



# Approaches used in conducting health outcomes and effectiveness research

Aurelio José Figueredo\*, Lee Sechrest

*Evaluation Group for Analysis of Data, Department of Psychology, University of Arizona, Tucson, AZ 85721, USA*

## Abstract

Over the past several decades, a number of approaches (e.g., decision analysis, meta-analyses, clinical trials, analysis of claims data, longitudinal observational studies including those done through patient outcomes research teams, etc.) have been used to conduct outcomes, effectiveness, and appropriateness research. Each of these approaches has varying degrees of comparative advantage and disadvantage with respect to the other. As knowledge of outcomes and effectiveness increases, and as new issues emerge as subjects of research, these approaches may or may not be adequate to generate the necessary information and level of confidence in findings that are desired. What are the strengths and weaknesses of each approach, and what gaps exist in our methodological armamentarium? How do existing methods need to be strengthened? What is the most appropriate application of specific research methods to particular problems? What is the appropriate balance of use of the different available approaches? What types of new methodologies need to be developed to further the field of outcomes and effectiveness research? © 2001 Elsevier Science Ltd. All rights reserved.

## 1. A multiplicity of methods

The aim of outcomes research is to produce a persuasive conclusion about the effectiveness and utility of some diagnostic procedure or treatment. No single outcome study may address the entire array of issues surrounding effectiveness and utility; indeed, that degree of comprehensiveness is unlikely. Still, we believe that outcomes research must contribute to the resolution of uncertainty pertaining either to the effectiveness or utility of a medical procedure or to both. That is not to gainsay the usefulness of other research findings, e.g., on the acceptability in the medical community of some procedure or on the number of procedures needed to be done per week in order for a piece of equipment to be financially affordable. Such findings are useful and perhaps even relevant in some ways to ultimate decisions about implementing conclusions from outcomes studies, but they do not resolve uncertainty about effectiveness or utility. It will be recognized by some readers that we are here following the Shannon–Weaver definition of ‘information’ as reduction of uncertainty.

Many methods exist by which one might address questions about effectiveness and utility of medical procedures. The problem for the would-be investigator is to choose among the methods available. In many cases the choice

will be constrained by reality: some methods will be impossible to implement, some may be deemed too expensive, and so on. In at least some cases, though, the investigator may have a choice and, we would argue, should opt for the design that produces the greatest reduction in uncertainty for the least cost, whether the latter is assessed solely in economic terms or incorporates social costs such as organizational disruption, investigator stress, and the like. It should be thought, however, that uncertainty, like so many other attributes, is in the eye of the beholder, and what reduces uncertainty for one eye may not do so for another more skeptical, more jaundiced, or more persuaded of another truth already. We will return to the calculus of uncertainty in a moment, but let us first make clear what we regard as methodological options available to the investigator wanting to assess effectiveness or utility of a procedure. Or, as will be seen, we want to lay out an array of options involving both research methods and sources of information. Our purpose in so doing is to make clear that we, as researchers, choose among methods for a purpose, with an audience in mind, or at least we should choose in that way.

The discussion that follows focuses on issues of effectiveness of procedures or medical interventions, which is reasonable given that the conference for which this paper was originally prepared was on outcomes. Still, many other factors than specific medical interventions may influence outcomes and are deserving of research in their own right.

\* Corresponding author. Tel.: +1-520-621-7444; fax: +1-520-621-9306.  
E-mail address: ajf@u.arizona.edu (A.J. Figueredo).

Our purpose here is meant to show the range of methods available and the implicit, as well as explicit, choices that are involved in choosing one rather than any of the others. The ideas are applicable to research questions of almost any kind.

Although we will use the term ‘effectiveness’ generically to refer to the purely medical information about a procedure, whether diagnostic or therapeutic, except where context will make it evident we are doing otherwise, the distinction between efficacy and effectiveness must be preserved. The distinction proposed so nicely by Cochrane (1972) is critical to arguments that must be resolved between advantages of different kinds of research and research settings and is pertinent to the discussions to follow concerning generalizability (or external validity) of findings (see Cook & Campbell, 1979). We do note that the efficacy–effectiveness distinction is not absolute; the terms refer to ends of a continuum. On one end is an assessment of outcome of application of a procedure under tightly controlled, experimentally ideal, conditions, usually requiring substantial restrictions on samples, context, and variations in treatment. On the other end of the continuum is an assessment of application of a procedure as it would actually be employed in the field, usually with very few restrictions on samples, context, and treatment variations.

### 1.1. *Clinical experience*

At the only-in-some-ways simplest level of a research design, we could ask clinicians—or even patients!—about the effectiveness or utility of some medical procedure. In fact, we do just that for some procedures at the individual practitioner or patient level. If a physician says, ‘I get better results with a nasal spray than with oral doses’, we are likely to accept that judgment. If a patient says, ‘I don’t know, but I just seem to get more relief from aspirin than from ibuprofen’, we are not inclined to challenge that patient’s conclusion by citing a bunch of research studies.

The problem is that clinical experience will be somewhat unpredictably persuasive, and it is unlikely to be persuasive at all when the stakes are high. Clinical experience is not a basis on which truly important decisions—large dollar costs, potential for malpractice—are going to be made. The power of clinical experiences of even a large panel of practitioners to reduce uncertainty is limited. Clinical experience is likely to be taken seriously only when the issues themselves are not so serious: uncertainty is itself small in magnitude or the matter is not of great consequence, e.g., as in the choice between highly similar drugs.

### 1.2. *Case studies*

Every week the *New England Journal of Medicine* (NEJM) publishes a case study, presumably because something is to be learned from a case study well done. Indeed, case studies may be highly persuasive on certain narrow issues (Sechrest, Stewart, Stickle & Sidani, 1996). One in particular is refutation of an alleged universal. For example,

the claim that some condition is invariably fatal is successfully countered by a single case of survival *if* the case is sufficiently well documented and *if* it is representative of the population for which the universal is alleged. It is unlikely, though, that a single case could reduce by any useful amount the uncertainty surrounding most medical controversies. For example, one case of successful imaging of abdominal pathology by MRI would not do much to reduce more general doubts about the usefulness of MRI for such purposes. Such a single case would be likely to be dismissed as an anomaly.

### 1.3. *Study of medical records*

Another possible method of obtaining information about the effectiveness of a medical procedure would be to examine a series of medical records for evidence of its use that could then be related to case outcomes. If a certain procedure were regularly associated with favorable outcomes, especially if those outcomes were more favorable than those achieved in cases in which the procedure was not employed, then one might think the procedure effective. Such a conclusion requires, though, defending the propositions that anticipated case outcomes did not determine use of the procedure and that any comparison cases were equivalent to the cases in which the procedure was used except for use of the procedure. An experimental procedure might appear to be harmful if it were employed on cases not responding to other treatments. A new imaging procedure might appear useful if it were used with patients who had already gone through more standard procedures that had already established the diagnosis. The proposition that comparison cases were equivalent to experimental cases except for the procedure is more easily stated than defended. Portacaval shunt surgery for bleeding esophageal varices looked good until it was established that patients were excluded from that treatment because they were too sick to withstand the surgery. It took some time to determine that pancreatic cancer patients had greater coffee consumption precisely because the medical problems of comparison cases with gastric conditions had caused them to give up coffee (Feinstein, Horwitz, Spitzer & Batista, 1981).

The study of medical or administrative records for evidence of effectiveness of medical procedures is often useful. Whether it is probatory is at issue here. When it is used in situations of high uncertainty or in which decisions are critical, the approach may not yield results sufficiently persuasive to be useful. The method is not strong in the face of strong critics and doubts.

### 1.4. *Prospective observation*

Still another method of assessing effectiveness of a procedure is to try out the procedure on a series of cases and see how well it works. In fact, this approach to assessing effectiveness of procedures is so common in practice that it is probably not often recognized as ‘a method.’ For the most

part, prospective observation is used informally. Physicians may ‘try out’ a new drug or a new diagnostic test of some sort to ‘see if it works.’ Probably a lot of medical decisions, e.g., to shift from one drug to another, are made on the basis of such informal testing. The possibilities of bias, however, are so great as to rule out the use of any such ‘evidence’ in more critical decision-making.

### 1.5. *The experiment or randomized clinical trial*

The last method to be discussed here for producing evidence of effectiveness or utility of some procedure is the deliberate experiment or, as such experiments are often called in a medical context, the clinical trial. In a well-designed clinical trial, one has an experimental group that receives the procedure in question under strictly controlled conditions and a comparison group that can be assumed to be equivalent to the comparison group in all relevant ways save for the application of the procedure. Many possible ‘designs’ can be specified for a clinical trial, e.g., pre-test post-test, post-test only, completely randomized, or randomized after blocking (matching) on some variable, and so on, but they are all aimed at providing the strongest possible basis for the inference that any final differences between groups are attributable to the experimental procedure and to no other factor. Thus, experiments have considerable potential to reduce uncertainty. But, as we will point out later, experiments do not always reduce uncertainty, and they may reduce some uncertainties but not others.

Experiments are our most powerful tool for producing information, but they have several drawbacks that leave alternative strategies still attractive. If they did not have drawbacks, there would be no need for this chapter to exist, for one would merely need to insist: when in doubt, experiment. Experiments are, however, often difficult to plan and implement, slow to produce results, expensive to mount, restrictive in the interpretations they warrant, and, occasionally unethical. Sometimes, in fact, experiments are simply impossible, e.g., if the critical variable of causal interest cannot be manipulated or controlled. They can be frustrating and disappointing, too, since even after all the expense and effort put into them, they may still produce equivocal results.

## 2. **The calculus of uncertainty**

We have stated that the aim of outcomes research—or other modes of inquiry if we wish to restrict somewhat the use of the term research—is to reduce uncertainty about effectiveness of some medical procedure. We need to admit at the outset that the aim of some research is actually to *increase* uncertainty, and much other research has that aim without any intent. Rein and White (1977) have referred to this phenomenon as ‘muddying the waters.’ Often the medical community, or a part of it, is reasonably certain about what to do for a particular problem. For many years, the ‘Trendelenberg position’ (feet elevated, head

down) was firmly recommended as the proper way of handling emergency cases; certainty was high. Then research appeared that suggested that the Trendelenberg position was wrong; certainty was decreased, the waters were muddied. Ultimately, further research indicated the advisability of some elevation of the legs, but without lowering the head. Transurethral prostatectomy was ‘known’ to be the best surgical option until recently, when evidence cast doubt on the certainty of that notion (Wennberg, Roos, Sola, Schori & Jaffe, 1987).

