SAMHSA / CSAP WORKPLACE MANAGED CARE
FINANCIAL / COST RESEARCH EVALUATION GUIDE

January, 1999

Ronald J. Ozminkowski, Ph.D.
Senior Economist and Research Manager
The MEDSTAT Group, Inc.
Ann Arbor, Michigan

Ron Z. Goetzel, Ph.D.
Vice President, Consulting and National Practice
The MEDSTAT Group, Inc.
Washington, D.C.

Rodney L. Dunn, M.S.
Statistician and Project Manager
The MEDSTAT Group, Inc.
Ann Arbor, Michigan

This paper was supported under Contract No. 277-93-2027 through the Division of Workplace Programs (DWP), Center for Substance Abuse Prevention (CSAP), Substance Abuse and Mental Health Services Administration (SAMHSA). The opinions expressed in this document are the authors' and do not necessarily represent the opinions of The MEDSTAT Group, Inc., DWP, CSAP, or SAMHSA.

TABLE OF CONTENTS

CHAPTER 1

EXECUTIVE SUMMARY AND RECOMMENDATIONS

Introduction

The MEDSTAT Group, Inc. is pleased to provide this guide in support of a new research initiative sponsored by the Division of Workplace Programs in the Center for Substance Abuse Prevention (CSAP), Substance Abuse and Mental Health Services Administration (SAMHSA). The Workplace Managed Care Initiative involves a set of cooperative agreements between CSAP and nine grantees who are evaluating the impact of employer-sponsored, managed care programs focused on prevention and early intervention programs directed at substance abuse problems. Many of those evaluations will include efforts to estimate the impact of prevention and early intervention programs on medical expenditures, mental health care expenditures, and other financial outcomes. This guide is intended to facilitate the conduct of financial impact studies.

The processes used to estimate the financial and health-related program impacts are broadly referred to as cost-benefit or cost-effectiveness analysis. This guide defines these terms and describes how to conduct cost-benefit analyses (CBA) or cost-effectiveness analyses (CEA) for workplace sponsored substance abuse prevention and early intervention programs. The conceptual and logistical issues involved in CBA or CEA are described in Chapters 2 - 4. Several recommendations for conducting CBAs or CEAs are also provided. Some standard definitions of important outcome measures are provided in the Glossary, to help guide data collection and analysis. Finally, an extensive reference list is provided.

Summary and Recommendations

Cost-benefit analysis (CBA) and cost-effectiveness analysis (CEA) are two techniques which may be used to estimate the economic value of resources used to deliver substance abuse prevention and early intervention programs in managed care settings. CBA and CEA are systematic approaches to valuing the costs and benefits of such interventions and their alternatives over time. Most analysts differentiate between CBA and CEA by noting that CBA focuses largely on the monetary gains and losses of a program, while CEA focuses largely on non-monetary gains or losses.

The systematic valuation of all measurable costs and benefits of a program and its alternatives helps to:

1. Illustrate the relative impact of the program given the resources spent on it;
2. Aid decisions to adopt, expand, reduce, or cancel the program;
3. Facilitate comparisons between the program and its likely alternatives; and
4. Facilitate broader comparisons between the program and others with widely varying goals and objectives (e.g., between medical programs and those dealing with housing, welfare, other social services, or other corporate programs).

CBA and CEA involve comparing the costs of a program and its alternatives to the benefits of those programs or alternatives. Differences in costs between programs may also be compared to differences in benefits or effectiveness between programs, to find the most economically attractive one. The methods for making these comparisons are noted in Chapter 3.

CBA and CEA are most likely to be useful for evaluating the impact of a substance abuse program or for comparing the relative strengths and weaknesses of multiple programs when these guidelines are followed:

1. Conduct the analysis from multiple perspectives, including clients, caregivers, employers, managed care organizations, and society (i.e., others who are directly or indirectly influenced by the program). When time and money are limited, focus first on clients, those who pay for the program, and other important stakeholders. The societal perspective is likely to produce the most generalizable results, but it will also be the most expensive perspective to incorporate.
2. If time, data, or budget constraints do not allow all relevant perspectives to be included, note the implications of these constraints in the project report.
3. Focus the CBA or CEA on inherently valued outcomes. These are the outcomes that matter most to clients and stakeholders. These can usually be identified by a review of program- or discipline-based theory, the relevant economic or health services literature, case histories, or interviews with providers, other caregivers, clients, or treatment experts.

