Searching the Biomedical Literature: Research Study Designs and Critical Appraisal

PETER W CALLAS

Two essential issues to consider when assessing the validity of research studies are the strengths and weaknesses of the study design and quality of methodology. This paper reviews study designs commonly used in clinical research, including case reports, cross-sectional studies, case-control studies, cohort studies, randomized controlled trials, reviews, and meta-analyses. It concludes with an outline for assessing study quality.

ABBREVIATIONS: AIDS = acquired immunodeficiency syndrome.

INDEX TERMS: case-control studies; cohort studies; cross-sectional studies; epidemiologic studies; randomized controlled trials.

Clin Lab Sci 2008;21(1):42

***Peter W Callas PhD** is **research associate professor** in the Department of Medical Biostatistics at the University of Vermont, Burlington VT.*

***Address for correspondence:** Peter W Callas PhD, research associate professor, Medical Biostatistics, University of Vermont College of Medicine, Hills Building, 105 Carrigan Drive, Burlington VT 05405. (802) 656-3195, (802) 656-0632 (fax). pcallas@uvm.edu.*

***Frances Delwiche MLIS MT(ASCP)** is the **Focus: Information Literacy** guest editor.*

The Focus section seeks to publish relevant and timely continuing education for clinical laboratory practitioners. Section editors, topics, and authors are selected in advance to cover current areas of interest in each discipline. Readers can obtain continuing education credit (CE) through P.A.C.E.® by completing the continuing education registration form, recording answers to the examination, and mailing a photocopy of it with the appropriate fee to the address designated on the form. Suggestions for future Focus topics and authors, and manuscripts appropriate for CE credit are encouraged. Direct all inquiries to the Clin Lab Sci Editorial Office, IC Ink, 858 Saint Annes Drive, Iowa City IA 52245. (319) 354-3861, (319) 338-1016 (fax). ic.ink@mchsi.com

LEARNING OBJECTIVES

1. Describe what is meant by "strength of evidence" when referring to research study designs.
2. Describe each of the following study designs: case report/case series, cross-sectional study, case-control study, cohort study (prospective and retrospective), and randomized controlled trials.
3. Discuss the strengths and weaknesses of each of the above study designs.
4. Define selection bias, recall bias, interviewer bias, and information bias.
5. Explain the major issues that should be considered when conducting a critical appraisal of a research study.

A search of the primary biomedical literature frequently brings up a number of research studies on the topic of interest. Much to the searcher's consternation, there will often be multiple studies that arrive at very different conclusions. Which is correct? Could two conflicting studies both be correct?

To answer these questions, the reader needs a sound basis on which to judge the quality of each study, for in reality, all published research studies are not equally valid. Some may be very strong, while others may have methodological weaknesses that render their findings questionable or even useless. In evaluating a research study, two essential issues must be considered: the particular study design used and its inherent strengths and weaknesses, and the quality of the methodology employed. This paper will review the various study designs commonly used in clinical research, and conclude with an outline for assessing the quality of individual studies.

STUDY DESIGNS

Most papers encountered in the biomedical literature can be classified as one of three types: clinical research, basic science research, or non-research. Clinical research is any research involving human subjects. Basic science research consists of laboratory ("bench" or "test tube") studies and animal studies. Basic science research is valuable because it can safely and ethically study topics not yet understood to the level where they can be applied to humans, but because it is not yet applicable to humans, basic science research is sometimes of

questionable relevance. Non-research papers include commentaries, news, and correspondence. These forms of publication, too, play an important role. For example, letters to the editor can be helpful in elucidating methodological issues with published research studies. This article will focus primarily on clinical research, since this is most relevant for clinical practice.

Table 1. Study designs used in clinical research

Design	Strengths	Limitations
Case report/ case series	Useful for generating hypotheses	No comparison group
	Relatively quick and inexpensive	
Cross-sectional	Relatively quick and inexpensive	Causal direction unknown
		Affected by duration of outcome
Case-control	Useful for rare outcomes	Difficult to select unbiased control group
	Relatively quick and inexpensive	Possible information bias since outcome has already occurred
Cohort	Directly measures incidence of outcome	Loss to follow-up
	Risk factors measured prior to outcome	Can be very time-consuming and expensive
Randomized controlled trial	Groups differ only on randomization variable	Ethical concerns
		Lack of compliance
		Sometimes less generalizable
		Can be very time-consuming and expensive

Most clinical studies are conducted to examine the relationship between one or more independent variables (risk factors, exposures, treatments, therapies, interventions) and one or more dependent variables (outcomes such as illness, death from a specific cause, death due to all causes, recovery, cure, or quality of life). Depending upon the rigor of the research study design, the relationship found between the independent and dependant variables is considered more or less valid. A study with high validity means that the reported relationships are likely to be true, while studies of questionable validity may have other explanations for the reported findings. Table 1 presents the most common types of clinical research, ordered from weakest to strongest in terms of the strength of evidence they provide.

