17 Clinical Investigation

Susan L. Furth · Jeffrey J. Fadrowski

Clinical Research Question

How can we best evaluate, treat, and assess long-term risks for children with kidney disease? Who is at risk of developing end-stage renal disease (ESRD) in childhood or young adulthood? Clinicians are often faced with questions such as these with uncertain answers in the practice of pediatric nephrology. Parents ask, "Why did my child get this disease?" "What is the most effective method to treat this condition?" "What's the prognosis of this condition in my child?" Frequently, these answers are not known, and these questions are the inspiration for high-quality clinical research. The first step in developing a valuable clinical study is determining whether the initial query can be translated into a good research question.

Hallmarks of a Good Research Question

A good research question gives *useful* information, is interesting to the researcher, builds on what is known, and can be answered with available resources. Research is a labor of love, demanding attention to detail, perseverance, honesty, and imagination. Developing a good research question is an iterative process. One needs input from knowledgeable colleagues and collaborators. The researcher must become thoroughly familiar with what is already known about the topic by reviewing the literature and consulting with experts in the area. Investigating what is already known has several benefits. First, it can reveal that the candidate research question has already been answered adequately. Second, learning what is already known provides insight into potentially useful methods for addressing a research question. For example, previous studies may demonstrate good ways to measure a variable of interest or provide background information for determining sample size. Third, a literature review may suggest ways to frame the research question at hand. For example, a literature review may reveal that particular risk factors are consistently associated with a disease process, and an intervention to modify these risk factors may form a sound basis for a clinical trial.

Finally, a good research question needs to be answerable with available resources. These include subjects available for study, technical expertise of the research staff, and the time and money that can be devoted to the project. Once a question is framed, the researcher needs to outline the study protocol or methods, which include specifying the recruitment method, number of subjects and how they will be recruited, how each variable will be defined, and the plan for data analysis. A poorly designed study is worse than no study at all because, like imprecise measurements and an improper analytic plan, it can also lead to false conclusions.

Steps in Refining a Good Research Question

A good research question usually begins with a broadly stated concept. The initial question is then made more specific by identifying independent and dependent variables. Often, research questions are concerned with causal relationships. *Independent variables* are those conceptualized to be causes; *dependent variables* are those conceptualized to be effects. The research question can be modified to ask about the role of multiple potential causes in leading to the specific outcome. A simple research question asks whether x (independent variable) causes y (dependent variable). More complex research questions could assess the relative importance of x and other variables (e.g., a and b) as causes of y. A different research question might ask how strongly x predicts y in one population versus another.

The next step is to translate a research question into a hypothesis. In our simple example, the researcher may hypothesize that x causes y. In the actual research project, the information collected is examined to determine whether it is reasonable to conclude that x does cause y. In examining the data, the investigator tests the *null hypothesis* that x does not cause y versus the *alternative hypothesis* that it does.

Scientific Method

A study's potential value is determined by the relevance of the research question. Its ultimate worth is determined by

the study methods. Methodologic issues concern the study design, subject selection, data collection techniques, and the analytic plan. Subsequent sections in this chapter discuss each of these aspects. As a foundation, this section describes the concepts of inference, generalizability, and validity.

Inference

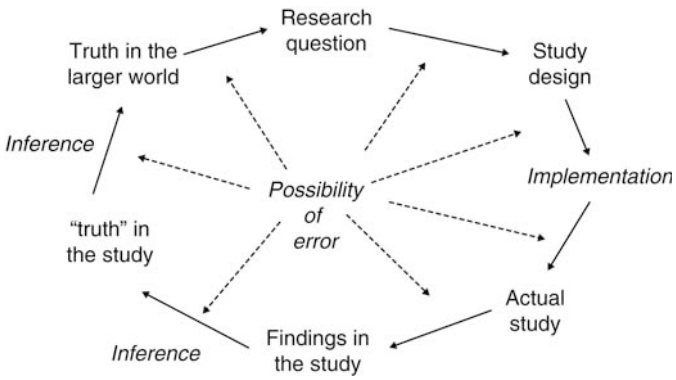
As *Fig. 17-1* shows, scientific research begins with the research question. It then moves (clockwise in the figure) to the controlled arena of the study design and then through the implementation of the actual study and findings. Inferences from the findings in the study approximate the "truth in the study." From these "truths" we attempt to infer the applicability of the findings to general clinical practice. Researchers describe and explain reality by sampling a portion of it, measuring characteristics of the sample, analyzing the measurements, and interpreting the results. Errors in the design or implementation of the study can lead to false conclusions. The strength of inference depends largely on the research methods used in the recruitment of study subjects (sampling) and in the choice and integrity of the study design. At any stage in study design and implementation, bias can occur. Bias is the result of any process that tends to produce results that depart systematically from the truth. Three broad categories of bias include selection bias, measurement bias, and confounding bias. Selection bias arises if the manner in which subjects with the outcome of interest and the comparison group were selected yields an apparent association when, in reality, exposure and disease are not associated. Measurement bias occurs when the methods of measurement systematically differ between groups. Confounding occurs when a factor is associated both with the exposure and the outcome and the effect of the main exposure under study is confused or distorted by the confounder.

Statistical inference depends on the methods used to define and sample the population. The researcher uses inferential statistics to *extrapolate* the sample findings to the larger population of children from which the sample was drawn. Inferential statistics assume that the studied sample is drawn by probability methods and can be used to make inferences about the larger population. The size of a probability sample determines the certainty of inferences from it. All other things being equal, the larger the sample size, the greater the certainty of inferences to the population.

The researcher's ability to make a *causal inference* from study results depends largely on issues of study design. Study designs are *observational* when the investigator does not manipulate the risk factor but merely selects children with and without disease and compares them in terms of the risk factor(s). Study designs are *experimental* when the investigator not only observes but actually manipulates the *relationship* between two variables. Observational designs provide somewhat weaker evidence of causation, because they fail to rule out explanations other than association between the variables studied. Experimental designs can provide much stronger evidence for causation. In an experimental study, the investigator controls the independent variable, which is the factor hypothesized to produce change in a dependent variable. In an experimental study, subjects are randomized to receive or not receive the independent variable. The goal of the process of randomization is to produce

Figure 17-1

The role of inference in drawing conclusions from clinical research studies.



study groups that are "balanced" in terms of other factors that could influence the dependent variable. Unfortunately, experimental designs often are not feasible, ethical, or desirable. Epidemiologic studies of disease preclude manipulation of risk factors in humans. Health services researchers studying the public health impact of changes in health policy rarely can control these changes.

Within the broad categories of observational and experimental designs, there are many variations. These variations, distinguishing characteristics, primary uses, strengths, and weaknesses are discussed in the section Study Design. Study design also influences the validity of study results.

Validity

Validity is the extent to which study findings correctly reflect and explain reality (Fig. 17-2). Validity can also be thought of as accuracy. Systematic bias undermines validity. As a research question's relevance increases, so does the need for validity. The clarity and rigor of study design as well as the careful implementation of the research plan increase the likelihood that inferences from the study are valid.

Cook and Campbell (1) define several *aspects* of validity. *Statistical validity* is the correctness of study conclusions regarding the existence of a relationship between two variables. A study lacks statistical validity when it concludes there is no relationship between variables when in fact there is one or when it concludes there is a relationship when in fact there is none. Statistical validity is jeopardized most often by inadequate sample size and by improper use of statistical tests.

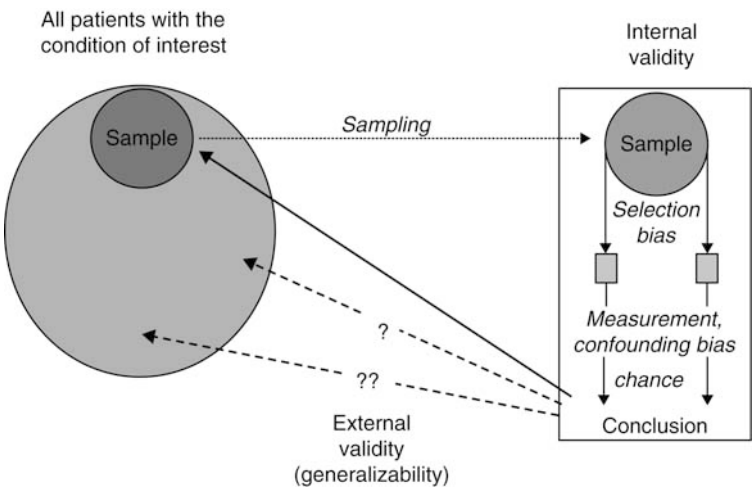
If there is a relationship between two variables, *internal validity* is the correctness of conclusions about whether the relationship between the independent and dependent variables is causal. Internal validity is jeopardized when a study design fails to control for factors that could confound the hypothesized causal relationship.

Finally, given a causal relationship between the independent and dependent variables, *external validity* is the correctness of generalizing to other persons, settings, and times. Poor choice of a study population and inadequate research procedures are common challenges to external validity.

Reliability. While validity is the degree to which data measure what they were intended to measure, reliability is the extent to which repeated measurements at different times and places, or by different people, are reproducible. An instrument measuring an assay may be accurate if, on average, the measures vary around the true value, but may not be reliable, because the measures are widely scattered around the true value. Reliability and precision are related concepts. A precise measure is one that has nearly the same value each time it is measured. It is very reliable. Precision can be described statistically using the standard deviation of repeated measures. The standard deviation

Figure 17-2

External and internal validity in experimental designs (from Fletcher RH, Fletcher SW, Wagner EH. Clinical Epidemiology: The Essentials, 3rd edn. Baltimore, Williams & Wilkins, 1996:12, with permission).



divided by the mean is the coefficient of variation. Small coefficients of variation imply precise measurements.

In summary, the scientific method involves extrapolating inferences from a study situation to the larger world. The value of research depends on the validity of such inferences, which in turn is determined by the researcher's choice of reliable methods. The following sections explain the strengths and weaknesses of the study designs available.

