Critical Appraisal of Randomized Clinical Trials: Can We Have Faith in the Conclusions?

Martina Mittlböck

Core Unit for Medical Statistics and Informatics, Section of Clinical Biometrics, Medical University of Vienna, Austria

Key Words

Validity, internal and external · Reliability · Randomization · Blinding · Intention-to-treat

Summary

Randomized clinical trials (RCTs) are the most appropriate research design for studying the effectiveness of a specific intervention. Its results are considered as the highest 'level of evidence'. Published reports on RCTs have already succeeded in a peer review process, but still there can be undetected major deficiencies of the study that may question the reported outcome. It is still up to the readers to assess the quality of publications and to question if the published results apply to their patients. The major points of such a critical appraisal process are reviewed and discussed with a focus on breast cancer studies.

Introduction

Nowadays, clinical practitioners are usually overloaded with information from the literature. They have to appraise whether a reported piece of research is worthwhile to be used in their own clinical decision-making. A lot of the research papers have flaws (even after peer review), but many of these deficiencies will be negligible and will have almost no effect on the conclusions to be drawn from a study [1]. Thus, the important question is not whether there are defects, but whether those defects matter. It is up to the readers to use their critical appraisal skills to detect the flaws of a study and to decide how they affect the usefulness of the paper. The critical read-

Schlüsselwörter

Validität, interne und externe · Reliabilität · Randomisierung · Verblindung · Intention-to-treat-Prinzip

Zusammenfassung

Randomisierte klinische Studien (RCTs) sind am besten geeignet, um die Wirksamkeit von Interventionen zu untersuchen. Ihre Ergebnisse werden als höchste Evidenzstufe betrachtet. Publikationen von RCTs haben bereits erfolgreich einen Peer-Review durchlaufen, trotzdem kann man nicht ausschließen, dass noch bedeutende unentdeckte Mängel in der Studie vorhanden sind. Nach wie vor obliegt es den Lesern, die Qualität der Publikation zu beurteilen und zu fragen, ob die publizierten Ergebnisse auf ihre Patienten anwendbar sind. Die wichtigsten Punkte einer solchen kritischen Abschätzung werden besprochen und mit einem Schwerpunkt auf Brustkrebsstudien diskutiert.

ing of a scientific article and the evaluation whether its results can be applied to patients is a fundamental skill that all clinicians should have [2].

The evidence of studies depends on their design. The so-called 'levels of evidence' from the Oxford Centre for Evidence-Based Medicine [3] include, from high to low evidence, (1) randomized clinical trials (RCTs), systematic reviews on RCTs [4, 5] and 'all-or-none' case series; (2) cohort studies; (3) case control studies; (4) case series; and (5) expert opinions. As RCTs are the most appropriate research design for studying the effectiveness of a specific intervention or treatment [6], the critical appraisal at hand focuses on RCTs [1, 7–14] only.

Questions and Criteria for Critical Appraisals

Are the Results of the Study Valid?

The first part of a critical appraisal process deals with the internal validity or accuracy of the results. It assesses whether the reported effects represent the correct direction and magnitude. That is, do the results represent an unbiased estimate of the treatment effect or have they been influenced in some systematic fashion to lead to false conclusions [12]?

The main five questions [12] to assess validity are:

Was the Assignment of Patients to Treatment Randomized?

An essential part of most prospective RCTs is the random allocation of patients to the treatment groups so that accidental or even intentional biases are avoided. Proper randomization ensures that the next treatment allocation is predictable neither for clinicians and study personnel, nor for patients. Pseudo-randomization by birth date, e.g. group A for even and group B for uneven days, is predictable and thus not recommended. A coin toss meets this criterion but has the disadvantage that randomizations are not traceable later. Often, sealed envelopes containing random treatment sequences are used, and it can be traced later whether the envelopes have been opened in the assigned order. More complex randomization methods additionally ensure that important prognostic factors are balanced between treatment groups, e.g. age distribution, sex or tumor staging. This stratified randomization can be performed either by sealed envelopes or by computer programs (e.g. with the 'Randomizer' [15]). Sealed envelopes can only be used for cases with few strata.