In fact, the very existence of variations in outcomes of medical procedures is a cause for uncertainty, i.e., until those variations can be accounted for. If 70% of patients do well with a given drug regimen, but 30% do not, then that variation is a basis for uncertainty as to how to treat any individual patient. If, however, the variation can be accounted for, e.g., by some patient characteristic that is easily assessed, then uncertainty, at least for the 70% who do respond, is diminished.

Uncertainty should be considered multifaceted, or at least potentially so. One can be uncertain whether a treatment works at all or whether it will work in a particular case. One can be confident that a diagnostic procedure yields correct results but uncertain whether the results are worth the cost. One can be certain that a procedure is effective if applied at the right time but uncertain as to when to apply it. One may be certain that a procedure is effective in the hands of an expert but uncertain whether it can be effectively employed by less well qualified clinicians. Good information may reduce substantially any one aspect of uncertainty and leave all others unchanged.

Uncertainty is also dimensional rather than categorical. We are uncertain in degrees, and uncertainty is reduced by degrees when we know outcomes of research. An uncertainty function may very well reach asymptote well short of zero. That is to say, clinicians, like all of us, have to live with a fair degree of uncertainty. The aim of outcomes research is to reduce uncertainty to a point as close to asymptote as possible.

Because uncertainty is multifaceted, though, it has to be recognized that a strategy for reducing uncertainty of one kind may not only leave other kinds unchanged but actually increase them. For example, in order to reduce uncertainty about the theoretical limits of effectiveness of a procedure, it may be necessary to test it under such restrictive conditions that one is left less certain than ever about the generalizability of the conclusion concerning effectiveness.

Some decisions that must be made for or with patients are fraught with consequences; others are not. Some interventions may have substantial potential differences in outcomes; other interventions may have only small potential differences. Some interventions may differ greatly in cost, others only slightly. Some interventions carry great risks of side effects, others almost none. We may assume, then, that some decisions are critical, make a lot of difference, and others are not so critical in the sense of not mattering a lot.

Degree of uncertainty experienced by the clinician may be unrelated to the criticality of the decision to be made. Thus, a physician may be quite uncertain whether aspirin or acetaminophen would be more effective for a patient's headache but may not care a lot because either would give reasonable relief. A cardiologist may be uncertain whether to recommend angioplasty or coronary artery surgery and may care a lot about the decision. Going from a CAT scan to MRI involves a lot of extra cost. Measuring blood pressure in a supine as well as a sitting position would not add much to cost of diagnostic studies.

Judgments about the nature of evidence needed to reduce uncertainty, and therefore about the specifications for research designs, properly should depend on some multiplicative function of uncertainty and criticality of the decision. We should not require great rigor for answers that do not matter much or that do not reduce uncertainty by much. Think, for example, of the casual way in which most academicians go about selecting textbooks for their courses. No one demands a large scale, rigorous test to determine which of two or more textbooks would be most effective. That is probably because, first, we believe—with very little justification—that our very unrigorous method of inspecting books personally will result in a correct choice, and, second, we probably also realize at some gut level that the textbook does not make all that much difference anyway. On the other hand, a physician who has to respond to a cancer patient who asks whether post-surgical chemotherapy will really be helpful and worth the bad times wants more than just an opinion or an uncontrolled observational study done by enthusiasts for a new drug.

We are not proponents of weak methodologies, but we do know that strong methods are usually costly in several senses. What we need a better sense of is how the nature of our uncertainty and our decision options dictate, or should dictate, the methods of inquiry and the research designs we choose. We believe, in fact, that more careful thought about uncertainties and the quandaries they put us in will make it evident that our methods are almost always weaker rather than stronger than what is needed.

### 2.1. *The lens model*

The lens model of Egon Brunswik (1952) is a useful metatheoretical framework for dealing with the quantification of uncertainty in a functional context. This model distinguishes between: (1) the arrays of input converging upon and (2) the arrays of output emanating from the decisionmaker, which can be represented as an intervening pair of focusing lenses placed between these two arrays (Petronovich, 1979). In the study of biomedical decisionmaking, the input elements represent the various sources of information available to the decisionmaker (e.g., the physician or patient), and the output elements represent the various means of intervention available to the decisionmaker. In a Brunswikian analysis, the objective is to identify how the

multiple alternative sources of information are integrated and utilized by the decisionmaker to select among the multiple alternative means of intervention. This cannot be adequately accomplished without explicit reference to both: (1) the relative trustworthiness of the different sources of information and (2) the relative effectiveness of the different means of intervention. This is done by assessing the ecological validities of both: (1) the sources of information in relation to the distal or objective state of affairs that they presumably indicate in the external world and (2) the means of intervention in relation to the distal or objective achievements that they presumably produce in the external world.

In the Brunswikian framework, a distinction is also made between these objective 'ecological' validities and the subjective 'functional' validities manifestly attributed to them by the decisionmakers. An optimal decisionmaker can thus be characterized as one whose functional validities match the ecological validities for both the available sources of information and the available means of intervention. An analysis of the functional validities implied by biomedical decisions actually observed is thus purely descriptive, whereas an analysis of the corresponding ecological validities may be viewed as more normative. A complete normative analysis, however, must also include consideration of the relative utilities of the ultimate biomedical outcomes, or distal achievements. It is not sufficient that a certain outcome be attainable, it must also be desirable. Value judgments are therefore required to determine whether a certain biomedical strategy, of either diagnosis or treatment or both, is truly optimal.

Our review can therefore be organized along two Brunswikian and one extra Brunswikian concept: (1) the sources of biomedical information, (2) the means of biomedical intervention and (3) the utilities of biomedical outcomes. These correspond to the operations of biomedical (1) diagnosis, (2) treatment and (3) evaluation. Throughout the review, there will also be the explicit distinction between descriptive and normative analysis at all three levels. Although these issues are obviously not independent of each other, they will be considered separately for purposes of exposition, as prefatory to the needed synthesis. Both the ecological and functional validities of the sources of information available will be considered first in the analysis of biomedical diagnosis, and those of the means of intervention will be considered next in the analysis of biomedical treatment. Finally, the relative utilities of outcomes will be considered in the analysis of biomedical evaluation.

## 3. **Diagnosis: the sources of biomedical information**

### 3.1. *Serial versus parallel models*

Two distinctly different, but not mutually exclusive, conceptual models exist for the biomedical decisionmaking

process. These can be roughly characterized as ‘serial’ versus ‘parallel’ models of decisionmaking. ‘Serial’ models represent the decisionmaking process as a sequence of successive steps in which discrete binary judgments (e.g., yes or no) are made upon each of multiple inputs or sources of information. These binary alternatives are often represented as branching points, leading to the graphical representation known as a ‘decision tree’. The interdependency between decision elements is modeled hierarchically in that each successive binary judgment affects all others that follow in the series by limiting the further alternatives available. This serial dependency thus adds information to the system by reducing uncertainty, much as in a game of ‘twenty questions’. Such ‘serial’ models represent the approach most often used by formal ‘decision analysis’ in developing decisionmaking algorithms, as well as by iterative procedures of ‘elimination by aspect’ (Tversky, 1972).

‘Parallel’ models, on the other hand, represent the decisionmaking process as a simultaneous process of integrating continuously graded judgments regarding the weighting of the multiple inputs or sources of information. These are often represented as converging arrows rather than branching points, leading to the graphical appearance of an optical lens focusing incoming rays of light. The interdependency between decision elements is modeled simultaneously in that the weighting of each input is affected by that of all the other inputs or sources of information available. This is the approach favored by Brunswikians for ‘policy capturing’, as well as by proponents of ‘compensatory models’ of decisionmaking. Although, more recently, there have been proposed both parallel models with binary inputs and serial models with graded inputs, as well as models that convert one type of input into the other (such as in AI and PDP research), most decisionmaking models have, historically, fallen into one of the two principal categories described above.

The basic difference between these two approaches seems to be that serial models are better suited to deal with information inputs that are individually reliable and collectively cumulative, or mostly complementary, whereas parallel models are better suited to deal with information inputs that are individually fallible but collectively semiredundant, and thus partially intersubstitutable. It is conceivable that mixed models of decisionmaking could be developed that would deal with both of these plausible contingencies by constructing a hierarchical series, where appropriate, of simultaneous parallelprocessing modules, where required. The deterministic and cumulative features of each hierarchical branching point required by a serial model could be approached by the enhanced reliability provided by the multiple redundancy in each parallelprocessing module.

Another possible contrast between these two alternative approaches is that, whereas parallel models might better characterize certain forms of implicit and intuitive individual decisionmaking, serial models might better characterize certain forms of explicit and formal institutional

decisionmaking, as well as that of any individuals who have internalized the bureaucratic ethos of industrial society.

### 3.2. *Descriptive versus normative models*

Either of these types of model, or the combinations thereof, can be applied to either a descriptive or a normative analysis by representing either the functional or the ecological validities of each of the inputs, or sources of information. A purely descriptive analysis can be performed exclusively upon the observed utilization by decisionmakers of the various sources of information available, such as different diagnostic tests, without reference to the relative validity, or scientific merit, of those sources of information. A purely normative analysis, on the other hand, can be performed exclusively upon the relative validity, or scientific merit, of those sources of information, without reference to the observed utilization of those inputs by decisionmakers. The great majority of ‘decision analysis’ models found in our computer search of the literature and review of published abstracts for the last 5 years were of this latter type. These models were virtually identical to the kinds of stochastic optimality models commonly applied in both economics and engineering. A few models, however, were of the former type and thus more psychological in character. An even smaller minority of models compared selected descriptive to normative results, usually in order to decry the lack of correspondence between the observed and presumably optimal decision rules in Brunswikian terms, between the functional and ecological validities of the information input array.

In principle, a normative decision analysis can also be performed exclusively upon either: (1) the optimal utilization of the sources of information available in the biomedical decisionmaking process, without reference to the effectiveness of the means of intervention ultimately selected by that process in producing the desired outcomes or (2) the effectiveness of the means of intervention available in producing the desired outcomes, without reference to the optimal utilization of the sources of information available in the decisionmaking process. A normative analysis, however, is of limited value if directed exclusively to the use of diagnostic information. In addition to applying estimates of diagnostic certitude, most ‘decision analysis’ models also make explicit use of estimates of relative treatment effectiveness and outcome utility. Thus, a specification of the optimal use of diagnostic information is usually deemed necessary but not sufficient to constitute a pragmatically useful or complete biomedical decision analysis.