Case-report/case series

The simplest studies are descriptions of a single case (case report) or a number of cases (case series) that were encountered in clinical practice or routine disease surveillance. Often the cases are presented in the literature because there is some aspect about them that is unusual.

Example: The June 5, 1981 issue of the *Morbidity and Mortality Weekly Report* reported that five young homosexual men in Los Angeles had been diagnosed with *Pneumocystis carinii* pneumonia.[1] This was noteworthy because this disease is extremely uncommon in young people who are otherwise healthy. The report led others to look for similar cases, and eventually led to the identification of acquired immunodeficiency syndrome (AIDS) as a newly discovered disease.

The major weakness of case reports and case series is that without any frame of comparison for the cases, the meaning

of any observed associations is unclear. Early case reports of what became known as AIDS mentioned that many of the patients had used recreational drugs, particularly amyl nitrite.[2] Without a comparison group of similar people without AIDS, it was impossible to know if amyl nitrite use was more common in AIDS patients, let alone causally related.

Though they rank low in terms of strength of evidence, case reports and case series are often the genesis of analytical studies using one of the more sophisticated study designs.

Cross-sectional studies

In cross-sectional studies (also know as "prevalence" or "survey" studies), the independent variable of interest (often a particular exposure) and the outcome are measured at the same point in time. A typical methodology may involve performing a diagnostic test, followed by completion of a questionnaire. The association between different amounts of the exposure and the likelihood of having the outcome is then computed.

Example: Wells and others[3] found that injection drug users had a higher prevalence of hepatitis antibodies than did subjects who were not injection drug users. This study is cross-sectional because the risk factor (injection drug use) and outcome (hepatitis A seropositivity) were determined at the same point in time.

There are two major drawbacks to cross-sectional studies. First, since exposure and outcome are measured at the same time, it is unknown which came first. In the above example, it was impossible to tell if injection drug use was preceded by or followed hepatitis A infection. Thus, it cannot be determined if the exposure caused the outcome or vice versa. Or the association could be noncausal, with the exposure and outcome related only because both are associated with some other factor, such as high risk sexual activity.

The second drawback to cross-sectional studies is that because they detect outcomes that exist at a particular time in the surveyed population (i.e., the prevalence of the outcome), they cannot tell if an exposure leads to increased risk or increased duration of the outcome. For example, insulin use is associated with being diabetic, not because insulin use increases the risk of diabetes but because it increases survival with diabetes.[4]

Case-control studies

Case-control studies start with a group of people known to have the outcome of interest ("cases"). A comparison group of people without the outcome is then assembled ("controls"). Finally, the exposure history of both groups is compared and the results analyzed for any association between the exposure and the outcome. Past exposure is determined through interviews, questionnaires, medical record reviews, laboratory tests for biomarkers, or similar methods.

Example: A case-control study was conducted to compare several serum markers on their ability to detect gestational diabetes at several points during pregnancy.[5] The cases were 35 women with confirmed gestational diabetes. Two control groups were used. The first consisted of 37 women who had abnormal one-hour post-glucose loading test glucose levels but no gestational diabetes, and the second was 73 pregnant women with normal one hour post-glucose loading test results. Comparison of the three groups showed that decreased first trimester levels of sex hormone-binding globulin were most strongly associated with being a case. The authors remark that future research using a prospective cohort design (described below) could investigate how well first trimester levels of sex hormone-binding globulin predict the development of gestational diabetes.

Case-control studies are well suited for rare outcomes. Unlike studies that need to follow very large numbers of subjects to obtain a few cases with the outcome, case-control studies can continue to enroll cases until they have enough to ensure adequate statistical power for the study objectives.

A difficulty with case-control studies is that it can be very challenging to create a valid control group. The controls should represent the population from which the cases arose with regard to past exposure history, but in practice they are often not entirely representative. Selecting an appropriate control group has been called "one of the most difficult problems in epidemiology".[4] If an unrepresentative control group leads to incorrect study findings, this is considered a form of *selection bias*, wherein "bias" means results that differ systematically from the truth.