Were All Patients Who Entered the Trial Properly Accounted for and Attributed at Its Conclusion?

Was follow-up complete? Every patient who entered the trial should be included in the final analysis and thus contribute to the conclusions. If a substantial number of patients are lost to follow-up, then the validity of the trial may be questioned as missing patients may behave differently than the patients remaining in the study (informative missingness). For example, patients may either not return to follow-up recalls because of adverse outcomes of the treatment or because they are doing so well that they do not want to waste their time by returning to the follow-up visits. The larger the proportion of missing values, the higher is the risk that informative drop-outs may change the results of the study.

Were all patients analyzed in the groups to which they were randomized? The so-called 'intention-to-treat' (ITT) principle requires that randomized patients remain in their randomized group for analysis even if the scheduled treatment was only applied in parts or, even worse, if no treatment at all or a completely other treatment (than randomized) was applied. This might be surprising, but if patients do not properly take their randomized medication or receive a different one, then there

will usually be prognostically relevant reasons. For example, if patients are too frail for the scheduled treatment, then the exclusion of such non-compliant patients restricts the analysis to those who may be destined to have a better outcome and destroys the unbiased comparison provided by randomization. ITT is recommended for studies designed to show differences. However, in non-inferiority and equivalence trials, the application of the ITT principle could become problematic, as treatment differences under ITT are often underestimated and showing non-inferiority or equivalence may become easier.

Additional analyses may include per-protocol analyses, i.e. only patients are included who are treated with the randomized treatment and as described in the protocol. An as-treated analysis contributes patients to the treatment actually received and not necessarily randomized. Both the per-protocol and the as-treated analysis are vulnerable to bias due to excluded patients and patients changing groups, respectively. A comparison of the results of all three analysis strategies may shed light on this issue.

Were Patients, Their Clinicians and Study Personnel 'Blind' to Treatment?

Patients, clinicians or other study personnel may intentionally or unintentionally change their attitude and/or behavior in a systematic way by knowing the applied treatment. For example, special attention may be unintentionally given to patients under new therapy by clinicians, special expectations (or fear) of the patients may be connected to the experimental therapy, or the treatment knowledge affects a subjective assessment of the outcome. Such behavior introduces systematic bias and distorts the results. This is closely related to the well-known 'placebo effect', i.e. patients show a stronger 'treatment effect' under placebo treatment compared to untreated patients. Thus, treatments should be blinded to all involved persons (including the statistical analyst), whenever possible. For a drug study, smell, taste, shape and color of the drugs and frequencies of applications should be identical. However, specific side effects could still unmask the drugs. Unblinding of single patients should be possible for the responsible physician at any time in case of side effects or adverse events. However, any unblinding has to be documented and should be described, e.g. in the CONSORT-flowchart [16].

Were the Groups Similar at the Start of the Trial?

If the sample size is sufficiently large and the randomization method is proper and robust, then the make-up of the treatment and control groups should be quite similar, selection bias should be prevented, and the only difference should be intervention versus control treatment [14]. Still, random treatment assignment does not guarantee that the groups are equivalent at baseline. Any differences in baseline characteristics are, however, the result of chance rather than bias [17]. Important demographic and clinical characteristics of the study groups at baseline (start of the study) are frequently summarized in a

Table 1. Results from a hypothetical study with two groups and dichotomous outcome (complete tumor remission ‘yes’ or ‘no’)

Treatment group	Success proportion	Estimations (95% CI)	
		n = 100 per group	n = 500 per group
<i>Absolute difference</i>			
Standard	0.2		
New	0.3	0.1 (−0.019; 0.219)	0.1 (0.047; 0.153)
<i>Relative risk (chance)</i>			
Standard	0.2		
New	0.3	1.5 (0.916; 2.456)	1.5 (1.203; 1.870)

