

REVIEW ARTICLE

Study Design in Medical Research

Part 2 of a Series on the Evaluation of Scientific Publications

Bernd Röhrig, Jean-Baptist du Prel, Maria Blettner

SUMMARY

Background: The scientific value and informativeness of a medical study are determined to a major extent by the study design. Errors in study design cannot be corrected afterwards. Various aspects of study design are discussed in this article.

Methods: Six essential considerations in the planning and evaluation of medical research studies are presented and discussed in the light of selected scientific articles from the international literature as well as the authors' own scientific expertise with regard to study design.

Results: The six main considerations for study design are the question to be answered, the study population, the unit of analysis, the type of study, the measuring technique, and the calculation of sample size.

Conclusions: This article is intended to give the reader guidance in evaluating the design of studies in medical research. This should enable the reader to categorize medical studies better and to assess their scientific quality more accurately.

Dtsch Arztebl Int 2009; 106(11): 184–9
DOI: 10.3238/arztebl.2009.0184

Key words: study design, quality, study, study type, measuring technique

Medical research studies can be split into five phases—planning, performance, documentation, analysis, and publication (1, 2). Aside from financial, organizational, logistical and personnel questions, scientific study design is the most important aspect of study planning. The significance of study design for subsequent quality, the reliability of the conclusions, and the ability to publish a study are often underestimated (1). Long before the volunteers are recruited, the study design has set the points for fulfilling the study objectives. In contrast to errors in the statistical evaluation, errors in design cannot be corrected after the study has been completed. This is why the study design must be laid down carefully before starting and specified in the study protocol.

The term "study design" is not used consistently in the scientific literature. The term is often restricted to the use of a suitable type of study. However, the term can also mean the overall plan for all procedures involved in the study. If a study is properly planned, the factors which distort or bias the result of a test procedure can be minimized (3, 4). We will use the term in a comprehensive sense in the present article. This will deal with the following six aspects of study design: the question to be answered, the study population, the type of study, the unit of analysis, the measuring technique, and the calculation of sample size—, on the basis of selected articles from the international literature and our own expertise. This is intended to help the reader to classify and evaluate the results in publications. Those who plan to perform their own studies must occupy themselves intensively with the issue of study design.

Question to be answered

The question to be answered by the research is of decisive importance for study planning. The research worker must be clear about the objectives. He must think very carefully about the question(s) to be answered by the study. This question must be operationalized, meaning that it must be converted into a measurable and evaluable form. This demands an adequate design and suitable measurement parameters. A distinction must be made between the main questions to be answered and secondary questions. The result of the study should be that open questions are answered

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

Zentrum Präventive Pädiatrie, Zentrum für Kinder- und Jugendmedizin, Johannes Gutenberg-Universität Mainz: Dr. med. du Prel, M.P.H

and possibly that new hypotheses are generated. The following questions are important: Why? Who? What? How? When? Where? How many? The question to be answered also implies the target group and should therefore be very precisely formulated. For example, the question should not be "What is the quality of life?", but must specify the group of patients (e.g. age), the area (e.g. Germany), the disease (e.g. mammary carcinoma), the condition (e.g. tumor stage 3), perhaps also the intervention (e.g. after surgery), and what endpoint (in this case, quality of life) is to be determined with which method (e.g. the EORTC QLQ-C30 questionnaire) at what point in time. Scientific questions are often not only purely descriptive, but also include comparisons, for example, between two groups, or before and after the intervention. For example, it may be interesting to compare the quality of life of breast cancer patients with women of the same age without cancer.

The research worker specifies the question to be answered, and whether the study is to be evaluated in a descriptive, exploratory or confirmatory manner. Whereas in a descriptive study the units of analysis are to be described by the recorded variables (e.g. blood parameters or diagnosis), the aim in an exploratory analysis is to recognize connections between variables, to evaluate these and to formulate new hypotheses. On the other hand, confirmatory analyses are planned to provide statistical proofs by testing specified study hypotheses.

