Key Issues in Outcomes Research

Evaluating the Evidence
Is There a Rigid Hierarchy?

P. Michael Ho, MD, PhD; Pamela N. Peterson, MD, MSPH; Frederick A. Masoudi, MD, MSPH

Health outcomes research applies a wide range of methods to identify optimal approaches to determine optimal approaches to determine the effects of healthcare interventions and policies. Critical evaluation skills are necessary to navigate the evidence and identify studies that should influence clinical decision making and policy. A hierarchical “pyramid of evidence” that emphasizes randomized controlled clinical trials (RCTs) has been promulgated as the approach to judging study design and quality. Whereas this hierarchy may be suitable for the evaluation of efficacy, it may be inappropriate for many health outcomes research questions. This article examines the relevance of this construct to questions beyond those of therapeutic efficacy and discusses the essential role of study designs beyond RCTs. The strengths and weaknesses of study designs commonly encountered in the medical literature are reviewed. The article concludes with 2 case studies that apply the concepts reviewed and illustrate the need to match the study design with the research question. Of note, the concepts of internal validity, external validity, confounding, and bias are central to the discussion of the strengths and weaknesses of study designs in the medical literature. Readers not familiar with these concepts should refer to the Appendix in the online-only Data Supplement for clarification.

The Pyramid of Evidence: A Useful Construct?

Traditionally, the design of a study has been considered a principal barometer of the validity of its findings. In this construct, different study designs are considered in the context of a pyramid of evidence, in which studies most susceptible to threats to internal validity reside at the bottom and those least prone reside at the top (Figure). This hierarchy is widely used in discussions about the quality of medical studies and is integrated into the grading of evidence in practice guidelines. However, although the pyramid of evidence is undoubtedly well suited for questions of therapeutic efficacy, it is of limited to no value for many other questions addressed in the medical literature, such as the effectiveness of therapies (ie, how well a treatment works in practice); the association between exposures that cannot be controlled and outcomes; or the safety or harm of a therapy or intervention.

For some questions, RCTs are not feasible or are simply impossible. When RCTs are not available to answer a given question in a specific population, which is frequently the case, readers of the medical literature must determine the extent to which the study design implemented was appropriate for the question asked rather than implementing rigid constructs of study quality.

Observational effectiveness studies provide insights into how an intervention or treatment works in a population representative of clinical practice. Effectiveness studies assess the extent to which trial findings can be extended to patients seen in clinical practice by evaluating the intervention in populations not included in the efficacy studies. For example, the initial clinical trials of β-blockers for acute myocardial infarction (AMI) were conducted in narrowly defined patient populations with AMI. However, cohort studies suggest similar benefits of β-blockers in populations not resembling those in the clinical trials, including older patients and those with potential relative contraindications. Reliance on RCTs alone would limit our understanding of the manner in which evidence from efficacy trials translates into clinical practice. Thus, effectiveness studies are an important complement to RCTs in addressing questions about the benefits of therapy.

RCTs do not provide information on the manner in which results are implemented in clinical care. Specifically, although RCTs can establish what should be done, other study designs are necessary to identify underuse, overuse, and misuse, all of which have the capacity to erode the potential benefits of evidence-based therapy. An example of the power of observational designs is illustrated in a study of the use of antiplatelet and antithrombotic agents for acute coronary syndromes. In an observational analysis of patients with non–ST-segment elevation myocardial infarction (non-STEMI), Alexander and colleagues found that excessive medication doses were administered to 33% of patients treated with heparin, 14% of those treated with low-molecular-weight heparin, and 27% of those treated with glycoprotein IIb/IIIa inhibitors. Among patients older than 75 years treated with glycoprotein IIb/IIIa inhibitors, excessive dosing was the rule (64.5% of patients). Perhaps not surprisingly, excessive dosing was associated with significantly
higher risks of major bleeding and was estimated to account for 15% of major bleeding episodes. Because dosing protocols are strictly specified in RCTs, this harmful practice could not have been identified without a subsequent observational study focused on practice patterns in “real-world” settings.

For questions evaluating exposures or risk factors that cannot be controlled, nonrandomized designs are again the study design of choice. For example, a recent analysis from the INTERHEART study found an independent graded relationship between waist-to-hip ratio and risk of myocardial infarction even after adjustment for body mass index using a case-control study design. In addition, several studies have demonstrated a graded inverse relationship between renal function and patient outcomes after hospitalization for acute coronary syndrome. In these scenarios, it is not possible to randomize patients to waist-to-hip ratios or degrees of renal dysfunction.