The treatment effects are represented by absolute risk (chance) difference and relative risk (chance) and their corresponding 95% CI. In the first case, a sample size of 100 patients per group is assumed; in the second case, the results for 500 patients per group are shown.

table. This gives additional information on the comparability of the groups. Despite many warnings of their inappropriateness [16–19], significance tests of baseline differences between groups are still common in an RCT. Even for a statistically non-significant comparison group, differences can be clinically relevant, especially if the sample size is small. Thus, the description and critical discussion of the baseline characteristics for each group is usually preferable to statistical tests.

In some studies, imbalances in participant characteristics (prognostic variables) are adjusted for by using some form of multiple regression analysis [16, 20]. In RCTs, the decision to adjust should not be determined by whether baseline differences are statistically significant [21]. Ideally, adjusted analyses should already be specified in the study protocol.

Aside from the Experimental Intervention, Were the Groups Treated Equally?

It is important that the groups are treated completely equally except for the different treatment under investigation. This can be guaranteed if contacts of health workers and study personnel with the patients are blind with respect to the treatment. Treatment applications and follow-up schedules should be identical.

After having considered these five points, one has to decide whether the results of the study can be trusted or whether it is likely that biases may have invalidated the findings of the study. Thus, the next question is: *Is it worth continuing?* The final assessment of validity is never a clear ‘yes’ or ‘no’ decision and has to remain subjective, at least to some extent [12].

What Are the Results?

The reliability of the results can be assessed by posing the following two questions [13]:

How Large Was the Treatment Effect?

The outcome measured in a study can differ with respect to the measuring scale. Often, a dichotomous or binary outcome

is chosen, a so-called ‘yes’ or ‘no’ outcome, e.g. remission of the tumor after preoperative chemotherapy or no remission. The risk difference between two groups can be expressed as absolute risk difference, as relative risk, or as odds ratio. In the example of tumor remission, one is not interested in the risk but rather in the chances of an increased remission rate. The underlying statistical concept is the same for assessing risks or chances. For example, if the proportion of a complete remission is 30% (0.3) under the new treatment and 20% (0.2) under standard treatment, then group difference expressed as absolute risk (chance) difference yields 0.1 (= 0.3 – 0.2): The new treatment increases the chance for a complete remission by 10 percentage points (table 1). Equality of both groups would result in a value of zero. On the other hand, the group differences expressed as relative risk (chance) give 1.5 (= 0.3/0.2): The new treatment increases the healing probability by 50%. For the relative risk, a value of 1 would mean no treatment difference. The choice of the appropriate measure (here absolute risk difference or relative risk) usually depends on the interpretation of the effect and should be fixed during the planning phase of a trial.

If the outcome measure is metric (e.g. bone density after 1 year) and normally distributed, then the difference between groups is described by the difference of the means.

In breast cancer studies, the primary outcome is often ‘time to event’, e.g. time to death or time to disease recurrence. Differences between groups in the time to event, so-called survival curves, are often described by the survival probability at a specific time, e.g. 5-year survival. However, treatment differences are quantified in most cases by a hazard ratio, e.g. the risk for an event is 1.5 times higher under standard treatment than under experimental treatment. A hazard ratio of 1 expresses no group difference. The estimation of a single hazard ratio assumes that this hazard ratio is equal over the whole time period under observation. For example, the hazard ratio is the same at, let’s say, 1 year but also at 3 and 5 years. This is an assumption that has to be checked before quantifying the treatment difference with a single number.

How Precise Was the Estimate of the Treatment Effect?