4. Collect data for the analysis at the most detailed level possible, given the budget for the project. This will allow analysts maximum flexibility.
5. Establish contact with a data expert who knows the nuances of the data sets to be used in the CBA or CEA.
6. Fully document each data source and generate data quality reports that show the validity of the data to be used in the analysis.
7. Identify outlier data values and sort them by source. Then confer with the data expert to determine whether outlier data are valid.
8. Document methods used to impute data and create new variables for analysis.
9. Identify potential data problems likely to affect the analyses. Confer with the data expert as necessary to avoid these problems.
10. Anticipate data needs by referring to program theory, other discipline-based theory, the relevant literature, and interviews with stakeholders and experts.
11. Protect the confidentiality of all data and confer frequently with the custodian in charge of maintaining confidentiality to avoid breaches.
12. Obtain data on the opportunity cost of resources used to produce the program and consume its services. Opportunity costs reflect the value of resources in alternative, next-best uses outside the intervention program. Examples of methods used to estimate opportunity costs are noted in Chapter 3 and in Hargreaves, et al. (1998).
13. Measure opportunity costs for staff labor, volunteer services, overhead and fringe benefits, telephone, fax, e-mail, copy services, consultants' costs, computer and other equipment purchase or rental, supplies, software licenses or access fees, capital expenditures (e.g., mortgage or lease for buildings and land), and clients' and caregivers' time for travel, waiting, and service use.
14. If data on both charges for and payments for medical expenditures are available but the budget will allow only one to be chosen, focus on payments for services. These are likely to better reflect the opportunity costs of medical care resources.

15. Adjust opportunity cost measures for changes in prices due to inflation by using an index constructed from the Gross Domestic Product Implicit Price Deflator or from changes in per capita personal health expenditures in the U.S. economy over time. GDP Implicit Price Deflator values can be found on the web site maintained by the Bureau of Economic Analysis (http:\\www.bea.doc.gov/bea/dn/dpga.pdf).
16. Discount the cost, benefit, and effectiveness measures to account for the differences in values incurred in the base year versus those costs, benefits, or effectiveness values incurred in later years. Repeat analyses using multiple discount rates that are chosen from the multiple perspectives to be accommodated in the CBA or CEA (Krahn and Gafni, 1993).
17. Try to measure both the positive and negative consequences of participating in the intervention. If all consequences cannot be measured, focus first on those of most importance to clients and other major stakeholders, and those expected from theory.
18. Extremely long-term costs and benefits may be excluded from analysis unless data are available that allow reliable estimates of these costs and benefits to be generated. An example of an extremely long-term consequence of program participation is the cost of treating non-substance abuse-related diseases that result from living many years longer than would be the case if participation in the substance abuse program had not taken place.
19. Choose a sound research design to base estimates of program impact (see Appendix 2).
20. When randomized designs are not feasible or do not work, use a pre-post, quasi-experimental design with comparison groups, along with multivariate statistical analyses, to adjust for threats to the validity of the CBA or CEA (Ozminkowski and Branch, 1997).
21. Consider using two-part regression models to estimate program impacts in non-randomized settings in which many participants incur zero values of major outcomes of interest.

