The Methods of Comparative Effectiveness Research

Harold C. Sox[1] and Steven N. Goodman[2]

[1]Department of Medicine, The Dartmouth Institute for Health Policy and Clinical Practice, Dartmouth Medical School, Hanover, New Hampshire 03755; email: hsox@comcast.net

[2]Department of Medicine and Health Research and Policy, Stanford University School of Medicine, Palo Alto, California 94305; email: steve.goodman@stanford.edu

Annu. Rev. Public Health 2012. 33:425–45

First published online as a Review in Advance on January 3, 2012

The *Annual Review of Public Health* is online at publhealth.annualreviews.org

This article's doi:
10.1146/annurev-publhealth-031811-124610

0163-7525/12/0421-0425$20.00

Keywords

propensity score, instrumental variable, sensitivity analysis, treatment-response heterogeneity, clinical prediction rule

Abstract

This review describes methods used in comparative effectiveness research (CER). The aim of CER is to improve decisions that affect medical care at the levels of both policy and the individual. The key elements of CER are (*a*) head-to-head comparisons of active treatments, (*b*) study populations typical of day-to-day clinical practice, and (*c*) a focus on evidence to inform care tailored to the characteristics of individual patients. These requirements will stress the principal methods of CER: observational research, randomized trials, and decision analysis. Observational studies are especially vulnerable because they use data that directly reflect the decisions made in usual practice. CER will challenge researchers and policy makers to think deeply about how to extract more actionable information from the vast enterprise of the daily practice of medicine. Fortunately, the methods are largely applicable to research in the public health system, which should therefore benefit from the intense interest in CER.

CER: comparative effectiveness research

PCORI: Patient-Centered Outcomes Research Institute

IOM: Institute of Medicine

Confounding: when a potential cause-outcome relationship is distorted by a second cause whose effect is falsely attributed to the first cause

INTRODUCTION

Clinical researchers are excited about comparative effectiveness research (CER) because of its potential to address research questions that have long been ignored. Although CER studies had appeared over the past 50 years, a program focused on CER did not exist until recently. The American Reinvestment and Recovery Act of 2008 allotted $1.1 billion to support CER. The Patient Protection and Affordable Care Act of 2010 (PPACA, the health reform legislation) created the Patient-Centered Outcomes Research Institute (PCORI). The estimated annual revenue of this public-private organization will be $500 million derived principally from a per-capita tax levied on insurance companies and the Medicare program (45, 46). The Affordable Care Act directs PCORI to set national priorities for CER topics and formulate a "research project agenda."

In 2009, the Institute of Medicine (IOM) (37) defined CER as follows:

> [T]he generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care. The purpose of CER is to assist consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and population levels. (p. 41)

The definition of CER in the Patient Protection and Affordable Care Act [Pub. L. 111–148 (2010)] is similar:

> The term "comparative clinical effectiveness research" means research evaluating and comparing health outcomes and the clinical effectiveness, risks, and benefits of 2 or more medical treatments, services, and items described in subparagraph (B). (p. 630)

The key elements of CER are (*a*) direct comparisons of active treatments; (*b*) study patients, clinicians, and interventions that are representative of usual practice; and (*c*) a focus on helping patients, clinicians, and policy makers to make informed choices. These elements have several implications for the principal methods of CER, which are randomized trials, observational research, and decision analysis.

First, of necessity, observational research will have a prominent place in the PCORI research project agenda. Because differences in treatment effects will be smaller in head-to-head comparisons of active treatments than in placebo-controlled trials, studies will require larger study populations, longer follow-up, or both. These requirements are more easily met with observational studies that use large data sets collected over the course of usual practice.

Second, the results of CER will often sacrifice internal validity for external validity. A central goal of CER is to obtain results that typical clinicians can confidently apply to typical patients (i.e., external validity). The best way to assure external validity is to study interventions, clinicians, and patients that are typical of community practice, which will often mean using data that were not gathered for research purposes and are therefore plagued with missing data and subject to confounding.

Third, CER seeks the information that patients, physicians, and policy makers need to make informed decisions when the choice is between alternatives with uncertain outcomes. The two principal elements of decision making under uncertainty are probabilities and preferences. Obtaining this information in representative practice settings will be difficult unless somehow integrated into the routines of daily practice.

CONTENT AND AUDIENCE

Why should public health practitioners be interested in CER? The IOM and PPACA definitions of CER reflect a medical model, in which individual patients and their physicians make decisions. Interventions directed at the population-level causes of disease appear to be outside the scope of CER, at least as defined in federal law. Decisions about the deployment of resources take place in three domains: the care of individual patients, policy making about clinical care, and public health policy. Most of the elements of decision making are the same in these three domains. Despite this commonality, public health decision making and policy making do differ from individual decision making. The target of public health decision making and public policy is the population (which means less room for considering individual characteristics and preferences), and the decision makers are public officials rather than clinicians and their patients,. In public health practice, the interventions are directed less at the biological mechanisms of disease and more at the behavioral and social conditions that result in disease, unlike policy making and clinical decisions.

That said, many domains overlap. Many therapeutic or preventive treatments are applied to large populations, and clinical policies in these realms have population-level impacts. In prevention, questions about prostate-specific antigen and mammography rely on evidence from CER and are, in many respects, public health issues. Drug safety presents similar concerns; policy decisions about Vioxx and Avandia—which are prescribed by doctors and taken by individuals—were based largely on public health considerations, and the United States Food and Drug Administration (USFDA) leaders have outlined their conception of it as a public health agency (31). Finally, decisions by insurance entities, particularly the Centers for Medicare and Medicaid Services, about which medical procedures to reimburse typically can rely on CER and affect public health both directly and indirectly by shifting resources from one domain of health care to another. Thus, many clinical decisions and policies that depend on CER do not affect only public health, but policy makers often invoke public health considerations to explain their decisions. Finally, the essentials of decision making in the medical care system and the public health system are similar. Decision makers strive to be evidence based. They use similar research methods (1, 10, 22). More importantly, both must make decisions despite imperfect evidence and with uncertainty about the outcome of the decisions. Although the principal focus of this article is CER in a medical setting, we do try to point out applications and caveats for public health decision makers.

A GRAND SYNTHESIS

Figure 1 puts each of the topics of this review into its place in the flow of activities that starts with research and ends with information designed to assist with decisions about patient care (see the figure caption for details). Each element of this review has a place on **Figure 1**.

OBSERVATIONAL STUDIES

An observational study uses data from patient care as it occurs in real life. Sources of data include patients' medical records, insurance claims, and surveys. Here is a hypothetical example: researchers would obtain medical records of patients who had a diagnosis of early-stage prostate cancer. They would search each record for items on a prespecified list of demographic and clinical data, including treatment and vital status 10 years later, enter their findings in a data base, and do a statistical analysis to identify the predictors of survival. Perhaps, after adjustment for differences in age and comorbidity, radical prostatectomy was associated with better survival than was radiation therapy. In reporting this result, the authors would discuss the possibility that radical prostatectomy causes better survival than radiation therapy. They would have to address the possibility that more patients died with radiation therapy because physicians persuaded sicker patients to have radiation therapy and healthier patients to have radical prostatectomy. In this example, being older and sicker confounds the relationship between treatment and outcome because it affects the choice of treatment and also affects treatment outcome. Maybe radical prostatectomy causes better prostate cancer outcomes, but one cannot be sure. This example introduces the topic of confounding by indication, whereby characteristics that

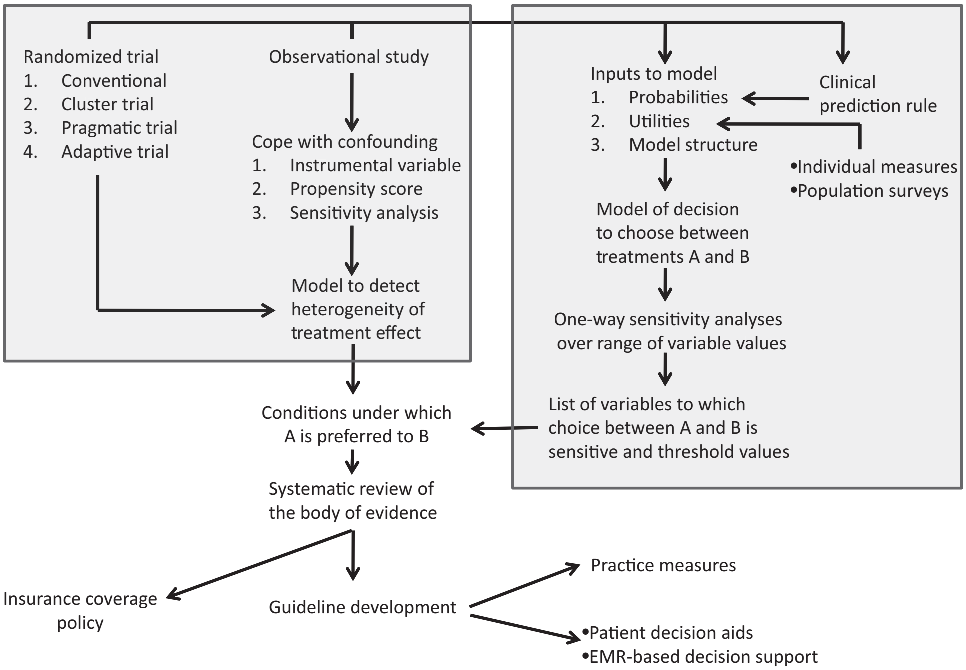
Randomized trial
1. Conventional
2. Cluster trial
3. Pragmatic trial
4. Adaptive trial
Observational study
Cope with confounding
1. Instrumental variable
2. Propensity score
3. Sensitivity analysis
Model to detect heterogeneity of treatment effect
Inputs to model
1. Probabilities
2. Utilities
3. Model structure
Clinical prediction rule
•Individual measures
•Population surveys
Model of decision to choose between treatments A and B
One-way sensitivity analyses over range of variable values
List of variables to which choice between A and B is sensitive and threshold values
Conditions under which A is preferred to B
Systematic review of the body of evidence
Insurance coverage policy
Guideline development
Practice measures
•Patient decision aids
•EMR-based decision support