The true population difference between treatment groups can never be known; all we have is the estimate provided by a rigorous controlled trial, and the best estimate of the true treatment effect is that observed in the trial [13]. However, the estimate of the treatment difference is a point estimate and it is unlikely that it exactly agrees with the true treatment difference. But it is likely that the area around the estimate covers the true population effect with high probability. This area can be estimated by a confidence interval (CI). In medical research, the 95% CI is frequently given in publications, although any other percentile (e.g. 90 or 99%) could be used as well. There is a close connection between p-values and CIs. For example, if a 95% CI ($100-\alpha\%$) for the absolute risk does not include zero (= no difference), then the corresponding p-value is smaller than the significance level of 5% ($\alpha\%$). If the CI includes zero, then the p-value is larger than 5%. For the 95% CI of the relative risk, one has to look if it includes 1, which then indicates no difference. In some cases, the estimates of the absolute difference and of the relative risk can differ with respect to their main message of a treatment difference. However the type of estimate has to be determined in advance and stated in the study protocol. It is not allowed to choose the estimator depending on the data and the outcome. The width of the CI depends on the sample size of the trial – the higher the number of observations in the groups, the more information is available for the estimation, the more precise will be the estimate and the smaller its CI. For example, let us consider the example of the estimated remission rates of 30 and 20% for the new and standard treatment, respectively. If the number of patients is 100 in each group, then the 95% CI for the absolute risk (chance) difference of 0.1 is (–0.019; 0.219), and for the relative risk (chance) of 1.5, the 95% CI is (0.916; 2.456) (table 1). For 500 patients per group, the absolute risk difference and relative risk estimates would be more precise and the 95% CIs would shorten to (0.047; 0.153) and (1.203; 1.870), respectively. The larger the sample size, the narrower is the CI; but *when is the sample size big enough?*

Usually, a proper sample size calculation has to be made before starting a trial and the underlying assumptions should be described in the paper. The significance level in a medical research setting is usually 5% and should be two-sided (in almost all cases). The power is the probability to detect a group difference of a specific size if it really exists and is usually chosen to be 80–90%. The group difference used for sample size calculation should be the smallest clinically relevant difference worth to be detected to change the clinical routine so that the new treatment becomes standard. If no details on the sample size calculation are given, one can look at the width of the CI; it is directly related to the sample size.

The CI also helps to interpret negative studies in which the authors have found that the experimental treatment is no better than the control therapy. If the upper boundary of the CI is a value that is still clinically important, e.g. 2.456 for the rela-

tive risk (chance) 95% CI for 100 patients per group, then the study has failed to exclude an important treatment effect and the number of patients included is too small. Such studies fail to prove treatment differences, but also fail to prove that there is no difference.

It is also worth noting that, if a published study with extremely small sample size shows a significant difference, then there will be a non-negligible chance that the significantly better treatment is in reality worse [22]. This is the most severe error in statistical testing and is called type III error.

When size and precision of the treatment effect of the published study is assessed, one can raise the final question whether the published results can also be applied to a specific patient.

Will the Results Help Me in Caring for My Patients?

This last step in a critical appraisal process is called external validity or applicability and can be assessed by answering the following three questions [13]:

Can the Results Be Applied to My Patient Care?

It is a natural question whether a specific patient should be treated with the significantly better experimental treatment from a recent publication. If the patient at hand meets all inclusion and exclusion criteria of the study, then the published results should be applicable. Otherwise, the patient would not have been eligible for the study and judgment is required [13] whether the study results may still be applicable. Assume that the in- and exclusion criteria are nearly met, e.g. the patient may be 2 years too old to be included in the study, then it may still be reasonable to generalize the study results to this patient. Another approach is to ask whether there is some compelling reason why the results should not be applied to the patient.

Reports about subgroup analyses: Special care has to be given to findings about subgroups. Often the treatment effects are also shown for subgroups, which may help to find patient groups with different behavior and treat them in the best possible way. Sometimes it appears that subgroups of patients benefit from treatment and others do not. In statistical terms, this is called an interaction between the therapy and the variable defining the subgroups. In such a situation an explanation has to be found as to whether there is some powerful biological reason behind the different effects in the subgroups.