Study Design

Observational Studies

There are four major types of observational studies: case series, cross-sectional, case-control, and cohort. Observational designs are weaker than experimental designs in establishing causation, but they are useful when it is not feasible to manipulate the independent variable. Studies of disease etiology usually are observational. In these types of observational studies, risk factors or exposures are the independent variables, and disease is the dependent variable.

Case Series

In a *case series* study, a sample of cases is chosen, and the presence of the risk factor is measured. A case series study is easy to conduct and is useful as a preliminary study to reinforce anecdotal evidence, to generate hypotheses, or to establish variable distributions in planning future research. Case series can sometimes identify previously unrecognized constellations of symptoms or morbidities attributed to exposures to drugs or toxins. An example of a case series is the report by Furth et al. on diabetes associated with the use of tacrolimus in pediatric renal transplant recipients (2). The authors identified a number of pediatric transplant recipients treated with tacrolimus who developed diabetes. The authors summarized the case histories and reviewed existing literature regarding diabetes associated with immunosuppressive therapy in adult transplant recipients. A case series such as this can provide useful information for the clinician. However, as a method to determine the risk associated with a particular factor, this design is extremely weak because there is no means of comparison. Even if a risk factor is highly prevalent among the cases (in this example, all the cases of diabetes posttransplant had been treated with tacrolimus), there is no way of knowing whether the risk of the disease is greater with exposure to tacrolimus than with exposure to other immunosuppressive medications – for example, cyclosporine or steroids. The case series design cannot provide an estimate of risk.

Cross-Sectional Design

A cross-sectional study is one in which the disease and risk factors are measured at the same time in a sample of subjects. Subjects can be categorized as either having or not having the risk factor. Within each group, the presence of the disease can be determined. Analytically, the association between a particular risk factor and the disease is measured as the relative prevalence of the disease among those with, versus those without, the risk factor. The cross-sectional study design is superior to the case series in that it provides a means for comparison. Cross-sectional studies are relatively economic, easy to conduct, and allow simultaneous examination of multiple risk factors.

An example of a cross-sectional design is an analysis of hemoglobin levels among the first 340 children enrolled in the Chronic Kidney Disease in Children Prospective Cohort Study (CKiD) (3). Although the CKiD Study is a "cohort study," because hemoglobin in the present analysis was described in each subject at *only one time point*, the first study visit, it would be considered a cross-sectional design. As the CKiD Study progresses, serial hemoglobin measurements among the CKiD participants will be available, allowing for cohort study designs (see "*Cohort Design*"). *Table 17-1* from this cross-sectional study describes the prevalence of anemia. The prevalence of anemia increased as the severity of chronic kidney disease increased.

Cross-sectional studies have a number of limitations. Studying prevalent patients runs the risk of missing those patients who were "cured" or who died soon after

Table 17-1

Prevalence ratio of anemia by CKID stage in cross-sectional study of children with chronic kidney disease

KDDQI CKD stage[a]	*n*	% Anemic	Prevalence ratio	95% CI
2	39	21	1.0	–
3	217	39	1.9	1.0–3.6
4	82	73	3.6	1.9–6.7

[a]Based on GFR measured by plasma disappearance of iohexol (3)

developing the disease. Also, because in cross-sectional analyses the presence or absence of two factors is assessed at the same time, it is not possible to attribute causality.

Case-Control Studies

A case-control study starts with the identification of persons with the disease or other outcome variables of interest in the population at risk (*Fig. 17-3*). A suitable control group of persons without the disease or outcome is also selected from the population at risk. This is pictured on the right side of *Fig. 17-3*. To examine the possible relation of one or more exposures to the given disease or outcome, the researcher then looks back in time to compare the proportions of the cases and controls exposed and not exposed to the risk factor in question.

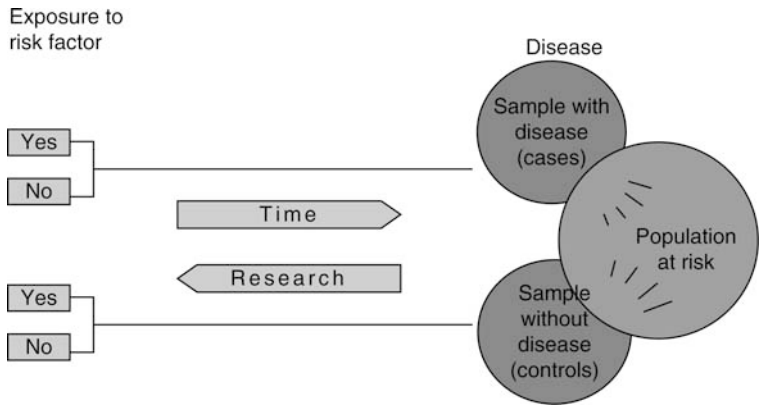
The case-control design has several advantages. It provides stronger evidence of causation than the cross-sectional design. In a cross-sectional study, outcomes and exposures are assessed simultaneously, and the investigator must infer cause and effect relationships because the temporal sequence cannot be established. In a case-control design, an attempt to establish a temporal relationship between the outcome and exposure is made by starting with a population of persons with and without the outcome and then working backwards to examine suspected exposures. Thus, compared to the cross-sectional design, the investigator is more confident that the exposure of interest came before the outcome, not as a result of the outcome. As compared to other study designs, case-control studies are efficient in the study of rare diseases or those with long latent periods between exposure and outcome. Whereas cross-sectional or cohort designs would require a large number of subjects and time to identify risk factors for a rare disease, a well-designed case-control study can identify similar risk factors with comparatively fewer subjects and much less time and expense. Also adding to the efficiency of the case-control design, several potential risk factors for a disease or outcome can be examined simultaneously.

A case-control study by Fored et al. confirmed the association of acetaminophen and aspirin with chronic renal failure (4). Adult Swedish patients with early-stage chronic renal failure were identified as cases ($N = 918$) from monthly reports of serum creatinine measurements from medical laboratories. Controls were randomly selected throughout the ascertainment period from the Swedish Population Register ($N = 980$). Aspirin and acetaminophen were used regularly by 37 and 25%, respectively, of the patients with renal failure and by 19 and 12%, respectively, of the controls. Regular use of either drug in the absence of the other was associated with an increase by a factor of 2.5 (odds ratio via logistic regression) in the risk of chronic renal failure from any cause.

Case-control studies yield an odds ratio as an estimate of relative risk. This measure is calculated by dividing the odds that a patient was exposed to a given risk factor by the odds that a control was exposed to the risk factor. It can also be obtained from logistic regression analysis. Logistic regression allows the investigator to obtain the odds ratio for a given risk factor independent of other potential risk factors or confounders using the technique of adjustment. Odds ratios are generally a good approximation of relative risk if the outcome is rare.

Figure 17-3

Design of a case-control study (from Fletcher RH, Fletcher SW, Wagner EH. Clinical Epidemiology: The Essentials, 3rd edn. Baltimore, Williams & Wilkins, 1996:213, with permission).



As with any study design, the case-control method has limitations. Case-control studies allow for the study of only one disease at a time, as opposed to cross-sectional or cohort studies. This design does not allow for the measurement of incidence, prevalence, or excess risk. Case-control studies are also subject to error, or bias, which can threaten the validity of the study. Selection and information biases, the two major categories of bias, are possible in the case-control design. Selection bias arises if the manner in which cases and controls were selected yields an apparent association when, in reality, exposure and disease are not associated. For example, cases, by definition, include only individuals who have been identified as having the disease and who are available for study. Those who have not been diagnosed, have been misdiagnosed, or have died are excluded. If diagnosis or availability is related to the exposure being studied, the sample of cases will be biased.

Avoiding selection bias can be even more challenging in the selection of controls. The control group must be comparable to the cases. They should not be chosen in such a way that important differences between cases and controls exist that might influence exposure history and thus limit the inferences derived from the study. A number of strategies exist to select a control group that is at risk for the disease and otherwise representative of the same population as the cases. These include sampling cases and controls in the same way (e.g., from the same clinical setting), matching controls to cases on key variables related to the disease (e.g., age), using multiple control groups, and using population-based samples of both cases and controls (e.g., using disease registries).

Information bias occurs when the case and control groups differ in terms of the quality of the data collected to measure risk factors. The retrospective approach to measuring an exposure in the case-control design introduces the possibility of differential recall between the cases and controls. Cases may have been asked more often about the presence of a given exposure and/or may be more circumspect in their recall of such exposures. This introduces *recall bias*, a form of information bias because of better, and sometimes exaggerated, recollection of exposures by cases as compared to controls.

It can be difficult for an investigator to remain objective in collecting exposure information. In interviewing subjects and in reviewing records, there may be a tendency to look more carefully or evaluate evidence differently for cases than for controls. Strategies for dealing with this problem include the use of objective measures and to ensure that the individuals collecting data are unaware of the subject's group status (*blinding/masking*). The more subjective the method for measuring the exposure, the more important it is to mask the observer. Blinding as to the specific exposure being studied or study hypothesis is useful and also can be used to attempt to control recall bias.

Nested case-control studies and nested case-cohort studies are alternative case-based hybrid designs that have many advantages. A nested case-control study involves selecting all cases and control subjects from a known cohort. In this design, the controls are free of the outcome or disease. Nested case-control studies eliminate the problem of recall bias, because the exposure information is obtained before the outcome has developed (cohort design). Also, the temporal sequence between exposure and outcome is defined. This design is also much more economical and efficient; the entire cohort need not be analyzed for a given exposure (e.g., via a laboratory specimen). Nested case-cohort studies also use the selection of cases and controls from a known cohort. However, in this design, controls are randomly selected from the initial cohort irrespective of outcome. This design permits the delineation of *relative risk* for an exposure.