For example, a novel research tactic, called Small Area Analysis (SAA), does not fit neatly into our classificatory scheme. In SAA, microgeographic regions are compared on their mean levels of utilization of specific medical treatments. Where systematic differences are found, the analysis proceeds beyond the mere description of the heterogeneous

patterns of medical practice and into a causal analysis. Hypotheses are formulated involving the existence of initial objective differences in the patient population between microgeographic regions. If no relevant demographic differences are found, such as in the prevalence of the particular disorders requiring treatment, it follows that variations in the subjective functional validities of whatever decision-making algorithms are being applied cannot reflect the constant and objective ecological validities presumably underlying universal standards of medical practice. A comprehensive SAA goes further, however, to determine whether the ultimate medical outcomes of the arbitrarily different treatment regimes differ significantly between the areas. What is revealed by this step is not only whether different subjective decisions are made based upon the same objective informational inputs, but whether the different decisionmaking algorithms manifested in selecting interventions ultimately lead to significantly different medical outcomes. Only then can a pragmatic assessment be made of the relative appropriateness of the different strategies of treatment. The following section will therefore focus on the available means of intervention in the analysis of biomedical treatment.

#### 4. Treatment: the means of biomedical intervention

The estimates of treatment effectiveness used by most decision analysis models are obtained from empirical research. Sometimes, the results of a single efficacy study, such as a randomized clinical trial, are subjected to a decision analysis by the primary researchers. More often, pooled or composite estimates of treatment effectiveness are obtained from the corpus of biomedical literature by secondary decision analysts. When the results are equivocal, or the estimates deemed unreliable, decision analysis models are subjected to a 'sensitivity analysis' to determine how critical the precise model predictions are to the particular parametric assumptions made. The quality of the data regarding the relative effectiveness of the alternative means of intervention is therefore central to biomedical decisionmaking. The sources of scientific information available for determining these crucial parameters therefore warrant a more detailed scrutiny.

##### 4.1. Sources of data

A distinction is required between sources of data and types of research designs. Sources of data have to do with where our observations or numbers come from. Research designs are the strategies of obtaining or analyzing data in such a way as to make legitimate the causal interpretations that are wanted as the end result of the research. Data may come, for example, from existing medical records, from data sets collected for other purposes, from previously completed studies, or they may be collected *de novo*. In some cases new variables may be created from recorded

information not remotely related to the aims of the research at hand, e.g., analyses of personal diaries or family photographs to plot the progression of disability. Research designs are the basis for algorithms by which observations (numerical data) are allocated to particular statistical procedures aimed at determining the likelihood that variations in observations are attributable to some intervention. For example, in the case of a randomized experiment the algorithm is simple: determine what experimental arrangement gave rise to the number and then assign that number directly to that arrangement. In the case of a quasi-experiment, the algorithm may be more complicated: e.g., determine the value for the observational unit before the intervention and subtract that number from the number observed after intervention and then assign that difference (gain) score to the arrangement giving rise to the observation. Different strategies may result in algorithms that 'correct' observed numbers in diverse ways or that assign them to subunits for analysis on the basis of diverse characteristics, e.g., sex, age, comorbidity, and so on.

We are here more concerned with research designs and strategies for analysis than with sources of data *per se*. It is true, though, that sources of data and research designs are substantially confounded along the major fissure of randomized versus non-randomized designs. In general, reliance on existing records, for example, results in data that require correction or allowance for differences that might have arisen for reasons other than the intervention of interest. Randomized experiments, conversely, always require collection of data *de novo*. It is true, though, that occasionally one may encounter an 'as if random' (Lord, 1963) arrangement that was made for administrative reasons or that resulted from some more or less naturally occurring event and that may be explored by use of existing records. A clinic might have closed, for example, with its caseload being split between two other facilities in such a way that no bias would seem likely, and if the two receiving facilities handled cases in interestingly different ways, one might interpret differences in outcomes fairly directly and confidently. Doctors' strikes in Los Angeles and Israel appeared to have affected dispositions of sets of cases that by chance would have been handled differently but for the strikes, and retrospective analyses of records shed light on the value of medical interventions (see Greenberg, 1983; Slater & Ever-Hadani, 1984).

The point to be made here is that data from almost any source could, however correctly or incorrectly, be subjected to the algorithms and analyses of almost any design. One could, as an instance, obtain data on biliary tract surgery done in different ways at two clinics and simply compare outcomes. To do so would be to treat the data as if they had resulted from a randomized experiment. Similarly, one could take the data from a randomized trial, break them out by patient characteristics such as comorbidity, and analyze them by 'correcting' for any differences; that would represent the strategy of a quasi-experiment. Or

one might obtain from nursing home records the daily incidence of incontinence per patient over the previous 6 weeks (retrospective) and then initiate an intervention with a random half of patients followed by another 6 weeks of data collection (de novo and prospective). The resulting data might be treated as a simple randomized experiment using post-intervention data only, as a pre-, post-treatment analysis of change scores, or as two time series, one interrupted and one not.

Space does not permit elaboration on problems involving quality of data, but they are many and disturbing. Many of the measures used in health outcome studies are sensitive to the effects of many variables other than treatment and, hence, are not completely generalizable measures of treatment effects, and other variables may be inherently unstable. Much more attention needs to be paid to measurement problems of all kinds.

#### 4.2. *Strategies for generating data*

We believe that four basic strategies exist for generating data relevant to arguments about treatment effectiveness—or any other causal argument. That is, one wishes to make the argument that a given treatment if applied in the presence of a particular condition will result in a dependable change in that condition. (For the most part, in medical/health research one wishes to demonstrate improvement, but the task of arguing cause and effect is the same whether one is arguing for salubrious or deleterious effects.) One may, at the simplest level, note the results of application of a treatment and argue that they are so obviously contrary to theoretical or intuitive expectations that they compel acceptance of the cause-effect relationship. Second, one may observe a systematic relationship between the amount or level of the treatment applied and the outcome—the dose response curve—and therefore argue that the treatment must have produced the outcome. Third, one may observe a case (patient, subject, etc.) in one condition, apply the treatment, and then observe that the condition has changed, i.e., pre-condition minus post-condition equals treatment effect. Finally, one may observe a case to which a treatment has been applied and observe another case to which no treatment has been applied and argue that any difference must have been due to the treatment.

But what of randomized experiments, case-control studies, Non-equivalent comparison groups, regression-discontinuity analyses, and so on? Those and other ‘design’ features of studies are all specific research tactics aimed at enhancing the interpretation, i.e., persuasiveness, of one or the other of the preceding strategies for developing data. A randomized experiment, for example, is an attempt to buttress the comparison group strategy so as to make stronger the argument that the difference between groups is really attributable to the treatment. The addition of a pre-test so that the design becomes a randomized trial *with* a pre-post comparison is aimed at even further strengthening the argu-

ment. Similarly, designs in which treatment dosages are varied deliberately, even with reversals, are attempts to strengthen raw, naturalistic observation.

From a more systematic perspective, elaborations on the four basic strategies described here are attempts to deal in one way or another with ‘threats to validity’ of conclusions we draw from our studies (Cook & Campbell, 1979). Each of the four strategies can be implemented in diverse ways, nearly all of which will leave the conclusions from research vulnerable to one or more challenges to the causal interpretation of the relationship of the treatment to the outcome (threats to internal validity). Specific research tactics, e.g., blinding of observers to conditions, random assignment to groups, are directed at reducing vulnerability to one or more of the critical challenges (threats). Specific tactics may weaken one or more challenges in varying degrees and, hence, leave the causal interpretation relatively stronger.

The strategy of arguing from the obviousness of observed outcomes is not likely to be persuasive very often, in part because neither theory nor intuition provides sufficiently firm grounding for judging outcomes. Most often, arguing from the obviousness of observed outcomes is likely to be in the context of case studies or very small samples, adventitiously noted, and in exploratory stages of work. Many of the case studies in the NEJM depend for their interpretations on the argument that it is obvious that they should not have turned out as they did. Fleming’s discovery of the antimicrobial effects of penicillin may be another example: in theory, the cultures in the dish should have been thriving. Surgeons may be led to look carefully at certain cases because no reason exists why they should have died, perhaps as in early observations of halothane anesthetic. Ordinarily, though, this strategy of arguing from expectations will not be productive.

Correlations between administration of treatments and outcomes can be determined either when one or a small number of cases get several treatments whose outcomes can be observed or when a larger number of cases get treatments that vary in some way that can be associated with outcomes. The multiple treatments (doses) per case instance requires fairly quick response to treatment and equally quick decay of treatment effects or else successive doses may build up or interfere with each other. Some treatments (e.g., surgeries, electroconvulsive shock) are irreversible, i.e., one cannot observe a return to baseline, and others (e.g., surgery) are not only usually irreversible, but they can only be given once and in one standard dosage. Many treatments would be very difficult to quantify in our present state of knowledge, and, therefore, dose-response relationships cannot readily be estimated. The interpretation of observed correlation between treatment and outcome is enhanced by such research tactics as blinding and double-blinding, use of Non-reactive outcome measures, alternating treatments, systematic treatment variation, and so on. The results of the observational strategy should be highly persuasive under optimal conditions, but those may not be easily achieved.

The strategy of observing changes in cases from before to after treatment is widely employed; it has the advantage of simplicity and directness. What could be more straightforward than measuring some characteristic of a person before treatment, administering the treatment, and then measuring the characteristic of interest after treatment? Problems arise from the vulnerability of such simplistic approaches to a wide range of threats to validity of any causal interpretation of observed differences. Changes from pre- to post-measures may be observed because cases are changing anyway, because instruments change, because cases are expected to change (e.g., patient expectancies), because of secular historical events, and so on (Cook & Campbell, 1979). Therefore, such studies usually require design refinements such as double pre-testing (Boruch, 1974), quantification of treatment to reveal dose-response relationships, and monitoring for secular events and trends.

The best single step to strengthen interpretation of pre-, post-data is likely to be addition of a comparison group that did not undergo the treatment of interest. Comparison groups can add greatly to the persuasiveness of conclusions. In fact, comparison groups are often essential. The critical issue in employing them has to do with the likelihood that comparison groups differ from experimental groups in no relevant way other than in not having had the treatment of interest. That comparability is reasonably well assured when sample sizes are not truly small and when cases are assigned randomly to treatments, i.e., when the design is a true or randomized experiment. Problems arise when treatment decisions cannot be said to have been made randomly, for then the experimental and comparison groups may be different in some other relevant way, causing the research to be referred to as a 'Non-equivalent comparison (control) group design' (Cook & Campbell, 1979), which often poses formidable analytic challenges (Reichardt, 1979). When the assumption of equivalence cannot be defended, and that is usually difficult to do when the design is not randomized, comparison group designs are subject to various threats to internal validity. On the other hand, not all threats to validity are especially plausible, and sometimes a combination of good theory, large differences, and the implausibility of some Non-equivalencies may make comparison group designs quite persuasive. Often, it must be recognized, no other design choice that would be better is at the same time possible or feasible.