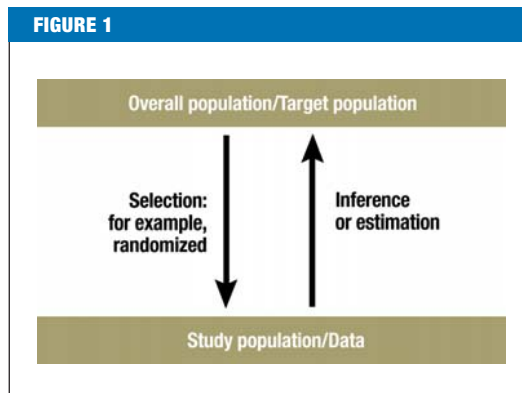
The question to be answered also determines the type and extent of the data to be recorded. This specifies which data are to be recorded at which point in time. In this case, less is often more. Data irrelevant to the question(s) to be answered should not be collected for the moment. If too many variables are recorded at too many time points, this can lead to low participation rates, high dropout rates, and poor compliance from the volunteers. The experience is then that not all data are evaluated.

The question to be answered and the strategy for evaluation must be specified in the study protocol before the study is started.

Study population

The question to be answered by the study implies that there is a target group for whom this is to be clarified. Nevertheless, the research worker is not primarily interested in the observed study population, but in whether the results can be transferred to the target population. Accordingly, statistical test procedures must be used to generalize the results from the sample for the whole population (*figure 1*).

The sample can be highly representative of the study population if it is properly selected. This can be attained with defined and selective inclusion and exclusion criteria, such as sex, age, and tumor stage. Study participants may be selected randomly, for example, by random selection through the residents' registration office, or consecutively, for example,



Connection between overall population and study population/data

all patients in a clinical department in the course of one year.

With a selective sample, a statement can only be made about a population corresponding to these selection criteria. The possibility of generalizing the results may, for example, be greatly influenced by whether the patients come from a specialist practice, a specialized hospital department or from several different practices.

The possibility of generalization may also be influenced by the decision to perform the study at a single institution or site, or at several (multicenter study). The advantages of a multicenter study are that the required number of patients can be reached within a shorter period and that the results can more readily be generalized, as they are from different treatment centers. This raises the external validity.

Type of study

Before the study type is specified, the research worker must be clear about the category of research. There is a distinction in principle between research on primary data and research on secondary data.

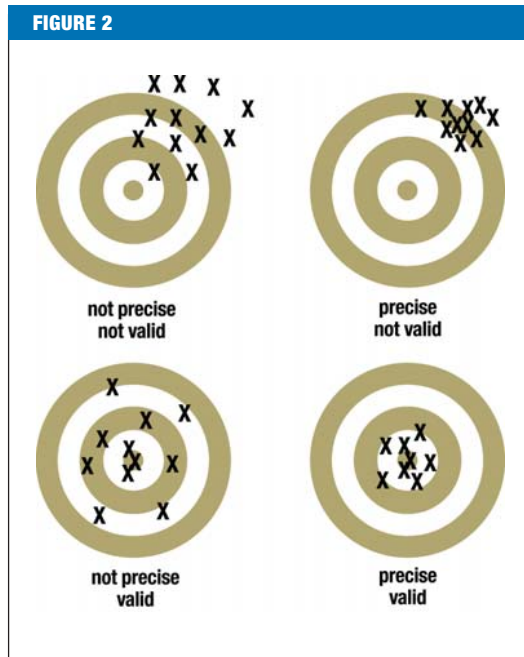
Research on primary data means performing the actual scientific studies, recording the primary study data. This is intended to answer scientific questions and to gain new knowledge.

In contrast, research on secondary results involves the analysis of studies which have already been performed and published. This may include (renewed) analysis of recorded data, perhaps from a register, from population statistics, or from studies. Another objective may be to win a comprehensive overview of the current state of research and to come to appropriate conclusions. In secondary data research, a distinction is made between narrative reviews, systematic reviews, and meta-analyses.

The underlying question to be answered also influences the selection of the type of study. In primary research, experimental, clinical and epidemiological research are distinguished.

Experimental research includes applied studies, such as animal experiments, cell studies, biochemical and physiological investigations, and studies on

Portrayal of the terms reliability (precision) and validity (trueness) using a target



material properties, as well as the development of analytical and biometric procedures.

Clinical research includes interventional and non-interventional studies. The objective of interventional clinical studies (clinical trials) is "to study or demonstrate the clinical or pharmacological activities of drugs" and "to provide convincing evidence of the safety or efficacy of drugs" (AMG, German Drugs Act §4) (5). In clinical studies, patients are randomly assigned to treatment groups. In contrast, noninterventional clinical studies are observational studies, in which patients are given an individually specified treatment (6, 7).

Epidemiological research studies the distribution and changes with time of the frequency of diseases and of their causes. Experimental studies are distinguished from observational studies (7, 8). Interventional studies (such as vaccination, addition of food additives, fluoride addition to drinking water) are of experimental character. Examples of observational epidemiological studies include cohort studies, case control studies, cross-sectional studies, and ecological studies.