Because of the high risk for spurious findings, subgroup analyses are often discouraged. Post hoc subgroup comparisons (analyses done after looking at the data) are especially likely not to be confirmed by further studies, so that such analyses do not have great credibility [16, 20, 23]. On the one hand, the probability increases that effects in subgroups are found by chance alone (multiplicity problem) so that false-positive results may be easily found. On the other hand, the number of

observations decreases in the subgroups so that false-negative results may result due to decreased power (less information available). Despite that, there is still a strong temptation to make subgroup analyses in order not to waste potential information and to come to preliminary conclusions/information. Oxman and Guyatt [23] give some recommendations for the decision whether an effect in a subgroup may be real or not:

- Is the magnitude difference clinically important?
- How likely is it that the observed effect is just due to chance, where the number of investigated subgroups and the size of the p-value have to be considered?
- Did the hypothesis precede rather than follow the analysis (e.g. is it already described in the study protocol)?
- Was the subgroup analysis one of a small number of hypotheses tested?
- Was the difference suggested by comparisons within rather than between studies? Comparisons of effect sizes between studies entails a high risk compared with inferences made on the basis of within-study differences. Patients may be randomized to the treatments, but they are not randomized to one study or another and results can thus be biased.
- Was the difference consistent across studies?
- Is there indirect evidence that supports the hypothesized difference? Indirect evidence may include information from animal studies or analogous situations in human biology.

Were All Clinically Important Outcomes Considered?

All important outcomes should be addressed by the study, e.g. a report on favorable effects of treatment on one outcome may be judged differently by a harmful effect on other outcomes. For breast cancer patients, a prolongation of disease-free survival may be differently judged if the treatment has severe side effects or if the quality of life is heavily affected by the new treatment.

Furthermore, improvement of the outcome should be beneficial to the patients, e.g. a considerable remission of the tumor from preoperative chemotherapy should also have an effect on disease-free or overall survival or at least on an increase in the quality of life by a smaller proportion of breast ablations.

Are the Likely Treatment Benefits Worth the Potential Harm and Costs?

If the study results can be generalized to patients and its outcomes are important, the next question is whether the probable treatment benefits are worth the effort [13]. Let us consider again the example of preoperative chemotherapy. The proportions of complete remission are 30% (0.3) under the new

treatment and 20% (0.2) under standard treatment, resulting in a relative risk (chance) of 1.5, which sounds quite impressive. The ‘number needed to treat’ (NNT) is the number of patients to be treated with the new therapy in order to prevent in expectation one negative event or, in our example, to achieve one additional complete remission compared to standard therapy. With complete remission proportions of 0.3 and 0.2, the NNT is 10 ($= 1/(0.3 - 0.2)$). So, 10 patients have to be treated with the new therapy to have in expectation one complete remission more than if these 10 patients were treated by standard therapy. If the proportions for a complete remission had been 0.03 and 0.02, respectively, the resulting relative risk would also have been 1.5, but the NNT would now be 100 ($= 1/(0.03 - 0.02)$). This has to be kept in mind when balancing the benefits and the harms (side effects, reduced quality of life, etc.) of the new treatment.

Conclusions

When conducting a randomized clinical trial, special attention has to be given to the design of the study to avoid any bias from the beginning [24]. Statistics and its proper application is an essential part of well-conducted trials, from study planning to study analyses and interpretation [25]. The best statistical analysis strategy cannot outweigh badly planned and conducted studies.

Nowadays, the CONSORT statement [16] gives general guidelines on how to prepare manuscripts on RCTs for publication so that the most relevant information is included and the reader is able to understand what has been going on. Medical journals have improved their peer review process in the last years and nowadays a medical-statistical review is nearly always mandatory for high-quality journals. Still, it cannot be excluded that flaws may be published undetected and it is still up to the readers to assess publications for themselves.

From the homepage of the Centre for Evidence-Based Medicine [3] checklists for the critical appraisal of publications can be downloaded. They also offer a software tool (CATmaker) that helps creating Critically Appraised Topics (CAT) about therapy, diagnosis, prognosis, etiology/harm and systematic reviews of therapy.

