REVIEW ARTICLE

Randomized Controlled Trials

Part 17 of a Series on Evaluation of Scientific Publications

Maria Kabisch, Christian Ruckes, Monika Seibert-Grafe, Maria Blettner

SUMMARY

Background: In clinical research, randomized controlled trials (RCTs) are the best way to study the safety and efficacy of new treatments. RCTs are used to answer patient-related questions and are required by governmental regulatory bodies as the basis for approval decisions.

Methods: To help readers understand and evaluate RCTs, we discuss the methods and qualitative requirements of RCTs with reference to the literature and an illustrative case study. The discussion here corresponds to expositions of the subject that can be found in many textbooks but also reflects the authors' personal experience in planning, conducting and analyzing RCTs.

Results: The quality of an RCT depends on an appropriate study question and study design, the prevention of systematic errors, and the use of proper analytical techniques. All of these aspects must be attended to in the planning, conductance, analysis, and reporting of RCTs. RCTs must also meet ethical and legal requirements.

Conclusion: RCTs cannot yield reliable data unless they are planned, conducted, analyzed, and reported in ways that are methodologically sound and appropriate to the question being asked. The quality of any RCT must be critically evaluated before its relevance to patient care can be considered.

► **Cite this as:**
Kabisch M, Ruckes C, Seibert-Grafe M, Blettner M: Randomized controlled trials: part 17 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2011; 108(39): 663–8. DOI: 10.3238/arztebl.2011.0663

Interdisziplinäres Zentrum Klinische Studien (IZKS), Universitätsmedizin der Johannes-Gutenberg-Universität Mainz: Dipl.-Biomath. Kabisch, Dipl.-Math. Ruckes, Dr. med. Seibert-Grafe

Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI), Universitätsmedizin der Johannes-Gutenberg-Universität Mainz: Prof. Dr. rer. nat. Blettner

Clinical research lays the groundwork for progress in medicine and is an indispensable prerequisite for evidence-based medicine. Randomized controlled clinical trials (RCTs) are the gold standard for ascertaining the efficacy and safety of a treatment. RCTs can demonstrate the superiority of a new treatment over an existing standard treatment or a placebo. In clinical research RCTs are used to answer patient-related questions, and in the development of new drugs they form the basis for regulatory authorities' decisions on approval. Alongside meta-analyses, high-quality RCTs with a low risk of systematic error (bias) provide the highest level of evidence (1, 2).

The aim of this article is to provide an introduction into the methods and quality requirements of RCTs in order to help the reader understand and evaluate publications that present the results of such studies. Since RCTs are by definition interventional, often investigating drugs or medical devices, ethical and legal aspects will also be discussed.

The discussion here corresponds to expositions of the subject in numerous textbooks (3–5) but also reflects the authors' own experience of planning, conducting and analyzing RCTs. To aid understanding, some methodological issues are illustrated by reference to a published trial, the ALIFE study (Anticoagulants for LIving FEtuses). The fundamental principles of methodology and statistical analysis for all studies, including RCTs, have been expounded in earlier articles in this journal's series on evaluation of scientific publications (6–11).

The results of the ALIFE study were published in the New England Journal of Medicine in April 2010 (12) and presented in the "Studies in Focus" series of the German-language edition of *Deutsches Ärzteblatt* in July 2010 (13). In this study, women who had had two or more miscarriages were assigned randomly to one of three treatment groups: aspirin plus heparin, aspirin alone, or placebo. The primary objective of the study was to investigate the efficacy of the different treatments as shown by the rate of live births.

Objectives

The basis of every RCT is the study protocol that describes the medical/scientific background, the risk:benefit assessment, the study design, the study methods, and the overall planning, conduct and

analysis (14). The primary study question, i.e., the primary objective, results from the medical/scientific rationale for the study.

To answer the primary study question, a primary endpoint is required. This is a parameter measured or observed that is recorded at a defined time and can be assumed to reflect the effect of a treatment. The endpoint may be clinical, e.g., the live birth rate in the ALIFE study.