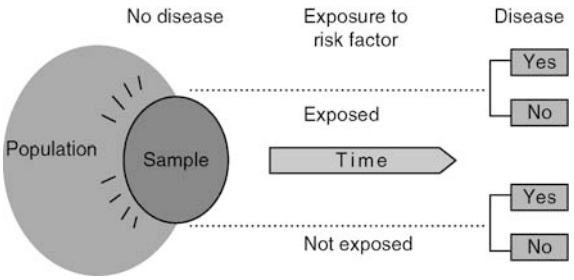
Cohort Design

Various names, including prospective, follow-up, and longitudinal, have been used to label cohort studies in the past, reflecting the temporal sequence of exposure and disease in this category of observational studies. The word *cohort* originated from the Latin word *cohors*, describing a group of warriors that marched together. Clinical investigators have adapted this term to a specific type of research study: a group of individuals free of the disease(s) of interest is assembled, their risk status is determined, and the group is followed over time to measure the incidence of disease. Comparison of the incidence of disease (or rate of death from disease) between those with and without the exposure of interest permits measurement of the association between the risk factor and the disease (*Fig. 17-4*).

The significance of the cohort design has been emphasized by the wealth of scientific data obtained from famous cohorts such as the Framingham Study or the Physicians Health Study. In the field of pediatric nephrology, the Chronic Kidney Disease in Children Prospective Cohort Study (CKiD) is ongoing and designed to improve the understanding of the occurrence and progression of chronic kidney disease and its multiple complications (5). Pediatric nephrology has previously benefited from studies

Figure 17-4

Design of a cohort study (from Fletcher RH, Fletcher SW, Wagner EH. Clinical Epidemiology: The Essentials, 3rd edn. Baltimore, Williams & Wilkins, 1996:102, with permission).



using the cohort design. For example, Wong et al. used the cohort design to demonstrate a 17-fold increase in the risk of hemolytic uremic syndrome (HUS) associated with antibiotic use in children with *Escherichia coli* 0157:H7 diarrhea infections (6). In this study, children with *E. coli* 0157:H7 were followed to assess risk factors for the development of HUS.

The cohort design has an obvious niche in clinical research. Ethical and practical considerations often do not allow for randomization of individuals to an exposure of interest. Cohort designs allow for the examination of exposure and disease associations under such circumstances.

Cohort studies can be classified as concurrent or nonconcurrent. In a concurrent cohort study (also referred to as a prospective or longitudinal study), the clinical investigator identifies the population and collects extant exposure information and then follows the cohort to a designated point in the future. Nonconcurrent cohort studies (i.e., retrospective, historical, and nonconcurrent prospective) require the investigator to identify a cohort that has been delineated in the past, along with information regarding the exposure(s) of interest. This population can then be followed for the development of a given disease in the more recent past, the present, or into the future.

Traditionally, the outcome of interest in cohort studies is the ratio of the incidence of disease in those with the exposure divided by the incidence of disease in those without the exposure. This can be interpreted as the *relative risk* for disease in many cases. When calculating and interpreting risks in the cohort design, the absence of randomization must be taken into account. Because the investigator is merely observing the exposure and not controlling for it via randomization, subjects with and without the risk factor might differ in terms of other characteristics that are related to the disease. If the characteristic is related to both the exposure being evaluated and the disease, it can lead to a misleading association between the exposure and the disease. Such a characteristic would be a *confounder*.

To avoid misinterpreting such an association, the investigator must measure potential confounders and adjust for them in the analysis. Multivariable analyses are examples of statistical tools used to adjust for confounders. However, unsuspected confounders might still jeopardize the validity of conclusions.

For example, Wong et al. (6) used a multivariate logistic-regression analysis to account for potential confounders in the association of antibiotic use for *E. coli* 0157:H7 associated diarrhea and HUS. Adjustments were made for the initial white blood cell count and the day of illness on which the initial stool culture was obtained for analysis. These factors had been previously associated with increased risk of HUS. A higher initial white blood cell count could be a potential confounder, for example, because it is associated with an increased risk of HUS (the outcome) and might make the physician more likely to prescribe antibiotics (the exposure), thus potentially falsely linking antibiotic usage with HUS. After adjustment for these factors, the multivariate analysis revealed a persistent association, reassuring the discerning reader.

Cohort studies have several advantages. Because risk factors are measured before disease, the temporal sequence of risk and disease is established, and the potential for biased risk measurement is avoided. Several diseases or outcomes can be measured, and disease occurrence can be measured in terms of incidence, not just prevalence. Cohort studies often require large sample sizes and are unsuitable for studying rare diseases. Large sample size and long follow-up periods can make cohort studies costly. A nonconcurrent cohort design can reduce cost, but it decreases the investigator's control over subject selection and risk factor measurement.

Experimental Studies

In an experimental study, the investigator controls the independent variable, also known as the intervention or treatment, and then observes the effect on an outcome or series of outcomes. The classic form of experimental design is the clinical trial.

Randomized Controlled Clinical Trials

Most investigators, clinicians, and patients are familiar with the "gold-standard" of experimental designs, the

randomized blinded controlled clinical trial. Such trials are considered "gold-standard" because the rigid design helps minimize the influence of confounding variables and bias, allowing the true effect of the intervention to be elucidated.

The IgA Nephropathy Study is an example of a randomized controlled clinical trial. In this study, eligible patients younger than 40 years with immunoglobulin A (IgA) nephropathy on kidney biopsy, estimated glomerular filtration rate $\geq$50 mL/min/1.73 m^2, and evidence of moderate to severe proteinuria are randomly assigned to receive alternate day prednisone, fish oil, or placebo (7). The goal of this study is to determine the relative benefits of fish oil or alternate day prednisone on the progression of IgA nephropathy. In this and other controlled clinical trials, randomization is the key feature of its experimental nature. Participants are *randomly* assigned to the group that will receive the intervention (the intervention, treatment, study or experimental group) or the group that will not receive the intervention (the control or placebo group) (*Fig. 17-5*). Through randomization, all potential confounders, both those recognized by the investigator and those that are not suspected, are likely to be balanced between the study groups. In other words, the three groups in this study are considered to be the same except for the treatment they receive. Differences in rates of progression of kidney disease in the three groups are attributable to the intervention, because the effect of confounders has been ruled out by the balance achieved by randomization. Therefore, experimental designs offer stronger evidence of causality than do observational designs.

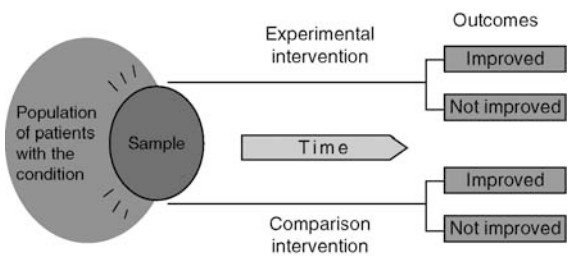
In an experimental study, it is important to ensure that subject assignment is truly random. This can be achieved by using random numbers, either through computer-assisted assignment or manually, with a table of random numbers. Sometimes *blocking* is used in conjunction with random assignment. A block of subjects is simply a set number of consecutive study enrollees. Within each block, a predetermined number of subjects are randomly assigned to each study group. For example, if the block size is set at six and two study groups of equal size are desired, then three subjects in each block of six are randomly assigned to one group and three to the other. Blocking is useful when study enrollment is expected to be prolonged. Over extended periods, both study procedures and outside conditions can change. Blocking ensures that the study groups will be balanced with regard to such changes. For example, given the prolonged enrollment period due to the relative rarity of IgA nephropathy in children and young adults, block randomization was used in the IgA Nephropathy Study.

Experimental studies, like observational studies, are subject to measurement bias. Research staff should be masked or *blinded* to the subject's group assignment during data collection, especially if any outcome measures are not strictly objective. In the IgA study, for example, both the research staff and the study subjects are unaware of which treatment they receive. In one arm of the study, placebo capsules are made to look exactly like fish oil capsules. In another arm, placebo tablets are distributed that are identical to prednisone. If outcomes are measured, without subject or staff knowledge of the subject's group status, they are less likely to be influenced by expectations about potential differences between treatment and control group outcomes. Studies in which both the participant and the staff are unaware of the treatment group status are known as "double-blinded." There may be situations in which blinding the clinical staff and patients is difficult or impossible. Clinical trials that are not blinded are known as open-label trials. Such designs are common in cancer clinical trials, as treatments being compared are often complex, with different side effect profiles and delivery protocols, and thus masking is not feasible.

Clinical trials typically provide the necessary medical evidence to bring a new treatment into use, and for drugs, such trials are referred to as "phases." A *phase I* trial is most often the first stage in testing a new drug in humans, and may include healthy participants and/or patients. In these trials, information on the distribution, metabolism, excretion and side effects of the drug is investigated. The optimal dose of the drug to deliver is also investigated. Such studies are typically not randomized or blinded.

Figure 17-5

Design of a randomized trial (from Fletcher RH, Fletcher SW, Wagner EH. Clinical Epidemiology: The Essentials, 3rd edn. Baltimore, Williams & Wilkins, 1996:139, with permission).



Phase II clinical trials are designed to evaluate the effectiveness and short-term safety and side effects associated with the drug. These trials are generally carried out in persons having the disease of interest. *Phase III* trials are expanded trials to prove the efficacy of a drug and provide the information necessary for physician labeling of the drug. These studies are most often randomized and blinded. Once completed, the pharmaceutical manufacturer may apply to market the drug for the indication(s) studied in the phase III trial. *Phase IV* clinical trials are postmarketing studies (post licensure) designed to delineate more information about the drug's risks, benefits, and optimal use in "real-world" conditions.

Alternative Designs

It is not always feasible to conduct a traditional randomized controlled clinical trial for every treatment. In such scenarios, multiple alternative experimental study designs are available and may provide useful information regarding the treatment under consideration. A few examples of such alternative designs follow. *Cross-over trials* randomly assign half of the study population to start with the control period and then subsequently switch to active treatment; the other half are on the opposite schedule. Such trials allow each participant to serve as their own control, allowing for increased statistical power and thus fewer participants. However, these studies generally take a longer period of time. The analysis and interpretation of the results may also be complicated if the treatment effects are thought to persist for a period of time even after the intervention has been ceased. A before-after trial compares the outcomes of different types of treatments in a group or groups of interest by taking advantage of calendar time. In such a study, outcomes in individuals receiving one type of treatment during a given period are compared with individuals at a subsequent time, who have received a different treatment. Although economical to perform, the results of such designs may be more prone to error as it may be difficult to know and/or control factors independent of the treatment that may have also influenced the outcome of interest.