#### 4.3. Data from ongoing medical record systems

Many different medical records are kept that might be usefully exploited for study of medical outcomes. Pharmacies keep records that might be examined to determine drug usage by patients treated under different regimens. One might, for example, determine the number of refills for prescribed analgesics in groups of patients being treated in different ways for low back pain. Hospital records will provide information on a wide range of services delivered,

charges levied, complications, lengths of stay, and so on. Medical insurance plans can be especially useful sources of information. In their study of prostatectomy, Wennberg et al. (1987) used claims data to show that important differences in death rates exist between hospitals and that reoperation is more likely to be needed following transurethral than open surgery.

Ongoing medical records, including claims data, have several important advantages, as Wennberg et al. (1987) note. If records are well kept and if they are computerized, as most are likely to be these days, they can provide nearly immediate information, often at low cost. They can also make it possible to follow patients over long periods of time. Moreover, ongoing records are not likely to be biased with respect to the aims of researchers, e.g., pharmacies do not selectively record prescription refills, and second surgeries are not more likely to result in claims for some kinds of patients than for others. It is also possible by relying on ongoing records to assemble data on far larger numbers of cases than could be acquired on any other basis, thus providing for great statistical power and the opportunity to study relatively rare events. In their prostatectomy study, Wennberg et al. (1987) were able to obtain data on 4570 patients aged 65 and older who had undergone prostate surgery in a period of roughly 4 years in one of 16 hospitals.

The problems with ongoing record systems, however, are important and result in limitations, often severe, on the conclusions they will justify. To begin with, no one record system may contain all the information desirable for a given study, and matching records across systems may be difficult at best or impossible altogether. Pharmacy records, for example, are unlikely to contain any information at all on patient diagnosis, and few record systems will provide more than crude information on outcomes. Almost none, for instance, will contain information on functional status—except in rehabilitation units, for which functional status is the outcome measure. Wennberg et al. (1987) were not able to obtain satisfactory information from claims files in order to adjust for any possible initial differences between patients in comorbid factors. Thus, the finding that death rates following prostatectomies are higher in small hospitals requires, for any causal interpretation, i.e., small hospitals provide less adequate surgical services, the assumption that small hospitals do not get generally older, sicker patients.

#### 4.4. Sources of bias

Upon undertaking any investigation, one should try to discern possible sources of bias in any existing data or to anticipate those sources in data to be collected. If one knows that data may well be biased, one can either derive conclusions cautiously in light of that bias or, perhaps, take action to reduce or avoid the bias. Obviously, if one knows that bias exists and can estimate its extent, the data can be corrected for it statistically.

Bias often arises from reactivity of measures (Webb,

Campbell, Schwartz, Sechrest & Grove, 1981). Reactivity occurs when the act of measuring something changes the value of the measure. For example, patients who know that their pain is being assessed in relation to disability determination may respond differently from patients who know that their pain is being assessed to determine effectiveness of an analgesic. A physician who believes that he is being observed from behind a one-way mirror may behave differently from one who believes he is not being watched. Webb et al. (1981) offer many suggestions about how reactivity may be dealt with, but a strong, general recommendation is to use multiple sources of information that do not all share the same sources of reactivity, i.e., method variance.

When one wishes to make substantive generalizations across studies that apply different methods to measure related outcomes, the multitrait–multimethod approach (Campbell & Fiske, 1959) is often useful. This approach uses the evidence of converging results across different methods to identify common characteristics that are not limited to any single method of measurement, as well as of converging results across different traits to identify the systematic biases that may be introduced by any given measurement operation regardless of what characteristic is being measured. It also uses the evidence of diverging results across different traits, regardless of the measurement operations used to assess them, to discriminate between discrepant characteristics that may otherwise be confounded by shared method bias. Recent advances in confirmatory factor modeling have greatly aided in the multivariate analysis of complex multitrait–multimethod data (Ferketich, Figueredo & Knapp, 1991; Figueredo, Ferketich & Knapp, 1991).

Other sources of bias include systematic error, e.g., as might arise from always giving one's institution the benefit of any doubt about a value, deliberate error, poorly calibrated instruments, and insufficiently trained staff. Any source of bias, reactivity or error, is a research hazard when the bias is more likely in one group or at one point than another. An instrument that systematically reads 'too high' will not necessarily lead to erroneous conclusions if it is consistent throughout a project, but if an instrument drifts from a pre-test to a post-test period or if an instrument used in one office is less well calibrated than a similar instrument in another office, errors might be made that would affect conclusions about relative effectiveness of treatments.

#### 4.5. Choice of outcomes

The choice of outcome measures, or endpoints, is not always obvious and is often difficult or even arbitrary. Survival, for example, is often taken as the ultimate outcome in treatment of serious illnesses, but current medical knowledge often permits sustaining of life for long periods of time but with minimal apparent additional gratifications. Adjustments to survival time to represent quality of life have been suggested in the form of the

Index of Well-Being (see Kaplan, 1985) and Quality Adjusted Life Years (Weinstein & Stason, 1977). When outcomes are very different, as might be the case for survival time and for Quality Adjusted Life Years, aggregating outcomes across studies as is required for a meta-analysis may result in a nearly meaningless metric. On the other hand, confining outcome measurement to physiological parameters in order to achieve uniformity of meaning may be equally meaningless (Kaplan, 1990).

Particularly troublesome is indication of opportunistic selection of outcome measures to capitalize on observed differences, e.g., choosing patient satisfaction for one treatment or one study and a physiological parameter, e.g., reduction in blood pressure, in another. A particularly interesting example of shifting frame of outcome reference may be seen in the report of a recent consensus panel on adjuvant therapies for breast cancer. Chemotherapy was recommended because it seemed to be associated with reduced reoccurrence, although no evidence could be found for increased survival time. Radiation therapy was recommended because, although it does not appear to reduce reoccurrence, evidence suggests that it may result in increased survival time.

Measured outcomes should, like other variables, be considered as representative of a population (of some sort) of variables to which one might generalize. Even survival, for example, is not the final, gold standard outcome. Consider the possibility that a treatment might guarantee 5 years of coma for a patient. When we elect survival as an outcome, we really—implicitly—have in mind that the patient will also be sentient, will have moments of pleasure, will experience a favorable ratio of positive to negative emotions, and so on. We probably would measure all those and other things if we could. Similarly, an outcome such as reduction in tumor size is of interest only to the extent that we can generalize from the outcome to a range of other characteristics suggestive of improved condition. Care should be taken in choosing outcome measures and constructing instruments for them to consider the probability that any observed effects would generalize beyond the specific outcomes and instruments employed.

#### 4.6. Measurement problems

Research design problems cannot, or at least should not, be considered separately from measurement problems, for any of the latter may have profound effects on the former. Nothing like the full array of measurement issues can be considered here: validity of measures, metric properties of scales, sensitivity of measures, reliability, and so on. Validity is a necessary property of any measure so obvious as scarcely needing mention; we are not interested in measures that do not tap the characteristics with which we are concerned. Not all measures are necessarily valid, but invalid measures are likely to be exposed early on. Obviously, we also want measures likely to be sensitive to interventions

we are testing, to some extent a matter of validity. Mortality may not be a sufficiently sensitive measure to be useful in most research contexts. We need to know more about the metrics in which our variables are registered, especially in light of growing recognition that many, perhaps most, of our observed variables are imperfect representations of the latent variables that are of real interest.

Reliability of measurement, however, requires special mention here because it is directly tied into research design problems. The critical reliability issue is that no intervention can have an effect on an unreliable dependent variable. Put another way, to the extent that a variable is unreliable, it can be considered random and, therefore, by definition, independent of any other variables. Unreliable dependent variables will, then, attenuate estimates of the effects of interventions (Sutcliffe, 1980). Any such attenuation will, in turn, decrease the statistical power of an experiment—or other type of study—and require a larger sample size in consequence.

Reliability issues are also of importance in relation to independent variables, including the intervention. Understandably if, let us say, a measure of attitudes is to be used to predict response to treatment, and the attitude measure is not reliable, we may mistakenly conclude that attitudes are not related to treatment outcome. What is not so widely recognized, though, is that the concept of reliability applies to the intervention itself (Sechrest, West, Phillips, Redner & Yeaton, 1979). We are inclined to think of an intervention in binary terms, 1 and 0: someone either got the intervention or did not. Some interventions are not, however, dependably (reliably) delivered, i.e., some persons may get more of the intervention than others, get it in more complete form, get it at more nearly the right time, and the like. Such variability means that the representation of the intervention as 1 and the Non-intervention as 0 is an unreliable index of the treatment condition of interest. That, then, means that the effect of the intervention may be misestimated. The great relevance of the concept of reliability to independent variables is apparent when we note that some variables may be dependent variables under some circumstances, thus creating concern about their reliability, but independent variables under other circumstances. Concern for reliability of a variable should not be different simply because a variable changes positions in some theoretical or analytic scheme. If we do not know that our measures are unreliable and by about how much, we will not be able to make good decisions about crucial aspects of research.

#### 4.7. *Individual studies: causal inference and generalization*

For an individual study, two related questions may be raised regarding the scientific quality of the results. The first is the issue of internal validity. This reflects how ‘well-controlled’ a study is in terms of precluding the possibility of plausible alternative causal hypotheses (i.e., ‘confounds’) by experimental manipulation and thus

permitting unequivocal causal inference. Internal validity is thus a property of the ‘ideal experiment’, and is the favored model for a randomized clinical trial, or ‘efficacy’ study. Although rigid experimental control does indeed facilitate causal inference, it leaves open the question whether the demonstrated ‘efficacy’ extends anywhere beyond the idealized conditions of the study. This question of generalizability of results is called the issue of external validity. The problem here is that the more tightly controlled, or internally valid, an individual study is, the less likely it is that it will constitute a representative sample of any natural population. This means that it may lack external validity for its intended ‘universe of generalization’. Unfortunately, since the real world is often full of confounds, the more externally valid an individual study becomes, the more internal validity, and thus unequivocal causal inference, may suffer as a consequence. Thus, ironically, the more an individual study measures realworld ‘effectiveness’, under more natural conditions, the less it becomes able to support claims of ‘efficacy’, in principle.

The concept of external validity is usually applied to generalization across populations of patients, physicians, and peripheral conditions of treatment. By applying it to the definition of the treatment itself, however, we may assess what has been called the construct validity of treatment. Construct validity, consisting of convergent and discriminant validity, is a psychometric concept that addresses critical questions of measurement. It asks whether an instrument measures what it is intended to measure (convergent validity) and not something entirely different (discriminant validity). These psychometric criteria can be fruitfully applied to biomedical treatments. We may ask, generalizing from Campbell and Fiske (1959), whether an intervention operates as it is supposed to operate (convergent validity) and not through some alternative mechanism (discriminant validity). The answer can be determined by comparing results across either naturally occurring or artificially generated variability in treatment administration. The essential features of the treatment can be defined as those on which the desired outcome can be seen to depend critically. Non-essential features of the treatment can, therefore, be identified as those that do not significantly affect the outcome. Construct validity is then seen as a kind of external validity, or generalizability, across systematic variation in the very definition of the treatment.