A subsequent article will discuss the different study types in detail.

Unit of analysis

The unit of analysis (investigational unit) must be specified before starting a medical study. In a typical clinical study, the patient is the unit of analysis. However, the unit of analysis may also be a technical model, hereditary information, a cell, a cellular structure, an organ, an organ system, a single test individual (animal or man), or specified subgroup or the population

of a region or of a country. In systematic reviews, the unit of analysis is a single study. The sample then includes the total of all units of analysis. The interesting information or data (observations, variables, characteristics) are recorded for the statistical units. For example, if the heart is being investigated in a patient (the unit of analysis), the heart rate may be measured as a characteristic of performance.

The selection of the unit of analysis influences the interpretation of the study results. It is therefore important for statistical reasons to know whether the units of analysis are dependent or independent of each other with respect to the outcome parameter. This distinction is not always easy. For example, if the teeth of test persons are the unit of analysis, it must be clarified whether these are independent with respect to the question to be answered (i.e. from different test persons) or dependent (i.e. from the same test person). Teeth in the mouth of a single test person are generally dependent, as specific factors, such as nutrition and teeth cleaning habits, act on all teeth in the mouth in the same way. On the other hand, extracted teeth are generally independent study objects, as there are no longer any shared factors which influence them. This is particularly the case when the teeth are subject to additional preparation, for example, cutting or grinding. On the other hand, if the observations are on tooth characteristics developed before extraction, these characteristics must be regarded as dependent.

Measuring technique

The term "measuring technique" includes the use of measuring instruments and the method of measurement.

Use of measuring instruments

Measuring instruments include instruments which specifically record measuring data (such as blood pressure or laboratory parameters), as well as data collection with standardized or self-designed questionnaires (for example, quality of life, depression, or satisfaction).

During the validation of a measuring instrument, its quality and practicability are evaluated using statistical parameters. Unfortunately, the nomenclature is not fully standardized and also depends on the special area (for example, chemical analysis, psychological studies with questionnaires, or diagnostic studies). It is always the case that a measuring instrument of high quality should be of high precision and validity.

Precision describes the extent to which a measuring technique consistently provides the same results if the measurement is repeated (9). The reliability (or precision) provides information on the precision or the occurrence of random errors. If the precision is low, the correlation coefficients are low, measurements are imprecise and a larger sample size is needed (9). On the other hand, the validity (accuracy of the mean or trueness) of a measuring instrument is high if it measures exactly what it is supposed to measure. Thus the

validity provides information on the occurrence of systematic errors (10). Whereas the precision describes the difference (variance) between repeated measurements, the validity reflects the difference between the measured and true parameter (10). *Figure 2* portrays the terms, using a target as a model.

Reliability and validity are subsumed in the term accuracy (11, 12). The accuracy is only high when both the precision and the validity are high. *Table 1* summarizes the important terms to validate a measurement method.

The problem is not only that the measurements may be invalid or false, but also that the measurements may lead to erroneous conclusions. External and internal validity can be distinguished (13). External validity means the possibility of generalizing the study results for the study population to the target population. The internal validity is the validity of a result for the actual question to be answered. This can be optimized by detailed planning, defined inclusion and exclusion criteria, and reduction of external interfering factors.

Measurement plan

The measurement plan describes the number and time points of the measurements to be performed. To obtain comparable and objective measurements, the measurement conditions must be standardized. For example, clinical study measurements such as blood pressure must always be performed at the same time, in the same room, in the same position, with the same instrument, and by the same person. If there are differences, for example in the investigator, measuring instrument, analytical laboratory or recording time, it must be established that the measurements are in agreement (10, 13).

The type of scale used for the recorded parameter is also of decisive importance. Putting it simply, metric scales are superior to ordinal scales, which are superior to nominal scales. The type of scale is so important, as both descriptive statistics and statistical test procedures depend on it. Transformation from a higher to a lower scale type is in principle possible, although the converse is impossible. For example, the hemoglobin content may be determined with a metric scale (e.g. as g/dL). It can then be transformed to an ordinal scale (e.g. low, normal and high hemoglobin status), but not conversely.