In a confirmatory study hypotheses are formulated *a priori* according to the primary study question. If the primary objective of the trial is to demonstrate the superiority of a new treatment over an existing treatment or placebo, then the initial assumption (null hypothesis) is that the two treatments do not differ in efficacy. Based on statistical analysis the null hypothesis can be retained or must be rejected in favor of the alternative hypothesis. The alternative hypothesis is assumed when a statistically significant difference is ascertained between the two treatments. (A detailed description of methods for statistical evaluation is given in an earlier article in this series [15].)

The primary study question is accompanied by one or more ancillary study questions, i.e., secondary objectives. The secondary endpoints investigate other effects of the treatment, e.g., the occurrence of adverse events or the influence on biomarkers. In the ALIFE study, the secondary endpoints included the rate of miscarriage, the premature birth rate, and the rate of maternal thrombopenia.

From the statistical viewpoint it is vital to distinguish between the primary and secondary study questions, because the number of study subjects depends solely on the primary endpoint (16). Study planning includes calculation of the number of subjects necessary for detection by statistical analysis of a minimally relevant difference in efficacy, from the clinical viewpoint, between the treatments. The number of patients is therefore crucial for the statistical power of a study. (Sample size calculation is described in detail in a previous article in this series [17].)

In the ALIFE study a difference of 15% in live birth rate was assumed between the combination of aspirin plus heparin and aspirin alone or placebo. In order to demonstrate the postulated positive effect of the combination therapy, 364 women were enrolled in the trial.

Study design

In trials with randomized and controlled design (e.g., a two-armed study with parallel groups), the effects of the study treatment (intervention) are compared with those of a control treatment and the patients are randomly assigned to the two groups. The patients in the control group receive either another treatment or a placebo. The ALIFE trial is a three-armed parallel group study to establish whether the combination treatment or the monotherapy improve the live birth rate compared with placebo. The use of placebos in clinical trials is ethically justified provided that no standard treatment is available. If comparison with placebo is indispensable for methodological reasons, it can be justified as long as patients will not be harmed (18). That is the case, for example, if the study is of only short duration or if the severity of disease permits postponement or interruption of treatment.

As in any study of human subjects, the study population of an RCT must be clearly defined. Precise inclusion and exclusion criteria are elaborated to ensure that only eligible patients are recruited. The study participants must be homogeneous with regard to demographic characteristics, disease state, and possibly even comorbidity and comedication.

To ensure "fair" comparison between the treatments, the different study groups must be truly comparable. This can be achieved by standardization of, for example, the time(s) of intake of the study medication and the methods used to measure clinical parameters, but most important for comparability is randomization of the participants.

Randomization

In RCTs the patients are randomly assigned to the different study groups. This is intended to ensure that all potential confounding factors are divided equally among the groups that will later be compared (structural equivalence). These factors are characteristics that may affect the patients' response to treatment, e.g., weight, age, and sex. Only if the groups are structurally equivalent can any differences in the results be attributed to a treatment effect rather than the influence of confounders. If the confounders are known, structural equivalence of the patient groups can be attained by stratified randomization *(Box)*.

In the ALIFE study the patients were assigned to the three treatment groups with a randomization ratio of 1:1:1. They were randomized taking account of the prognostic factors of age (<36 years or ≥36 years) and number of miscarriages (2 or ≥3), and because the study was multicentric they were stratified by study center. If patients were allocated to treatment groups by conscious or unconscious selection for prognosis-related characteristics, rather than randomly, this could lead to biased treatment comparison and distorted results (selection bias).

The assignment to study groups must not be in any way predictable. Predictability of group allocation is avoided by ensuring the study staff are unaware to which treatment the next patient will be allotted. Alternating assignment to the different treatments is not truly random.

Blinding

Bias is avoided not only by randomization but also by blinding. A study may be double blind, single blind, or open.