22. Collect as much data as possible on reasons why clients choose to participate or not in the intervention. These data can be used in subsequent analyses to adjust for selection bias (see glossary) when randomized designs are not feasible or have broken down (Heckman and Smith, 1995).
23. If two-part regression models are estimated, choose the propensity scoring approach as a defense against selection bias if randomization is not feasible or has broken down. If a one-part model is estimated, other techniques can be chosen. These techniques are described in Chapter 4 and referenced in the glossary.
24. Collect data on as many subjects as possible, to increase the power of the analysis. Randomization test techniques may be used if sample sizes are too low for conventional statistical tests.
25. In a CBA, choose the net present value (NPV) as the measure of program impact. The NPV is defined as the difference in the inflation-adjusted, discounted benefits and costs of program participation. If a benefit / cost ratio or return on investment ratio is requested, supplement it with the NPV to provide a more complete picture of the impact of the program (Warner and Luce, 1982).
26. In a CEA, show the total costs and total values of the effectiveness measures for each program alternative, as well as the cost-per-unit-of-effectiveness measures. This will allow the incremental value of each program to be illustrated.
27. Use sensitivity analyses to deal with any uncertainties about ways to measure costs or benefits, appropriate discount rates, varying perspectives, and other factors of importance to the analyst and stakeholders (Warner and Luce, 1982; Weinstein, et al., 1996).
28. Also use sensitivity analyses to show program impacts for important subgroups (e.g., for males vs. females, for those with less severe vs. more severe substance abuse problems, etc.).
29. If an intent-to-treat design (see glossary) is used, structure the analyses to deal explicitly with those who crossover from the treatment groups to the comparison groups of interest. (See Gibaldi and Sullivan (1997) and Sclar, et al., (1998) for a discussion and examples of intent to treat designs.)
30. Present a complete picture of the value of the substance abuse program. Note the

context in which it is offered, its goals and objectives, and the multiple perspectives used for the CEA or CBA. Also note details about the research design used in the analysis, details about how data were collected and analyzed, the results from the sensitivity analyses, and the impact of any time, data, or budget constraints that limited the scope of the analysis.

31. Comment on the consequences of acting on the results of the CBA or CEA. Explain what may happen to all affected parties if resources are taken from one program to enhance another. Also discuss any ethical issues that would arise from acting on the results of the CBA or CEA.
32. Discuss the threats to the validity of the CBA or CEA (see Chapter 4, the glossary, and the reference by Cook and Campbell, 1979, for good discussions of validity threats).
33. Present results with pictures as well as words and focus on concepts and outcomes that are most important to clients and major stakeholders.
34. Offer stakeholders the chance to review the agenda for the CBA or CEA and its results. Revise the research agenda or refine the analyses as necessary to address their perspectives to the extent that time, data, and budget allow.

It is worth noting that CBA and CEA are not intended to be the only tools for making programmatic decisions. There may be other political, ethical, or business reasons to adopt, expand, reduce, or cancel programs, or to favor one program over another. All of these reasons should be considered when making decisions about the fate of one program versus another.

Finally, CBA and CEA often show how many added benefits or costs accrue to the group of persons who are the subject of the intervention of interest. These techniques are not designed to show impacts for any particular individual, and the appropriate treatment of any given individual may or may not be the same as the treatment which is deemed most cost-beneficial or cost-effective for the employer, the managed care organization, or society.

CHAPTER 2
DEFINING CBA AND CEA

Cost-benefit and cost-effectiveness analyses are systematic approaches to valuing, over time, the costs and benefits of an intervention and its alternatives. These techniques are designed to estimate the total value of all of the resources used to produce and consume the intervention, and to weigh these resource estimates against the consequences of the investment in the intervention. When conducted properly, the results of a CBA illustrate (to the extent possible with the input data) how many dollars would be gained or lost from the intervention, compared to its alternatives. The results of a CEA show how many units of health or other measures would be gained or lost, and how many dollars would be spent on the intervention and its alternatives.

CEA and CBA are often conducted from a variety of stakeholder perspectives and with many differing assumptions about which costs and benefits to count and how to estimate them. The results are then used to show multiple audiences how much "bang for their buck" is associated with the intervention of interest. The results may also illustrate how that "bang" varies by audience or when legitimate differences of opinion are addressed about which benefits and costs are important or about how best to estimate them.

In CEA and CBA, the value of the resources used to produce and consume the intervention of interest and its alternatives are referred to as "costs, and the consequences of participating in the intervention or its alternatives are referred to as "benefits." Most researchers differentiate between CBA and CEA according to how the benefits of the intervention and its alternatives are estimated. In CBA the benefits of the intervention and its alternatives are denoted in discounted, inflation-adjusted dollar terms. This leads to a relatively straightforward estimate of the value of the intervention and its alternatives. Discounted, inflation-adjusted costs can be subtracted from discounted, inflation-adjusted benefits to yield a net present value figure that shows how many of today's dollars may be gained or lost by each program. From a limited cost-benefit perspective in which only one program may be chosen, the program with the highest net present value "wins" because it is expected to yield the best economic return on its investment.