Another concern is that since the outcome is already known at the time the study begins, determining past exposures in an unbiased manner can be difficult. Cases may recall past exposures differently than controls ("recall bias"), interviewers may ask cases and controls about past exposures in a different way ("interviewer bias"), or researchers extracting information from medical records may search for or document information differently for cases and controls. If information obtained differently for cases and controls leads to incorrect study findings, this is called *information bias*.

In spite of these potential sources of error, a well-conducted case-control study can be a very efficient way to investigate relationships between risk factors and outcomes, especially for rare outcomes.

Cohort studies

Cohort studies start with either separate exposed and unexposed groups of subjects or a single group of subjects (a "cohort") which is then divided into groups based on exposure status. In either approach, the subjects are followed over time, and the incidence of the outcome is compared for the exposed and unexposed groups. Sometimes cohort studies are referred to as prospective or longitudinal studies.

Example: The Nurses' Health Study began in 1976, when 121,700 female nurses in the US provided baseline information on health status and numerous potential disease risk factors.[6] This study has examined many possible risk factor/outcome associations, including oral contraceptive use and breast cancer, vitamin D intake and hypertension, genetic polymorphisms and endometrial cancer, and phobic anxiety and coronary heart disease. A recent analysis focused on the relationship between plasma C peptide concentration and cognitive function among a subset of the nurses (n = 718) who had plasma C peptide levels measured and did not have diabetes.[7] Cognitive function was measured at baseline and again after two years of follow-up. Women with higher levels of C peptide were found to have lower cognitive function at baseline (a cross-sectional analysis) and greater cognitive decline over follow-up (a cohort analysis), suggesting that increased C peptide levels may be related to cognitive impairment in nondiabetics.

Because incidence rates of the outcome for exposed and unexposed subjects are directly computed in cohort studies, and because exposure is measured before the outcome has occurred, cohort studies can provide stronger evidence that an observed association is truly causal. However, cohort studies also have some potential difficulties. Since subjects are followed over time, some may be lost before the end of the study. If those lost differ from those not lost for both exposure and outcome status, study findings may be biased. Another problem is that these studies can be very expensive and time-consuming to conduct, particularly for chronic diseases that may require many years to develop or rare conditions that require a large number of subjects to obtain enough cases of the outcome to be able to compare the exposed and unexposed groups.

A variation on the prospective cohort design described above is the historical or retrospective cohort study, in which the exposed and unexposed groups are assembled based on information available about past exposures and then "followed" to the present. For instance, industrial hygiene records may allow estimation of worker exposure to asbestos many years ago, and health records may be used to determine who developed lung cancer since then. Such studies are less time-consuming than prospective cohort studies, but are dependent upon accurate exposure records being available. Additionally, data are often lacking other information that may help explain observed associations (e.g., were the workers who were exposed to asbestos also more likely to smoke cigarettes?).

All of the study designs discussed thus far are known as observational studies because the researchers are simply observing and measuring what occurs naturally. Clinical research can also use an experimental design, in which there is an intentional manipulation of the independent variables.

Randomized controlled trials

Randomized controlled trials are conducted by randomly assigning study participants to one of two or more exposures, and then following the subjects over time to determine the outcome of interest.

Example: A randomized controlled trial was conducted to study if all pregnant women should be screened for gestational diabetes, or only those judged to be at high risk.[8] All consenting women at a particular obstetrics clinic were randomized at the first visit to one of two groups: selective screening, in which screening was only performed if the women had one or more known risk factors for gestational diabetes, or universal screening. All women were followed to the end of pregnancy. For the universal screening group, significantly more cases of gestational diabetes were detected, diagnosis of gestational diabetes was significantly earlier, and pregnancy outcomes were better, leading the authors to conclude that universal screening for gestational diabetes is superior to screening based on the presence of risk factors.

Randomized controlled trials most closely follow the model of basic science research in which everything is controlled except the exposure of interest. Thus they provide the strongest evidence that any observed relationship is true. So why aren't randomized controlled trials always used for clinical research?

First, there are ethical concerns. It is unethical to randomize people to exposures known to be harmful. Observational designs must therefore be used for examining known risk factors such as cigarette smoking, pesticide exposure, poverty, lack of exercise, exposure to lead paint, and poor air quality. Randomized controlled trials, however, are ideal for studying potential prevention or therapeutic interventions. Here the key word is "potential." If the intervention is not thought to at least be possibly beneficial, it may be unethical to randomize subjects to receive it. But, if an intervention is known to be beneficial, it is not ethical to randomize subjects to *not* receive it. So, careful ethical consideration of the proposed intervention and alternatives is necessary when proposing a randomized controlled trial.