In a double-blind study neither patient nor study physician knows to which treatment the patient has been assigned. Double-blind studies are advantageous if knowledge of the treatment might influence the course and therefore the results of the study. Thus it is

particularly important that the study physician is blinded to treatment if the endpoints are subjective. Blinding of patients to their treatment is important, for example, if their attitude could potentially affect their reliability in taking the test medication (compliance) or even their response to treatment.

If only one party, either patient or study physician, is blinded to the treatment, the study is called single blind; a study with no blinding is described as open. The highest possible degree of blinding should be chosen to minimize bias.

Analysis population

The data subjected to statistical analysis in RCTs are those gathered from patient populations defined in the study protocol. The primary population for analysis is the so-called intention-to-treat (ITT) population, comprising all randomized patients. In analysis according to the ITT principle, patients are allocated to the group to which they were randomized, thus retaining the advantages of randomization such as structural equivalence. Because the ITT population includes all patients who were randomized, the data for analysis include some patients whose treatment was interrupted, prematurely discontinued, or did not take place at all. The analysis strategy for ITT data is therefore conservative, i.e., the treatment effect tends to be underestimated (19), regardless of whether the primary endpoint represents an improvement or a deterioration. Many studies define a modified ITT (mITT) population, which may for example comprise the patients who received at least a defined amount of study treatment.

An alternative strategy is to restrict analysis to the data from the per-protocol (PP) population. Patients in whom study conduct deviated from the protocol are excluded from analysis. These so-called protocol violations include, for example, failure concerning the application of inclusion or exclusion criteria and incorrect administration of the study treatment. In analysis according to the PP principle, patients are allocated to the treatment groups depending on the treatment they actually received. Because the PP population includes only those patients who completed the study according to the protocol, the results may be distorted in favor of the investigational intervention (19).

To assess the robustness of the study findings, PP evaluation is carried out as a sensitivity analysis if the ITT population is the patient population for the primary efficacy analysis (16). If the results of PP and ITT evaluation of the primary endpoint are very similar, they can be regarded as reliable. Should this not be the case, the possible reasons for the discrepancy between the results of the ITT and PP analyses must be discussed in the results section of the publication.

The data of the ALIFE study, particularly the primary endpoint, were statistically evaluated on the basis of the ITT population. The rates of live births in the three treatment groups did not differ significantly *(Table 1)*. Analysis according to the PP principle confirmed this finding. Neither aspirin and heparin combined nor aspirin alone were demonstrated to have a greater effect than placebo on the live birth rate.

Quality standards and legal requirements in Germany

Clinical trials have to be performed according to national and international regulations. The Declaration of Helsinki, first formulated by the World Medical Association in 1964 and revised several times in the intervening years (20), lays down fundamental ethical principles for research on human beings. Trials investigating drugs and medical devices have to comply with the relevant German laws for drugs—the German Medicines Act (AMG; for German text see Bundesgesetzblatt I p. 2262)—and the GCP regulation (GCP-Verordnung [21]), and for devices the Medical Devices Act (MPG; for German text see Bundesgesetzblatt I p. 983), revised in March 2010. The GCP regulation, which came into force in 2004, made adherence to good clinical practice (GCP) a legal requirement in Germany (21). GCP Guideline ICH-E6 of 1997 forms the basis for European Directives 2001/20/EG and 2005/28/EG, on which in turn the GCP regulation is based (14). The aim of GCP is to protect study participants and ensure high quality of study data.

In 2004 the International Committee of Medical Journal Editors made registration of a clinical trial in a public registry a precondition for its publication (22). The professional code of conduct for physicians in Germany demands that every study in human subjects be submitted to the responsible ethics committee for approval. Drug trials and most studies of medical

BOX

Stratified randomization

If the stratification factors sex (male, female) and age (<18 years, ≥18 years) are to be considered and 150 patients are to be randomized in a ratio of 1:1 into the active treatment and placebo groups (2×75 patients), then randomization has to be performed for each separate subgroup (stratum). Two stratification factors, each with two values, yield four strata (male and <18 years, male and ≥18 years, female and <18 years, female and ≥18 years).