#### 4.8. *Case study*

Obviously, no one is going to propose reliance on a case study for determination of effectiveness or utility of a medical procedure. It is, however, instructive to ask why not as a stimulus to thinking and planning for more persuasive approaches to answering the questions. Although limitations of case studies are often couched in terms of limited generalizability, that may often be no more of a problem—sometimes even less—than the lack of any sound basis for a

causal interpretation. The problem begins with the fact that most cases are *selected* exactly because they are unusual in some respect. NEJM cases, for example, are never of the variety ‘an ordinary case of diabetes treated in the standard way with no remarkable consequences.’ Hence, case studies that one might want to adduce in favor of some causal conclusion have usually been noticed because they appear to pertain to the conclusion. In life, however, the improbable happens now and again: people abruptly get well, people abruptly die, people behave in unusual ways. In larger scale studies, it is less probable that many people might just spontaneously have recovered from an unusual disease. Norman Cousins’s recovery from ankylosing spondylitis was remarkable enough but still not fully persuasive of the power of his hypothesis about the effects of humor on illness. Had 15 or 20 people recovered in a similar way, medical authorities would have been forced to take notice. Case studies are useful enough, but they are too readily challenged to be useful for probatory purposes.

#### 4.9. Case control

Case control methodology results in a variant of what is more generically known as the ‘Non-equivalent comparison group design’ (Cook & Campbell, 1979). The essence of case control methodology involves identifying ‘cases’ of some sort, e.g., cases subjected to some diagnostic procedure or cases treated in some one way, and then identifying ‘control’ cases as similar as possible in all save the critical variable, the test or therapy. If cases differ systematically from controls in outcomes, the inference that the difference is caused by the difference in the intervention is attractive. The problem is that the inference requires the assumption that cases and controls were really equivalent in all important ways other than the intervention. That can often be a very large assumption.

One attraction of case control methodology is that, by relying on extant cases and data, it is likely to be far less costly than prospective experiments. A second attraction is that case control methodology is usually applied retrospectively, i.e., to cases and controls already completed, thus providing information very quickly in comparison to the much longer perspective of a clinical trial. A third attraction is that case control methodology does not require any administrative changes, does not result in any logistical problems, and so on. It is simple and straightforward, completely Non-disruptive. It does not even require a human subjects clearance since it focuses on changes in procedures done without the aim of experimentation. All these attractions are counterbalanced by potential problems stemming from uncertainty that cases and controls can justifiably be assumed to be equivalent.

Case control methods result inevitably in some variant on the Non-equivalent comparison group research design (see Reichardt, 1979). That is, a comparison group must be constructed in such a way as to support the argument that

cases are as they are because of the presence or absence of some critical variable in which they differ from the comparisons. In the context of this paper, cases must be thought to differ from comparisons in, and only in, the nature of the treatment they will have received; only if that is the case can it be argued completely persuasively that differences in outcomes are the result of differences in treatment. Large problems stem from the fact that it is often difficult to make the crucial assumption that cases and comparisons differ only in treatment. Non-equivalent comparison groups leave studies vulnerable to several threats to internal validity that are often highly plausible and even more often at least arguably plausible to those who would doubt the findings of treatment differences.

The problems start with the inescapable fact that for some reason cases will have had treatments different from those given to comparisons. At the most obvious level, then, case control methods are vulnerable to selection biases. That is, cases and comparisons may differ in treatments because of other ways in which they differ, and some of those ways may be relevant to outcomes. If, to suggest one example, women are selected to have hysterectomies because they appear strong enough to endure the surgery and comparisons are less likely to have the surgery because they seem more fragile, then outcomes would not necessarily be expected to be the same even if both groups were treated the same way. Exercise stress tests are likely to be given only to relatively strong post-MI cases and, hence, the use of that test may be spuriously linked to a positive outcome. When cases and comparisons are identified within the same time frame and other treatment parameters, one must be intractably suspicious of unrecognized differences between groups. If patients treated during the same time interval, in the same or highly similar places, by intersubstitutable physicians are, nonetheless, treated quite variably, then one must suspect that some variation characteristic of patients but not necessarily explicated is responsible for treatment variation.

The persuasiveness of case control methodology depends on the proposition that cases did not become cases and controls did not become controls for reasons related to outcome status. Suppose, for example, that a new treatment for a medical condition became available and began to be widely used in a medical center in June, 1989. A case control study might be developed by comparing cases treated after June, 1989, by the new treatment with controls treated previously by a different method. If the cases had more favorable outcomes, that might be taken as evidence for effectiveness of the new treatment. If, on the other hand, the time of initiating the new treatment coincided fairly closely with change in diagnostic criteria or improvement in diagnostic methods, the apparent advantage of the new treatment might be entirely specious. Instances of artifact have abounded in the literature involving case controls. The problem is that we do not yet have a method for identifying likely problems in its application.

If cases are cases because a new treatment has only just become available and, therefore, could not have been used with comparisons, the obvious selection bias may be less plausible. Case comparisons are often constructed by identifying cases treated from the time a new therapy became available and controls treated by previously available methods. If the new treatment was not available, bias in assigning it could not have operated. Unfortunately, that version of the case comparison method is vulnerable to other critical threats that manifest themselves over time. First, diagnostic acumen may increase over time, leading to earlier identification of cases and, therefore, more favorable intermediate outcomes. That very phenomenon has been suggested as an explanation for otherwise seemingly favorable results of cancer treatment (Feinstein, Sossin & Wells, 1985). Second, changes in ancillary or supportive treatment may occur over time: nursing services may improve or deteriorate, better antimicrobials may reduce secondary infections, physicians may be more experienced, and so on.

#### 4.10. *Randomized clinical trial*

The clinical trial, as it is called in biomedical research, or the randomized experiment, is, when well-implemented, unquestionably our strongest research tool in terms of its warrant for causal inference, i.e., internal validity. The qualification ‘when well-implemented’ is advisable, for randomization does not *assure* anything. Randomization itself is a process, and like any other process it can go awry; investigators are wise to devise a method of randomization that is infallible and then to monitor it anyway. Unless samples are very large, randomization may not necessarily produce equivalence of groups. In one experiment involving 10,000 cases assigned randomly to two conditions, randomization did not produce initial equivalence (Dales, Friedman & Collen, 1979). Randomization does not ensure that treatments are carried out properly, that outcome measures are adequate, that data are analyzed properly, nor protect against.

We have already mentioned the problems of generalizability that may arise from the restrictive experimental requirements imposed by formal experimentation. Cases are usually selected on the basis of exclusion and inclusion criteria that assure reasonable homogeneity within groups, hence reducing error terms for purposes of statistical testing, and that assure that all cases are really suitable candidates for the treatment(s). These restrictions may mean, however, that the results are not fully generalizable to everyday medical care. Similar restrictions limiting generalizability may stem from standardized treatment doses (e.g., Seltzer, 1972), using treatment protocols that may not be followed in extra-experimental settings, close monitoring of cases (e.g., for side effects), and so on.

Randomized experiments do not guarantee unambiguous results and firm conclusions. For one thing, randomized experiments as planned do not always end up as randomized

experiments in realization. Attrition, for example, is a virtually ubiquitous problem in any field experiment of consequence, and attrition is very often differential across groups. If even one subject is lost from one group, the study is, strictly speaking, no longer randomized. Obviously, such a defect would not be cause for great concern in most circumstances, but it is difficult to know just when to draw the line, when attrition is large enough that randomization must be regarded as seriously suspect. Even if attrition is not different across groups, it may occur for different reasons and, therefore, jeopardize analyses and interpretations. Empirical data of any kind are almost always messy, and that applies to experimental data as well as to any other kind. Outcome measures are variable within groups. Some people may appear to have benefitted greatly from a treatment and others not at all, not to mention the subgroup that might actually have been harmed. Results have to be tested against some error term that may itself be messy. For example, results obtained by one investigator may be significant and those by another not for no reason other than differential precision of the two experiments. Determining whether the error term in any one experiment may have been ‘too large’ and, consequently, implicated in failure to find significance is often nearly impossible.

The combination of these and other problems has, of late, cast more doubt on the value of single experiments than was likely the case 25 or more years ago. Standards of evidence now appear often to demand multiple studies, multiple sources of data, diverse investigators, and the like. These multiple lines of evidence then require methods of synthesis, meta-analysis currently being the most popular. We turn to that topic now.

#### 4.11. *Aggregation of individual studies: quantitative reviews*

It has been argued that the severe and conflicting demands of internal and external validity, including construct validity, are both too high and too contradictory to expect any single study to simultaneously satisfy. An alternative analytic strategy is to use the results of many converging studies to compensate for the individual deficiencies of any single one. Although some largescale program evaluations are able to assess effectiveness across a wide range of treatment settings, population demographics, etc., most individual field trials are very limited in this regard. Given adequate internal validity of individual studies, the ‘heterogeneity of irrelevancies’ (Cook, 1990) between experiments can be used to address the wider issue of external validity. The question then becomes whether these heterogeneities are indeed irrelevant, representing random sampling error, or are instead relevant, representing systematic variance. This decision may affect whether a treatment is deemed merely unreliable, or whether lawful structural parameters or boundary conditions are discovered that moderate treatment effectiveness.

#### 4.12. *Small area analysis*

One way of tapping this natural ‘heterogeneity of irrelevancies’ (or ‘relevancies’, as the case may be) is the approach of Small Area Analysis (SAA) discussed above. This involves conducting new studies with that explicit intent. SAA studies begin with an observation, however arrived at, that variations in the frequency of some medical procedure or practice are larger than one would imagine they ought to be and then: (1) the determination that variations are not sufficiently accounted for by differences in actual health problems across the areas and (2) determination whether variations in practice are reflected in commensurate variations in outcome. Condition 1 is required in order to make a SAA worthwhile. If, for example, it were discovered that tonsillectomies were more frequent in Minnesota than in Arizona, one would scarcely be surprised. If, by contrast, hip replacement surgery were much more frequent among the elderly in one popular retirement area than another, that might be worth further investigation.

If outcomes are commensurate with variations in practice, given that variations cannot be accounted for by actual health problems, it follows that one may argue that the case for the variation with the better outcomes is supported. If more reliance on tonsillectomies is associated with better health outcomes, then recommendations about increasing rates of tonsillectomy might be warranted.