Calculation of sample size

Whatever the study design, a calculation must be performed before the start of the study to estimate the necessary number of units of analysis (for example, patients) to answer the main study question (14–16). This requires calculation of sample size, exploiting knowledge of the expected effect (for example, the clinically relevant difference) and its scatter (for example, standard deviation). These may be determined in preliminary studies or from published information. It is generally true that a large sample is required to discover a small difference. The sample must also be

TABLE 1

Summary of important terms to validate a measurement method

Term	Concept
Reliability	Precision
Validity	Trueness Accuracy of the mean
Accuracy	Accuracy Reliability and validity

large if the scatter of the outcome parameter is large in the study groups. Sample size planning helps to ensure that the study is large enough, but not excessively large. The sample size is often restricted by the available time and/or by the budget. This is not in accordance with good scientific practice. If the sample is small, the power will also be low, bringing the risk that real differences will not be identified (16, 17). There are both ethical problems—stress to patients, possibly random allocation of therapy—and economic problems—financial, structural, and with regard to personnel—which make it difficult to justify a study which is either too large or not large enough (16–19). The research worker has to consider whether alternative procedures might be possible, such as increasing the time available, the personnel or the funding, or whether a multicenter study should be performed in collaboration with colleagues.

Discussion

Planning, performance, documentation, analysis, and publication are the component parts of medical studies (1, 2). Study design is of decisive importance in planning. This not only lays down the statistical analysis, but also ultimately the reliability of the conclusions and the significance and implementation of the study results (2). A six point checklist can be used for the rapid evaluation of the study design (*table 2*).

According to Sackett, about two thirds of 56 typical errors in studies are connected to errors in design and performance (20). This cannot be corrected once the data have been collected. This makes the study less convincing. As a consequence, the design must be precisely planned before starting the study and this must be laid down in the study protocol. This requires a great deal of time.

In the final analysis, studies with poor design are unethical. Test persons (or animals) are subjected to unnecessary stress and research capacity is wasted (21, 22). Medical studies must consider both individual ethics (protection of the individual) and collective ethics (benefit for society) (22). The size of medical studies is often too small, so that the power is also too small (23). For this reason, a real difference—for example, between the activity of two therapies—is either unidentified or only described imprecisely (24). Low power is the result if the study is too small, the

TABLE 2

Checklist to evaluate study design

Item	Content/information
Question to be answered	<ul style="list-style-type: none"> ● Is the question clearly defined?
Study population	<ul style="list-style-type: none"> ● Information on <ul style="list-style-type: none"> – recruitment (type, area, time) – sociodemographic information on test persons (for example, age, sex, illness) – inclusion and exclusion criteria – period of follow-up observation
Type of study	<ul style="list-style-type: none"> ● Research on secondary data ● Research on primary data (actual trials) <ul style="list-style-type: none"> – Experimental studies – Clinical studies – Epidemiological studies
Unit of observation	<ul style="list-style-type: none"> ● Technical model (for example, a prosthesis, material in dentistry, a blood sample) ● Hereditary information ● Cell ● Cell system ● Organ (for example, heart or lung) ● Organ system (for example, cardiovascular system) ● Single test subject (animal or man) ● Selected patient group (for example, hospital group, risk group) ● Population (for example, from a region)
Measuring technique	<ul style="list-style-type: none"> ● Use of measuring instruments (=validation) <ul style="list-style-type: none"> – Reliability – Validity ● Measurement plan <ul style="list-style-type: none"> – Time points – Number of investigators – Standardization of measurement conditions – Type of scale
Calculation of sample size	<ul style="list-style-type: none"> ● Was the sample size calculated? ● If yes, what were the conditions? <ul style="list-style-type: none"> – Type of test – Level of significance – Power – Clinically relevant difference – Scatter/variance

difference between the study groups is too small, or the scatter of the measurements is too great. Sterne demands that the quality of studies should be increased by increasing their size and increasing the precision of measurement (25). On the other hand, if the study is too large, unnecessarily many test persons (or animals) are exposed to stress and resources (such as personnel or financial resources) are wasted. It is therefore necessary to evaluate the feasibility of a study during the planning phase by calculating the sample size. It may be necessary to take suitable measures to ensure that the power is adequate. The excuse that there is not enough time or money is misplaced. The power may be increased by reducing the heterogeneity, improving measurement precision, or by cooperation in multicenter studies. Much more new knowledge is won from a single accurately performed, well designed study of adequate size than from several inadequate studies.

Only adequately planned studies give results which can be published in high quality journals. Planning errors and inadequacies can no longer be corrected once the study has been completed. It is therefore advisable to consult an experienced biometrician during the planning phase of the study (1, 16, 17, 18).