Figure 1

A grand synthesis. The schematic starts with the principal forms of clinical research, observational studies, and experiments, which are enclosed in the left-hand gray box. The results are inputs into a decision model (*right-hand gray box*). The flow through each gray box leads to methods for identifying the clinical factors that predict a response to an intervention and thence to a list of these factors, which are the ultimate product of CER. The next step is a systematic review of the body of evidence depicted in the left-hand box to find the recurring themes and predictors. The systematic review feeds into the organizations that produce practice guidelines, practice measures, and insurance-coverage decisions. Abbreviations: CER, comparative effectiveness research; EMR, electronic medical record.

influence the choice of treatment also influence outcome.

In observational studies that compare two or more treatments, confounding by indication is especially important. Solving this problem would remove a major roadblock to using observational studies to achieve the goals of CER. Physicians choose a treatment in part because the patient's clinical features match up with the established indications for the treatment. In principle, the best way to address confounding by indication is to compare treatment outcomes in subpopulations in which patients had the same reason for choosing a treatment but had different treatments (e.g., a hypertensive patient who develops heart failure might choose from among several drugs that would treat both conditions). This strategy would require the physician to record the reasons for choosing a treatment. Barriers to this solution include the need to elicit cooperation of busy clinicians and the implicit assumption that the physician's report was complete and accurate.

Data quality will be a problem with many observational CER studies. Public health researchers and epidemiologists usually work with data sets that were obtained prospectively for research. The NHANES (National Health and Nutrition Examination Survey) and the Framingham Study are examples, as would be observational studies designed to test a specific hypothesis. Typically, low-quality data—missing data, missing patients, and nonstandardized data definitions—are relatively minor problems. CER studies of patient care in representative patients and clinical settings will use data obtained during daily practice and recorded in an electronic medical record. Currently, the Veterans Administration health care system is the best example of a large health care system whose clinical sites share the same medical record system. Consortia of large health care organizations have developed data-sharing networks (6), which have received substantial support from the American Recovery and Reinvestment Act of 2009 (65). Electronically stored data from daily practice are likely to be of low quality. Physicians differ from each other in the data set that they obtain for a problem such as chest pain, and individual clinicians typically do not obtain exactly the set of data from patient to patient. Between-clinician and within-clinician variation in phrasing a question means that the meaning of the data they do record may vary. Data quality will be a large problem with observational studies of patients in the community.

Coping with Confounding: Propensity Score Analysis

Propensity score analysis (**Figure 2**) is one approach. Its aim is fundamentally the same as asking clinicians to state the reason for choosing a treatment, but it approaches the analytic problem in a manner that offers both practical and theoretical advantages. Propensity scores are used mainly in contexts where the effects of binary treatment choices (i.e., treatment 1 or treatment 2) are being assessed, although they have been extended to multiple treatments and continuous exposures (62). Propensity score analysis occurs in two stages. The first stage involves developing a statistical model for the chance of the patient receiving a particular treatment, given all their measured characteristics (9). This model produces the propensity score, which is the estimated probability, or propensity for a patient with those characteristics to have received a given therapy. This sets the stage for comparing treatment outcomes in similar patients, where "similar" means having a similar probability of receiving a specified treatment.

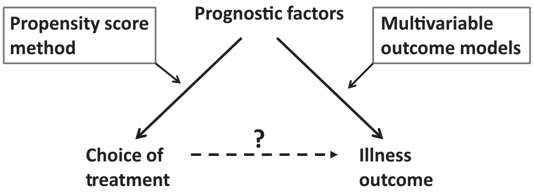

Figure 2

Controlling confounding by adjusting for prognostic factors. The figure has been modified slightly from a slide kindly provided by Alan Brookhart, Ph.D.

The second stage of a propensity score analysis involves developing a model to predict the clinical outcome. Those potential factors that are bundled into the propensity score are controlled for in this step. The most common such model is nonparametric; patients are stratified into quantiles of the propensity score (usually numbering 5–8) and then the treatment effect is estimated within each strata and averaged across strata. With this approach, the treatment effect can easily be summarized as a relative risk, risk difference, or odds ratio.

When applicable, this approach has several advantages over multivariable modeling. A subtle advantage comes from observing how the propensity score is distributed between the two treatment groups. Once the patients are stratified into their propensity score quantiles, it may become apparent that there are very few treated patients in the lower quantiles and very few untreated patients in the upper quantiles. This distribution means that the data set contains very little information about the treatment effect from patients with extreme propensities to be treated. These patients are often excluded from the analysis, which results in more reliable inferences but ones that are restricted to patients with intermediate treatment propensities. Thus, propensity score analysis provides an easy way to visualize the types of patients for which the data are too meager to draw conclusions. This kind of stratified analysis also allows one to see immediately if the treatment effect varies among patients with different propensity scores. If it does, averaging treatment effects across all patients may be unwise. Thus, with stratified propensity

Propensity score: a quantitative measure of the tendency to take an action (e.g., prescribe a medication)

Sensitivity analysis: modeling the influence of an outcome predictor by assigning it an extreme value and recalculating the outcome

score analyses, the investigator can see where the data are sufficient to allow robust inferences and which patients receive most of the treatment effect. Knowing whether an average effect across a population fairly summarizes the response to treatment is a very desirable feature for CER because a heterogeneous response provides opportunities to tailor the treatment to the patient's characteristics and preferences.

A practical advantage of propensity score analysis arises when there are few outcome events, which occurs quite commonly with serious harms of treatment (9). A cardinal rule of statistics is that one must have more outcome events than predictors, and robust inferences require many times more events than predictors. By bundling many potential predictors into one propensity score, almost all the statistical information in the data can be used to estimate the relationship of the treatment to the outcome instead of wasting some of that information by estimating individually the effects of many potentially confounding variables whose effect is of interest only in the aggregate.

The propensity score can also be included in a traditional regression model together with the treatment variable. This approach is more efficient when the model fits well but at a cost of losing the robustness of the nonparametric summary and the insights achieved by forcing the analyst to examine the propensity score distribution.

A final subtle and underappreciated difference between propensity and traditional regression approaches is that they are actually estimating different kinds of causal effects. The propensity score estimates an effect—the "counterfactual" effect—that might be regarded as most relevant to public health. It estimates the effect of the entire population receiving the same treatment compared with that same population receiving the alternative treatment. In contrast, traditional regression approaches estimate the treatment effect on an individual, controlled for all their other characteristics. In practice, these effect estimates usually turn out to be quantitatively close. In situations when they are not, the propensity score analysis is a better estimate of population effect.

The main limitations of propensity analyses are that their use becomes more complex when treatment choices are not binary, and this method also shares the limitation with standard regression modeling that it cannot control for unmeasured confounders. Also, when particularly strong confounders are included in the propensity score, the building of the propensity model requires experience and sophistication.

In summary, propensity score analyses are well suited to CER because they attempt explicitly to model the process by which patients are selected for therapy. They also focus statistical attention on the treatment effect, which minimizes loss of information from estimating separate confounding effects that are peripheral to the purposes of the analysis. They can provide insight into subgroups of patients to whom the average effects do not apply (because their effect is different) or cannot apply because the data in those subgroups are too sparse to estimate an effect.

Coping with Confounding: Sensitivity Analysis

Sensitivity analysis is a general technique that can be applied to the problem of confounding. The basic principle is straightforward: Alter a model for predicting an outcome by changing its structure or the value of a variable and recalculate the predicted outcome. If the outcome does not change much, the variable (or the structural element) probably does not explain variation in the outcome. Similarly, if the coefficient of an independent variable does not change over a wide range of values assigned to another variable, the variable is not an important determinant of the importance of the independent variable.