On the other hand, if variations in outcome are not commensurate with variations in practice, then the worth of the more frequent, more expensive, or more risky medical practice is called into question. If greater frequency of total hip replacement is not associated with better outcomes than those achieved by a more conservative strategy, then the advisability of the surgery is doubtful. Or at least the advisability of what would appear to be excess surgery is doubtful. SAA cannot answer questions about the management option that is less extreme. That is, even the frequency of total hip replacement surgery in the lower frequency area might be greater than advisable, but that could not be determined from SAA unless a small area with an even lower frequency could be found.

#### 4.13. *Data synthesis*

Another way of proceeding is the approach of meta-analysis. This approach involves searching the literature, usually exhaustively, for the results of studies not explicitly designed for this analytic strategy. Implicitly, the hope is that the heterogeneities in the literature will constitute a representative sample of the heterogeneities in the world, or at least those of some intended ‘universe of generalization’. The results of a meta-analysis will therefore have greater external validity than any individual study that went into it. The main concern should be the presence of systematic bias in the literature that would compromise that claim. An important issue then becomes that of the optimal

selection of studies for meta-analysis. One approach is to screen studies carefully for internal validity, on the assumption that there can be no question of generalizability where the basic veracity of the causal inferences themselves are in question. This approach, however, may introduce systematic bias reflecting the specific criteria used for the selection of studies. An alternative approach is to not screen studies for internal validity at all, on the assumption that any confound present in an individual study is not likely to be present throughout the entire sample of studies and therefore will not support invalid generalizations anyway.

As discussed above, related to external validity is the issue of construct validity, this time as applied to the generalizability of results across different outcome measures. It is often the case that measures of biomedical outcomes across the individual studies sampled by a meta-analysis are highly idiosyncratic. The question then becomes whether to ‘lump’ or ‘split’ such results in meta-analysis, depending on whether the differences between measures are conceived of as ‘irrelevant’, being different indicators of the same hypothetical construct, or indeed ‘relevant’, being indicators of entirely different constructs and thus representing conceptually distinct outcomes. Whether to split or lump is, of course, an empirical question that could be addressed by either: (1) a single large experiment in which common factor models for the different outcome measures are tested or (2) a more sophisticated meta-analysis in which structural models regarding either the conjoint or the differential causation of the putatively different outcomes are evaluated.

Meta-analysis is not a method of collecting data. Rather, it is a method for systematic synthesis of data already available. Obviously, the conclusions of a meta-analysis cannot be better than the data base that exists, although the conclusions may be considerably beyond those permitted by any one study. A properly done meta-analysis provides a basis for making a decision between two or more alternative dispositions of cases. A meta-analysis may or may not provide much guidance on how to get to the point of making decisions between alternatives. If data are available though, a meta-analysis may yield information specific to particular characteristics of cases such as alternatives dependent on case severity or comorbidity.

Meta-analysis, like the randomized experiment, is a potentially powerful tool when appropriately used. We doubt that a swamp of murky data can be made to yield much even by the most elaborate meta-analytic techniques. A report that recording errors constitute from 1 to 40% of data in studies is disturbing (Rosenthal, 1984). A large and reasonably good data base can, however, be rendered clearer and much more useful. The key to better meta-analytic conclusions lies in better original studies as input.

Meta-analysis is a methodology still under development but with substantial advances having been made in recent years (Hedges & Olkin, 1985). Some substantial disagreements still exist, a critical one having to do with criteria for inclusion of studies in analyses. Chalmers, Levin, Sacks,

Reitman, Berrier & Nagalingam, 1987a,b, and Peto (1987) have been among the leaders in biomedical research insisting that only results from clinical trials should be entered into meta-analyses. Results may well differ as a function of data sources. Brown (1988) found that effect sizes in control groups studies of diabetes care were smaller than those from one group pre-test–post-test studies. On the other hand, Stevenson and Black (1988) found in an analysis of studies of effects of paternal absence on sex-role development in children that effect sizes were larger in published papers than in dissertations and conference papers. Dickersin, Chan, Chalmers, Sacks & Smith (1987) investigated findings in clinical trials completed but never published and found that they were smaller than for published studies, indicative of publication bias. What is one to make of this? If the aim of a meta-analysis is limited to determining whether a treatment is effective or not, some variation in effect size is tolerable. On the other hand, if one needs to enter an effect size into some equation predicting cost or another utility measure, the value of the coefficient can make considerable difference. If one wants merely to summarize findings from over a set of clinical trials, excluding other types of studies is justified; if, though, one is concerned that bias may have resulted in serious distortions in outcomes through selective effects on publication, then examining the unpublished literature may be imperative.

Meta-analysis is a variant on standard research methods, but one in which the individual study is usually the unit of analysis, i.e., each study is a case. It follows that many, perhaps most, of the recommendations that one would make about research methods apply as well to the meta-analysis. Three examples should suffice here, but they only illustrate the general approach. First, meta-analysts should show appropriate concern for reliability of their methods. If they claim to be retrieving an entire body of literature, they should show evidence of the thoroughness of their procedures. In coding characteristics of studies, they should demonstrate agreement between raters. Orwin and Cordray (1985), for example, showed agreement as low as .4 for some variables coded for the same studies by two groups. In two meta-analyses with ten studies in common, two of the ten were classified as randomized in one analysis and nonrandomized in the other. Second, meta-analysis requires a plan for statistical analysis, just as do other studies. *All* analyses across study variables are nonexperimental, and that must be allowed for, e.g., by allowance for multiple comparisons (as with Bonferroni adjustments in alpha), better prior theory, and so on. Third, most meta-analyses are statistically underpowered since the number of cases (studies) on any one topic is likely to be limited. Meta-analyses may be underpowered even for main effects, but they are particularly underpowered to detect subgroup differences, e.g., between studies with good and poor methodology, between interventions implemented one way versus another. Investigators should give more considera-

tion to issues of statistical power in meta-analyses and limit their ambitions accordingly.

Meta-analysis is, as stated, a potentially powerful and useful research tool. But like most power tools, it can easily be misused. Chalmers and his associates (Chalmers et al., 1987a,b) have found that meta-analyses of the same topic by different investigators not only may not agree, but, in a disturbing number of instances, they may even reach opposite conclusions. What kind of policy decisions does that lead to?

Decision analysis also is not a method of collecting data but is a method for synthesizing data already available. Decision analysis, unlike meta-analysis, synthesizes across types of data to provide support for a serial decision process. That is, decision analysis should permit informed choice as one goes through a series of decisions that culminate in terminal decisions. Meta-analysis might even be the source of information at one or more decision points, e.g., whether to make use of one treatment or another. Unlike meta-analysis, decision analysis does not result in any new information. It is a way of systematizing information already available.

#### *4.14. The role of theory in demonstrating effectiveness*

It is evident, at a purely descriptive level, that means of intervention based on clear predictions from established theories command more credibility than those from more questionable points of view. As a result, in practice, if not in principle, the effectiveness of those means of intervention are commonly deemed to require less empirical demonstration than otherwise (see Lipsey, 1990). An intervention with a plausible mediating mechanism or proposed mode of action is less likely to occasion the skepticism of the biomedical community than one derived from purely empirical data or one that cites a disreputable causal theory. One example of this theoretical doubt is the case of acupuncture. Skepticism in Western medicine and consequent demand for proof was widespread as long as the mode of action was attributed to the channeling of ‘chi energy’, as claimed by the traditional practitioners. This skepticism abated somewhat as more credible mediating mechanisms, such as endorphins and enkephalins, were dragged into it. The same thing happened with the phenomenon of ‘mesmerism’ in psychology, which long suffered from association with the theory of ‘animal magnetism’ and the related theatrics of early proponents. This doubt abated when it was redubbed ‘hypnosis’ and adopted by more reputable theorists who unceremoniously doffed the mysticism, flowing robes, and magnets.

Although as fellow Western scientific materialists, we may sympathize with the above particulars, we must concede that sometimes we permit strong theory to serve as a proxy for external evidence, a point worthy of more critical examination. Although radical empiricist would quickly condemn such a practice, some justification can be offered for this trust in theory. In the more modern

hypotheticodeductive view of science, a theory is corroborated by its correct empirical predictions. A well-supported causal theory, therefore, to some extent stands behind a new and unequivocal prediction. This circumstance arguably brings to bear some of the weight of evidence supporting that theory to the credibility of the novel prediction. The danger of this practice, however, is in the historical role of theory as belief and sociocultural construction, which may be partially irrational and not based on an adequate foundation of empirical evidence.

## 5. Evaluation: the utilities of biomedical outcomes

### 5.1. *The basis for assignment of utilities*

In marked contrast to estimates of relative treatment effectiveness, the estimates of the relative utilities of biomedical outcomes used by decision analysis models are only sometimes empirical in derivation, and then variably so. One common index of utility is monetary cost (Kaplan, 1985). Whereas few would recommend relying on money as the sole measure of the concept of utility, and thus (by default) the operational definition, most would recognize it as one important element in a multiple operationalization.

Often, any nonmonetary outcome utilities are arbitrarily assigned by the decision analysts themselves, as when there is a clear and uncontroversial choice between utilities of 1 and 0, such as life (presumably in perfect health) and death. Perhaps more often they are developed by consensus among a panel of expert judges. Technically, these judgments qualify as somewhat empirical estimates, but they may fail to hold up under critical scrutiny on methodological issues, such as that of parametric generalizability beyond the group of panel members sampled. Less common are more psychological studies in which there is a reasonable attempt to obtain utility estimates from a representative sample of some well-defined population. Least common, and perhaps fortuitously so, are the purely armchair philosophical remarks of ethicists and such on the morally right and proper way to establish said utilities. These latter have, however, correctly pointed out that current applications of expected utility theory uncritically incorporate certain fundamental ethical assumptions, such as the ‘happiness of the greatest number’, lending a deceptive air of rationality by association to what are perhaps arbitrary and problematic ethical principles.

The quantifying of utilities is the area of biomedical decisionmaking that is least sophisticated in either methodological rigor or just plain clear thinking about the nature of the problem. For example, assuming that adequate techniques are available for sampling utilities from a duly constituted representative sample, what exactly is the population of interest, or intended ‘universe of generalization’? In assigning unidentified ‘utilities’, may we assume complete confluence of interest between patients and physicians, or between

either and, say, insurers? What about hospitals or regulatory agencies? If not, to which utilitarian interests do we give primacy? For whom do the decision analysts work? One possible solution that has been suggested to this dilemma is to specify precisely what utilities are being estimated and what decision rules are being applied. A more critical multiplist alternative (Cook, 1990; Shadish, 1990) would be separately to simulate model predictions using the utilities obtained from different subject populations. Such substitution of values would constitute a kind of representative sensitivity analysis to the issue of what set of stakeholder interests are being considered. These are not merely academic protestations of armchair ethicists but issues absolutely crucial to the scientific enterprise of valid psychometric measurement. If they are to be applied as influential coefficients in our quantitative models, we need to know just what those numbers mean. If we do not know with exactness what our assigned ‘utilities’ represent, the whole result becomes uninterpretable.