In CEA the benefits are usually denoted in non-dollar terms. In this scenario the obvious advantage of CEA is that benefits which are difficult or impossible to monetize can be considered in the cost-effectiveness calculation (e.g., in terms of a percentage drop in substance abuse prevalence). Here, the program with the lowest incremental cost-effectiveness ratio (e.g., with the lowest estimate of dollars per percentage drop in prevalence, relative to other alternatives) may be the economically preferable one.

A disadvantage of CEA is that it can be difficult to aggregate the variety of cost-effectiveness ratios that result when many types of non-monetary benefits are associated with the programs to be evaluated. Many researchers try to avoid this problem by focusing the benefit calculation on changes in health status and the preferences that people assign to various health states. This information can be used in a form of CEA known as cost-utility analysis, to estimate benefits in terms of quality-adjusted life years gained from each program of interest. Under this scenario, the program with the lowest cost per quality adjusted life year (QALY) gained is the economically preferable one (Russell, et al., 1996). Methods for estimating QALYs are noted in Drummond, et al. (1997).[1]

While the cost-utility analysis approach can indeed be helpful by accounting for the preferences for the variety of health states that may result from an intervention, it may not really solve the dilemma of how to deal with multiple benefit types. Many types of benefits may be important in addition to better health status. These include improvements in productivity, fewer disciplinary actions and associated arbitration, reduced employee turnover, fewer marital or family problems, less litigation, fewer encounters with the criminal justice system, and lower levels of stress at home and at work. These outcomes may be correlated with health status changes, but they are not totally interchangeable with health status, and many quality of life and work environment measures cannot easily be recast as measures of quality-adjusted life years (Revicki, 1993). Thus, these measures should not be ignored in the cost-effectiveness calculation. Because the cost-utility analysis approach may not fully account for all of the relevant types of non-monetary benefits, we do not advocate cost-utility analysis as the sole analytic approach. Rather,

[1] Cost-utility analysis is less frequently used for studies of mental health programs, partly because it is difficult to measure quality of life reliably and translate that measure into preference states. Hargreaves, et al. (1998) do not recommend cost-utility analyses for mental health programs, for this reason.

CUA should be supplemented with CEA, to include all of the relevant benefit measures in the analysis. Later we note methods for dealing with multiple types of effectiveness measures.

An Alternative View

Like others, we differentiate between CBA and CEA in terms of how benefits are measured. While this appears to be the majority approach, it is not the only way to differentiate between these two types of analysis. For example, Donaldson (1998, page 391) argues that:

"[Cost benefit analysis] seeks to answer the question, 'is it worth achieving this [i.e., the program's] goal?' or 'how much more or how much less of society's resources should be allocated to achieving this goal...? It involves interpersonal comparisons of preferences because allocation of more resources to one group implies a redistribution whereby that group gains but another group loses those resources." In contrast, Donaldson argues that cost-effectiveness analysis seeks to answer a different, but related question: "Given that it has been decided that a goal is to be achieved, what is the least-cost way of doing so?"

Donaldson's argument appears to be a minority view, and its merits are not central to this paper. However, one of his recommendations is very important. Specifically, once the cost-benefit or cost-effectiveness calculation has been made, it is wise to consider the potential impact of redistributions in resources that are implied by the CBA or CEA and suggested by any recommendations coming from those analyses. In other words, if the CBA or CEA suggests that one program is preferred over another, policy makers may decide to expand the more attractive program and reduce or cancel less attractive alternatives. Before this decision is made, it is wise to consider the business, political, social, and ethical consequences of moving resources from one program to another, in addition to the economic consequences of doing so.