Issues with Analysis in Experimental Designs

It is not uncommon in clinical trials to have patients who were assigned to one group switch to a different group. For example, a patient assigned to receive the active treatment under study may discontinue the treatment. Alternatively, a patient assigned to the control group may end up receiving active treatment. To avoid introducing bias in the results, it is common practice in clinical trials to analyze the results by "*intention-to-treat*." In such conservative analyses, every patient is grouped according to his/her original randomization assignment when analyzing the results, regardless of whether the patient actually received the assigned treatment or not. Intention-to-treat analyses may dilute the effect of the treatment of interest, but more importantly, likely minimize the introduction of bias into the study which may lead to erroneous conclusions.

To maximize the yield of arduous and costly clinical trials, investigators may perform *subgroup analyses*, defined as an evaluation for treatment effects within a subset of patients. For example, in the IgA Nephropathy Study, baseline proteinuria was found to be associated with "time to failure," defined as a decline in estimated GFR to $\leq$60% of baseline. The authors performed a subgroup analysis comparing the effects of the treatments and placebo among those with more severe proteinuria, defined as a first morning baseline urine protein/creatinine ratio between 1 and 3; no significant difference between the treatment groups or the placebo group in the time to failure was observed (7). Related to subgroup analyses, *post hoc analysis* refers to examining the data after the study has concluded for associations that were not specified a priori. Although subgroup analyses and post hoc analyses may provide additional useful information, due to analytic challenges, particularly in the area of sample size, results may be misleading. Given these statistical limitations, it is recommended that all hypotheses and intended analyses be stated prior to the initiation of the study, which may help in planning the design of the study, including the sample size. In the event that post hoc analyses are performed, they should be clearly labeled as such so that the reader is able to identify the potential limitations to any conclusions derived from such analyses. A recent report in the New England Journal of Medicine reviews the challenges of subgroup analyses, and provides guidelines for their use within the Journal (8).

In summary, many research designs are available to the investigator. No single design is best for all research questions. Although experimental designs are superior to observational designs in addressing threats to internal validity, they are not always feasible or ethical. The most appropriate design for a given research question is the design that maximizes internal validity within the constraints of the research environment.

Important Issues in Carrying out a Research Plan

Selection of Subjects

In any research study, one would like to extrapolate the findings to all patients with the condition of interest. The study population is the group that is meant to represent the target population from which a sample is drawn. Sampling decisions involve defining the study population and sample.

Defining the Target Population

Although there is no one single ideal target population, the investigator needs to consider the ramifications of one definition versus another. If the investigator were interested in studying risk factors for a specific disease, the target population could be defined as all children with this disease or a subset of them (e.g., children of a certain age). The broader the target population, the greater the generalizability of the study findings. On the other hand, the increased heterogeneity of a broadly defined target population could introduce variability among subgroups in terms of the importance of risk factors. For example, a particular characteristic could be a major risk factor in some population subgroups but not in others. Assessing the importance of risk factors within subgroups requires a larger study sample and perhaps a more complex sampling design.

Defining the Study Population

A practical consideration in defining the target population is availability of the population for study. The investigator could have all children available seen in a particular clinical setting. Insofar as children seen in this setting are representative of the target population, the clinical site would be a good choice for study; the experience of its enrollees could be considered *generalizable* to the target population. If children enrolled in the clinical setting differ systematically from the target population, *sampling bias* is introduced. For example, tertiary care pediatric nephrology centers might be more likely to serve children with advanced *stages* of the kidney disease or more severe or complicated cases. Studying only these cases may introduce bias toward only studying the most complex forms of a particular disease. Sampling bias impairs the generalizability of study findings. Representativeness, therefore, is a prime consideration in defining a study population. Investigators should evaluate the representativeness of candidate settings and the likely implications of potential biases. One possible approach to this in studies of patients with kidney disease is to compare the characteristics of participants in a study to known characteristics of the larger population to whom one would like to generalize the results.

Defining the Sampling Scheme

Just as we generalize from the study population to the target population, we generalize from the study sample to the study population. *Sample statistics* are measures that pertain to the samples that are studied. A sample mean, for example, is the sample's average score on a particular measure, and a sample standard deviation expresses the variability of the sample scores. The sample statistics are the investigators best estimates of the *population parameters*. The sample mean is the best estimate of the population mean; the sample standard deviation is the best estimate of the population standard deviation.

Extending beyond inference to hypothesis testing, sample statistics of the association between variables are the best estimates of these associations in the target population. The association between a hypothesized risk factor and the occurrence of disease in the study sample (perhaps measured by odds ratios or relative risks) is the investigator's best estimate of the association between the risk factor and disease in the target population.

Probability theory is the rationale for extrapolating inferences from a study sample to the reference population. A *probability sample* is one in which every subject, or element, in the study population has a known probability of being selected. A *nonprobability sample* is one in which the probabilities of selection are unknown. It is legitimate to extrapolate from a sample to its population only if probability sampling has been used.

There are several types of probability sampling. In *simple random sampling*, each element has an equal chance of being selected. In *systematic sampling*, each element in the population is assigned a consecutive number, and every *nth* element is sampled. Systematic sampling is easy to use, but it will generate a biased sample if the sampling fraction (e.g., every tenth case) is the same as some periodicity in the ordering of cases in the population. For example, if every tenth patient is sampled in a clinic where ten patients are seen each session and the most complex cases are scheduled first, then the sample will contain either all complex cases or no complex cases, depending on the first element drawn. *Stratified random*

sampling is useful when one believes that population subgroups differ in important ways. The population is divided into the subgroups, or *strata*, of interest. Simple or systematic random samples are then drawn from each stratum. *Cluster sampling* is useful when it is difficult or costly to sample elements in a population individually. Instead of elements, groups of elements are sampled. For example, in a study of school children, the investigator could take a probability sample of classrooms and then study all the students within each selected classroom. The selected classrooms, in combination, must be representative of the overall population. As with stratified sampling, formulas for calculating variance must be modified, and consultation with a statistician is recommended.

Nonprobability sampling techniques include *convenience sampling*, *quota sampling*, and *purposive sampling*. A convenience sample is one that is most readily obtained without the use of random sampling. A quota sample is a convenience sample drawn to assure specified numbers of subjects in specified strata, without the use of random sampling. A purposive sample is one in which subjects are selected because they are judged to be representative of the population of interest.

Probability sampling is preferred but not always possible. In clinical research, the investigator is often limited to a particular clinic population. If a clinic population is believed to be representative and if it is larger than the number of subjects needed for study, the investigator should use a probability sampling technique to draw the study sample.

An example of probability sampling using stratified sampling techniques can be seen in a recent survey study of adult and pediatric nephrologists (9). The authors created a survey containing ten case vignettes to assess whether increased experience with pediatric patients influenced nephrologists' recommendations for peritoneal or hemodialysis in otherwise identical patients with ESRD described in the vignettes. Because the authors wanted the survey respondents to represent the population of U.S. adult and pediatric nephrologists, they randomly selected a representative sample of nephrologists in five geographic regions of the United States. Each randomly selected nephrologist was mailed a survey containing ten case vignettes to assess what factors affected the nephrologists' dialysis recommendations.

Determining Sample Size

In any study, several factors determine the required sample size. This section describes those that come into play in several common types of investigations. Detailed sample-size formulas and tables are beyond the scope of this chapter, but several excellent references are listed in Suggested Reading. Briefly, to estimate sample size, the researcher needs to set the acceptable level of α (probability of type I error), β (probability of type II error), and determine the effect size that one is likely to see. In determining sample size, the probability of making a type I error, α, is usually set at 0.05. This is the probability of concluding that an association between two variables exists when it does not. β error is the probability of concluding that no association exists, when in fact it does. The reader will be more familiar with β error in terms of its relationship to "power." The power of a study is equal to 1-β. In many studies, β is customarily set at 0.2 for a power of 80%. If β is set to 0.1, the power of the study is 90%.

In addition to specifying α and β, the researcher must also determine an estimate of the response to treatment in one of the groups (in a clinical trial) or the rate of occurrence of disease (in a cohort study). The effect size is an estimate of how much better than the comparison group you expect a treatment group to be in a clinical trial or how increased the risk of a particular disease is in the setting of a particular risk factor (in a cohort study). An illustration of estimated sample sizes for given α, β, and effect sizes is shown in *Table 17-2* for a study comparing differences in proportions in two groups (10).

Table 17-2 illustrates how varying the acceptable levels of α, β, and effect size influences sample size. For example, if we designed a study to determine whether a new drug could "cure" 40% of patients compared to an old drug that "cured" 10% of patients, we would have 90% power to see such an effect with 95% confidence in a total sample of 96 patients (see row 2 in *Table 17-2*). In contrast, to obtain a significant result documenting a smaller effect size from the old drug cure rate of 10% to a new drug cure rate of 20% with the same $a = 0.05$ and 90% power, we would need to study 572 patients.

Attrition of Study Subjects

Sample size calculations determine the number of subjects needed at the study's conclusion. In determining the number of subjects to enroll, the investigator must estimate attrition rates and enroll a sufficiently large sample to compensate for study dropout.

Even if probability sampling is used to define the study, subject attrition could produce a biased sample at the study's conclusion. An attrition rate of more than 25% is cause for concern. In data analysis, subjects completing the study should be compared with those who drop out to

Table 17-2

Example illustrating how α, β, and effect size affect sample size

Confidence 1-α (%)	Power 1-β (%)	Effect size (%)	Total sample No.
95	80	10→40	76
95	90	10→40	96
95	90	10→20	572
95	99	10→40	156

Adapted from sample size calculator (Statcalc) in Epi-info Stat Calc (10)

determine whether the two groups differ in a clinically significant way. Such differences must be considered in interpreting the study findings.