Sensitivity analysis is also used to test the importance of a hypothetical unmeasured confounder in a multivariable model of an outcome. The analyst inserts a variable into the data set and into the model to represent an unmeasured confounder. In the case of a dichotomous variable (e.g., male versus female gender), the analyst would assign one extreme value to the variable (everyone in the data set is female) and calculate the model. Then the analyst would assign the other extreme value (everyone is male) and recalculate the model. If the coefficient (e.g., weighting factor) of each variable in the model does not change, an unmeasured confounder would not change the model even under extreme assumptions about its value (59).

Coping with Confounding: Instrumental Variables

An "instrument" is something that is an external cause of the intervention or exposure but is by itself unrelated to the outcome. The perfect instrument is randomization itself, which affects the treatment given but has no relation to outcome except through the treatment. An instrumental variable has this same characteristic, but instead of a physical randomization process, one can imagine that the variable represents a "natural" randomization process (11, 12). What we often call "natural experiments" typically have an associated instrument, such as a policy change, a geographic difference, or a natural phenomenon that separates people seemingly randomly into two groups. But true randomization creates groups that either receive a therapy or not. In contrast, two groups with different values of the instrumental variable might have different chances of receiving the intervention, e.g., an 80% chance versus a 40% chance. Comparing outcomes between groups who have different values of the instrumental variable is then like comparing groups that are randomized to receiving an intervention with different probability. This probability is unaffected by individual characteristics, in contrast to the situation in typical clinical practice, where individual characteristics are an important driver of what treatment is given. If such an instrumental variable exists, it provides the opportunity to analyze an observational study as though it were a randomized trial, controlling for both measured and unmeasured confounders.

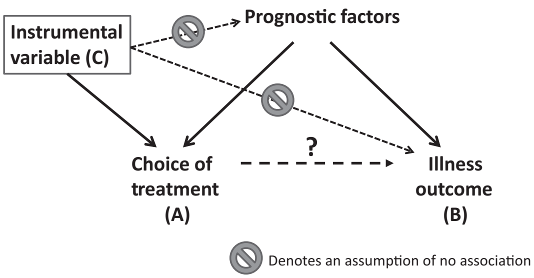
Instrumental variable (C)
Prognostic factors
?
Choice of treatment (A)
Illness outcome (B)
Denotes an assumption of no association

Figure 3

Instrumental variable analysis. The figure has been modified slightly from a slide kindly provided by Alan Brookhart, Ph.D.

Erratum

An instrumental variable analysis divides the population into subgroups according to the value of the instrumental variable. The rates of treatment in these subgroups are compared to see if the treatment rates differ. If the treatment rates do differ, but the outcome in each subgroup is not different, then the treatment cannot cause the outcome. Conversely, if outcomes differ among the groups, we can with some confidence causally attribute that difference in outcomes to the intervention (**Figure 3**).

This method has long been used in economics research but saw comparatively little application in biomedical research until 1994, with the publication of a now-classic article by McClellan and colleagues (51). McClellan and colleagues studied a question that was unlikely to be tested in a randomized study: the extent to which invasive interventions (cardiac catheterization, angioplasty, and bypass surgery) affect the mortality rate from acute myocardial infarction. They hypothesized that distance from a patient's home to the treating hospital would be independent of patient characteristics and outcomes. Their instrumental variable was the distance between home and a hospital with (versus without) invasive cardiac testing and treatment facilities. Their choice fulfills the definition of an instrumental variable because distance from cardiovascular specialty services is associated with the use of those services, but geographic distance per se is not causally linked to the outcome of heart disease. They divided patients into those cared for at nearby (versus distant) hospitals. The clinical characteristics of the two groups were quite similar, a result that is

Instrumental variable: A variable that by itself is unrelated to outcome but that determines the chance that an intervention is used consistent with their hypothesis that the instrumental variable divided the population into two prognostically equivalent groups. Patients relatively near a hospital with catheterization and invasive treatment were much more likely to receive those interventions, but distance to the hospital was not associated with better survival, a result that could rule out a causal relationship between invasive treatment and survival.

Instrumental variable analysis is the only method (aside from actual randomization) that is potentially unaffected by unmeasured confounding, because, like randomization, the receipt of treatment is related to an essentially random process. However, such analyses are always subject to questions about how "good" the instrument is, i.e. whether the outcome is truly independent of the value of the instrumental variable. For example, the assumption in the McClellan et al. study that distance from a hospital was not related to cardiovascular outcomes would not hold if patients more likely to die from heart attacks, in ways we could not measure, deliberately chose to be closer to more fully equipped hospitals.

So, as with true randomization, in which investigators must attest to the fidelity of the randomization and blinding process, the users of instrumental variable analysis must present a convincing case that the instrumental variable is not in some way selecting for patients at different risk for the outcome, independent of treatment. This typically cannot be proven empirically in an individual study. Therefore, one is left with a justification based on expert opinion rather than hard evidence for the internal validity of the contrast between the groups with different values of the instrumental variable. Instrumental variable analysis can also hide important heterogeneity of treatment effects (2), just as sometimes occurs in clinical trials.

Instrumental variable analysis is a potentially powerful method for analyzing observational data, especially data derived from natural experiments. Its strongest application is to show that an intervention does not affect outcomes, which is important information in a practice environment plagued by uncritical adoption of new tests and treatments. The shortcomings are the difficulty of verifying the principal assumptions of the method, as illustrated in **Figure 3**, and of discovering valid instrumental variables.

RANDOMIZED TRIALS

Confounding, the scourge of observational research, is usually of minor concern in randomized trials. However, the three essential elements of CER will, each for different reasons, require larger trials than heretofore necessary. First, direct comparisons of active treatments will usually result in similar outcome rates in each arm. Second, the desire for study populations (doctors, patients, institutions) that are typical of usual practice will usually require trials to be performed in community settings, in which adherence to study protocols may be difficult to achieve. Deviations from study protocols may cause more variation, wider confidence intervals, and therefore larger studies to achieve adequate power. Third, the search for clinical characteristics that define patient groups with stronger or weaker responses to treatments will require comparing the outcomes of treatments in subgroups defined by clinical characteristics. Drawing statistically valid conclusions in subgroups requires larger study populations. Problems with the conduct of randomized trials will also affect their ability to answer CER questions.

Cluster Randomized Trials

When an IOM committee established priorities for CER, comparison of methods to improve translation of CER results into practice led the list of high-priority topics (37). Cluster trials meet the requirements of comparing interventions to improve care remarkably well. In cluster randomized trials, randomization is by group, typically a practice site, rather than by individual patient. In a cluster trial, each site typically implements only one intervention. This feature solves several logistical problems that would otherwise plague efforts to compare interventions to change the delivery of care

(57). First, implementing two interventions side by side in the same site creates chaos at every level. Logistics aside, implementing a single intervention at a site improves external validity because it is usual practice in the real world. As a general principle, comparisons of interventions that are implemented at a unit level should be randomized at the level of that unit. Second, enrolling patients and obtaining informed consent are much easier to integrate into routine care if everyone receives the same intervention. Third, comparing two interventions will usually be easier logistically if only one is available at a site, if only because it reduces the burden of informing patients about the choices. Fourth, using the site's administrative leaders to help implement the intervention is easier when they must organize the staff to implement a single intervention rather than two or more. A single intervention in the organization and delivery of care is difficult enough to implement successfully without doubling the burden. Fifth, contamination of one arm of a trial by the intervention in the other arm is unlikely in a cluster trial. Sixth, institutional committees for protecting human subjects of research often waive the requirement for documenting individual consent to participate when everyone at a site receives the same intervention (20, 63). This practice arouses less controversy when the intervention is a quality of care improvement (49). In short, when the intervention alters patient care workflow or when contamination of one arm by another is likely, a cluster trial is the solution.

A cluster randomized trial of three interventions to reduce colonization with methicillin-resistant *Staphylococcus aureus* (MRSA) illustrates the essential features of the method (57). Forty-five Hospital Corporation of America hospitals were randomized to receive one of the interventions. The randomization was stratified by intensive care unit workload and by frequency of MRSA colonization. All units in a hospital received the same intervention. The interventions used the hospital's in-service education, compliance, and adherence monitoring programs. Regular, on-site teams implemented the intervention, rather than especially created teams. The primary outcome is the number of unit patients with a positive culture for MRSA. The institutional committee for the protection of human subjects of research waived the requirement for individual consent, citing minimal risk to the patient. The total cost of the study is less than $2 million. The study is ongoing.

Several features of cluster randomization require attention. First, unlike trials in which randomization is by patient within sites, in cluster trials, patients within a site are not independent of one another. They have in common the characteristics of patients attracted to the site, they all experience the clinical routines of practice at that site, and they receive care from the same physicians. Therefore, the between-patient variation at one site is less than if patients were independent of one another, as assumed in the standard method for calculating confidence intervals. Special statistical methods, such as hierarchical models, must be used to adjust for the tendency of outcomes to cluster by site so that confidence intervals are not inappropriately narrow. Such methods widen the confidence intervals, which means that, for a given number of patients, the statistical power of cluster trials is lower than it would be if patients were randomized individually. The power of cluster trials is often determined more by the number and variety of participating sites than by the number of participating patients. Enrolling many patients is easier with cluster trials but less important than enrolling enough sites.