In general, RCTs are conducted to assess the efficacy of treatment. Although safety end points are invariably included, the sample sizes of RCTs are calculated on the basis of estimates of the impact of therapy on efficacy end points. Because adverse events generally occur less frequently than efficacy end points, clinical trials typically have limited power to detect important safety signals. Furthermore, an intervention may result in an unforeseen event that is not included in the data collection of a RCT and thus is not detected.

Several study designs have been employed to identify important safety concerns. For example, case-control studies were instrumental to understanding the relationship between appetite-suppressant drugs and pulmonary hypertension and between aspirin and Reye syndrome in children. Meta-analyses of clinical trials, by increasing the numbers of patients exposed to therapy, can also play a role. For example, a recent meta-analysis of 42 clinical trials enrolling 27,847 patients with diabetes mellitus identified an increased risk of myocardial infarction and a trend toward a greater risk of cardiovascular death in those treated with the antihyperglycemic agent rosiglitazone. Although a subsequent analysis using alternative meta-analytic approaches and including a larger number of studies did not confirm these results, the original study illustrates how meta-analysis can be used to assemble markedly larger sample sizes, thus increasing the possibility of identifying adverse events of therapy.

Although RCTs are a critical component of the evidence base for contemporary medical practice, this design is not feasible or appropriate for all questions. Depending on the question posed, especially in situations in which an RCT is either not feasible or impractical, other study designs may be superior. Therefore, the appropriate use and the strengths and limitations of the wide range of available designs must be understood.

Overview of Study Designs

Research studies can be broadly categorized as experimental or observational. In experimental studies, the investigator controls the allocation of participants to particular study groups, which is often but not always performed randomly. In observational studies, the investigator “observes” phenomena or outcomes without specifically assigning a treatment.

Experimental Studies

Randomized Controlled Trials

RCTs are considered the criterion standard for the assessment of whether a treatment or intervention is efficacious—in other words, whether the treatment “works” under ideal conditions (Table). In RCTs, in which subjects are randomly allocated to receive active treatment or a control (placebo, no treatment, or even an active control), each patient has an equal chance of assignment to intervention or control group. Randomization produces comparable study groups in terms of measured and unmeasured variables apart from the intervention itself. Thus, confounding is not a threat to the internal validity of a properly conducted RCT, and any differences in outcome that occur between groups can be attributed to the intervention.

Despite this significant strength, RCTs also have important limitations. They are typically expensive, time consuming, and designed to answer a single question or small number of questions about treatment efficacy that are usually narrow in scope (eg, whether aspirin reduces the risk of second events in middle-aged patients after myocardial infarction). Thus, the RCT design may be impractical or even inappropriate to address questions beyond therapeutic efficacy in a well-circumscribed population.

Importantly, RCTs usually employ strict inclusion or exclusion criteria that are intended to maximize the internal validity of the study. However, these criteria intrinsically limit the extent to which the findings of an RCT apply to patients seen in clinical practice (ie, the extent to which the findings are “externally” valid). In fact, an inverse relationship often exists between internal and external validity in clinical trials. The investigator wants to demonstrate that the intervention/treatment works, which frequently requires that the study be performed in a well-defined population that has few competing risks and a low likelihood of dropping out of the study. Additionally, the intensive follow-up necessary for
Table. Strengths and Weaknesses of Study Designs in the Hierarchy of Evidence

<table>
<thead>
<tr>
<th>Study Design</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCTs</td>
<td>Internal validity</td>
<td>Not feasible in many circumstances</td>
</tr>
<tr>
<td></td>
<td>Control over intervention</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control over end points</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prospective cohort</td>
<td>Establishes sequence of events</td>
<td>Requires large sample sizes</td>
</tr>
<tr>
<td>studies</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Can assess several outcomes</td>
<td>Expensive</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control over subject selection</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control over measurement of variables and end points</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retrospective cohort</td>
<td>Maintains sequence of events</td>
<td></td>
</tr>
<tr>
<td>studies</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Study several outcomes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Less expensive and of shorter duration (than prospective cohort)</td>
<td></td>
</tr>
<tr>
<td>Case-control studies</td>
<td>Useful for rare outcomes</td>
<td>High confounding risk</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Short duration</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Small, inexpensive</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Can study several exposures</td>
<td></td>
</tr>
<tr>
<td>Cross-sectional</td>
<td>May study multiple outcomes and exposures</td>
<td></td>
</tr>
<tr>
<td>studies</td>
<td>Short duration</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Can provide an estimate of prevalence of an outcome</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Less feasible for rare predictors or outcomes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>High risk of confounding and bias</td>
<td></td>
</tr>
<tr>
<td>Case series</td>
<td>Convenient</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inexpensive</td>
<td></td>
</tr>
</tbody>
</table>

most study protocols often bears little resemblance to follow-up patterns in usual clinical practice. Thus, although rigorous RCTs establish the efficacy of a specific therapy in a limited patient population receiving close follow-up, they often leave remaining questions about the impact of therapy in populations not resembling those enrolled in the trial or of variations of the intervention that were not specifically studied. The need for more practical clinical trials—those that select clinically relevant alternative interventions applied to a diverse patient population in heterogeneous care settings and measuring a broad range of health outcomes—has been recognized.17