Next, the ability to detect differences in subjects randomized to different conditions depends on subjects complying with the condition to which they were randomized. For example, if subjects are randomized to be on a low fat diet for the next seven years, any effect of a low fat diet on health outcomes will be difficult to detect unless those subjects actually follow such a diet for the duration of the study. To increase the likelihood of compliance, there are often strict inclusion and exclusion criteria for enrollment into a randomized trial, which can then adversely affect the generalizability of findings from these select subjects.

Additionally, like prospective cohort studies, randomized controlled trials can be quite time-consuming and expensive to carry out, frequently requiring a very large sample size and lengthy follow-up to have any potential for definitively answering the question being studied.

So although randomized controlled trials are in theory the ideal study design for answering clinical questions, in practice they are not always feasible due to ethical, generalizability, or logistical constraints.

A variation on the randomized controlled trial is a study that is experimental but not randomized. For example, an educational intervention might be implemented for patients at one health clinic, with outcomes compared to patients at a health clinic which did not receive the intervention. This study is experimental because the researchers have intervened, but it is not randomized since patients are not randomly assigned to clinic, and clinics are not randomly assigned to receive or not receive the intervention. This design is not as strong as a randomized controlled trial since patients at different clinics may differ in ways other than just the educational intervention. This study design is sometimes referred to as quasi-experimental[9] or nonrandomized experimental.[10]

Almost all clinical studies use one of the designs outlined above. Strengths and limitations of these designs are summarized in Table 1.

Reviews and meta-analyses

Once a number of studies on a particular topic have been done, it can be quite informative to do a study of studies, looking for consistency of results across studies. This can be done qualitatively (reviews) or quantitatively (meta-analyses).

In a review, similarities and differences in findings are evaluated. Potential explanations for differences, such as different populations or measurement of variables, are explored. Consistent findings across several studies can increase confidence that these findings are true.

In a meta-analysis, results for different studies are statistically combined. If Study A found the risk of the outcome increased 2.5 times among those exposed to the risk factor as compared to those unexposed, and Study B found an increase of 3.5 times, an average of these two increases (weighted to account for different sample sizes for the studies) can be computed. Because the average is based on a combination of all the studies, it is a more precise estimate of the association. Care must be taken in such analyses since misleading results may be obtained if the studies being combined differ in significant ways.[11,12]

ASSESSING STUDY QUALITY

The primary question when reading a study is whether the observed findings are likely to be true or due to some other explanation. As noted, all studies are *not* equally strong. Some designs are inherently more likely to provide results with alternate explanations, and some use less valid methods for selection of subjects and measurement of variables.

A number of checklists and guidelines are available for evaluating research studies. A new edition was recently published of a very readable book on the subject.[13] The Journal of the American Medical Association published several articles between 1993 and 1996 on "Users' Guide to the Medical Literature", which are available on the website of the Centre for Health Evidence.[14] Similarly, a series of papers on "How to Read a Paper" appeared in the *British Medical Journal* in 1997. A compilation of these is available in book form.[15]

Briefly, the following items should be considered:

- Are the study objectives clearly stated? If the hypotheses being examined are unclear, the meaning of the findings will probably also be unclear.
- What study design was used? This will help in assessing the possible threats to validity (for instance, recall bias can occur in a case-control study but not a cohort study or randomized controlled trial).
- From what population were the study subjects selected? To whom might the results be generalized?
- How were the subjects selected from the population? If there is a comparison group (e.g., controls in a case-control study), might they differ in ways that could compromise the validity of the comparisons (that is, could there be selection bias)?
- How were the study variables measured? Were objective methods used? How probable is information bias (differences in outcome measurement for exposed and unexposed subjects or differences in exposure measurement for subjects with and without the outcome)?
- Is the sample size large enough to answer the study questions? Occasionally authors provide the rationale for the number of subjects used, but often they do not. Unfortunately, sometimes advanced knowledge of statistics is needed to evaluate sample size adequacy.
- Were the methods for statistical analyses clearly described and appropriate? Training in statistics can also be helpful for determining this, but no matter how complicated the analysis, the authors should provide a clear explanation of what was done.
- Were there any problems with low participation (due, for example, to high refusal rates or high loss during follow-up)? The higher the level of nonparticipation, the greater the possibility that those who did not participate differ in ways that affect the study validity.
- Could there be differences between study groups that might explain any observed associations? To follow up on an earlier example,[2] a case-control study did find an association between amyl nitrite use and AIDS.[16] Later research found that amyl nitrite use was strongly associated with amount of sexual activity, which was the actual risk factor for AIDS. A difference such as this in some other explanatory variable is called *confounding*, with the other variable (amount of sexual activity in this instance) known as a *confounder*. If there are potential confounders, the authors should discuss how they controlled for them in the design or analysis of the study.