Data Collection: Measurement

Decisions on what data to collect and how to do so begin with specifying the variables that need to be measured and operationally defining each. The investigator will need to evaluate the suitability of existing measures and determine whether to use an existing measure or develop a new one. The data sources for each variable must also be identified. Finally, the investigator should specify the level of measurement of each variable. An efficient way to document the data collection plan is to make a table with columns listing the variables to be measured, their definitions, the data source(s) for each variable, and level of measurement for each. This section describes issues pertaining to each of these tasks.

Identifying the Variables to Be Measured

Researchers are often tempted to collect as much information as possible. This can be costly, in terms of time, money, and data quality. The investigator should be able to justify each variable to be measured. Most important are the hypothesized independent and dependent variables. In addition, identified potential confounders should be mea\sured. Finally, data characterizing the study population and sample will be needed to describe the study's generalizability.

Types of Variables

Variables can be continuous or categorical. Age, height, and weight are examples of continuous variables. Categorical variables can be binary, with two possible outcomes such as male or female gender; nominal, i.e., nonordered categories such as race: White, African American, Asian, Native American or Pacific Islander; or ordinal, i.e., ordinal categories such as a pain rating scale: none, mild, moderate, severe.

Sources of Data

Study data can be collected from existing sources or can be generated specifically for the specific research hypothesis being tested, using surveys, interviews, or observations. Most studies combine both strategies.

An enormous variety of existing data sources is available, including medical records, vital records, national and local health surveys, and census data. Health programs often keep records of services provided, and billing records can be especially helpful. Examples of existing data sources in pediatric kidney disease include the United States Renal Data System (USRDS) (11) and data collected by the North American Pediatric Renal Trials and Collaborative Studies (NAPRTCS) (12). Existing sources can provide data for time periods and individuals otherwise unavailable to the investigator. The number of studies published by the NAPRTCS and their tremendous contribution to our understanding of clinical outcomes in pediatric kidney transplantation and dialysis illustrate this point. The chief disadvantage of existing sources gleaned from registry data is that the data are often not collected as systematically as in a prospective research study. Important data elements are sometimes missing. Incomplete data and inaccuracies are also possible with registry data.

Primary data collection is expensive and is limited to subjects available to the investigator. On the other hand, the investigator's control over data collection makes data quality more certain. Many primary data collection strategies are available, including mail surveys, mass-administered questionnaires, telephone and in-person structured and unstructured interviews, direct observation, and videotaping and audiotaping. Choosing a strategy should be based on the research question, the sensitivity of the data to be collected, the literacy of the population to be studied, and the resources available for the study. The key factor should be data quality – that is, *which* method will provide the most complete and accurate information within budgetary constraints.

Assessing Data Quality

One strategy to enhance data quality is to train research staff thoroughly. Often, data collection staff works

independently. To ensure that they follow study procedures, the protocol for data collection should be detailed in a study manual. Training sessions should be held to explain the study's aim to staff, as well as how each one's role fits into the big picture. Staff should be given ample time to practice their data collection skills. Once the fieldwork of the main study begins, staff should be encouraged to bring problems to the attention of supervisory staff. Such problems should be resolved in a timely fashion, and the resolution should be documented and added to the study manual. In this way, research staff are kept apprised of changes in the study protocol and are impressed with the importance of adhering to it.

After the instrument pretesting and staff training, the investigator should pilot test the data collection activities. A pilot test is a dress rehearsal of the activities for selecting study subjects, contacting them, securing informed consent, and collecting and processing data. Activities that do not work as planned should be modified and the pilot testing continued until the fieldwork procedures run smoothly.

Data quality should be monitored during the main study. Interviews and questionnaires should be reviewed as they are completed to allow recontacting subjects to correct errors. The reliability of subjective measures and those requiring special technical skill should be assessed. For studies with unsupervised interviewing of subjects, it is wise to validate a portion of the completed interviews handed in by fieldworkers. This can be done by recontacting a random sample of subjects and then asking them to verify their responses to a subsample of the interview questions.

Analytic Plan

Data Analysis

Investigators often defer considering data analysis until the data have been collected. This is a serious mistake. Study planning should include an analytic plan of the steps needed to answer the *study* questions once the fieldwork is completed.

Use of Statistical Tests

As noted earlier, statistical validity is the correctness of study conclusions regarding group differences and variable relationships. A key threat to statistical validity is the use of inappropriate statistical tests. Variable types and variable distribution, as well as the hypothesis being tested, determine the correct type of statistical test. The important properties of a variable's distribution are its location, spread and shape. Measures of location include the *median* (middle observation), the *mean* (arithmetic mean or average) and the *mode* (most frequent value). Measures of spread of a distribution include the *range*, which equals the maximum minus the minimum value, the *interquartile range*, which equals the 75th minus the 25th percentile, the *variance*, which equals the average squared deviation from the mean, and the *standard deviation*. The shape of a distribution can be described by its skewness i.e., symmetry, and its kurtosis i.e., "peakedness." *Parametric statistical tests* are based on assumptions about parameters of the population and are the most powerful tests available in situations in which these assumptions are met. *Nonparametric statistical tests* are based on fewer assumptions about the population, so they are appropriate in situations in which the assumptions underlying parametric statistics are not met.

Assumptions vary by statistical test. If a test is used in a situation that violates its assumptions, it will be inaccurate, leading to a misleading measure of statistical significance. This, in turn, will lead to an incorrect estimate of the likelihood of a Type *I* error.

In developing the analytic plan, the investigator should consider the assumptions of candidate tests in determining which ones to use. A discussion of specific statistical tests is beyond the scope of this chapter, but a framework for deciding which tests to use can be given. In this framework, three factors determine the type of test to use: the major analytic question to be answered, the levels of measurement used, and the number and independence of comparison groups.

In preparing the analytic plan, the investigator needs to translate the research question into analytic terms. Three major analytic approaches are to describe characteristics of the sampled population, to compare groups of subjects, and to measure associations among variables.

Where group comparisons are to be made, the appropriate statistical test is also determined by the number of groups to be compared (two vs. three or more) and by whether comparison groups are independent or matched. Thus, in a study of cases matched with sibling controls, it would be inappropriate to use a statistical test for independent groups.

As decisions about level of measurement and the selection of study groups are part of study planning, it is easy to see how these decisions are better informed if their ramifications for data analysts are considered. Level of measurement, study group formation, and data analysis

Table 17-3
Bivariates Statistical Tests

Level of measurement	Two groups		Three or more groups	
	Independent (unpaired groups)	Paired groups	Independent (unmatched groups)	Matched groups
Nominal dichotomy	Chi-square of Fisher's exact test	McNemar's test	Chi-square	Cochran's test
More than two categories	Chi-square	McNemar's test	Chi-square	Cochran's test
Ordinal	Mann-Whitney test	Sign test	Kruskal-Wallis one-way analysis of variance ANOVA	Friedman two-way ANOVA
		Wilcoxon matched-pairs signed-ranks test		
Interval	*t* Test for groups	*t* Test for pairs	One-way ANOVA	ANOVA for repeated measures

are all interrelated, and should all be considered part of study planning.

When the study's purpose is to assess the relationship between two continuous variables, the degree of concordance can be expressed as a simple *correlation coefficient*. A related approach is the use of a *kappa statistic* which expresses the degree of concordance beyond that due to chance. *Chronbach's alpha* expresses internal consistency among three or more variables that measure the same general characteristic.

When the study's purpose is to describe a population, the investigator makes inferences from sample statistics to population parameters. Sample proportions and measures of central tendency (mean, median, and mode) and dispersion (standard deviation, range) are used to estimate these parameters in the population. *Confidence intervals* can be constructed around proportions and means to express the certainty of the sample-based population estimates. When the study's objective is to compare two or more groups, sample group differences in proportions and means are used to estimate such differences in the population. Statistical tests of significance can be used to assess the certainty of sample-based inferences about group differences in the population. The appropriate statistical test depends on the number of groups compared, whether subjects in the groups are matched, and the level of measurement of the variable on which the groups are being compared. *Table 17-4* displays bivariate statistical tests commonly used in assessing the significance of group differences. For a variable with a normal distribution, a *t-test* compares two means, while *ANOVA* can compare means in three or more groups. *Chi-square* compares proportions.

When normality assumptions cannot be made, the *sign test* can be used for a single sample or paired sample to assess whether the medians of the sample and a reference, or two samples being compared differ. The *Wilcoxan signed rank test* can be used when the data are on an interval scale, and makes use of the magnitudes of the differences between measurements and a hypothesized location parameter. If the variable of interest is measured on an ordinal scale, the *Mann-Whitney test* can be used to assess whether the two populations have different median values.

When the research aim is to measure the association between variables, sample statistics again are used to estimate population parameters. The variables' levels of measurement determine the appropriate statistical measure of the strength of their association, the appropriate test of the statistical significance of the association, and the certainty of estimates of its strength. For continuous data, Pearson's correlation coefficient is used to measure association. For dichotomous variables, the *odds ratio* and *relative risk* (RR) are measures of the degree of association between two factors. Analytic studies in the medical literature often are designed to determine whether there is an association between exposure to a factor and development of disease. If there is an association, the question is how strong the association is. To assess the strength of the association, we measure the ratio of the incidence of disease in exposed individuals to the incidence of disease in nonexposed individuals. This ratio is called the *relative risk*. If the risk in exposed individuals is equal to the risk in unexposed, the relative risk is 1.0 and there is no association. If the risk in exposed individuals is greater than in unexposed (RR $>$ 1.0), then there is an association that may suggest that the exposure

Table 17-4
Sensitivity and specificity of urinalysis components (13)

Test	Sensitivity % (range)	Specificity % (range)
Microscopy: WBCs	73 (32–100)	81 (45–98)
Microscopy: bacteria	81 (16–99)	83 (11–100)
Leukocyte esterase	83 (67–94)	78 (64–92)
Nitrite	53 (15–82)	98 (90–100)
Leukocyte esterase or nitrite positive	93 (90–100)	72 (58–91)
Leukocyte esterase or nitrite or microscopy positive	99.9 (99–100)	70 (60–92)

confers risk. If the risk in exposed is less than in unexposed (RR < 1.0), the exposure may be protective against risk. The relative risk can only be calculated in a prospective study, as it requires incidence of disease. In a case-control study, since we do not know incidence, we cannot calculate the RR directly. Instead of the proportion of the exposed population who develop disease compared to the proportion of the unexposed population who develop disease, in case-control studies, we have the proportion of the cases who were exposed and the proportion of the controls who were exposed. In case-control studies we utilize the concept of odds to define the odds ratio, which approximates the relative risk if the incidence of disease is low. We compare the odds of a case having been exposed to a particular factor to the odds of a control being exposed to that factor. As in the case of relative risk, if the exposure is not related to the disease, the odds ratio will equal 1.0. If the exposure is positively related to the disease, the odds ratio will be greater than one.