The use of RCTs is not confined strictly to drugs or devices but can be useful to assess strategies of care. For example, several trials have compared strategies of care (early invasive versus early conservative) after acute coronary syndromes.18,19 In cases of nontherapeutic strategies (eg, patient education), trials may implement cluster randomization, in which the intervention is randomized according to units such as hospitals, clinics, or physicians rather than on the patient level.20 This approach is often used when researchers are concerned about “contamination” or when it is likely that patients within clusters are all likely to be influenced by the availability of the strategy regardless of whether it is intended to be applied to that patient. However, cluster randomization provides less precise effect estimates than individual-level randomization because of a loss of statistical efficiency with the former approach.20

Noninferiority Trials
Noninferiority trials are increasingly common variants of the RCT in cardiology. The objective of such trials is to demonstrate that a new treatment is no worse than an established treatment by a predefined margin.21 It is used in clinical scenarios when an efficacious therapy exists and a new treatment or intervention has important potential advantages over current standard therapy with respect to tolerability, safety, convenience, or cost. The existence of an established efficacious therapy precludes a placebo control on an ethical basis; thus, the established therapy must be used as a comparator.

The Valsartan in Acute Myocardial Infarction Trial (VALIANT) provides an example of this design.22 VALIANT was conducted to test the primary hypothesis that an angiotensin receptor blocker would result in better survival than treatment with an angiotensin-converting enzyme (ACE) inhibitor in patients with myocardial infarction complicated by heart failure or left ventricular systolic dysfunction. However, it also included an explicit assessment for the noninferiority of the angiotensin receptor blocker compared with ACE inhibitors, which at the time represented the established form of renin-angiotensin blockade in this patient population. After a mean follow-up of slightly >1 year, no difference was found in total mortality between valsartan and captopril treatment (hazard ratio = 1.00; 97.5% confidence intervals, 0.90 to 1.11). The upper limit of the confidence interval for this comparison was within the prespecified noninferiority margin of a hazard ratio of 1.13 (P = 0.004 for the null hypothesis of inferiority), suggesting a very small
likelihood that valsartan was >13% less effective than captopril on a relative basis. Furthermore, valsartan was generally better tolerated than captopril.

In the appraisal of noninferiority trials, one of the most important considerations is the prespecified margin of noninferiority. Although based on the impact of the established therapy on outcomes and statistical factors, the noninferiority margin is largely arbitrary. Greater margins result in smaller sample size requirements and increase the probability that the new treatment will be deemed noninferior to the established treatment, at the expense of the precision of the findings.

An additional consideration in the assessment of noninferiority studies is the choice of comparators. Some noninferiority trials have been criticized for using weak comparators, placing the new intervention being evaluated at an unfair advantage. Ultimately, if it is presumed that the noninferiority margin selected is reasonable and the comparator is fair, the demonstration of noninferiority permits the conclusion that “the new intervention is not worse than the established therapy by more than a pre-specified amount.”

The absence of a placebo arm in noninferiority trials has other important implications. First, an explicit comparison between the new therapy and placebo is not possible and thus must be assumed. Furthermore, the appropriateness of exposing patients to an experimental treatment that under the best circumstances will be found to be “no worse” than the standard of care has been questioned.

A brief mention of industry sponsorship of trials is warranted. Such sponsorship does not by any means invalidate the findings of a trial. On the contrary, many important clinical trials would not have been possible without manufacturer support. However, even contemporary manufacturer-sponsored cardiovascular trials are more likely to report positive results than those funded by nonprofit organizations. Furthermore, the choice of treatment, comparators, and end points in industry-sponsored studies is designated with the primary objective of achieving an indication for the use of the drug or device. The extent to which these important parameters are aligned with those dictated by larger societal interests may vary substantially.

**Quasi-Experimental Studies**

In some situations, an experimental design is not feasible because of ethical or logistical constraints. For example, it may be impractical to assess the impact of policy interventions or the publication of new study findings on outcomes because the exposure cannot be controlled. Similarly, a control group may not be available. In such circumstances, nonrandomized quasi-experimental studies are often the most appropriate, if not the only, design possible to assess causality. Among the many quasi-experimental designs, those commonly employed include (1) the pretest/posttest study without controls, (2) the interrupted time series design, and (3) the pretest/posttest study with a contemporaneous control.