Consideration of the above points, which are summarized in Table 2, will help in answering the two most important questions. What is your overall evaluation of the validity of the findings? Do the results appear to be true, or could they be due to bias in selection of study subjects, bias in measurement of study variables, or uncontrolled confounding? An additional point that can strengthen the evidence that the findings are correct is consistency with findings from other studies. One clinical study is rarely definitive, but consistency across studies that use different designs with different measurements in different populations will help lead to the conclusion that the findings are valid.

Warning! With a little practice, it becomes easy to criticize any study (e.g., "The findings might be due to uncontrolled confounding"). The critical issues are the likelihood that such flaws occurred and, if likely, the likelihood that these flaws will affect the findings enough to change the conclusions. Determining these requires a clear understanding of the principles of conducting research and of the subject matter being studied.

Table 2. Summary of issues for assessing study quality

- Are the study objectives clearly stated?
- What study design was used?
- From what population were the study subjects selected?
- How were subjects selected from the population?
- How were the study variables measured?
- Is the sample size large enough to answer the study questions?
- Were the methods for statistical analysis clearly described and appropriate?
- Were there any problems with low participation?
- Could there be differences between the study groups that might explain any observed associations?

Clin Lab Sci encourages readers to respond with thoughts, questions, or comments regarding this article. Email responses to ic.ink@mchsi.com. In the subject line, please type "CLIN LAB SCI 21(1) FO CALLAS". Selected responses will appear in the Dialogue and Discussion section in a future issue. Responses may be edited for length and clarity. We look forward to hearing from you.

REFERENCES

1. *Pneumocystis* pneumonia – Los Angeles. MMWR Morb Mortal Wkly Rep 1981;30(21):250-2.
2. Brennan RO, Durack DT. Gay compromise syndrome. Lancet 1981 Dec 12;2(8259):1338-9.
3. Wells R, Fisher D, Fenaughty A, and others. Hepatitis A prevalence among injection drug users. Clin Lab Sci 2006;19:12-7.
4. Gordis L. Epidemiology. 3rd ed. Philadelphia: Elsevier Saunders Company; 2004.
5. Smirnakis KV, Plati A, Wolf M, and others. Predicting gestational diabetes: choosing the optimal early serum marker. Am J Obstet Gynecol 2007;196:410.e1-410.e7.
6. Belanger CF, Hennekens CH, Rosner B, Speizer FE. The Nurses' Health Study. Am J Nurs 1978;78:1039-40.
7. Okereke O, Hankinson SE, Hu FB, Grodstein F. Plasma C peptide level and cognitive function among older women without diabetes mellitus. Arch Intern Med 2005;165:1651-6.
8. Griffin ME, Coffey M, Johnson H, and others. Universal vs. risk factor-based screening for gestational diabetes mellitus: detection rates, gestation at diagnosis and outcome. Diabet Med 2000;17:26-32.
9. Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin Company; 2002.
10. Hulley SB, Cummings SR, Browner WS, and others. Designing clinical research. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2007.
11. Shapiro S. Meta-analysis/shmeta-analysis. Am J Epidemiol 1994;140:771-8.
12. Rothman KJ, Greenland S. Modern epidemiology. 2nd ed. Philadelphia: Lippincott-Raven; 1998.
13. Gehlbach SH. Interpreting the medical literature. 5th ed. New York: McGraw-Hill; 2006.
14. Centre for Health Evidence: Users' guides to evidence-based practice. Available from http://www.cche.net/usersguides/main.asp. Accessed 2007 June 25.
15. Greenhalgh T. How to read a paper: the basics of evidence-based medicine. 3rd ed. Maldan, MA: BMJ Books/Blackwell Publishing Inc.; 2006.
16. Marmor M, Friedman-Kien AE, Laubenstein L, and others. Risk factors for Kaposi's sarcoma in homosexual men. Lancet 1982 May 15; 1(8281):1083-7.