Pragmatic Trials

Pragmatic trials (also called practical trials) are effectiveness trials; they study interventions in typical practice and in typical patients, another foundational feature of CER (48, 73). Because pragmatic trials are, by definition, intended to inform decision makers, they align with the goals of CER. They contrast with efficacy trials, which establish whether an intervention works under ideal circumstances and which typically exclude patients with comorbid conditions, advanced age, and other features. A classic example of a pragmatic trial was GISSI (30). In this unblinded study, 11,806 patients admitted at 1 of 176 coronary care units within 12 h of symptoms of a myocardial infarction were randomized to a clot lysing drug (streptokinase) or to usual care. The primary outcome, all-cause mortality at 21 days, was 10.7% in the streptokinase group and 13% in the usual care group.

In a pragmatic trial (48), study patients have the typical comorbid diseases of those who receive the compared treatments in usual practice, and their demographic characteristics closely resemble the typically treated patients (or the study may oversample key demographic groups). They compare active treatments provided in typical hospitals, outpatient settings, and practitioners. The researchers choose the outcome measures to meet the needs of decision makers, sometimes using decision analysis to identify the factors that should sway the decision (71). To minimize the costs of the trial, only the most pertinent data are obtained. Similarly, all-cause mortality may be the preferred outcome because linkage to national death indexes makes it possible to obtain long-term follow-up without incurring the costs of contacting the patient. The PRECIS (pragmatic—explanatory continuum indicator summary) framework is intended to help trialists assess where a proposed trial lies on the continuum between pragmatic and explanatory (71).

Pragmatic trials arguably combine the advantages of randomization (high internal validity) and observational research (high external validity). However, pragmatic trials have important shortcomings for achieving the goals of CER. They often do not measure key intermediate outcomes, such as adherence to the assigned treatment, that may help to understand the study results. For logistical reasons, they are often restricted to measuring easily ascertained and adjudicated outcomes, which means that some trials do not collect cause-specific outcome measures. Limiting data collection may mean that a trial will not measure baseline covariates that may help to define subgroups that are useful for personalized decision making. These shortcomings mean that pragmatic trials without such measurements may not adequately inform all decision makers (e.g., the patient versus the policy maker) and may fail to identify the means to individualize care or understand why an intervention succeeds or fails.

Adaptive Trials

Traditional randomized trials are not well suited to a dynamic environment in which accumulating evidence sends early signals about ultimate outcomes and new interventions enter the marketplace as the trial continues. The hallmark of traditional studies is that most if not all aspects of the design and analyses are prespecified and do not change during the trial. An adaptive trial's design changes in response to accumulating data. These changes can encompass almost any aspect of the design, from the number of arms, nature of treatments, sample size, and even outcome measures. The goal of such changes is to maximize the efficiency of the study and its relevance when completed.

The machinery of Bayesian analysis is usually the means to achieving the flexibility of adaptive trials. Reliance on traditional "frequentist" statistical approaches, in which a result either is or is not statistically significant, can lead to yes-no verdicts that oversimplify the results and ignore the pretrial odds that one treatment would be superior (26, 27). Adaptive trials address these shortcomings in a way that may potentially make better use of data and reach a useful answer more quickly (8). The Bayesian framework formally accounts for prior knowledge, including biological plausibility and prior studies, to estimate the pretrial odds that, for example, treatment A is superior to treatment B. As trial results accumulate, the odds that treatment A is superior are periodically recalculated with Bayes' theorem. When those odds reach a prespecified threshold, the trial of A versus B ends. If the odds that B is superior rise, the A arm may be dropped in favor of comparing B with a third treatment. Or the odds that A is superior may rise above a threshold and the trial ends. If a new treatment emerges, for which the pretrial odds of superiority are high, the winner of A versus B may be matched against it. If a subgroup appears to enjoy particular benefit, the trial can oversample patients in that subgroup, a key element for the purposes of CER, which is often to identify subgroups of patients who share the potential for greater-than-average benefit from an intervention. Other advantages for CER include the ability to consider multiple interventions, focusing experimental resources on the best interventions, and the flexibility to introduce new interventions when they first appear to have a performance advantage.

The trial community, including manufacturers, has been receptive to adaptive trials. The U.S. FDA has issued a guidance document for device trials (75) and more recently for adaptive drug trials (76), and reports of adaptive drug trials have been published (24, 43).

TREATMENT-RESPONSE HETEROGENEITY

Treatment-response heterogeneity (also called heterogeneity of treatment effects) refers to variation in the direction and/or magnitude of response to the same treatment (44). This heterogeneity can be a function of an individual's characteristics, or those of the disease, treatment, care setting, providers, or external factors. The search for treatment-response heterogeneity is an important element in CER for several reasons. First, one of the goals of CER is to identify ways to optimize treatment for an individual. Second, by design, CER will enroll representative community-based populations that can exhibit heterogeneity in the many aspects that affect the response to treatment. Third, because CER tends to make head-to-head comparisons of active therapies, the differences in average treatment effect may be small, shifting the emphasis to those factors that would magnify the treatment effect for individuals.

Two studies illustrate the basis of heterogeneous treatment response, which is nonrandom variation within a population. The usual ways to detect it are either with stratified analyses of subgroups or with multivariable modeling of response to treatment using treatment–clinical (or genetic) variable interactions to detect variables that define subgroups of patients that differ in response to treatment. These interactions are most plausible when prespecified. In one study, the authors reanalyzed the results of a randomized trial comparing coronary bypass surgery with medical management of chronic stable angina (17). The authors developed a model to predict five-year cardiovascular mortality and used it to estimate each trial patient's probability of an outcome event. They divided the trial participants into high-, medium-, and low-risk tertiles and compared the outcome rates within each tertile. Surgery was superior to medical treatment for high-risk patients and inferior in low-risk patients, and surgery and medical treatment were equal in intermediate-risk patients. In another example, the authors subdivided the sites of a placebo-controlled randomized trial of propranolol to prevent death after a myocardial infarction. Over all sites, propranolol was superior. In 21 sites (dominant centers), the mortality was higher with placebo, and in 10 sites (divergent centers), the mortality was higher with propranolol (35).

Treatment response heterogeneity can occur, but its frequency and clinical impact are not known. Its frequency will become clearer as researchers design studies with adequate power to detect it and confirm it in follow-up studies. Randomization with respect to treatment assignment is preserved in a subgroup analysis (e.g., by gender), so a difference in the treatment response within the subgroup (e.g., men versus women) is internally valid. However, attributing the difference in response to the defining subgroup characteristic (e.g., gender) is not valid because that characteristic itself is not randomly assigned and the subgroups could differ in other ways from each other, e.g. women in a study might be younger, or thinner, or have higher SES than the men (77).

Consider a hypothetical trial comparing treatment A to treatment B, which shows that treatment A was superior. Subgroup analysis shows further that patients taking a

Treatment-response heterogeneity (also heterogeneity of treatment effect): members of a population respond differently to a treatment because of measurable characteristics that distinguish them from others in the population particular vitamin responded better to treatment B than did patients who were not on the vitamin. Is this grounds for recommending the vitamin to everyone who gets treatment B? No. Those who take the vitamin might be different in other ways that affect their responsiveness to B, and those other ways cannot be changed by taking the vitamin. So taking vitamins is a marker of a person who might do better on B than those who do not take them, but giving everyone vitamins might have no impact on treatment B's effectiveness.

Differing inherent responsiveness to therapy is only one cause of heterogeneity of treatment effects. Treatment harms can lead to nonadherence or dropout or just offset the benefits in ways that vary across patients so that the balance of harms and benefits varies. In addition, competing risks (such as surgical death) can alter apparent treatment responsiveness (41). For example, in an elderly person with a slow-growing cancer, death from age-related causes (e.g., myocardial infarction or dementia) may reduce the benefit of cancer treatment because patients do not live long enough to experience the effects of treatment, which may take a long time to appear.

The analysis of subgroup treatment effects is fraught with the problems of any exploratory analysis: An effect is unlikely to be true even if the new evidence seems compelling. Statistical methods to address this problem include formal Bayes, with low priors put on exploratory interactions (8); empirical Bayes, which shrinks back subgroup estimates to a common mean (18, 29); false discovery rates, which is a frequentist correction for multiple comparisons and a close cousin to Bayesian procedures (7); and traditional frequentist multiplicity corrections (36). The most important factor by far is a strong, evidence-based biologic rationale for the subgroup effect, which implies high prior odds that it is true.

SYSTEMATIC REVIEWS

The systematic review is one of the most important methodological advances of the past 25 years (53). By linking systematic reviews to recommendations for practice, the U.S. Preventive Services Task Force and the Task Force on Community Preventive Services (1, 10, 22) have provided important leadership in the clinical and public health fields, respectively. The recent IOM report on standards for practice guidelines (38) is one of many authoritative sources to state that performing a systematic review is the essential first step in making recommendations for practice. Systematic reviews have become central to the process of establishing clinical policy (practice guidelines, practice measures), administrative policy (insurance coverage and benefit design), and public health practice.