In the pretest/posttest study design, single observations are made before and after the intervention. An example is the study evaluating the Guidelines Applied in Practice quality improvement initiative in Michigan for AMI care. In this study, AMI quality-of-care measures (eg, evidence-based medications and counseling) were assessed at baseline among eligible patients and then remeasured after introduction of clinical care tools designed to improve performance on these quality-of-care measures. Between baseline and remeasurement, significant improvements were found in quality of care during hospitalization and at hospital discharge, potentially attributable to the intervention. Because of the absence of an adequate control group, natural temporal trends or concurrent interventions could not be excluded as potential explanations for the findings.

A time series study design uses multiple observations before and after an event or intervention, allowing for an assessment of temporal trends. Stability of the outcome before and a marked change in the outcome after the introduction of the exposure under study provides greater confidence that the observed changes are related to the intervention itself. For example, a study performed in Canada evaluated rates of hospitalization for hyperkalemia before and after publication of the Randomized Aldactone Efficacy Study (RALES), which had established the efficacy of spironolactone therapy for patients with severe symptomatic systolic heart failure. Although rates of hospitalization for hyperkalemia were stable at several points in time in the 5 years before publication of RALES, hyperkalemia hospitalizations increased dramatically after publication of RALES and persisted at several points in time thereafter. One possibility to explain this finding, the indiscriminate use of spironolactone in patients at high risk for hyperkalemia, has been confirmed in other studies.

Similarly, the question of the impact of smoking bans on smoking-related morbidity and mortality is poorly suited to a randomized trial. However, quasi-experimental design has been implemented to assess the effects of these policy interventions on public health. A pretest/posttest design with a contemporaneous control was implemented to assess rates of hospital admissions for AMI before and after a public smoking ban in Helena, Montana, compared with the surrounding area. The investigators found that during the 6 months in which the ban was implemented, AMI admissions in Helena fell significantly from an average of 40 to 24. In the surrounding areas, such admissions increased from 12.4 to 18, although this change was not statistically significant. These findings have been subsequently replicated in other communities, lending strength to the hypothesis that smoking bans have a meaningful positive impact on population health.

As noted in the previous examples, a control group strengthens the conclusions of quasi-experimental studies. A control provides some evidence that changes occurring over time were not the result of natural temporal trends or of unmeasured events that occurred contemporaneously with the exposure under study. However, because such control groups are not selected randomly, associations between intervention and outcomes are at some risk for confounding or bias. Another potential limitation of quasi-experimental study designs is regression to the mean, a common statistical phenomenon in which extreme values on any measure at a point in time will probably be less extreme the next time it is
measured. This is particularly a problem if the groups selected for study are identified by their initial “outlier” status with respect to the outcome.

Nonexperimental/Observational Study Designs

In nonexperimental studies, the investigator plays no role in the assignment of subjects to their exposure status and observes the population either prospectively or retrospectively. A wide range of observational study designs are implemented in the medical literature, ranging from cohort studies to case series. As with quasi-experimental studies, because the assignment to the exposure in observational studies is not randomly allocated, confounding, especially unmeasured confounding, is a potential threat to internal validity.

Cohort Studies

Cohort studies follow groups of subjects over time and are often used to identify potential risk factors for outcomes or the impact of treatment in real-world populations. In prospective cohort studies, the investigator defines the sample population and measures exposure variables before the occurrence of outcomes. The Framingham Heart Study is a well-known example of the prospective cohort design that was originally initiated to identify patient characteristics that are risk factors for the development of cardiovascular disease. The investigators recruited 5209 men and women without cardiovascular disease, performed examinations and interviews to document postulated risk factors, and followed participants over time for the occurrence of cardiovascular disease events.

In retrospective cohort studies, the investigator defines the sample population and collects data about both exposures and outcomes that have occurred in the past. An example is a study evaluating the effectiveness of ACE inhibitors among older patients hospitalized with heart failure and left ventricular systolic dysfunction. The baseline characteristics of the patient population and the outcome had already been collected when the analysis was initiated. The investigators found that, after employing statistical techniques to account for measured differences between the treated and untreated patients, the prescription of ACE inhibitors was associated with a significantly lower risk of mortality after 1 year. Although previous RCTs had established the efficacy of ACE inhibitors in selected patients with left ventricular systolic dysfunction, this study provided evidence that the benefits of these drugs may extend to an older population with multiple comorbidities not resembling the subjects of the clinical trials.