That is the good news. The bad news is that, even given much clearer thinking on such thorny issues, there remain a number of quite formidable methodological obstacles to the very measurement of subjective utility. For example, perhaps the most common technique for the empirical estimation of relative utilities is called the ‘standard gamble’. Subjects are presented with a choice of mutually exclusive outcomes. These alternatives are then assigned complementary probabilities (such as  $p$  and  $1-p$ ) of occurrence. Subjects are then asked to choose between the two outcomes as weighted by these different likelihoods. The relative probabilities are then adjusted by the experimenter until the subjects are unable to make a clear choice of stochastic preference. As in microeconomics, this is called the ‘indifference point’. The probability coefficients assigned at this point of indifference are then used as quantitative indicators of the relative utilities of the two alternatives.

Although the standard gamble appears eminently rational, the whole procedure critically depends on the ability of subjects to first make and then apply accurate magnitude estimations of relative probabilities. Unfortunately, psychological research has historically shown that many people are notoriously poor judges of numerical probability. These findings have been used by some as an opportunity to cast doubt on expected utility theory as a whole and not just the methodology that is particular to the standard gamble. More recent evidence, however, indicates that there might be some hope for this method if the likelihood information were presented in the form of the relative frequencies of different occurrences rather than the complementary subjective probabilities of single events (Cosmides & Tooby, 1992), which is arguably a more natural heuristic for the processing of human numerical experience. Thus, the purported invalidation of this application of expected utility theory might be no more than a laboratory artifact of one specific method of presentation for the decision-making task.

Nevertheless, the alternative events provided by the standard gamble are often also hypothetical in that subjects have had no direct experience with the outcomes in question. Furthermore, it has been shown that the processing by naive subjects of this kind of hypothetical information is often quite variable, with performance being dramatically dependent upon the content of the task rather than the abstract logical form of the problem (Tooby & Cosmides, 1992), presumably activating different dedicated information-processing modules with distinct domain-specific operating characteristics. The implied probabilistic consent is therefore not adequately informed. Research has also shown that people who have actually experienced certain biomedical states, such as colostomy, assign very different utilities to these states than people who have not directly experienced them. Even given this information, are outcome-inexperienced subjects overestimating disutilities based on irrational fear and revulsion, or are outcome-experienced subjects underestimating disutilities based on self-defensive rationalization and resignation?

### 5.2. *Discrete versus graded utility functions*

Another interesting feature of the current use of expected utility theory is that it appears to have been applied exclusively to the evaluation of discrete (e.g., binary) outcomes. There is, however, no immediately obvious mathematical reason to preclude its application to continuously graded outcomes. A certain degree of graded effect could be proportionately evaluated by a utility coefficient in much the same way that an all-or-nothing dichotomy is evaluated. Since the probabilities of discrete outcomes can be modeled as mere special cases of regression coefficients for which the criterion variables are dichotomous (Cohen & Cohen, 1983), effect sizes from multiple regression analyses in general could perhaps be used to construct optimality models in which the graded levels of the dependent variables are weighted in such utility functions. The graded equivalent of a discrete decision tree is therefore a path analysis, in which the multiplication of successive path regression coefficients in the computation of indirect effects is the exact functional equivalent of that of a series of Bayesian conditional probabilities. Utility coefficients could be represented as additional causal pathways to a single 'utility' construct from all the dependent variables. Such an application of path analysis would also permit the combination of 'serial' and 'parallel' model components suggested earlier.

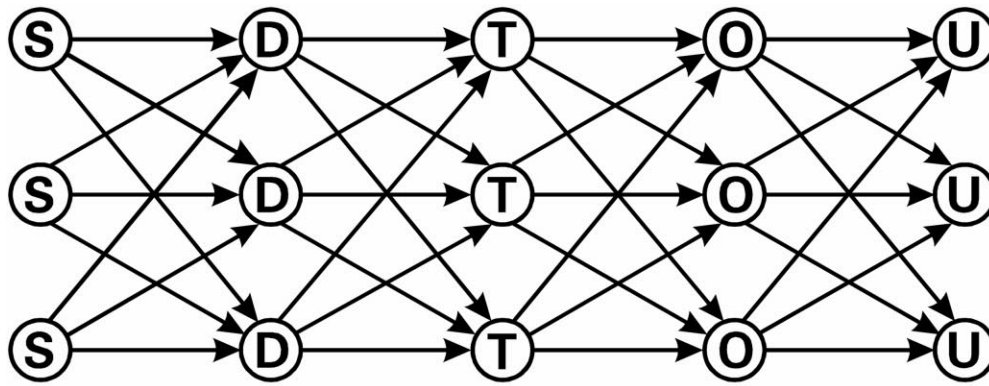
This kind of continuous information might be even more difficult to obtain than the utilities of discrete outcomes. However, since a substantial proportion of the published parameter estimates of the effectiveness of means of intervention are reported in a continuous, rather than a discrete, metric, this adaptation would make for a more powerful and potentially widespread application of the benefits claimed for expected utility theory.

## 6. Summary: the integration of results

In summary, we have reviewed the various approaches used in outcomes and effectiveness research in terms of the specific questions that each distinct approach addresses. We have broken down these specific questions by the different stages of the overall decisionmaking process that each of these distinct approaches model. An augmented Brunswikian Lens Model was used for this conceptual analysis. The three principal stages we identified were: (1) the sources of biomedical information, (2) the means of biomedical intervention and (3) the utilities of biomedical outcomes. Although some true alternatives were reviewed within each of these stages, the different approaches could be seen as mostly complementary and not in direct competition. They each contribute different facets to a complete understanding of the problem. Even within what were identified as bona fide alternatives, we concluded that a constructive synthesis of competing approaches was often both possible and desirable. Also discussed were possible extensions of existing models that were suggested by this synthesis.

The constructive, synthetic approach to this system of complementary relationships is illustrated in Fig. 1. One important implication of this representation is that the model parameter estimates at each of these successive stages, whether represented as Bayesian conditional probabilities or path regression coefficients, must be of comparable minimum quality. Since they will eventually be crossmultiplied, there is no ultimate advantage in estimating one component with extreme precision and another with inadequate reliability. For example, it does not help much in our decisionmaking if both the accuracy of a diagnostic test and the efficacy of a treatment are well characterized, but the utility of the final outcome is highly uncertain. Bringing the quality of parameter estimation at all stages of this process up to comparable standards is an important challenge for the future of outcomes research.

The path analytic approach we have outlined assumes that each measured outcome, for example, is a conceptually distinct entity, or hypothetical construct. If multiple measures are used to converge upon a single conceptual outcome, rather than correspondingly multiple outcomes, then common factor modeling should be applied to identify that latent common factor and, thus, reduce the complexity of the model. For example, if the multiple utilities obtained for different populations are not entirely discrepant, but seem to suggest some broader confluence of interest across stakeholders, then it is entirely appropriate (and perhaps politically astute) to extract a more general common factor possibly underlying all the specific utilities for different stakeholders. Any residual conflicts of interest between stakeholders can also be more readily identified by such a factoring procedure. By substituting latent variables for manifest variables in the path model, we may construct a factor analytic structural equations model (Bentler, 1989) for our Brunswikian analysis.



- (S) = TRUE PATIENT STATUS
- (D) = DIAGNOSIS (MEDICAL INFORMATION)
- (T) = TREATMENT (BIOMEDICAL INTERVENTION)
- (O) = BIOMEDICAL OUTCOME
- (U) = SUBJECTIVE UTILITY
  
- (S) → (T) **VALIDITY OF DIAGNOSIS**  
 = the degree to which a patient's true status produces a given diagnosis represents the validity of that diagnosis in reflecting the patient's condition.
  
- (D) → (T) **SELECTION OF TREATMENT**  
 = the degree to which a given diagnosis leads to a given treatment represents the influence of the diagnosis in the selection of that treatment.
  
- (T) → (O) **EFFECTIVENESS OF TREATMENT**  
 = the degree to which a given treatment produces a given outcome represents the effectiveness of that treatment in producing that outcome.
  
- (O) → (U) **UTILITY OF OUTCOME**  
 = the degree to which a given outcome produces subjective utility in a specific population represents the subjective utility of that outcome for that population

Fig. 1. Augmented Brunswikian path model for decision making analysis.

Another relevant implication of this Brunswikian, path analytic, and synthetic mode of thinking concerns the current controversy over Intention To Treat (ITT) analysis. An individual study may start out as a randomized clinical trial, with all the attendant merits and limitations described above, and end up, owing to both the ethics and pragmatics of implementation, as a Non-equivalent comparison groups design, with all the inherent difficulties also noted. In order to avoid those complications, ITT analysts insist that you ‘analyze what you randomize’, and thus preserve the idealized design characteristics of the original experiment. Opponents claim that ‘pretending’ that all patients actually received the intended treatment, especially where that is

known certainly *not* to be the case, is inherently ridiculous and tends to underestimate the ‘true’ effect of the treatment, where actually administered. Proponents counter that because one can never implement a medical treatment regimen with perfect certainty, it is better to estimate the effectiveness of a *policy* to administer that treatment, when considering the adoption of such a policy. Thus, the potentially fallible implementation of that treatment policy must be counted against it.

This debate is reminiscent of that over treatment efficacy versus treatment effectiveness. One group of researchers would like to know the potential effect of a particular treatment if flawlessly administered; the other group would like

to know what that putative effect might translate into in actual practice, ‘warts and all’. From a Brunswikian perspective, we see that these two questions are complementary and not competitive. In our path diagram, we merely substitute  $[R]$  (= Randomization) for  $[D]$  (= Diagnosis). Because the selection of treatment in a true experiment should be based on random assignment rather than on any information regarding patient status, there should be no causal effect,  $[S] \rightarrow [R]$ , as is shown for the corresponding pathway,  $[S] \rightarrow [D]$ . If patients were actually treated exactly as randomized, then the coefficient for the casual pathway,  $[R] \rightarrow [T]$ , would be a perfect 1.0, thus leaving the causal pathway,  $[T] \rightarrow [O]$  as the only remaining free parameter (given that  $[U]$  (= Utility) is not assessed). This coefficient, according to certain researchers, is the one of primary interest.