	Active treatment	Placebo
Male and <18 years	10	10
Male and ≥18 years	16	17
Female and <18 years	24	23
Female and ≥18 years	25	25
Total	75	75

TABLE1

Results of the ALIFE study (adapted from [12])

	Aspirin plus Heparin	Aspirin alone	Placebo	p-value
Intention-to-treat population n	123	120	121	
Live births n (%)	67 (54.5)	61 (50.8)	69 (57.0)	0.63
Relative risk (95% CI)	0.96 (0,76–1,19)	0.89 (0.71–1.13)	1.00	
Absolute difference in live birth rates (95% CI) %	-2.6 (-15.0–9.9)	-6.2 (-18.8–6.4)		

Relative risk and absolute difference were calculated for the comparisons between aspirin plus heparin and placebo and between aspirin alone and placebo. The p-value applies to all treatment group comparisons.
95% CI, 95% confidence interval

TABLE 2

Minimal requirements for a publication reporting a randomized controlled trial (adapted from [23])

Study design	Description of study design (e.g., parallel group comparison)
Study population	Specification of inclusion and exclusion criteria for patients
Treatments	Detailed account of treatments and their application in each intervention group and control group
Objectives	Precise formulation of primary and secondary objectives/study questions
Endpoints	Clear definition of primary and secondary endpoints
Sample size	Description of how the required number of study participants was determined
Randomization	Description of type of randomization of patients to treatment groups (e.g., stratified randomization)
Blinding	Specification of degree of blinding (e.g., double blind)
Analysis population	Number of patients analyzed in each treatment group and definition of population for analysis (e.g., ITT)
Results	Presentation of the results for all primary and secondary endpoints for each treatment group
Adverse events	Details of all major adverse events for each treatment group
Interpretation	Interpretation of the results, taking into account the study question, possible causes of bias, the current state of knowledge, and other researchers' publications on the same topic
Generalizability	Discussion of the applicability of the study results to general patient care

devices require not only approval from the local ethics committee but also from regulatory bodies at the federal level (Federal Institute for Drugs and Medical Devices [BfArM] or Federal Institute for Vaccines and Biomedicines, Paul-Ehrlich-Institut [PEI]). The applications have to be accompanied by the study protocol, the information to be supplied to the patients, the consent form for participation, and confirmation that adequate insurance has been arranged.

Trials of drugs and medical devices also have to be registered with state authorities. There are legally defined obligations to report suspected unexpected serious adverse reactions or early termination of a study, and the final study report must also be submitted. The Federal Data Protection Act (BDSG; for German text see Bundesgesetzblatt I p. 2814) and the AMG obligate researchers to pseudonymize all person-related data that are gathered, documented, stored, and analyzed in the course of a clinical trial. In other words, information revealing the identity of a patient (name or initials) must be replaced by a code. Only patients who have agreed in advance to the recording, storage, processing and dissemination of their data may participate in a clinical study.

Discussion

Any publication of an RCT must lucidly describe the planning, conduct, and analysis of the study. The CONSORT statement provides a minimum set of recommendations for reporting RCTs (23). The most important aspects that have to be described in the publication are listed in *Table 2*. The progress of patients through an RCT and the numbers of patients whose data were analyzed can be depicted in a flow diagram *(Figure)*.

The study results and their interpretation must be discussed in detail in the study report and any subsequent publication, and the limitations of the methods used should be described, all with reference to the study design, the recent literature, and the current state of knowledge. Critical discussion plays a decisive part in clinical evaluation of the results. In the publication of the ALIFE study, the findings were compared with those of other RCTs investigating the effects of heparin on reduction of miscarriages and inconsistencies were discussed. Ultimately, the available study data did not justify the recommendation of anticoagulants for women with recurring miscarriages.