Systematic reviews have great influence in part because the best systematic review authors take great pains to avoid bias. Systematic reviews have become a subdiscipline of clinical epidemiology, with standards set by, among others, the Cochrane Collaboration (33) and, more recently, the IOM report (39). Adherence to high standards is essential to sustain public confidence. Those undertaking a systematic review for the first time should follow the standards carefully and read some exemplary systematic reviews (56) before beginning.

The goal of a systematic review is an unbiased, clear description of a body of evidence (53). Because systematic reviews carry so much weight in present-day policy making, the review team should include individuals with prior experience in conducting systematic reviews and should avoid those with conflicts of interest. Review and comment are important at every stage of a systematic review; authors should obtain advice from stakeholders when designing the study protocol and should seek public comment on the study protocol and the results (39).

Two key features apply equally to the conduct of a systematic review of evidence bearing upon public health policy and clinical practice. First, the conduct of a systematic review should follow a prespecified plan, beginning with an analytic framework (a flow diagram indicating the key steps at which evidence might inform choice), which helps to identify the key questions that the systematic review should address. Second, a qualitative evidence synthesis is the most important element of the systematic review. It describes the body of evidence on which the recommendation for action depends. It characterizes the diversity of study designs and study populations and the patterns of strengths and weaknesses within the body of evidence. Standardized ways to characterize a body of evidence qualitatively is the most important methodological gap in systematic reviews.

Whether to combine the results of studies by doing a quantitative literature synthesis (meta-analysis) is the pivotal decision in the conduct of a systematic review. It is important because readers want a straightforward summary of a body of evidence, and a meta-analysis serves this need by generating a pooled relative risk. However, meta-analysis is a simplifying technique that can be misused and misinterpreted. Performing a meta-analysis of a heterogeneous body of evidence is questionable practice. Although statistical methods can be helpful, the best way to characterize the heterogeneity of the body of evidence is the qualitative literature synthesis. Here is a point at which public health systematic review practice should diverge from clinical systematic review practice. The body of evidence in most clinical systematic reviews is composed of randomized trials, a relatively homogeneous and bias-free study design. Most studies of the effectiveness of public health interventions are observational, which may include the gamut of study designs, all of which are strongly susceptible to bias and confounding. A meta-analysis may amplify the effects of these threats to internal validity. Public health systematic reviews should depend on a qualitative literature synthesis and avoid meta-analysis.

As currently defined, CER will make demands on meta-analysis methodology. First, the search for treatment-response heterogeneity will generate lists of predictors of response to treatment (**Figure 1**). Combining these lists from different CER studies will require meta-analysis methods. Second, as already noted, observational research, which will play an important role in CER, is susceptible to bias and confounding, effects that meta-analysis can amplify. Third, meta-analysis of studies of diagnostic tests is an unsolved problem. Many of the high-priority research questions in the IOM report address the comparative effectiveness of imaging tests. Methods to combine the results of studies of test sensitivity and specificity have not advanced in several decades despite dissatisfaction with current methods. Fourth, cross-sectional CER studies will generate clinical prediction rules for estimating pretest probability, and longitudinal studies will generate risk-assessment tools to estimate the probability of future events. Before clinicians can take full advantage of the broad base of study populations in a body of risk-prediction studies, researchers must develop informative methods for displaying (and perhaps combining) the risk-prediction tools from these studies.

DECISION ANALYSIS

Background

Decision analysis is emerging as a central discipline in CER, largely because it addresses a foundational objective of CER—identifying the best decision for a particular patient—and because it rests on axioms and theorems derived from first principles of logic. Decision analysis can also help to identify the targets of research by showing which variables are the most important in decision making and calculating the expected value of information. Finally, many cost-effectiveness analyses are based on decision models. Decision analysis is a highly evolved, specialized discipline. This brief account focuses on the basic principles, which are covered in several textbooks (67, 80), how clinicians use the results of decision analyses, and how CER will generate the data needed for decision analyses.

The Basic Principles

When people use subjective judgments and feelings to make decisions, as they usually do in

Utility: in medical decision analysis, a quantitative measure of preference for a health state obtained by using the standard reference gamble

everyday life, their mental processes are either intuitive or logical. When people use quantitative methods to help make difficult decisions, they must express these subjective judgments and feelings as numbers.

Two important elements of everyday decision making are usually subjective. One is uncertainty about whether an event will occur or has occurred. Representation of uncertainty by a probability is a basic principle of decision analysis. Probability can be subjective or frequentist. In the former, a person picks a number between 0 and 1.0 to represent her uncertainty about future events ("what is the likelihood that this person will develop heart disease?") or present states ("what is the likelihood that this person has heart disease now?"). Subjective estimates of probability are subject to systematic errors (74). The frequentist interpretation of probability requires empirical observations about the incidence of events or the prevalence of disease in groups. As an expression of population averages, frequentist probability best applies to an individual as a starting point for subjective estimates based on the characteristics of the individual (74).

The second subjective element is a person's feelings about present or potential future health states. Before making a decision that could result in better health, worse health, or the status quo, people should examine their feelings about these states relative to one another. Techniques exist to assign numbers to these subjective relative feelings. The standard reference gamble, which has the strongest theoretical basis, involves determining when a gamble between two outcomes seems equivalent to an intermediate health state. The time–trade-off method asks the patient to name the shortest lifetime in perfect health that she would accept in exchange for her life expectancy in an intermediate health state. The results with either method can be surprisingly reproducible (54) but are sensitive to the framing of the description of the health states (52).

Expressing feelings in terms of numbers is important because probability and preferences are the key elements in expected utility decision making. Von Neumann & Morgenstern in *Theory of Games and Economic Behavior* (78) showed that the patient should prefer the decision option with the highest expected utility to be consistent with his preferences for the health states that he might experience. This strong claim, which is based on first principles, is a compelling theoretical argument for using decision analysis to make difficult decisions. Expected utility is the product of the probability of an outcome multiplied by the utility of the outcome. The latter is usually expressed in quality-adjusted life years, which are the product of the patient's life expectancy in a health state multiplied by the patient's utility for that state.

Sensitivity analysis is a key technique in decision analysis. Each parameter in a decision model has an assigned value, typically the most likely one. Sensitivity analysis consists of calculating the expected outcome after substituting first the lowest and then the highest values of a parameter. If the outcome does not change, the parameter does not drive the decision.

How Is Decision Analysis Used? How Will It Be Used?

Decision analysis is perfectly suited for individual decision making, with its emphasis on individual probabilities and utilities. A decision analysis created for a single patient is unusual because creating a decision tree, searching the literature for evidence-based probabilities, and eliciting a patient's utilities are very time-consuming. Still, clinicians frequently make decisions that are influenced by a decision analysis. This apparent paradox refers to the use of decision analysis in the development of clinical practice guidelines. For example, decision analyses helped expert panels to formulate guidelines for screening for breast cancer (50), colorectal cancer, cervical cancer (19), and sore throat (13, 42). Current diagnostic practices for suspected pulmonary embolism (60, 61) and suspected chronic stable ischemic heart disease (81) have been shaped by clinical prediction rules to estimate pretest probabilities, the sensitivity and specificity of diagnostic tests, and Bayes' theorem. So, decision analysis helps to shape standard practice without being used explicitly in individual cases. Another new direction is the incorporation of computer-based decision aids into electronic health record systems. With decision models and clinical prediction rules for estimating probabilities retrievable from the electronic medical record, individual decision analysis is becoming feasible in daily practice, especially with computer aids for estimating utilities (54, 70).

CER and Decision Analyses: Clinical Prediction Rules

Clinical prediction rules form a bridge between subjective and objective probability because they use empirical observations to estimate the probability that corresponds to an individual's clinical findings (58). They predict the patient's present, unknown state (diagnostic prediction rule) or a future event (prognostic prediction rule). In either case, the prediction rule is based on an empirical study of which findings are the best predictors. The starting point is a study population with a common clinical finding, such as chest pain. The researchers obtain a standard data set that includes candidate predictors of the outcome. The study design is either cross-sectional (for a diagnostic prediction rule, the outcome is the patient's present state) or cohort (for a prognostic prediction rule, the outcome is a future event). In both cases, a multivariable analysis identifies the findings that are most strongly predictive of the outcome and assigns to each a weight that reflects how well it discriminates.

How are clinical prediction rules used? A diagnostic prediction rule produces an estimate of a patient's pretest probability of the target condition (e.g., coronary artery disease in a patient with chest pain). Using Bayes' theorem and the pretest probability, it is easy to calculate the post-test probability if the test's sensitivity and specificity are known. A prognostic prediction rule estimates the probability that the patient will experience a future event, an important input to a decision analysis.