In prospective cohort studies, the investigator has control over what exposures and outcomes are measured, whereas this flexibility is not possible in retrospective cohort studies because the measurements have already been performed. For the same reason, however, retrospective cohort studies are usually less expensive to perform, require less time to assemble the cohort, and can realistically study extremely large samples (eg, >17,000 patients were included in the study of ACE inhibitors in older patients with heart failure). A further disadvantage of prospective cohort studies is that patient consent is generally required, which can lead to selection bias or a bias commonly known as the Hawthorne effect. The Hawthorne effect occurs when people who know that they are being observed (such as during a research study) temporarily change their behavior or performance. Neither type of cohort study is ideal when the outcome is rare unless a large sample size is available.

In some cases, a question will be asked within the context of a prospective cohort that the study was not designed a priori to answer. For example, numerous studies have used the Framingham cohort data to address questions that extend beyond the initial intent of the prospective cohort. Such studies share the strengths and limitations of the retrospective cohort design.

Cohort studies have specific strengths and limitations (Table). Both prospective and retrospective designs establish the temporal relationship between exposures and outcomes, thus ensuring that the measurement of the exposure is not biased by the outcome and reducing the likelihood that an association is “effect-cause.” In both types of cohort studies, confounding must be considered as a possible threat to validity. When the study of ACE inhibitors cited above is considered, for example, it is possible that unmeasured factors that motivated the decision not to prescribe ACE inhibitors were predictive of mortality (“confounding by indication”). Statistical methods such as multivariable regression are typically employed to “control” for the influence of potentially confounding variables. Propensity scores represent the likelihood that the subject would receive the treatment, given the subject’s characteristics. They are used as a means of addressing confounding by balancing measured covariates between treated and untreated groups. However, they cannot account for unmeasured confounders.

Case-Control Studies

In case-control studies, the sample is defined by the presence or absence of the outcome of interest, in contrast with cohort studies, in which the sample is defined by the presence or absence of an exposure of interest. After identifying a group of cases and controls, the investigator looks back in time to identify potential exposures of interest. They are used as a means of addressing confounding by balancing measured covariates between treated and untreated groups. A recent study by Go and colleagues implemented a case-control design to study the extent to which the use of preventive cardiovascular medications influence the presentation of coronary artery disease. The investigators assembled groups of patients within a large health system whose first clinical presentation was either AMI (cases) or stable angina (controls). The use of statins and β-blockers (the exposures) was assessed in all subjects with the use of retrospective pharmacy data. After controlling for differences in cardiovascular risk factors, recent use of statins (odds ratio = 0.45) and β-blockers (odds ratio = 0.26) was associated with lower likelihoods of presenting with an AMI. These findings suggest that these medications may protect patients with underlying coronary artery disease from higher-risk clinical presentations.
The selection of an appropriate control group is a particularly important challenge in the conduct of case-control studies. Spurious associations may be found if the control group has an exaggerated exposure status (either too high or low) that does not represent the level of the exposure in the general population without the disease. Ideally, the control group is selected from a population similar to cases in all respects except for the presence of the outcome or from a representative population of all persons without the disease. In the case of the study of drug therapy on initial coronary presentation, both the case and control subjects were individuals without prior coronary artery disease enrolled in the same health system.

Case-control studies are characterized by specific strengths and limitations (Table). This design is particularly well suited for studying rare events. Furthermore, they are typically inexpensive and can generate results in a timely manner. Case-control studies have several limitations. First, such studies are particularly susceptible to unmeasured confounding. Second, if ascertainment of exposure is based on a subject’s recollection of past events, exposures may be misclassified if recollection is inaccurate (so-called “recall” bias), although the use of objective sources of information to assess exposure can ameliorate this concern. Finally, because the study groups are determined on the basis of the outcome, the prevalence of the outcome cannot be determined in a case-control study. For these reasons, case-control studies are typically considered hypotheses generating, although such studies have occasionally changed practice.

A variant of the case-control design can be conducted within randomized trials or cohort studies (“nested case-control” studies). An example is a recently published study evaluating the association between valvular regurgitation and past exposure to dopamine agonists in a population-based cohort of patients in the United Kingdom prescribed antiparkinsonian drugs. Cases with new valvular regurgitation and controls without valvular regurgitation matched by age, sex, and year of entry into the cohort were identified, and the investigators retrospectively ascertained exposure to antiparkinsonian drugs. The rate of valvular regurgitation was increased with current use of ergot-derived dopamine agonists but not other dopamine agonists.

The results of nested case-control studies are less prone to bias than case-control studies because the data are typically collected prospectively, retaining some of the advantages of cohort studies. In addition, the control group is obtained from the same cohort, minimizing the potential for selection bias.