Although RCTs are the gold standard with regard to level of evidence, their generalizability, i.e., the extent to which their results can be extrapolated to the wider patient population (external validity) is often questioned, because standardized and controlled study conditions do not adequately reflect clinical reality. Moreover, the patients selected for a study are not necessarily representative, in that those seen in routine daily practice will often have numerous comorbidities and comedications. After marketing approval of a new treatment, phase-IV studies are carried out to establish its efficacy and safety in a larger and more heterogeneous population; as a rule these studies are RCTs.

Epidemiological studies, e.g., cohort studies, are particularly suitable for detection of infrequent adverse effects.

Conclusion

RCTs are the best type of study for determining whether there is a causal relationship between intervention and effect (24). Recent discussions in the scientific community and the new Law on the Reorganization of the Pharmaceutical Market (AMNOG; for German text see Bundesgesetzblatt I p. 2262), which regulates the use of drugs and medical devices, clearly show that RCTs are still the standard for demonstrating efficacy and safety so that a new treatment can be approved for use in patients. However, it seems clear that post-marketing studies comparing new and established treatments are still required.

The IZKS Mainz is supported by the grant "Clinical Trial Centers [Klinische Studienzentren], no. FK 01KN1103, IZKS Mainz" from the Federal Ministry of Education and Research.

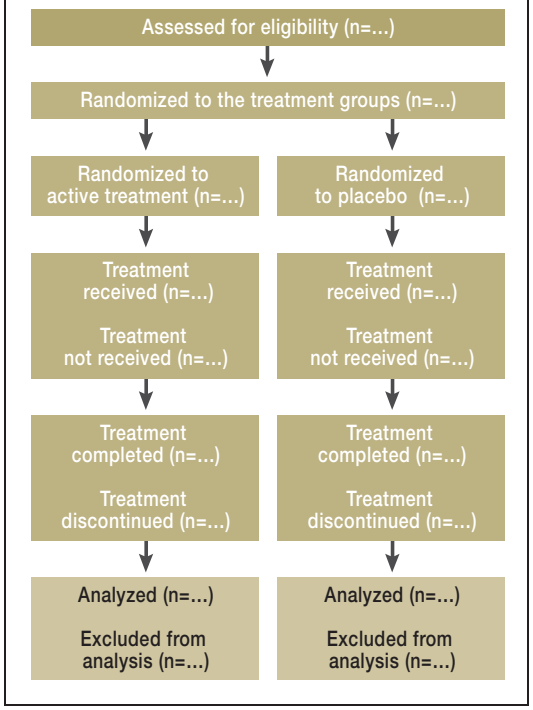
Acknowledgment
The authors are grateful to Daniel Wachtlin of the Interdisciplinary Center for Clinical Trials (IZKS) Mainz for helpful discussions.

Conflict of interest statement
The authors declare that no conflict of interest exists.

Manuscript received on 23 February 2011, revised version accepted on 28 June 2011.

Translated from the original German by David Roseveare.

FIGURE



Patient flow in a randomized controlled trial (adapted from [23])