Conflict of interest statement

The authors declare that no conflict of interest exists according to the guidelines of the International Committee of Medical Journal Editors.

Manuscript received on 30 November 2007, revised version accepted on 8 February 2008.

Translated from the original German by Rodney A. Yeates, M.A., Ph.D.

REFERENCES

1. Altman DG, Gore SM, Gardner MJ, Pocock SJ: Statistical guidelines for contributors to medical journals. *BMJ* 1983; 286: 1489–93.
2. Schäfer H, Berger J, Biebler K-E et al.: Empfehlungen für die Erstellung von Studienprotokollen (Studienplänen) für klinische Studien. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* 1999; 30: 141–54.
3. Altman DG, Machin D, Bryant TN, Gardner MJ: *Statistics with confidence*. 2nd edition Bristol: BMJ Books 2000; 173.
4. DocCheck- Flexikon: Thema: Studiendesign. <http://flexikon.doccheck.com/Studiendesign>.
5. Schumacher M, Schulgen G: *Methodik klinischer Studien, Methodische Grundlagen der Planung, Durchführung und Auswertung*. 2. Aufl., Berlin, Heidelberg, New York: Springer 2007; 1–28
6. Moher D, Schulz KF, Altman D, for the CONSORT Group: The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomized Trials. *Ann Intern Med* 2001; 134 : 657–62.
7. Beaglehole R, Bonita R, Kjellström T: *Einführung in die Epidemiologie*. Bern: Verlag Hans Huber 1997; 53–84.
8. Fletcher RH, Fletcher SW, Wagner EH, Heerting J: *Klinische Epidemiologie. Grundlagen und Anwendung*. Bern: Verlag Hans Huber 2007; 1–24 und 349–78.
9. Fleiss JL: *The design and analysis of clinical experiments*. New York: John Wiley & Sons 1986: 1–32.
10. Hüttner M, Schwarting U: *Grundzüge der Marktforschung*. 7. Aufl., München: Oldenburg Verlag 2002; 1–600.
11. Brüggemann L: *Bewertung von Richtigkeit und Präzision bei Analysenverfahren*, GIT Labor-Fachzeitschrift 2002; 2: 153–6.
12. Funk W, Dammann V, Donnevert G: *Qualitätssicherung in der Analytischen Chemie: Anwendungen in der Umwelt-, Lebensmittel- und Werkstoffanalytik, Biotechnologie und Medizintechnik*. 2. Aufl., Weinheim, New York: Wiley-VCH 2005; 1–100.
13. Liener GA, Raatz U: *Testaufbau und Testanalyse*. 2. Aufl., Weinheim: Psychologie Verlags Union 1998; 220–71.
14. Altman DG: *Practical Statistics for Medical research*. London: Chapman and Hall 1991; 1–9.
15. Machin D, Campbell MJ, Fayers PM, Pinol APY: *Sample Size Tables for Clinical Studies*. 2. Aufl., Oxford, London, Berlin: Blackwell Science Ltd. 1987: 296–9.
16. Eng J: *Sample size estimation: how many individuals should be studied?* *Radiology* 2003; 227: 309–13.
17. Halpern SD, Karlawish JHT, Berlin JA: *The continuing unethical conduct of underpowered clinical trials*. *JAMA* 2002; 288: 358–62.
18. Krummenauer F, Kauczor H-U: *Fallzahlplanung in referenzkontrollierten Diagnosestudien*. *Fortschr Röntgenstr* 2002; 174: 1438–44.

19. Altman DG: Statistics and ethics in medical research, misuse of statistics is unethical. *BMJ* 1980; 281: 1182–4.
20. Sackett DL: Bias in analytic research. *J Chronic Dis* 1979; 32: 51–63.
21. May WW: The composition and function of ethical committees. *J Med Ethics* 1975; 1: 23–9.
22. Palmer CR: Ethics and statistical methodology in clinical trials. *JME* 1993; 19: 219–22.
23. Moher D, Dulberg CS, Wells GA: Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 1994; 272: 122–4.
24. Faller, H: Signifikanz, Effektstärke und Konfidenzintervall. *Rehabilitation* 2004; 43: 174–8.
25. Sterne JAC, Smith GD: Sifting the evidence—what's wrong with significance tests? *BMJ* 2001; 322: 226–31.

Corresponding author

Dr. rer. nat. Bernd Röhrig
MDK Rheinland-Pfalz, Referat Rehabilitation/Biometrie
Albiger Straße 19 d
55232 Alzey, Germany
Bernd.Roehrig@mdk-rlp.de