Cross-Sectional Studies
In cross-sectional studies, both exposures and outcomes are measured at a single point in time, and the prevalence of the outcome is compared among those with and without exposure. For example, Hanna et al assessed the association between exposures of hyperlipidemia and lipid-lowering drug therapy and the outcome of atrial fibrillation among patients with left ventricular systolic dysfunction from a multicenter registry. They found that the prevalence of atrial fibrillation was much lower among patients with hyperlipidemia taking statin medications compared with patients with hyperlipidemia but not taking statins or patients without hyperlipidemia. The investigators concluded that lipid-lowering drugs were associated with lower prevalence of atrial fibrillation after accounting for multiple other factors.

Advantages of cross-sectional studies are that multiple exposures and outcomes can be measured simultaneously, disease prevalence can be estimated, and studies can be performed often with relatively few resources in a short period of time. Conversely, cross-sectional studies have important limitations in addition to the potential threats to internal validity posed by confounding or bias. Because exposure and outcome are measured simultaneously, a temporal sequence between exposure and outcome cannot be established. In the aforementioned study, for example, it is possible that because patients with atrial fibrillation were more likely to see a physician, they were more likely to have received treatment for hyperlipidemia. This is not a particular issue if the exposure is a fixed characteristic (eg, patient sex). Because of their limitations, cross-sectional studies typically form the foundation for more definitive studies.

Case Series
A case series is a descriptive study of a relatively small number of people with a disease or condition, all of whom experienced the same exposure (ie, no control group without the condition or not receiving the treatment of interest is available). An example is a study evaluating the benefit of percutaneous alcohol septal ablation for reducing left ventricular outflow tract gradients in patients with hypertrophic obstructive cardiomyopathy. The authors found that this procedure significantly reduced outflow tract gradients and exercise capacity in a series of 50 patients undergoing the procedure.

Case series generally provide weak evidence of causality because they are particularly prone to bias and confounding. Furthermore, the absence of a control group is a critical limitation. However, case series are useful in several situations, including the discovery of new diseases or rare manifestations of disease and the detection of unexpected benefits or risks of a treatment. The findings from a case series are used to generate hypotheses for testing with the use of more rigorous study designs that include a comparison group; however, in occasional circumstances in which the relationships found are extremely compelling, or when identifying a treatment that appears to provide benefits for rapidly fatal conditions, case series may have an important impact on medical practice.

Other Study Designs

Meta-Analyses
Meta-analyses are used to aggregate data from multiple studies, most often when the results of individual studies have been contradictory or inconclusive or when an outcome of interest was too rare in individual studies to generate firm conclusions. Combining findings from multiple studies may provide enough end points or outcomes to answer the question of interest. Although meta-analyses may include observational studies, they are frequently used to summarize data from experimental studies. Systematic reviews, like
Cost-effectiveness analyses, use standardized methods to search for and appraise the available studies to reduce bias. However, systematic reviews do not employ quantitative methods to summarize the results and thus may not provide the same strength of evidence as meta-analyses.47

Meta-analyses can be employed to assess questions of efficacy, safety, or both. A recent meta-analysis evaluated the efficacy and potential harm of antifibrinolytic agents among patients undergoing cardiac surgery.48 Despite the number of trials on antifibrinolytic agents, each of the individual studies had small sample sizes, and questions remained about their efficacy as well as the potential for harm. By combining 138 studies, the authors had enough end points and found that these agents reduced blood loss as well as transfusions without an increase in adverse outcomes.

Several issues related to the selection of the studies included are relevant to the evaluation of meta-analyses. The first pertains to the methodological quality of the studies included. Because the quality of the output can be no better than the inputs (ie, “garbage in” produces “garbage out”), poor-quality primary studies fundamentally undermine a meta-analysis. Separately from study quality, heterogeneity across individual studies or the extent to which results differ among the component studies should also be considered. Some authors argue that meta-analyses pooling significantly heterogeneous studies have limited validity.47 Another issue is publication bias, which is a form of selection bias whereby results that are more interesting, significant, of higher methodological quality, or from prominent groups are more likely to be included. Typically, because positive studies tend to be reported more frequently than negative studies, meta-analyses confined to published results may result in biases in favor of treatment. Publication bias is known to have an important impact on the results of meta-analyses.49

Cost-Effectiveness Studies
Although identifying the incremental efficacy or effectiveness of a treatment or strategy is useful, it does not provide a perspective on cost. The interest in understanding the incremental costs of achieving particular benefits in formal cost-effectiveness studies has grown proportionate to the dramatic growth in healthcare expenditures. Cost-effectiveness studies, which apply decision-analytic methods to compare the costs per unit of effectiveness of different treatment options, provide perspectives on the comparative advantages and disadvantages of these options.46

Cost-effectiveness is particularly relevant to expensive interventions such as implantable cardioverter-defibrillator therapy. In an analysis using data from 6 clinical trials, investigators estimated the costs associated with primary prevention implantable cardioverter-defibrillator therapy.50 Using base-case assumptions, the cost-effectiveness of implantable cardioverter-defibrillator therapy compared with control ranged from $34 000 to $70 200 per quality-adjusted life year gained. In sensitivity analyses, in which the parameters of the base-case scenario were altered over a range of assumptions, the cost-effectiveness ratio remained below $100 000 per quality-adjusted life year presuming that implantable cardioverter-defibrillator therapy resulted in lower mortality for at least 7 years.