REFERENCES

1. Harbour R, Miller J: A new system for grading recommendations in evidence based guidelines. BMJ 2001; 323: 334–6.
2. Phillips B, Ball C, Sackett D, Badenoch D, Straus S, Haynes B, Dawes M (2001): Oxford Centre for Evidence-based Medicine Levels of Evidence. www.cebm.net/levels_of_evidence.asp
3. Meinert CL: Clinical Trials: Design, conduct, and analysis. Oxford University Press: New York 1986.
4. Shein-Chung C, Jen-Pei L: Design and analysis of clinical trials: concepts and methodologies. John Wiley & Sons: New Jersey 2004.
5. Schumacher M, Schulgen G: Methodik klinischer Studien: Methodische Grundlagen der Planung, Durchführung und Auswertung. Berlin, Heidelberg: Springer-Verlag 2008.
6. Röhrig B, du Prel, Blettner M: Study Design in medical research: Part 2 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009; 106(11): 184–9
7. Röhrig B, du Prel du J, Wachtlin D, Blettner M: Types of study in medical research: part 3 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009; 106(15): 262–8.
8. Spriestersbach A, Röhrig B, du Prel J, Gerhold-Ay A, Blettner M: Descriptive statistics: the specification of statistical measures and their presentation in tables and graphs: part 7 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009; 106(36): 578–83.
9. Sauerbrei W, Blettner M: Interpreting results in 2×2 tables: part 9 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009; 106(48): 795–800.
10. Victor A, Elsäßer A, Hommel G, Blettner M: Judging a plethora of p-values: how to contend with the problem of multiple testing: part 10 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010; 107(4): 50–6.
11. Zwiener, I, Blettner M, Hommel G: Survival analysis: part 15 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2011; 108(10): 163–9.
12. Kaandorp SP, et al.: Aspirin plus heparin or aspirin alone in women with recurrent miscarriage. NEJM 2010; 362: 1586–96.
13. Heinzl S: Können ASS und Heparin wiederholte Fehlgeburten verhindern? Dtsch Arztebl 2010; 107(27): A-1355.
14. ICH E6: Guideline for good clinical practice. London UK: International Conference on Harmonization 1996; CPMP/ICH/135/95.
15. du Prel J, Röhrig B, Hommel G, Blettner M: Choosing statistical tests: part 12 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010; 107(19): 343–8.
16. ICH E9: Statistical Principles for Clinical Trials. London UK: International Conference on Harmonization 1998; CPMP/ICH/363/96.
17. Röhrig B, du Prel JB, Wachtlin D, Kwiecien R, Blettner M: Sample size calculation in clinical trials: part 13 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010; 107(31–32): 552–6.
18. ICH E10: Choice of Control Group and Related Issues in Clinical Trials. London UK: International Conference on Harmonization 2000; CPMP/ICH/364/96.
19. Ellenberg JH: Intention-to Treat Analysis. In: Redmond C, Colton T (eds): Biostatistics in Clinical Trials. New Jersey: John Wiley & Sons: 2001.
20. World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. 2008. www.wma.net/en/30publications/10policies/b3/17c.pdf.
21. Verordnung über die Anwendung der Guten Klinischen Praxis bei der Durchführung von klinischen Prüfungen mit Arzneimitteln zur Anwendung am Menschen (GCP-Verordnung – GCP-V). GCP-Verordnung vom 9. August 2004 (BGBl. I S. 2081), die zuletzt durch Artikel 4 der Verordnung vom 3. November 2006 (BGBl. I S. 2523) geändert worden ist.
22. Clinical trial registration: a statement from the International Committee of Medical Journal Editors (ICMJE). Lancet 2004: 364: 911–2.

23. Schulz KF, Altman DG, Moher D: CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010; 340: c332.

24. Windeler J: Bedeutung randomisierter klinischer Studien mit relevanten Endpunkten für die Nutzenbewertung. In: Diskussionsforum zur Nutzenbewertung im Gesundheitswesen: Begriffsdefinitionen und Einführung. 2007. www.gesundheitsforschung-bmbf.de/_media/DLR_Nutzenbewert_07–11–22_Druckversion.pdf.

Corresponding author
Prof. Dr. rer. nat. Maria Blettner
Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI)
Universitätsmedizin der Johannes-Gutenberg-Universität Mainz
Obere Zahlbacher Str. 69
55131 Mainz, Germany
blettner-sekretariat@imbei.uni-mainz.de

KEY MESSAGES

- In clinical research, randomized controlled trials are the gold standard for demonstrating the efficacy and safety of a new treatment.
- Randomized controlled trials cannot yield robust data unless they are planned, conducted, and analyzed in ways that are methodologically sound and appropriate to the question being asked.
- Methods to avoid bias, such as randomization and blinding, can help to prevent distortion of the study results.
- The robustness of the results is tested by statistical analysis of the data from patient populations defined *a priori*.
- The quality of a randomized controlled trial depends crucially not only on adherence to methodological standards but also on strict compliance with the protocol regarding the clinical conduct of the study.