CHAPTER 3

PURPOSE AND MAJOR COMPONENTS OF CBA-CEA

The systematic valuation of the measurable costs and consequences of a program and its alternatives can be used for several purposes:

1. To illustrate the impact of program alternatives given the resources to be spent on them;
2. To aid decisions to adopt, expand, reduce, or cancel a program;
3. To facilitate comparisons across programs with similar goals (e.g., to facilitate a cross-site evaluation of substance abuse prevention and early intervention programs); and
4. To facilitate broader comparisons (e.g., between prevention and treatment programs, or between programs with widely different goals).[2]

Programs with acceptable or better net present values or those with better dollar-per-level-of-effectiveness values may be recommended or expanded, while others may be reduced in scope or canceled. Alternatively, CBA and CEA can be used to evaluate slightly different modifications on the same theme, to help decide whether the way Program A was designed and implemented is superior in some respects to the design or implementation of Program B. Recommendations to fine-tune the programs may then be offered. Finally, by focusing on net dollars gained or lost, or on the non-monetary benefits gained or lost, CBA and CEA can help make comparisons across widely disparate program types. Such programs might include those devoted to prevention versus others devoted to treatment well after a disease has been established, or those devoted to Disease A versus others devoted to Disease B.

Regardless of the reason for conducting the CBA or CEA, it is worth keeping in mind that these techniques are only tools, and CBA and CEA were never intended to be the only decision-making tools worth using. There may be other political, ethical, social, or business reasons to

[2] For example, CEA could be used to help compare the value of programs directed at reducing caregiver burden, those directed at improving parenting skills, and those directed at improving work performance.

make programmatic decisions. In addition, the value of CBA or CEA as decision-making tools depends heavily on how well each of the following components are addressed in the analysis:

1. Deciding what questions to address and which hypotheses to test;
2. Estimating and discounting costs for each programmatic alternative;
3. Estimating and discounting monetary benefits;
4. Estimating and discounting any non-monetary benefits;
5. Combining results from components 2-4 into useful metrics to facilitate comparisons across programs;
6. Performing sensitivity analyses to deal with uncertainties and test assumptions;
7. Presenting results to aid effective decision-making; and
8. Recognizing the limitations of the analysis.

Each of these components is addressed below. Additional information about some of these components vis a vis substance abuse prevention may also be found in Stewart, et al., 1998.

Component 1: Deciding What Hypotheses / Questions to Address

Each of the nine grantees under the CSAP Workplace Managed Care Initiative has already written a detailed proposal describing interventions designed to address substance abuse prevention and early intervention in a managed care setting. Those proposals described research questions to be addressed in a program evaluation context. Each grantee has already described the nature of an intervention of interest and the value to be gained by evaluating the impact of that evaluation. This is evidenced by the fact that Federal funding has been granted in each instance, after a detailed, critical review by outside experts. Thus, in this paper we do not intend to address the details of the specific research questions to be studied by these grantees. Rather, the focus of this section is on deciding how to choose which hypotheses to test and which questions to address, given a context in which interventions are already underway.

Useful cost-benefit or cost-effectiveness analyses address this issue by noting one or more perspectives that legitimately guide the analyses. The implications of those perspectives are also described. In the context of the CSAP Workplace Managed Care Initiative, several relevant perspectives should be considered. These include the perspectives of 1) clients, 2) families, 3) caregivers, 4) insurance providers or financiers of care and 5) society at large.

Weinstein (1995) notes that, from a theoretical perspective, “the ‘cost’ that appears in the numerator of the cost-effectiveness ratio of a program or intervention should be the net burden of the program on the constrained budget in question” (page 81). Thus, the key to identifying the appropriate perspective(s) for the cost-benefit or cost-effectiveness analysis is to first identify which party’s costs are constrained and affected by those prevention, early identification, early intervention, and treatment activities. At a minimum, those who pay the bill for the intervention should have their perspective addressed. This usually includes clients, managed care organizations, other third party payers, and employers who offer or pay for the health promotion, primary care, or mental and behavioral health care services that are being evaluated.

For most health care interventions, however, those who pay for it are not the only ones affected by it. Anyone who is directly or indirectly affected by the intervention should have their perspective incorporated into the analysis. For substance abuse prevention and early intervention programs, the perspectives of family members, caregivers, or others (e.g., coworkers and managers) should also be addressed in this societal view.