Because cost-effectiveness analyses are by their nature hypothetical, concerns have been raised about the arbitrariness of the underlying assumptions. Readers of such studies should consider a number of attributes, including the perspective used, the comparators employed, the time horizon of the analysis, the sources of costs, the use of discounting, and the use of sensitivity analyses to assess the robustness of the findings across a wide range of reasonable assumptions.46 Expert recommendations on guidelines for conducting cost-effectiveness analysis,51 the calculation of the cost-effectiveness ratio,52 and appropriate assessment of quality-adjusted life years53 are available to facilitate the assessment of cost-effectiveness studies.

Qualitative Studies
Although frequently overlooked in discussions of the taxonomy of study designs, qualitative studies generate narrative stories, explanations, typologies of phenomena, and conceptual frameworks.54 They are often used in areas in which data or knowledge is inadequate or in which conventional theories may not be applicable. Qualitative studies can generate ideas and identify relevant variables to be studied subsequently in quantitative studies, or they can be used in a complementary fashion to yield findings that are broader in scope and richer in meaning.

In a novel application of qualitative methods, Bradley and colleagues55 conducted in-depth interviews in 8 hospitals that varied widely in terms of changes in the use of β-blockers for their patients hospitalized with AMI in the 3 years preceding the study. Hospitals with greater improvements in β-blocker use uniquely demonstrated several characteristics, including shared goals for improvement, substantial administrative support, strong physician advocates for β-blocker use, and the implementation of credible data feedback. Although this study could not conclusively prove that these face-valid characteristics were the cause of improvements in β-blocker use, it identified potential elements of successful programs to improve AMI care.

Case Studies
The following 2 case studies are intended to illustrate the strengths and weaknesses of the traditional hierarchy of evidence and of the various study designs. The first, which involves hormone replacement therapy (HRT), illustrates the relevance of the hierarchy of evidence for a question focused on treatment efficacy. The second, which involves improving door-to-balloon times for STEMI, demonstrates the failings of this traditional hierarchy of evidence and highlights the role of other study methods when the question of interest does not focus on treatment efficacy.

Example 1: HRT
In the 1990s, use of HRT was widely believed to confer protection from cardiovascular events on the basis of consistent findings of benefits in prospective cohort studies.56–59 For example, in the Nurses’ Health Study, a prospective cohort study of 59 337 women with 16 years of follow-up,
HRT was associated with a significantly lower risk of major coronary heart disease events. As a result of this body of evidence, HRT was widely prescribed for this indication.

However, long-awaited RCTs completely changed the understanding of the relationship between HRT and cardiovascular outcomes. The Women’s Health Initiative, a RCT, found an excess coronary heart disease risk with estrogen plus progestin in healthy postmenopausal women (hazard ratio, 1.29; 95% confidence interval, 1.03 to 1.69) compared with placebo. In fact, excess risks were found per 10,000 person-years that were attributable to estrogen plus progestin, including more coronary heart disease events, strokes, pulmonary emboli, and invasive breast cancers. Subsequent RCTs corroborated the finding of no cardiovascular benefit with HRT. These RCTs have widely reversed the clinical practice of prescribing HRT for cardiovascular protection.

It is likely that, at least in part, confounding and bias in the observational studies explain the marked differences in conclusions between the early studies that suggested a benefit of HRT and subsequent RCTs. In the observational studies, patients were not randomly allocated to treatment. In general, women who were willing to take HRT likely had healthier lifestyles, and women with healthier lifestyles have reduced risk of cardiovascular disease. Therefore, a healthier lifestyle may have confounded the association between current HRT use and lower cardiovascular events seen in the observational studies. Furthermore, women taking HRT in the observational studies may have been more likely to take medications in general and comply with other healthcare recommendations; in addition, this healthy adherer effect has been associated with improved outcomes even among placebo-treated patients. Therefore, adherent patients may have been more likely to enroll in the observational studies of HRT, and this selection bias could have accounted for the association between current HRT use and cardiovascular outcomes. These 2 threats to internal validity, confounding and bias, highlight the fact that for questions on efficacy of therapy, randomized studies remain the ideal study design. The case of HRT demonstrates that adhering to the traditional pyramid of evidence is very important to overcome the limitations inherent in observational studies when efficacy is assessed.