Acknowledgement

The author wishes to thank Harald Heinzl for his valuable comments and suggestions.

References

- 1 Crombie IK: The pocket guide to critical appraisal. London, BMJ Publishing, 2005.
- 2 Williams R: Linking research to practice; in Williams R, Baker LM, Marshall JG (eds): Information searching in healthcare. New York, Slack, 1992.
- 3 Oxford Centre for Evidence-Based Medicine: Levels of Evidence; downloaded from <http://www.cebm.net>; accessed in September 2008.
- 4 Pritchard KI: The Early Breast Cancer Trialists Collaborative Group (EBCTCG) Process: Is it still relevant in 2006? *Breast Care* 2006;1:338–339.
- 5 Senn H-J: Significance and usefulness of guidelines. *Breast Care* 2006;1:220–222.
- 6 Chmura-Kraemer H, Kraemer-Lowe K, Kupfer DJ: To Your Health: How to Understand What Research Tells Us About Risk. Oxford, Oxford University Press, 2005.
- 7 Ajetunmbola O: Making Sense of Critical Appraisal. London, Hodder Arnold, 2002.
- 8 Bucher HC, Guyatt GH, Cook DJ, Holbrook A, McAlister FA, for the Evidence-Based Medicine Working Group: Users' guides to the medical literature: XIX. Applying clinical trial results. A. How to use an article measuring the effect of an intervention on surrogate end points. *JAMA* 1999;282:771–778.
- 9 Elwood JM: Critical appraisal of epidemiological studies and clinical trials. Oxford, Oxford University Press, 1998.
- 10 Greenhalgh T: How to read a paper. London, BMJ Publishing, 2004.
- 11 Straus SE, Richardson W, Glasziou P, Haynes RB: Evidence-Based Medicine: How to Practice and Teach EBM. Edinburgh, Churchill Livingstone, 2005.
- 12 Guyatt GH, Sackett DL, Cook DJ: Users' guides to the medical literature. II: How to use an article about therapy or prevention. A. Are the results of the study valid? *JAMA* 1993;270:2598–2601.
- 13 Guyatt GH, Sackett DL, Cook DJ: Users' guides to the medical literature. II: How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? *JAMA* 1994;271:59–63.
- 14 Hill A, Spittlehouse C: What is critical appraisal? What is ...? *Series* 2001;3:1–8.
- 15 Randomizer; www.meduniwien.ac.at/msi/biometrie/randomizer.htm; accessed September 2008.
- 16 The CONSORT Group: Consolidated standards of reporting trials; <http://www.consort-statement.org>; accessed in September 2008.
- 17 Altman DG, Doré CJ: Randomisation and baseline comparisons in clinical trials. *Lancet* 1990;335: 149–153.
- 18 Senn S: Base logic: tests of baseline balance in randomized clinical trials. *Clin Res Reg Affairs* 1995;12:171–182.
- 19 Roberts C, Torgerson DJ: Baseline imbalance in randomised controlled trials. *BMJ* 1999;319:185.
- 20 Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gøtzsche PC, Lang T, for the CONSORT Group: The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;134:663–694.
- 21 Assmann SF, Pocock SJ, Enos LE, Kasten LE: Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000;355:1064–1069.
- 22 Heinzl H, Benner A, Ittrich C, Mittlböck M: Proposals for sample size calculation programs. *Methods Inf Med* 2007;46:655–661.
- 23 Oxman AD, Guyatt GH: A consumer's guide to subgroup analysis. *Ann Intern Med* 1992;116:78–84.
- 24 Fletcher RH, Fletcher SW: *Klinische Epidemiologie: Grundlagen und Anwendung*. Bern, Verlag Hans Huber, 2007.
- 25 Draxler W, Mittlböck M: Basic principles in the planning of clinical trials in surgical oncology. *Eur Surgery* 2006;38:27–32.