Studies of the combined and relative impacts of multiple independent variables, or the effect of an independent variable after controlling for other factors, will require multivariable analytic tests. The appropriateness of a multivariable technique is determined by the levels of measurement of the independent and dependent variables. Multiple linear regression analysis can be used to assess association between a putative exposure and a continuous outcome while adjusting for other possibly confounding factors. Multiple logistic regression analysis can be used to assess the association between a putative risk factor and a binary outcome measure, while adjusting for other potential confounding factors. These measures can be used to calculate an adjusted relative risk, or an adjusted odds ratio. The references cited at the end of this chapter describe the applicability and interpretation of the most commonly used multivariable statistical tests.

In addition to describing association between two variables, and assessing the risk associated with an exposure and an outcome, another common research question in the medical literature includes assessing the time from a particular exposure until an outcome such as death, or hospitalization or transplantation. A commonly used statistical tool to assess the time to an event is survival analysis, or *Kaplan Meier analysis*. Kaplan Meier analyses can be used to compare survival between two treatment modalities. When adjusting for other potential confounding variables in a survival or time to event analysis, Cox *proportional hazards methods* can be used, yielding a hazard ratio which can be thought of as comparable to a relative risk.

Statistical Significance and Confidence Intervals

Most readers of the medical literature will be familiar with the term *statistical significance*, which is most often referred to in clinical reports as a "p value $<.05$." This highly sought after result of a statistical test refers to the probability of α, or a type I error. A p value of .05 in a study means that there is a 5% chance that the results seen in the study could have occurred by chance. However, the authors have concluded that this probability is low enough for them to accept the alternative hypothesis (that there is a real difference between groups) and to reject the null hypothesis (there is no difference between groups). It is important to note that the p value in a study result depends on the size of the observed difference between the groups in question and the size of the sample of patients studied. Standing alone, the p value does not convey any sense of the magnitude of the treatment effect seen in the study or the precision of the estimate of the size of the treatment effect. *Confidence intervals*, in contrast to p values, can convey this information in a more meaningful way.

For any estimated value, it is useful to have an idea of the uncertainty of the estimate in relation to the true

value it is trying to approximate. For example, if we designed a study to estimate the beneficial effect of a new lipid-lowering medication in chronic kidney disease in adolescents, we would try to recruit a large representative sample of hyperlipidemic adolescents and randomize them to treatment with a new lipid-lowering medication. From our study, we might want to estimate the magnitude of lipid level reduction associated with the new medicine. We might also want to use this estimate as an approximation for the "true" reduction in lipid levels that would be seen in the "universe" of pediatric patients with hyperlipidemia and chronic kidney disease. To estimate the "true" reduction in lipid levels (which can never be directly measured), we can generate a confidence interval around our estimate.

In any study, construction of a *confidence interval* around the point estimate gives us a range of values in which we can be confident that the true value resides. A confidence interval gives a sense of the estimates precision; it extends evenly on either side of the estimate by a multiple of the standard error (SE) of the estimate. In our example, our study might yield an estimate of the drop in serum low-density lipoprotein cholesterol levels of the treatment group of 30% with a standard deviation of that estimate of 20%. One could then use this estimate to generate a confidence interval around this estimate. In the medical literature, one will most often see references to 95% confidence intervals. The general equation for a 95% confidence interval is equal to the estimate ± 1.96 times the SE of the estimate. The factor 1.96 comes from the standard normal distribution, in which 95% of estimates would fall within ± 1.96 SEs of the mean. If one wanted to increase the probability of including the true estimate in the confidence interval, one could generate a 99% confidence interval, which would equal the estimate ± 2.56 times the SE of the estimate. Because the SE of an estimate is equal to the standard deviation of the population divided by the square root of the sample size, n, one can see that a larger sample size is needed in a study to generate a precise estimate of a treatment's effect. Given the same standard deviation, the SE in our study would be smaller if we studied 100 children compared to 20 children. The larger study would generate a narrower 95% confidence interval. The strict interpretation of a 95% confidence interval is that this is the range of values for the true population estimate that is consistent with the data observed in the study. In our hypothetical example, the smaller study might give us the opportunity to conclude that the new lipid lowering is associated with a 30% reduction in low-density lipoprotein cholesterol with a relatively broad 95% confidence interval of 21–39%, whereas the larger study yields a more precise estimate. The 95% confidence interval around the same point estimate of a 30% reduction is 26–34% in the study with the larger sample size.

Topics Related to Clinical Decision Making

Clinicians are routinely faced with patients with unknown diagnoses, and patients expect that the clinician will know how to efficiently and accurately diagnose the problem with which they are presenting. Diagnostic acumen relies in large part in understanding the "epidemiology" of the possible diagnoses, as well as the strengths and limitations of the various tests that might be performed to diagnose such diseases. The following section provides an introduction to concepts and terms that are related to the science of clinical decision making.

The Disease

When describing the burden of a disease in a population, the terms incidence and prevalence are often used. *Incidence* is defined as the number of new cases of a disease that occur in a population at risk for developing the disease during a specified period of time. *Prevalence* is defined as the proportion of persons present in the population currently affected by the disease at a specified point in time. Distinguishing between these terms is important when considering the burden of given disease. When prevalence is reported, people affected by the disease for varying amounts of time are included; these are not necessarily all new cases. Those with severe forms of the disease may have died and thus depending on the time specified in the definition, may not be included in the prevalence. The incidence and prevalence of end-stage renal disease (ESRD) are routinely reported by the United States Renal Data System (USRDS). In 2005, the incidence was 347 (new) cases of ESRD per million population per year. This was only 1% higher than the incidence in 2001. The prevalence of ESRD in 2005 (also referred to as point prevalence) was 1,569 cases per million population (11). Based on these definitions, we know that 347 (22%) of the 1,569 prevalent cases of ESRD per million population in 2005 were new cases.

The Diagnostic Test

When utilizing a diagnostic or screening test to aid in the care of a patient, a clinician must know "how good" the test is. Terms used to describe "how good" a test is

include sensitivity and specificity. The *sensitivity* of a test describes the ability of the test to correctly identify those that have the disease. A sensitive test has a low *false-negative rate*, meaning the test result will not frequently be negative in those who have the disease. The *specificity* of a test describes the ability of the test to correctly identify those who do *not* have the disease. A specific test has a low *false-positive rate*, meaning the test result will not frequently be positive in those who do *not* have the disease.

Table 17-4 from the American Academy of Pediatrics' Practice Parameter for the diagnosis of initial urinary tract infections in children examines the sensitivity and specificity of various components of the urinalysis (13). For example, the sensitivity of nitrite is reported as 53%. Thus, among children with a urinary tract infection, this test will be positive approximately 53% of the time. Knowing the relatively low sensitivity of the nitrite test, a clinician would not be reassured that a urinary tract infection does not exist if the result is negative. The specificity of the nitrite test is 98%. Thus, among children *without* a urinary tract infection, this test is negative approximately 98% of the time. The high specificity of the nitrite test informs the clinician that a low false-positive rate exists. As can be observed in the figure, and as often happens in clinical practice, using a combination of diagnostic tests can significantly increase the overall sensitivity and specificity, and thus the accuracy of the diagnosis.

Often in clinical medicine, clinicians are faced with a positive test result, and the next question asked is, "among patients with a positive test result, what proportion will actually have the disease?" This is known as the *positive predictive value* of the test. The *negative predictive value* of a test relays the probability that if the test is negative, the patient does not have the disease. It is important to remember that the predictive value of a test is affected by both the prevalence of the disease in the population being considered, and, if the disease in uncommon, the specificity of the test being used. Higher disease prevalence generally leads to an increase in the positive predictive value of a test. However, as most diseases are rare, tests with higher specificity likely have a greater impact on increasing the positive predictive value for a given test. These values can be easily calculated as demonstrated in *Fig. 17-6*.

Likelihood ratios (LR) are another measure of test performance that helps the clinician utilize the results of a given test diagnostically. When presented with a patient with particular signs and symptoms, the clinician has an initial assessment of the probability, (pretest probability) of a particular disease. A diagnostic test ordered to help determine the presence of a disease may be positive or negative. LRs help inform the clinician as to how much the test result should shift his/her initial assessment of the probability (pretest probability) of the disease being present or absent to the posttest probability. Strong, conclusive tests yield very big or very small LR's. Weak, inconclusive tests yield modest LR's, close to 1.0. For example, following the results of a positive test result, if the positive LR is found to equal 1, then there is no change in the likelihood of the disease. If the LR was 10 after the positive test result, then the posttest odds of the disease is equal to the likelihood ratio multiplied by the pretest odds. The odds of disease can be calculated from the probability as probability/1- probability. The higher the LR, the better the test is for ruling in a diagnosis. These ratios depend upon the validity of the test being ordered in distinguishing who has

Figure 17-6

Calculation of sensitivity, specificity, positive and negative predictive values.