Published standards describe good research practice for developing (in the training set) and validating (in the test set) clinical prediction rules (47, 79). To evaluate a clinical prediction rule, it must be tested in a fresh population or, using bootstrap methods, in many randomly chosen subsets of the training set (21, 32). Questions to ask of a clinical prediction rule follow: What is the range of probabilities corresponding to the range of possible findings (none of the key predictors to all of them)? What is the distribution of patients across the range of probabilities? Accounting for any differences in prevalence of the target condition (66), how closely do the probabilities in the test set match up with those in the training set (calibration)? How well does the rule discriminate between those who have the target condition and those who do not [discrimination, area under the receiver operating characteristic (ROC) curve]? Recently, researchers have compared discrimination or reclassification (14, 15, 40) with a prediction rule using clinical predictors of a future event alone or clinical predictors plus a test result (including genetic tests) (16, 55). The ultimate test of a clinical prediction rule is whether using it improves clinical outcomes, reduces costs, or both (58, 68, 69).

CER and Decision Analyses: Modeling Long-Term Outcomes

Decision analyses typically use life expectancy (defined as the average length of life remaining at a specified age) as an outcome. However, most randomized trials follow patients on treatment for a limited period, not until death. Decision analysts cope with this limitation by modeling lifetime outcomes.

The most commonly used approach is the declining exponential approximation to life expectancy (the DEALE) (3, 4). Empirically, the probability of being alive after a specified age follows a curve that is best approximated by a complex mathematical formula called the Gompertz Function. For patients with serious illness, the mortality rate is approximately constant over time, which means that the probability of being alive after an event can be represented by a declining exponential function. This approximation simplifies the calculation of life expectancy from postintervention survival statistics because life expectancy equals 1 divided by the patient's total mortality rate, which is the sum of the mortality rates from each of the patient's diseases (4, 67). The DEALE approximation assumes that the death rate from a disease is constant over time, which may not be true as a person ages.

A second approach to estimating life expectancy is to simulate a series of hypothethical patients as they pass through several health states to death using a Markov model (5, 64). The key elements of a Markov model are health states, transition probabilities between health states, and the cycle length (a period of time, such as one year). Uncertainty about whether a transition between health states occurs during a cycle is represented by the transition probability. The number of cycles until the hypothetical person enters the dead state is the length of life for that person. This process is repeated for many hypothetical people. The time to death will vary between simulated patients according to the play of chance; the average time to death is, by definition, the life expectancy.

COST-EFFECTIVENESS ANALYSIS

Collectively, the U.S. population is worried about the high cost of health care. Individually, people are worried that they will not have access to specific interventions because insurers are unwilling to pay for them. Reflecting this ambiguity, the United States has no societal consensus about the use of techniques for estimating the value of health care interventions. Therefore, methods such as cost-effectiveness analysis are not used for making decisions about federally insured health care. Nonetheless, information about the value of health care and public health interventions is important to many stakeholders in health care.

According to the Federal Panel on Cost-Effectiveness in Health and Medicine (25), cost-effectiveness is "a method designed to assess the comparative impacts of expenditures on different health interventions" (p. 26). Equivalently, it compares the health effects that result from alternate uses of a given amount of health care resources (23). Cost-effectiveness, like CER, always compares alternative choices. The appropriate measure, therefore, is incremental cost-effectiveness: the difference in costs due to using one intervention instead of another divided by the difference in their health outcomes.

CER studies can measure the costs of care associated with each of the compared interventions. These costs include direct costs, usually defined as the costs to produce the health care service. Most of these costs are incurred by the provider of the service who then is reimbursed at a rate that is negotiated with the payer. A few costs, such as transportation to the doctor's office, are paid by the patient. Indirect costs are defined as all other costs. Indirect costs are mostly opportunity costs: money not earned because the patient is receiving health care. The difference in cost attributable to the intervention (versus the alternative choice) is the numerator in the cost-effectiveness ratio.

The denominator of the cost-effectiveness ratio is the impact of the intervention on health. Cost-effectiveness analysis measures incremental effects, so the impact on health is measured by differences in quality-adjusted life years (QALYs) attributable to the intervention as compared with the alternate choice. The change in QALYs is the difference between the products of the life expectancies associated with two compared interventions and the patient's utility for the health states associated with those two life expectancies.

The usual measure of the health impact of medical care, the QALY, is the product of two numbers. One is years of life remaining (LE or life expectancy). The other (Q) is a measure of the quality of life, customarily on a scale of 0 to 1. Most cost-effectiveness analyses apply to populations. The health utility index (HUI) has been used widely to assess population-level utilities for different diseases (34, 72). Researchers interviewed individuals from the general population to assess their utilities for their state of health using the standard reference gamble. They also assessed functional status along seven dimensions of health (e.g., mobility, fertility). Using these variables, they developed a scoring system to predict a person's utility for a health state. The investigators administered the HUI instrument to patients with a disease to define the distribution of utilities of patients with the condition. The HUI can be used to estimate an individual's utility rather than assess it directly with the standard reference gamble.

The impact of an intervention on remaining life may be measured directly through a CER study that measures costs and outcomes. However, most cost-effectiveness analyses model costs and outcomes using data from the existing literature or, alternatively, from a CER study designed specifically to obtain the data needed for the model. Analysts seldom have the opportunity to use data obtained expressly for a decision model. CER will provide this opportunity.

Cost-effectiveness analysis is relevant to several settings. One is policy making for clinical care interventions, which would be applied at the population level using average figures for life expectancy, quality of life (utilities), and probabilities (and their distributions in the community). A second would be decision making for individuals, typically patients who are paying for their health care out-of-pocket and are concerned about getting value for their money. The parameters of the analysis would be the individual's probabilities, utilities, and costs. In practice, cost-effectiveness analysis for individual decision making seldom occurs, but it could become more common as payers shift the costs of health care to the individual. Finally, cost-effectiveness analysis can apply to interventions implemented at the population level outside the medical care setting (interventions in the public health system). The costs would be those incurred by the public health system, the probabilities would be the population average probability, and the utilities would be derived from the population through surveys or application of the HUI.

Cost-effectiveness analysis has been applied to public health interventions. In a review, Graham and colleagues (28) focused on several conditions, of which prevention of trauma is illustrative because the interventions were at the population level. Examples include air bags, seat belts, compulsory motorcycle helmet use, daytime running lights, a 55 miles per hour speed limit, and mandatory bicycle helmet use. Outcome measures were lives saved. No study used QALYs. The analyses took the societal perspective. Some studies used sensitivity analysis for uncertain variables in the model. The authors called for analysts to use more uniform methods to increase the credibility of analyses for public policy making. They pointed out that cost-effectiveness is one factor among many in public health decision making, a point that applies equally to decisions about clinical policy.

SYNTHESIS

The two principal forms of CER are randomized trials and observational research. Trials provide results that can often be interpreted as cause and effect, which helps to ensure that the programs they inspire will have the intended effects. Trials are expensive, and CER trials will be especially expensive because the comparison of active treatments requires large study populations to avoid false-negative and false-positive conclusions. Observational studies can provide large study populations quickly and at low cost, important factors for PCORI, an organization that needs to show value to the public in the few years before 2019, when it is scheduled to come before the U.S. Congress for a reauthorization vote. Because of confounding and lower-quality data, policy makers and clinicians will be much more cautious about interpreting the results of observational studies as cause and effect and correspondingly cautious about acting on the findings. Formulating a research project agenda that strikes the right balance between these two forms of research, each with its advantages and disadvantages, will be a high-stakes process of decision making under uncertainty, the recurring theme of this review. One partial solution is to support methodological research on how best to formulate policy that is based on imperfect evidence. This methodological research will be an important legacy of the current public support for CER. Fortunately, it will be a legacy that can benefit the public health system.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. Anderson LM, Brownson RC, Fullilove MT, Teutsch SM, Novick LF, et al. 2005. Evidence-based public health policy and practice: promises and limits. *Am. J. Prev. Med.* 28:226–30
2. Basu A. 2011. Estimating decision-relevant comparative effects using instrumental variables. *Stat. Biosci.* 3:6–27
3. Beck JR, Kassirer JP, Pauker SG, Gottlieb JE, Klein K, Kassirer JP. 1982. A convenient approximation of life expectancy (the "DEALE"). I. Validation of the method. *Am. J. Med.* 73:883–88
4. Beck JR, Pauker SG. 1982. A convenient approximation of life expectancy (the "DEALE"): II. Use in medical decision-making. *Am. J. Med.* 73:889–97
5. Beck JR, Pauker SG. 1983. The Markov process in medical prognosis. *Med. Decis. Mak.* 3:411–58
6. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, et al. 2011. Developing the Sentinel system—a national resource for evidence development. *N. Engl. J. Med.* 364:498–99
7. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57:289–300
8. Berry DA. 2006. Bayesian clinical trials. *Nat. Rev. Drug. Discov.* 5:27–36
9. Braitman LE, Rosenbaum PR. 2002. Rare outcomes, common treatments: analytic strategies using propensity scores. *Ann. Intern. Med.* 137:693–95
10. Briss PA, Zaza S, Pappaioanou M, Fielding J, Wright, et al. 2000. Developing an evidence-based guide to community preventive services—methods. *Am. J. Prev. Med.* 23:35–43
11. Brookhart MA, Rassen JA, Schneeweiss S. 2010. Instrument variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol. Drug. Saf.* 19:537–54
12. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. 2010. Confounding control in healthcare database research: challenges and potential approaches. *Med. Care.* 48(6 Suppl.):S114–20
13. Centor RM, Witherspoon JM, Dalton HP, Brody CE, Link K. 1981. The diagnosis of strep throat in adults in the emergency room. *Med. Decis. Mak.* 1(3):239–46
14. Cook N, Buring JE, Ridker PM. 2006. The effect of including C-Reactive protein in cardiovascular risk prediction models for women. *Ann. Intern. Med.* 145:21–29
15. Cook NR, Ridker PM. 2009. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Ann. Intern. Med.* 150:795–802
16. Cornelis MC, Qi L, Zhang C, Kraft P, Manson J, et al. 2009. Joint effects of common genetic variants on the risk for type 2 diabetes in US men and women of European ancestry. *Ann. Intern. Med.* 150:541–50
17. Detre K, Peduzzi P, Murphy M, Hultgren H, Thomsen J, et al. 1981. Effect of bypass surgery on survival in patients in low- and high-risk subgroups delineated by the use of simple clinical variables. *Circulation* 163:1329–38
18. Dixon DO, Simon R. 1991. Bayesian subset analysis. *Biometrics* 47:871–81
19. Eddy DM. 1980. *Screening for Cancer: Theory, Analysis, and Design*. Englewood Cliffs, NJ: Prentice-Hall
20. Edwards SJL, Braunholtz DA, Lilford RJ, Stevens AJ. 1999. Ethical issues in the design and conduct of cluster randomised controlled trials. *BMJ* 318:1407