The ongoing debate notwithstanding, the HRT story may be perceived as signaling the demise of observational studies. As a generalization, this conclusion is unfortunate; with respect to questions of therapeutic efficacy, RCTs are understandably considered superior study designs. However, as discussed previously, nonexperimental study designs are indispensable in answering many questions beyond efficacy.

**Example 2: Door-to-Balloon Times for STEMI**

The evolution of the understanding of how to improve the delivery of primary percutaneous coronary intervention to patients with STEMI illustrates the importance of study designs other than the RCT. Despite data that faster door-to-balloon times improve outcomes for patients presenting with STEMI, substantial variation exists in door-to-balloon times among hospitals, and many do not reach performance targets recommended in guidelines. Although it is clear that improvement is needed, little has been known about how to achieve improvement.

To address this gap in knowledge, Bradley and colleagues conducted interviews among top-performing hospitals in an attempt to understand what high performers were doing to achieve their results. Subsequently, these investigators assessed the relationship between these strategies and door-to-balloon times in a survey of 365 hospitals nationwide. They identified several strategies associated with significantly lower door-to-balloon times, including (1) having emergency medicine physicians activate the catheterization laboratory, (2) having a single call to a central page operator to activate the laboratory, (3) activating the laboratory before hospital arrival, (4) expecting staff to arrive in the catheterization laboratory within 20 minutes after being paged, (5) having an attending cardiologist always on site, and (6) using real-time data feedback. None of these strategies was employed by a majority of hospitals, suggesting that instituting these approaches might result in important improvements in care.

Other studies using various study designs have provided complementary data on improving reperfusion therapy for STEMI. In a before/after quasi-experimental design, Khot and colleagues found that door-to-balloon times dropped by 38 minutes after the institution of emergency department activation of the catheterization laboratory. In another quasi-experimental study, Carstensen and colleagues found that field triage and emergency department bypass were associated with shorter symptom onset to balloon times compared with a concurrent control group. A recent large-scale before-and-after quasi-experimental study suggested the effectiveness of a statewide system for providing reperfusion therapy in North Carolina. Finally, a cross-sectional study of patients presenting with STEMI suggested that suboptimal ECG interpretation and suboptimal understanding of evidence-based indications and contraindications to reperfusion therapy were associated with failure to provide reperfusion therapy, thus identifying possible targets for improvement initiatives.

Although the understanding of how to improve reperfusion therapy for patients with STEMI is not by any means complete, these examples demonstrate the power of many different study designs to address critical questions. Although RCTs were instrumental in establishing the knowledge that reperfusion therapy should be provided to patients with STEMI, quasi-experimental, cross-sectional, and qualitative studies have generated important insights into how this therapy can be delivered in a timely manner to as many eligible patients as possible.

**Conclusions**

A hierarchy of study designs has been proposed on the basis of strength of evidence that these studies provide for causal- ity, and experimental studies, especially RCTs, reside at the pinnacle. This hierarchy is most appropriate in evaluating the efficacy of a treatment or intervention. However, experimental studies are not always feasible or appropriate and are often not well suited to answer important questions, such as those on the safety and effectiveness of therapies in real-world
populations, the impact of risk factors on outcomes, or the effects of policy interventions. Ultimately, the interpretation of the medical literature requires not only the understanding of the strengths and limitations of different study designs but also an appreciation for the circumstances in which the traditional hierarchy does not apply and integration of complementary information derived from various study designs is needed.

Disclosures
Dr Masoudi has served on advisory boards for Takeda, NA; Amgen; and United Healthcare. Dr Masoudi has also received research support from Amgen, has contracts with the Oklahoma Foundation for Medical Quality and the Colorado Foundation for Medical Care, and is an Associate Editor for Journal Watch Cardiology of the Massachusetts Medical Society. The other authors report no conflicts.

References


28. Ho et al Evaluating the Evidence 1683


Evaluating the Evidence: Is There a Rigid Hierarchy?

P. Michael Ho, Pamela N. Peterson and Frederick A. Masoudi

*Circulation.* 2008;118:1675-1684
doi: 10.1161/CIRCULATIONAHA.107.721357

*Circulation* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2008 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the
World Wide Web at:
http://circ.ahajournals.org/content/118/16/1675

Data Supplement (unedited) at:
http://circ.ahajournals.org/content/suppl/2008/11/06/118.16.1675.DC1

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in *Circulation* can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to *Circulation* is online at:
http://circ.ahajournals.org//subscriptions/