	Population	
Test results	**With disease**	**Without disease**
Positive	True positive (TP) = have disease and test result positive	False positive (FP) = no disease and test result positive
Negative	False negative (FN) = have disease and test result negative	True negative (TN) = no disease and test result negatve

$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad \text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Positive predictive value} = \frac{TP}{TP + FP} \qquad \text{Negative predictive value} = \frac{TN}{FN + TN}$$

the disease in question from who does not; hence these ratios may be derived from sensitivity and specificity of the test (see below). LR's may be more useful than sensitivity and specificity in certain situations. LR's can be calculated for tests without dichotomous results such as those that are resulted as "positive, intermediate, or negative." The results of several diagnostic tests may be combined to provide a single LR. Finally, with some relatively simple calculations, the posttest probability of a disease can be calculated using the LR and the pretest odds of the disease. Formulas for calculating LRs follow.

$$\text{Positive LR}: \frac{\text{probability of an individual with the disease having a positive test}}{\text{probability of an individual without the disease having a positive test}}$$

$$\text{Negative LR}: \frac{\text{probability of an individual with the disease having a negative test}}{\text{probability of an individual without the disease having a negative test}}$$

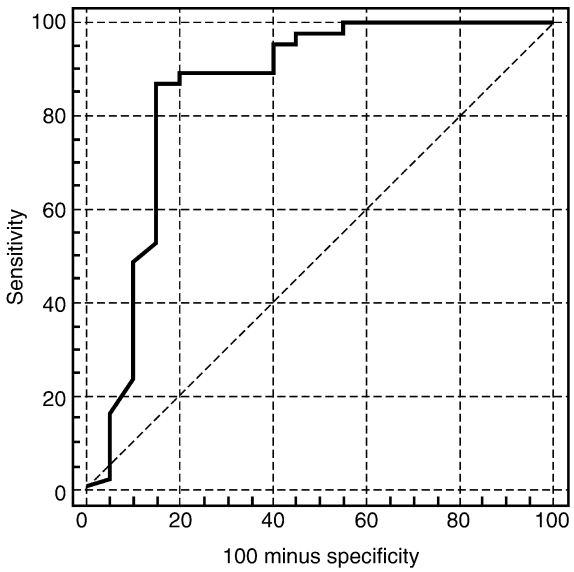
For dichotomous tests:
Positive LR: sensitivity/(1-specificity)
Negative LR: (1-sensitivity)/specificity

Receiver operator characteristic (ROC) curves, originally developed in the field of electronics, allow for a graphical display of the trade-off between sensitivity and specificity for diagnostic tests with ordinal or continuous results, in which several values of sensitivity and specificity are possible. Several cutoff points are determined and the sensitivity and specificity are determined at each point. The sensitivity (or true-positive rate) is graphed on the Y axis as a function of 1-specificity (the false-positive rate) on the X axis. Tests with values falling in the upper left corner of the graph are considered ideal (100% true positives and no false positives). If a test followed the diagonal line from the lower left corner to the upper right corner, it would be considered useless–on this diagonal line the true-positive rate equals the false-positive rate. The area under the ROC curve can range from 0.5 for a worthless test to 1.0 for a perfect test.

Filler et al. made use of both an ROC curve and LRs in a study examining nonminimal change nephrotic syndrome in children referred to a tertiary care medical center in Canada, as can be seen in the *Fig. 17-7* included from this study (14). The authors noted an increase in the incidence in focal segmental glomerulosclerosis (FSGS) over a 17 year study period. Based on the International Study of Kidney Disease in Children (ISKDC), the approach to a child with new-onset nephrotic syndrome has been to perform a kidney biopsy if the disease was unresponsive to a standard dose of corticosteroid therapy of at least 28 days in duration (15). Given the increased incidence of FSGS observed in their study, Filler et al. considered that kidney biopsies to distinguish FSGS from minimal change disease may need to be performed sooner after presentation with the nephrotic syndrome, and based on their data, investigated the ideal time to perform a kidney biopsy for the detection of nonminimal change nephrotic syndrome (i.e., FSGS). The clinical feature (as opposed to a "diagnostic" test) represented in the ROC curve is "time to remission after starting corticosteroid therapy." By plotting and comparing the various sensitivities and specificities of values for "time to remission," the authors concluded that the cutoff of 28 days was statistically the best point when a biopsy should be considered. At this cutoff point, the detection of "true positives" (nonminimal change nephrotic syndrome) was maximized, and the detection of "false-positives" (minimal change nephrotic syndrome) was minimized. The LRs for "time to remission" are also listed in the manuscript. The longer a patient took to enter remission, the

Figure 17-7

ROC curve for the detection of non-MCNS in relation to remission time (14). The graph plots the true-positive rate expressed as sensitivity (%) as a function of the false-positive rate (*100-specificity* [%]) at different cutoff points. Area under the ROC CURVE = 0.859; SE = 0.057; 95% confidence interval, 0.793 to 0.911.



higher the positive LR for the diagnosis of a nonminimal change pathology underlying the nephrotic syndrome.

Screening Programs

Screening programs are designed to detect or diagnose a disease as early as possible, in hopes of improving the prognosis. A common screening program is the use of mammography for the earlier detection of breast cancer. Detecting breast cancer early, before it is "symptomatic" or advanced, has been proven to improve outcomes and survival in older women. A sampling of the many factors that must be considered when evaluating the feasibility and effectiveness of a screening program follows.

Disease

Is there a preclinical phase of the disease – a time when the disease is present but clinical symptoms have not yet manifested? Does intervening earlier in the natural history of the disease make treatment easier and/or improve morbidity or mortality? Is the disease prevalence high enough to make a screening program cost-effective?

Test

Does a screening test exist with acceptable sensitivity and specificity for a screening program – are the false-positive and false-negative rates acceptable? Is the test acceptable to the population – will they consent to the test? Do the benefits gained from early diagnosis of the disease outweigh the cost of the test?

Person/Population

Does screening/early diagnosis improve outcomes for an individual and the population? Do those with earlier diagnosis of the disease comply with treatment recommendations and regimens – how many of those who screen positive receive a final diagnosis and treatment? Is there an improvement in the quality of life in those screened?

Ethics in Research

The foundation of medical research is to help improve the lives of patients. In keeping with this, it is important to be familiar with the various terms and concepts related to the responsible and ethical conduct of medical research.

Conflicts of interest may occur between a researcher's interest in advancing medical knowledge versus his/her self-interest in fame, prestige, academic or financial advancement. Transparency of contractual obligations and relationships is mandatory, and in some circumstances, these must be dissolved, or the researcher should not participate in a related project. For example, a conflict of interest would exist if a physician was paid by a pharmaceutical company as a medical consultant, and this physician also served as a primary investigator in a clinical trial of a medication produced by the same drug company. Conflicts of interest may also occur in the clinician-investigator role when a researcher is also the clinician for the research subject. In such situations, what may be best for the research project may not be best for the individual patient. In such conflicts, the physician is expected to do what is best for the patient.

Several types of scientific *misconduct* have been described. Scientists have a responsibility to report misconduct, and institutions have the responsibility to investigate the misconduct, and protect the person alleging the misconduct. *Fabrication/forgery* is the invention and reporting of data that does not exist from an experiment that was not performed. *Falsification/fraud* is the manipulation of research data such that what is reported misrepresents the actual findings. *Plagiarism* is the presentation of another person's words or ideas as one's own or without giving appropriate credit to the original author(s).

Research on *human subjects* requires special ethical considerations and protections. Respect for research participants requires investigators to obtain informed consent or assent, maintain privacy and confidentiality, and protect the vulnerable. *Informed consent* involves relaying an unambiguous description of the research project and allowing the subject to make an informed decision regarding participation. It must be clearly stated that the patient will be involved in a research project and participation is entirely voluntary. A clear description of the potential risks and benefits, and any compensation, should be provided. For pediatric patients participating in research, *assent* is also required. Children cannot legally give permission to participate in a research study, nor can they give "consent" as consent implies

full understanding. However, ethicists and medical and legal professionals have agreed that children should be routinely asked if they agree (assent) to participate in a research project, and their wishes should be respected. Other *vulnerable* populations that require special attention to ensure their safety include prisoners, pregnant women and their fetuses and embryos, and people with impaired capacity to make decisions, such as the mentally ill. During the consent process, *confidentiality and privacy* procedures utilized by the study should be outlined. The extent of confidentiality should be disclosed – the subject should understand who will and who will not have access to the data.

Evaluating the Literature: Rating the Strength of Scientific Evidence

Health care decisions should be based on research-based evidence. Whether the individual nephrologist is making a clinical decision or a national organization is developing clinical practice guidelines, efforts should be made to systematically assess the strengths of scientific evidence related to a particular clinical diagnostic or treatment plan. Guidelines first developed more than 20 years ago at the Department of Clinical Epidemiology and Biostatistics at McMaster University first introduced tools to allow clinicians to critically review original articles on etiology, diagnosis, prognosis, and therapy (16). In the following decade, the series was widely read and cited, was modified for use by the general public, and was published in clinical epidemiology texts (17). At the same time, clinicians at McMaster University and across North America continued to expand and improve the guidelines. Their focus has expanded to include clinicians' ability to access, summarize, and apply information from the literature to everyday clinical problems, transforming the Readers' Guides to Users' Guides (18–41).

Such systematic approaches have also been adapted to assess entire bodies of research on particular subjects. In 1999, the U.S. Congress directed the Agency for Health Care Policy Research and Quality to identify methods to assess health care research results. The results of that effort were published in a report entitled "Systems to rate the strength of scientific evidence." The goals of this project were to describe systems to rate the strength of scientific evidence, including evaluating the quality of individual articles that make up a body of medical evidence related to a particular disease, allowing for the most informed medical assessments and decision making. The report provides framework for the clinician regarding the evaluation of the various types of study design as described in this chapter (42).