21. Efron B, Tibshirani R. 1993. *Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC
22. Fielding JE, Teutsch SM. 2009. Integrating clinical care and community health: delivering health. *JAMA* 302:317–19
23. Garber AM, Sox HC. 2010. The role of costs in comparative effectiveness research. *Health Aff.* 29:1805–11
24. Giles FJ, Kantarjian HM, Cortes JE, Garcia-Manero G, Verstovsek S, et al. 2003. Adaptive randomized study of idarubicin and cytarabine versus troxacitabine and cytarabine versus troxacitabine and idarubicin in untreated patients 50 years or older with adverse karyotype acute myeloid leukemia. *J. Clin. Oncol.* 21:1722–27
25. Gold MR, Siegel JE, Russell LB, Weinstein MC, eds. 1996. *Cost-Effectiveness in Health and Medicine*. New York: Oxford Univ. Press
26. Goodman SN. 1999. Toward evidence-based medical statistics. 1: The *p* value fallacy. *Ann. Intern. Med.* 130:995–1004
27. Goodman SN. 1999. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann. Intern. Med.* 130:1005–13
28. Graham JD, Corso PS, Morris JM, Sequi-Gomez M, Weinstein MC. 1998. Evaluating the cost-effectiveness of clinical and public health measures. *Annu. Rev. Public Health* 19:125–52
29. Greenland S, Robins JM. 1991. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology* 2:244–51
30. Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI). 1986. Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. *Lancet* 327:397–402
31. Hamburg MA, Sharfstein JM. 2009. The FDA as a public health agency. *New Engl. J. Med.* 360:2493–95
32. Hastie T, Tibshirani R, Friedman JH. 2009. Model assessment and section. In *The Elements of Statistical Learning: Data mining, inference, and prediction*, pp. 219–57. New York: Springer Science + Business Media. 2nd ed.
33. Higgins JPT, Green S, eds. 2008. *Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series*. Chichester, UK: Wiley
34. Horsman J, Furlong W, Feeny D, Torrance G. 2003. The Health Utilities Index (HUI®): concepts, measurement properties and applications. *Health Qual. Life Outcomes* 1:54
35. Horwitz RI, Singer BH, Makuch RW, Viscoli CM. 1996. Can treatment that is helpful on average be harmful to some patients? A study of the conflicting information needs of clinical inquiry and drug regulation. *J. Clin. Epidemiol.* 49:395–400
36. Hsu JC. 1996. *Multiple Comparisons: Theory and Methods.* Boca Raton, FL: Chapman & Hall/CRC
37. Inst. Med. (IOM). 2009. *Initial National Priorities for Comparative Effectiveness Research*. Washington, DC: Natl. Acad. Press. **http://www.nap.edu/catalog/12648.html**
38. Inst. Med. (IOM). 2011. *Clinical Practice Guidelines that We Can Trust*. Washington, DC: Natl. Acad. Press. **http://www.iom.edu/Reports/2011/Clinical-Practice-Guidelines-We-Can-Trust.aspx**
39. Inst. Med. (IOM). 2011. *Finding What Works in Health Care: Standards for Systematic Reviews.* Washington, DC: Natl. Acad. Press. **http://www.iom.edu/Reports/2011/Finding-What-Works-in-Health-Care-Standards-for-Systematic-Reviews.aspx**
40. Janes H, Pepe MS, Gu W. 2008. Assessing the value of risk predictions by using risk stratification tables. *Ann. Intern. Med.* 149:751–60
41. Kent DM, Alsheikh-Ali A, Hayward RA. 2008. Competing risk and heterogeneity of treatment effect in clinical trials. *Trials* 9:30–36
42. Komaroff AL, Pass TM, Aronson MD, Ervin CT, Cretin S, et al. 1986. The prediction of streptococcal pharyngitis in adults. *J. Gen. Intern. Med.* 1:1–7
43. Krams M, Lees KR, Hacke W, Grieve AP, Orgogozo JM, et al. 2003. ASTIN Study Investigators. Acute stroke therapy by inhibition of neutrophils (ASTIN): an adaptive dose-response study of UK-279,276 in acute ischemic stroke. *Stroke* 34:2543–48
44. Kravitz RL, Duan N, Braslow J. 2004. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Q.* 82:661–87
45. Lauer M, Collins FS. 2010. Using science to improve the nation's health system. *JAMA* 303:2182–83

46. Lauer MS, Collins FS. 2010. NIH's commitment to comparative effectiveness research. *JAMA* 303:2182–83
47. Laupacis A, Sekar N, Stiell IG. 1997. Clinical prediction rules: a review and suggested modification of methodological standards. *JAMA* 277:488–94
48. Luce BR, Kramer JM, Goodman SN, O'Connor JT, Tunis SR, et al. 2009. Rethinking randomized clinical trials for comparative effectiveness research: the need for transformational change. *Ann. Intern. Med.* 151:206–9
49. Lynn J, Baily MA, Bottrell M, Jennings B, Levine RJH, et al. 2007. The ethics of using quality improvement methods in health care. *Ann. Intern. Med.* 146:666–73
50. Mandelblatt JS, Cronin KA, Bailey S, Berry DA, de Koning HJ, et al. 2009. For the Breast Cancer Working Group of the Cancer Intervention and Surveillance Modeling Network (CISNET). Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms. *Ann. Intern. Med.* 151:738–47
51. McClellan M, McNeil BJ, Newhouse JP. 1994. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA* 272:859–66
52. McNeil BJ, Pauker S, Sox HC Jr, Tversky A. 1982. On the elicitation of patients' preferences for alternate therapies. *N. Engl. J. Med.* 306:1259–62
53. Mulrow C. 1994. Systematic reviews: rationale for systematic reviews. *BMJ* 309:597–98
54. Nease RN, Kneeland T, O'Connor GT, Sumner W, Lumpkins C, et al. 1995 Variation in patient utilities for outcomes of the management of chronic stable angina: implications for practice guidelines. *JAMA* 273:1185–90
55. Paynter NP, Chasman DI, Buring JE, Shiffman D, Cook NR, et al. 2009. Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. *Ann. Intern. Med.* 150:65–72
56. Pignone M, Saha S, Hoerger T, Mandelblatt J. 2002. Cost-effectiveness analyses of colorectal cancer screening: a systematic review for the US Preventive Services Task Force. *Ann. Intern. Med.* 137:96–104
57. Platt R, Takvorian SU, Septimus E, Hickok SE, Moody J, et al. 2010. Cluster randomized trials in comparative effectiveness research: randomizing hospitals to test methods for prevention of healthcare-associated infections. *Med. Care* 48(6 Suppl.):S52–57
58. Reilly BM, Evans AT. 2006. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann. Intern. Med.* 144:201–9
59. Rothman KJ, Greenland S, Lash TL. 2008. *Modern Epidemiology*. Philadelphia, PA: Lippincott Williams and Wilkins. 3rd ed.
60. Roy PM, Durieux P, Gillaizeau F, Legall C, Armand-Perroux A, et al. 2009. A computerized handheld decision-support system to improve PE diagnosis: a RCT. *Ann. Intern. Med.* 151:677–86
61. Roy P-M, Meyer G, Vielle B, Le Gall C, Verschuren F, et al. 2006. For the EMDEPU Study Group. Appropriateness of diagnostic management and outcomes of suspected pulmonary embolism. *Ann. Intern. Med.* 144:157–64
62. Rubin DB. 1997. Estimating causal effects from large data sets using propensity scores. *Ann. Intern. Med.* 127:757–63
63. Sabin JE, Mazor K, Meterko V, Goff SL, Platt R. 2008. Comparing drug effectiveness at health plans: the ethics of cluster randomized trials. *Hastings Cent. Rep.* 38:39–48
64. Sonnenberg FA, Beck JR. 1993. Markov models in medical decision making: a practical guide. *Med. Decis. Mak.* 13:322–38
65. Sox HC. 2010. Comparative effectiveness research: an update. *Ann. Intern. Med.* 153:469–72
66. Sox HC, Hickam DH, Marton KI, Skeff KS, Sox CH, et al. 1990. Using the patient's history to estimate the probability of coronary artery disease: a comparison of referral and primary care practice. *Am. J. Med.* 89:7–14
67. Sox HC Jr, Blatt M, Higgins M, Marton KI. 1988. *Medical Decision Making*. Stoneham, MA: Butterworths
68. Sox HC Jr, Margulies I, Sox CH. 1981. Psychologically mediated effects of diagnostic tests. *Ann. Intern. Med.* 95:680–85
69. Stiell IG, McKnight RD, Greenberg GH. 1994. Implementation of the Ottawa ankle rules. *JAMA* 271:827–32