It may not be easy to identify everyone influenced by the success or failure of a substance abuse prevention and early intervention program. Moreover, even after these people are identified, it may not always be possible to incorporate their experiences into a CBA or CEA. Frequently, evaluation budgets and timelines are limited, requiring a narrower perspective. In addition, it may not be feasible to collect all of the data required to thoroughly examine each perspective. For example, it may be difficult or impossible to reach all affected health care providers, caregivers, or coworkers with surveys designed to elicit their experiences. If budget limitations or data sources do not allow all of the relevant perspectives to be included in the analysis, this should be clearly stated in the CBA or CEA report. The implications of the limited scope of the analysis should be described as well.

Once the perspective of the analysis has been decided upon, the research agenda for the CBA or CEA must be addressed. The research agenda-setting task for the CBA or CEA is probably its most important feature. Theory and program design should drive subsequent analyses, but limited resources will influence the research agenda too. Even in a situation where many perspectives can be addressed, it may be impossible to address all of the relevant outcomes the intervention is designed to influence. For example, health care claims data may be abundant for intervention subjects who are in non-capitated health care plans, thereby enabling estimates of the impact of the intervention on health care expenditures to be generated for them. On the contrary, claims data rarely exist or often do not include detailed information on expenditures for each health care service received by intervention subjects in capitated plans. Thus, it may not always be possible to estimate financial impacts for those subjects. (A few methods for addressing this problem are described in more detail later in this paper.) The limits of the research agenda should be noted in the CBA or CEA report, along with the implications of this limited scope.

When budget or data limitations make it impossible to address all of the research questions of interest, which ones should be chosen for analysis? Mohr (1992) suggests focusing on "inherently valued" outcomes. He defines inherently valued outcomes as those with either of two features: 1) If outcome X is attained and subsequent outcomes do not matter, then X is inherently valued. 2) Alternatively, if outcome X is attained and one can safely assume that unmeasurable (but desirable) outcome Y will also be attained at a satisfactory level, then X is inherently valued.

The identification of inherently valued outcomes will also help identify hypotheses to test in the CBA or CEA. The obvious example of such a hypothesis is that the inherently valued outcome is more likely to be achieved by intervention participants than non-participants. Inherently valued outcomes and associated hypotheses can be identified from several sources. These include program theory, economic or other discipline-based theory, the economic or health services literature on the subject, or meta-analyses of previous studies. Case histories or case studies may also be of value. Focus groups can also help identify inherently valued outcomes, as can Delphi techniques or interviews with program staff, clients or families, care providers, funders, other stakeholders, or experts in the substantive area. Consulting many of these sources will usually result in a useful research agenda for the CBA or CEA (Black, 1993; Hedrick, Bickman, and Rog, 1993; Luft, 1989; Ozminkowski and Branch, 1997; Rossi and Freeman, 1993).

Component 2: Estimating and Discounting Costs

The monetary value of all resources used to produce, deliver, and receive an intervention is referred to as its "cost." Economic theory suggests that the cost of interest to policy makers and other users of CEA or CBA is the "opportunity cost" of program resources. The opportunity cost is defined as the value of each resource in its next best use. For example, with regard to program staff the question is how much those staff would earn if used for another program. Similarly, how much would other program inputs (e.g., supplies, equipment, etc.) earn if used in their next best alternative program(s)? The answers to these questions are the elements that should be included on the cost side of the CEA or CBA.

In a perfectly competitive labor market, the wage or salary rate of program staff persons would equal their opportunity cost. Similarly, the prices paid for other supplies, equipment, etc. would equal their opportunity costs. However, Weinstein, et al. (1996) and others (e.g., Hargreaves, et al., 1998) note that the market for medical care is not perfectly competitive. This is because of the asymmetry between the knowledge levels of providers and clients or caregivers, and because the existence of insurance benefits dilutes the incentives that clients would otherwise have to “vote with their feet” if they did not think they were getting their money’s worth from their providers. Moreover, the deviation from perfect competition has lead to a number of regulatory pricing systems for medical care services (e.g., based on Diagnosis Related Groups for inpatient care or relative value scales for outpatient care). These systems may not reflect the opportunity cost of using those resources for any given program.

Because markets for medical care services are not perfectly competitive, it may be difficult or impossible to find accurate measures of opportunity costs. As a result, program accounting data may be the only available source of information about the cost of resources used for the intervention. If time or budget constraints do not allow other estimates of opportunity costs to be generated, this should be noted in the report. In addition, sensitivity analyses should be conducted to learn how the results of the CBA or CEA would vary if assumptions about the magnitude of these costs change.

