munro M, Langenberg P, et al: Surgical Treatments Outcomes Project for Dysfunctional Uterine Bleeding (STOP-DUB): Design and methods. Control Clin Trials 2003, 24:591–609.

24. Moher D, Pham B, Jones A, et al: Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998, 352:609–13.

25. Juni P, Altman DG, Egger M: Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ 2001, 323:42–6.

26. Schulz KF, Grimes DA: Generation of allocation sequences in randomised trials: Chance, not choice. Lancet 2002, 359:515–9.

27. Schulz KF, Chalmers I, Grimes DA, et al: Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. JAMA 1994, 272:125–8.

28. Gail M: Eligibility exclusions, losses to follow-up, removal of randomized patients and uncounted events in cancer clinical trials. Cancer Treat Rep 1985, 69:1107–13.

29. Fisher LD, Dixon DO, Herson J, et al: Intention to treat in clinical trials. In Statistical Issues in Drug Research and Development. Edited by KE Peace. New York, Dekker, 1990, pp 331–50.

30. Gail MH, Tan WY, Piantadosi S: Tests for no treatment effect in randomized clinical trials. Biometrika 1988, 75:57–64.

31. Gail MH, Wieand S, Piantadosi S: Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika 1984, 71:431–44.

32. Grimes DA, Schulz KF: An overview of clinical research: The lay of the land. Lancet 2002, 359:57–61.

33. 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.