70. Sumner W, Nease R, Littenberg B. 1991. U-titer: a utility assessment tool. *Proc. Annu. Symp. Comput. Appl. Med. Care*, pp. 701–5

71. Thorpe KE, Zwarenstein M, Oxaman AD. 2009. A pragmatic–explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *CMAJ* 180:1–10

72. Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. 1996. Multiattribute utility function for a comprehensive health status classification system: health utilities index mark 2. *Med. Care* 34:702–22

73. Tunis SR, Stryer DB, Clancy DM. 2003. Practical clinical rrials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA* 290:1624–32

74. Tversky A, Kahneman D. 1974. Judgment under uncertainty: heuristics and biases. *Science* 185:1124–31

75. U.S. Food Drug Adm. (FDA). 2006. *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials—Draft Guidance for Industry and FDA Staff*. Rockville, MD: Cent. Biol. Eval. Res. **http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071072.htm**

76. U.S. Food Drug Adm. (FDA). 2010. *Guidance for Industry Adaptive Design Clinical Trials for Drugs and Biologics*. Rockville, MD: Cent. Biol. Eval. Res. **http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf**

77. VanderWeele TJ, Knol MJ. 2011. Interpretation of subgroup analyses in randomized trials: heterogeneity versus secondary interventions. *Ann. Intern. Med.* 154:680–83

78. Von Neumann J, Morgenstern O. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton Univ. Press

79. Wasson JH, Sox HC, Neff RK, Goldman L. 1985. Clinical prediction rules. Applications and methodological standards. *N. Engl. J. Med.* 313:793–99

80. Weinstein MC, Fineberg HV, Elstein AS, Frazier HS, Neuhauser D, et al. 1980. *Clinical Decision Analysis*. Philadelphia, PA: Saunders

81. Zauber AG, Lansdorp-Vogelaar I, Knudsen AB, Wilschut J, van Ballegooijen M, Kuntz KM. 2008. Clinical guidelines evaluating test strategies for colorectal cancer screening: a decision analysis for the US Preventive Services Task Force. *Ann. Intern. Med.* 149:659–69

Annual Review of Public Health

Volume 33, 2012

Contents

Symposium: Comparative Approaches to Reducing Health Disparities

Health Disparities Research in Global Perspective: New Insights and New Directions
Shiriki Kumanyika . . . 1

Health Inequalities: Trends, Progress, and Policy
Sara N. Bleich, Marian P. Jarlenski, Caryn N. Bell, and Thomas A. LaVeist . . . 7

Conceptual Approaches to the Study of Health Disparities
Ana V. Diez Roux . . . 41

How Society Shapes the Health Gradient: Work-Related Health Inequalities in a Comparative Perspective
Christopher B. McLeod, Peter A. Hall, Arjumand Siddiqi, and Clyde Hertzman . . . 59

Disparities in Infant Mortality and Effective, Equitable Care: Are Infants Suffering from Benign Neglect?
Diane L. Rowley and Vijaya Hogan . . . 75

Clinical Care and Health Disparities
B. Starfield, J. Gérvas, and D. Mangin . . . 89

Epidemiology and Biostatistics

A Review of Spatial Methods in Epidemiology, 2000–2010
Amy H. Auchincloss, Samson Y. Gebreab, Christina Mair, and Ana V. Diez Roux . . . 107

Early Intervention to Reduce the Global Health and Economic Burden of Major Depression in Older Adults
Charles F. Reynolds III, Pim Cuijpers, Vikram Patel, Alex Cohen, Amit Dias, Neerja Chowdhary, Olivia I. Okereke, Mary Amanda Dew, Stewart J. Anderson, Sati Mazumdar, Frank Lotrich, and Steven M. Albert . . . 123

Preventability of Cancer: The Relative Contributions of Biologic and Social and Physical Environmental Determinants of Cancer Mortality
Graham A. Colditz and Esther K. Wei . . . 137

The Hurrider I Go the Behinder I Get: The Deteriorating International Ranking of U.S. Health Status
Stephen Bezruchka 157

Unintentional Injuries: Magnitude, Prevention, and Control
Sarah Stewart de Ramirez, Adnan A. Hyder, Hadley K. Herbert, and Kent Stevens 175

Environmental and Occupational Health

Community-Based Approaches to Controlling Childhood Asthma
Noreen M. Clark 193

Future Challenges to Protecting Public Health from Drinking-Water Contaminants
Eileen A. Murphy, Gloria B. Post, Brian T. Buckley, Robert L. Lippincott, and Mark G. Robson 209

Speed Limits, Enforcement, and Health Consequences
Rune Elvik 225

Toward a Systems Approach to Enteric Pathogen Transmission: From Individual Independence to Community Interdependence
Joseph N.S. Eisenberg, James Trostle, Reed J.D. Sorensen, and Katherine F. Shields 239

Public Health Practice

Advertising of Prescription-Only Medicines to the Public: Does Evidence of Benefit Counterbalance Harm?
Barbara Mintzes 259

Economic Evaluation of Pharmaco- and Behavioral Therapies for Smoking Cessation: A Critical and Systematic Review of Empirical Research
Jennifer Prah Ruger and Christina M. Lazar 279

Policies for Healthier Communities: Historical, Legal, and Practical Elements of the Obesity Prevention Movement
Samantha K. Graff, Manel Kappagoda, Heather M. Wooten, Angela K. McGowan, and Marice Ashe 307

Public Health and the Epidemic of Incarceration
Dora M. Dumont, Brad Brockmann, Samuel Dickman, Nicole Alexander, and Josiah D. Rich 325

Quitlines and Nicotine Replacement for Smoking Cessation: Do We Need to Change Policy?
John P. Pierce, Sharon E. Cummins, Martha M. White, Aimee Humphrey, and Karen Messer 341

Systems Science Methods in Public Health: Dynamics, Networks, and Agents
Douglas A. Luke and Katherine A. Stamatakis 357

Health Inequalities: Trends, Progress, and Policy
Sara N. Bleich, Marian P. Jarlenski, Caryn N. Bell, and Thomas A. LaVeist 7

Social Environment and Behavior

Health Disparities Research in Global Perspective: New Insights and New Directions
Shiriki Kumanyika 1

Health Inequalities: Trends, Progress, and Policy
Sara N. Bleich, Marian P. Jarlenski, Caryn N. Bell, and Thomas A. LaVeist 7

Conceptual Approaches to the Study of Health Disparities
Ana V. Diez Roux 41

How Society Shapes the Health Gradient: Work-Related Health Inequalities in a Comparative Perspective
Christopher B. McLeod, Peter A. Hall, Arjumand Siddiqi, and Clyde Hertzman 59

Disparities in Infant Mortality and Effective, Equitable Care: Are Infants Suffering from Benign Neglect?
Diane L. Rowley and Vijaya Hogan 75

Suicide Mortality in the United States: The Importance of Attending to Method in Understanding Population-Level Disparities in the Burden of Suicide
Matthew Miller, Deborah Azrael, and Catherine Barber 393

Health Services

From Small Area Variations to Accountable Care Organizations: How Health Services Research Can Inform Policy
Harold S. Luft 377

Suicide Mortality in the United States: The Importance of Attending to Method in Understanding Population-Level Disparities in the Burden of Suicide
Matthew Miller, Deborah Azrael, and Catherine Barber 393

The Medicalization of Chronic Disease and Costs
Kenneth E. Thorpe and Meredith Philyaw 409

The Methods of Comparative Effectiveness Research
Harold C. Sox and Steven N. Goodman 425

Clinical Care and Health Disparities
B. Starfield, J. Gérvas, and D. Mangin 89

Indexes

Cumulative Index of Contributing Authors, Volumes 24–33 447

Cumulative Index of Chapter Titles, Volumes 24–33 452

Errata

An online log of corrections to *Annual Review of Public Health* articles may be found at
http://publhealth.annualreviews.org/