In summary, the busy clinician can afford to be selective in reviewing the literature. In rating the strength of scientific evidence in evaluating a specific clinical problem or treatment, one needs to be selective. The simplest criteria for choosing which studies to read in detail or which to weigh heavily in evidence are clinical relevance and methodologic soundness. This chapter has introduced a simple framework for evaluating such features in the context of sound clinical research methodology and has outlined the most recent guidelines for assessing the strength of scientific evidence for making decisions in clinical care. These tools for systematic assessment of existing research can also guide the clinical investigator toward areas that require further study in which current evidence for treatment or outcomes is scant.

Suggested Reading

DeAngelis C. Introduction to Clinical Research. New York, Oxford University Press, 1990.

Hulley SB, Cummings SR. Designing Clinical Research, 2nd edn. Philadelphia, Lippincott Williams & Wilkins, 2001.

Clinical Epidemiology

Gordis L. Epidemiology, 2nd edn. Philadelphia, Saunders, 2000.

Fletcher RH, Fletcher SW, Wagner EH. Clinical Epidemiology: The Essentials, 3rd edn. Baltimore, Williams & Wilkins, 1996.

Lilienfeld AM, Lilienfeld DE. Foundations of epidemiology, 2nd edn. New York, Oxford University Press, 1980.

Study Design

Cook TD, Campbell DT. Quad-Experimentation: Design and Analysis Issues for Field Settings. Boston, Houghton-Mifflin, 1979.

Meinert CL. Clinical Trials: Design, Conduct and Analysis. New York, Oxford University Press, 1986.

Schiesselman JJ. Case-Control Studies: Design, Conduct, Analysis. New York, Oxford University Press, 1982.

General Statistic References

Dawson-Saunders B, Trapp RG. Basic and Clinical Biostatistics, 3rd edn. New York, Lange Medical Books-McGraw-Hill, 2001.

Fleiss JL. Statistical Methods for Rates and Proportions. New York, John Wiley & Sons, 1981.

Glaser AN. High Yield Biostatistics. Pennsylvania, Williams & Wilkins, 2001.

Hollander M, Wolfe DA. Nonparametric Statistical Methods, 2nd edn. New York, Wiley, 1999.

Kleinbaum DG, Kupper LL. Applied Regression Analysis and Other Multivariate Methods, 3rd edn. Pacific Grove, Duxbury Press, 1998.

Tabachnick BG, Fidel LS. Using Multivariate Statistics, 4th edn. Boston, Allyn and Bacon, 2001.

Zar JH. Biostatistical Analysis, 4th edn. Upper Saddle River, NJ, Prentice-Hall, 1999.

References

1. Cook TD, Campbell DT. Quasi-experimentation: design and analysis issues for field settings. Boston, Houghton Mifflin, 1979.
2. Furth SL, Neu AM, Colombani P et al. Diabetes as a complication of FK506 in pediatric renal transplant patients. *Pediatr Nephrol* 1990;10:64–6.
3. Fadrowski JJ, Pierce CB, Cole SR et al. Hemoglobin decline in children with chronic kidney disease: Baseline results from the chronic kidney disease in children prospective cohort study. *Clin J Am Soc Nephrol* 2008;3:457–62.
4. Fored CM, Ejerblad E, Lindblad P et al. Acetaminophen, aspirin, and chronic renal failure. *N Engl J Med* 200l;345:l80l–8.
5. Furth SL, Cole SR, Moxey-Mims M et al. Design and methods of the chronic kidney disease in children (CKiD) prospective cohort study. *Clin J Am Soc Nephrol* 2006;1(5):1006–15.
6. Wong CS, Jelacic S, Habeeb RL et al. The risk of the hemolytic-uremic syndrome after antibiotic treatment of Escherichia coli 0157: H7 infections. *N Engl J Med* 2000;342:1930–6.
7. Hogg RJ, Lee J, Nardelli N, Julian BA et al. Clinical trial to evaluate omega-3 fatty acids and alternate day prednisone in patients with IgA nephropathy: Report from the Southwest pediatric nephrology group. *Clin J Am Soc Nephrol* 2006;1:467–74.
8. Wang R, Lagakos SW, Ware JH et al. Statistics in medicine-reporting of subgroup analyses in clinical trials. *N Engl J Med* 2007;357 (21):2189–94.
9. Furth SL, Hwang W, Yang C et al. Relation between pediatric experience and treatment recommendations for children and adolescents with kidney failure. *JAMA* 2001;258:1027–33.
10. What is Epi info? Accessed 16 May 2008. Available: htrp://www.cdc.gov/epiinfo/.
11. U.S. Renal Data System. USRDS 2007 annual data report: Atlas of chronic kidney disease and end-stage renal disease in the United States. Bethesda, MD, National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, 2007.
12. North American Pediatric Renal Trials and Collaborative Studies. NAPRTCS 2007 annual report. Accessed 16 May 2008. Available at: https://web.emmes.com/study/ped/annlrept/annlrept.html.
13. Committee on Quality Improvement, Subcommittee on Urinary Tract Infection. Practice parameter: The diagnosis, treatment, and evaluation of the initial urinary tract infection in febrile infants and young children. *Pediatrics* 1999;103:843–52.
14. Filler G, Young E, Geier P et al. Is there really an increase in non-minimal change nephrotic syndrome in children? *Am J Kidney Dis* 2003;42(6):1107–13.
15. International Sudy of Kidney Disease in Children: The primary nephrotic syndrome in children. Identification of patients with minimal change nephrotic syndrome from initial response to prednisone. A report of the International Study of Kidney Disease in Children. *J Pediatr* 1981;98:561–4.
16. Department of Clinical Epidemiology and Biostatistics. McMaster University. How to read clinical journals. I. Why to read them and how to start reading them critically. *Can Med Assoc J* 1981;124:555–8.
17. Sackett DL, Haynes RB, Guyatt GH et al. Clinical Epidemiology: A Basic Science for Clinical Medicine, 2nd edn. Boston, Little, Brown, 1991.
18. Guyatt GH, Rennie D. Users' guides to the medical literature (editorial). *JAMA* 1993;270:2096–7.
19. Oxman AD, Sackett DL, Guyatt GH. Users' guide to the medical literature: I. How to get started. *JAMA* 1993;270:2093–5.
20. 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;270:2598–601.
21. 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.
22. Jaeschke R, Guyatt G, Sackett DL. Users' guide to the medical literature: III. How to use an article about a diagnostic test A. Are the results of the study valid? *JAMA* 1994;271:389–91.
23. Jaeschke R, Guyatt G, Sackett DL. Users' guide to the medical literature: III. How to use an article about a diagnostic test B. What are the results and will they help me in caring for my patients? *JAMA* 1994;271:703–7.
24. Levine M, Walter S, Lee H et al. Users' guide to the medical literature: IV How to use an article about harm. *JAMA* 1994;271:1615–9.
25. Laupacis A, Wells G, Richardson WS et al. Users' guide to the medical literature: V. How to use an article about prognosis. *JAMA* 1994;272:234–7.
26. Oxman AD, Cook DJ, Guyatt GH. Users' guide to the medical literature: VI. How to use an overview. *JAMA* 1994;272:1367–71.
27. Richardson WS, Detsky AS. Users' guide to the medical literature: VII. How to use a clinical decision analysis. A. Are the results valid? *JAMA* 1995;273:1292–5.
28. Richardson WS, Detsky AS. Users' guide to the medical literature: VII. How to use a clinical decision analysis. B. What are the results and how will they help me in caring for my patients? *JAMA* 1995;273:1610–3.
29. Hayward RSA, Wilson MC, Tunis SR et al. Users' guide to the medical literature: VIII. How to use clinical practice guidelines. A. Are the guidelines valid? *JAMA* 1995;274:570–4.
30. Hayward RSA Wilson MC, Tunis SR et al. Users' guide to the medical literature: VIII. How to use clinical practice guidelines.B. What are the recommendations and how will they help you in caring for your patients? *JAMA* 1995;274:1630–2.
31. Guyatt GH, Sackett DL, Sinclair JC et al. Users' guide to the medical litetature: IX. A method for grading health care recommendations. *JAMA* 1995;274:1800–4.
32. Naylor CD, Guyatt GH. Users' guide to the medical literature: X. How to use an article reporting variations in the outcomes of health services. *JAMA* 1996;275:554–8.
33. Dans AL, Dans LF, Guyatt GH et al. Users' guides to the medical literature: XIV. How to decide on the applicability of clinical trial results to your patient. Evidence-based medicine working group. *JAMA* 1998;279(7):545–9.
34. Richardson WS, Wilson MC, Guyatt GH et al. Users' guides to the medical literature: XV. How to use an article about disease probability for differential diagnosis. Evidence-Based Medicine Working Group. *JAMA* 1999;281(13):1214–9.
35. Berlin JA, Rennie D. Measuring the quality of trials: the quality of quality scales. *JAMA* 1999;282(11):1083–5.
36. McGinn TG, Guyatt GH, Wyer PC et al. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-based medicine working group. *JAMA* 2000;284 (1):79–84.
37. Giacomini MK, Cook DJ. Users' guides to the medical literature: XXIII. Qualitative research in health care B. What are the results and

how do they help me care for my patients? Evidence-Based Medicine Working Group. *JAMA* 2000;284(4):478–82.
38. Richardson WS, Wilson MC, Williams JW Jr et al. Users' guides to the medical literature: XXIV. How to use an article on the clinical manifestations of disease. Evidence-based medicine working group. *JAMA* 2000;284(7):869–75.
39. Shah NR. What is the best evidence for making clinical decisions? *JAMA* 2000;284(24):3127–8.
40. Altman DG. Poor-quality medical research: what can journals do? *JAMA* 2002;287(21):2765–7.
41. Guyatt. Users' guides to the medical literature: essentials of evidence-based clinical practice. Chicago, AMA Press, 2002.
42. West S, King V, Carey TS et al. Systems to rate the strength of scientific evidence. Evidence report/technology assessment No. 47 (Prepared by the Research Triangle Institute-University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011). AHRQ publication No. 02-E016. Rockville, MD, Agency for Healthcare Research and Quality. April 2002. Accessed May 2008. Available at: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hstat1.chapter.70996.