Thus, in a true experiment, no indirect connection will be found between  $[S]$  and  $[O]$ . If, however,  $[S]$  (or, indeed, any other factor) somehow directly influences  $[T]$ , as very often happens in a real study, then the fundamental reasoning behind our design is compromised. The coefficient for the causal pathway,  $[R] \rightarrow [T]$ , perforce falls below 1.0, and thus the product of the coefficients for the pathways,  $[R] \rightarrow [T]$  and  $[T] \rightarrow [O]$ , is reduced to a value below that of the direct effect,  $[T] \rightarrow [O]$ . Thus, if you ‘analyze what you randomize’, then this indirect effect is what is implicitly estimated by ITT analysis. If this indirect effect is incorrectly interpreted as the direct effect,  $[T] \rightarrow [O]$ , then it does, indeed, underestimate that effect, as the critics claim. On the other hand, there is no need to interpret that ITT estimate incorrectly. ITT analysts claim that it represents the effect of a given treatment policy (now substituting  $[P]$  for  $[R]$ ) which is not normally, meaning in applied medical practice, one of random assignment to treatments. ITT proponents claim that the ultimate indirect effect,  $[P] \rightarrow [O]$ , and not the hypothetical direct effect,  $[T] \rightarrow [O]$ , is what we should consider for decisions of medical policy. A full Brunswikian path analysis, however, can (and, indeed, *must*) provide both estimates and separately satisfy both interest groups.

### 6.1. Will our data ever be good enough?

That research, even good research, will provide a dependable basis for policy is not inevitable. We are not scientific nihilists, but we do think that research has fundamental limitations that must be taken seriously, especially by scientists who want very much to influence policy. We merely summarize the problems here, but they deserve extended consideration. First, all research findings reflect what Campbell (1986) calls ‘local molar’ conditions, i.e., specifics of time, place, circumstance, specific arrangements, and so on. Put another way, even when we have good findings, they may have limited generalizability, and the limits may not be easy to determine. Second, for a variety of reasons, including local molar conditions, dependable, replicable findings are difficult to produce. Small variations may magnify or may occur at critical points that tip the empirical scales one

way or another. Thus, we get disagreements between studies, laboratories, investigators. Things that seemed so yesterday in one journal are shown to be doubtful today based on some other paper. Third, the world itself is unstable. Good research may nail something down fairly well, but things change. Devine and Cook (1986), for example, found an advantage across a series of studies for psychoeducational interventions with surgical patients, but the advantage seemed to decrease over time, probably because the variance in hospital length of stay was being squeezed out of the system. A recent innovation in gall bladder surgery might make irrelevant a previous series of studies supporting some other method. Finally, research results, unfortunately, can just be wrong. We do not have any good way of estimating how many reported research findings are just plain wrong, but even if the proportion were fairly small, the possibility would still be limiting.

Our conviction is that the greatest limitation on science in relation to medical outcomes is the weakness of theory to guide planning for and integration of research. Good theory conscientiously applied can foster better research and better findings. It can also assist us in ‘smoothing’ across the seemingly random perturbations in observed results and reveal likely limits on broader generalizations. Good theory in the absence of good methodology is a misfortune, but good methodology in a theoretical vacuum is pointless.

## References

- Bentler, P. M. (1989). *EQS: structural equations program manual*, Los Angeles: BMDP Statistical Software, Inc.
- Boruch, R. (1974). *Double pre-testing as a strategy for dealing with regression artifacts in quasi-experimental designs*. Unpublished manuscript, Northwestern University.
- Brown, S. A. (1988). Effects of educational interventions on diabetes care: a meta-analysis of findings. *Nursing Research*, 37, 223–230.
- Brunswik, E. (1952). *The conceptual framework of psychology*. In *International encyclopedia of unified science (Vol. 1)*, Chicago: University of Chicago Press.
- Campbell, D. T. (1986). Relabeling internal and external validity for social scientists. In W. Trochim, *Advances in quasi-experimental design and analysis. new directions in program evaluation*, no. 31 (pp. 67–77). Beverly Hills: Sage Publications.
- Campbell, D. T., D. W., & Fiske, . (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Chalmers, T. C., Levin, H., Sacks, H. S., Reitman, D., Berrier, J., & Nagalingam, R. (1987a). Meta-analysis of clinical trials as a scientific discipline. I: Control of bias and comparison with large co-operative trials. *Statistics in Medicine*, 6, 315–325.
- Chalmers, T. C., Berrier, J., Sacks, H. S., Levin, H., Reitman, D., & Nagalingam, R. (1987b). Meta-analysis of clinical trials as a scientific discipline. II: Replicate variability and comparison of studies that agree and disagree. *Statistics in Medicine*, 6, 733–744.
- Cochrane, A. L. (1972). *Efficiency and effectiveness*, London: Nuffield Provincial Hospitals Trust.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression: correlation analysis for the behavioral sciences*. 2nd ed, New York: Erlbaum.
- Cook, T. D. (1990). The generalization of causal connections: multiple theories in search of clear practice. In L. Sechrest, E. Perrin & J.

- Bunker, *Research methodology: strengthening causal interpretations of nonexperimental data* (pp. 9–31). Washington, DC: DHHS, Agency for Health Care Policy and Research.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: design and analysis issues for field settings*, Boston: Houghton-Mifflin.
- Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides & J. Tooby, *The adapted mind: evolutionary psychology and the generation of culture* (pp. 19–136). New York, NY: Oxford University Press.
- Dales, L. G., Friedman, D. G., & Collen, M. F. (1979). Evaluating periodic multiphasic health checkups: A controlled trial. *Journal of Chronic Diseases*, 32, 384.
- Devine, E. C., & Cook, T. D. (1986). Clinical and cost-saving effects of psychoeducational interventions with surgical patients: a meta-analysis. *Research in Nursing and Health*, 9, 89–105.
- Dickersin, K., Chan, S., Chalmers, T. C., Sacks, H. J., & Smith Jr, H. (1987). Publication bias and clinical trials. *Controlled Clinical Trials*, 8, 343–353.
- Feinstein, A. R., Sossin, D. A., & Wells, C. K. (1985). The Will Rogers Phenomenon: stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. *New England Journal of Medicine*, 312, 1604–1608.
- Feinstein, A. R., Horwitz, R. L., Spitzer, W. O., & Batista, R. N. (1981). Coffee and pancreatic cancer: the problems of etiologic science and epidemiologic case-control research. *Journal of the American Medical Association*, 246, 957–961.
- Ferketich, S., Figueredo, A. J., & Knapp, T. R. (1991). Focus on psychometrics: the multitrait–multimethod approach to construct validity. *Research in Nursing and Health*, 14, 315–320.
- Figueredo, A. J., Ferketich, S., & Knapp, T. R. (1991). More on MTMM: the role of confirmatory factor analysis. *Research in Nursing and Health*, 14, 387–391.
- Greenberg, D. S. (1983). Health care in Israel. *New England Journal of Medicine*, 309, 681–684.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*, New York: Academic Press.
- Kaplan, R. M. (1985). Quantification of health outcomes for policy studies in behavioral epidemiology. In R. M. Kaplan & M. H. Criqui, *Behavioral epidemiology and disease prevention* (pp. 31–54). New York: Plenum.
- Kaplan, R. M. (1990). Behavior as the central outcome in health care. *American Psychologist*, 45, 1211–1220.
- Lipsey, M. W. (1990). Theory as method: small theories of interventions. In L. Sechrest, E. Perrin & J. Bunker, *Health services research methodology: strengthening causal interpretations of nonexperimental data* (pp. 33–52). Washington, DC: Agency for Health Care Policy and Research.
- Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris, *Problems in measuring change* (pp. 21–38). Madison, WI: University of Wisconsin Press.
- Orwin, R. G., & Cordray, D. S. (1985). Effects of deficient reporting on meta-analysis: a conceptual framework and reanalysis. *Psychological Bulletin*, 97, 137–147.
- Peto, R. (1987). An approach to overviews. *Statistics in Medicine*, 6, 233–240.
- Petrinovich, L. (1979). Probabilistic functionalism: a conception of research method. *American Psychologist*, 34, 373–390.
- Reichardt, C. S. (1979). The statistical analysis of data from nonequivalent group designs. In T. D. Cook & D. T. Campbell, *Quasi-experimentation: design and analysis issues for field settings* (pp. 147–205). Chicago: Rand McNally.
- Rein, M., & White, S. H. (1977). Can policy research help policy? *The Public Interest*, 49, 119–136.
- Rosenthal, R. (1984). *Meta-analytic procedures for social research*, Beverly Hills: Sage Publications.
- Sechrest, L., West, S. G., Phillips, M., Redner, R., & Yeaton, W. (1979). Some neglected problems in evaluation research: strength and integrity of treatments. In L. Sechrest, S. G. West, M. Phillips, R. Redner & W. Yeaton, *Evaluation studies review annual*, vol. 4 (pp. 15–35). Beverly Hills, CA: Sage.
- Sechrest, L., Stewart, M., Stickle, T., & Sidani, S. (1996). *Effective and persuasive case studies*, Cambridge, MA: Human Services Research Institute.
- Seltzer, H. S. (1972). A summary of criticisms of the findings and conclusions of the University Group Diabetes Program (UGDP). *Diabetes*, 21, 976–979.
- Shadish Jr, W. R. (1990). Critical multiplism: a research strategy and its attendant tactics. In L. Sechrest, H. Freeman & A. Mulley, *Health services research methodology: a focus on AIDS* (pp. 5–28). Washington, DC: DHHS, Agency for Health Care Policy and Research.
- Slater, P. E., & Ever-Hadani, P. (1984). Mortality in Jerusalem during the 1983 doctors' strike. *Lancet*, 2 (8362), 1306.
- Stevenson, M. R., & Black, K. N. (1988). Paternal absence and sex-role development: a meta-analysis. *Child Development*, 59, 793–814.
- Sutcliffe, J. P. (1980). On the relationship of reliability to statistical power. *Psychological Bulletin*, 88, 509–515.
- Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. H. Barkow, L. Cosmides & J. Tooby, *The adapted mind: evolutionary psychology and the generation of culture* (pp. 163–228). New York, NY: Oxford University Press.
- Tversky, A. (1972). Elimination by aspects: a theory of choice. *Psychological Review*, 79, 281–299.
- Webb, E., Campbell, D. T., Schwartz, R. D., Sechrest, L., & Grove, J. (1981). *Nonreactive measures in the social sciences*. 2nd edition, Boston, MA: Houghton-Mifflin.
- Weinstein, M. C., & Stason, W. B. (1977). Foundations of cost-effectiveness analysis for health and medical practice. *New England Journal of Medicine*, 296, 716.
- Wennberg, J. E., Roos, N., Sola, L., Schori, A., & Jaffe, R. (1987). Use of claims data systems to evaluate health care outcomes: mortality and reoperation following prostatectomy. *Journal of the American Medical Association*, 257, 933–936.