Hargreaves, et al., (1998) offers some suggestions for collecting data on cost elements and making assumptions about the value of various cost items in mental health studies. They also describe problems accounting for the value of clients’ time and the time of unpaid caregivers. It is typically very difficult or expensive to obtain survey or other data on wage or salary rates for clients to proxy the opportunity costs of their time. The same may be said for caregivers. Outside sources such as prevailing wage rates for home health or domestic workers or census data on mean or median wages at the zip code level are sometimes used to generate rough proxies for these costs. When detailed, person-level data are not available for each patient or caregiver, sensitivity analyses should be conducted to learn how the results of the CEA or CBA would change as estimates of the opportunity costs of their time vary.

Perspective and Cost Elements. The perspective chosen for the CEA or CBA will influence the cost elements to be counted in the analysis. For example, if the focus is strictly on the rate of return associated with the program, monetary costs and benefits associated with service provision may be collected. In contrast, if the focus is on changes in health status and family relationships, time and money costs for clients, family members, or other caregivers will be counted. If the focus is on the societal value of the intervention program, any costs imposed on others that are directly or indirectly affected by the program must be counted as well.

Often, the cost of producing or receiving a substance abuse prevention or early intervention program is based on the following items (Stewart, et al., 1998):

- Professional staff time;
- Consultants' time;
- Other persons' time (e.g., volunteers, non-professional staff);
- Overhead and fringe benefits;
- Other direct costs (e.g., telephone, fax, e-mail, travel, insurance);
- Computer and other equipment purchase or rental;
- The market value of donated equipment;
- Supplies;
- Software licenses;
- Capital (land and mortgage costs, and facility rental/lease costs); and
- Client and caregiver time for travel, waiting, and service use.

The particular cost elements that are important may vary according to the nature of the intervention. For example, telephone bills may be much higher for services that include telephone counseling, and rental costs may be much lower for more time limited services such as health fairs. A sample cost-allocation matrix for different types of prevention efforts is provided by Stewart, et al. (1998). In Appendix 1, we refine this cost-allocation matrix to account for multiple program years and describe how to estimate discounted program costs for each program.

Information on the cost elements must be collected for the cost side of the analysis, for every year in which the intervention program operates. Suggestions for estimating these costs are provided in Hargreaves, et al. (1998) and in data collection manuals produced by the CSAP Workplace Managed Care Coordinating Center (Bray, 1998; Bray and Zarkin, 1998a, 1998b). Note that some costs (particularly wages and property rental rates) tend to increase over time. Often, these changes occur as contracts expire or with the new calendar year. Wage rates may also change with personnel changes. Analysts should track these changes and account for them when estimating program costs.

Revenues received from clients or other intervention subjects represent a cost to them that should be counted if their perspective is included in the analysis. However, these revenues may also be counted on the benefits side of the cost-benefit equation, thus canceling each other out in the analysis.

Some argue that simple transfers of wealth from one group to another should be excluded from cost-benefit or cost-effectiveness analyses. This is suitable when the analysis is conducted from the societal perspective, because the movement of wealth from one group to another represents a cost to one group and a benefit to another that exactly cancels out in the analysis (Weinstein, et al., 1996). However, when a more limited perspective is adopted, transfers of wealth may not be totally offset in the analysis and should therefore be counted. An example is when prices are charged for program services but the analytical perspective is limited to the employer and does not include that of the patient or third party payer. In this case, the revenues obtained would be counted as benefits, but the charges made would not be counted as program-related costs.[3]

[3] Whether this scenario seems fair or equitable is an issue to be decided upon when the perspective of the analysis is chosen. As mentioned earlier in the discussion of Component 1, the perspective chosen should be dictated by the goals of the intervention or underlying program theory, and by a consideration of 1) who pays for the intervention, 2) who benefits from it, and the budget for the CBA.

Adjusting for Inflation. Many interventions last for several years. In such instances, program-related costs will be incurred in each year. In this scenario, cost estimates should be adjusted to account for the impact of inflation, which erodes purchasing power over time (Stewart, et al., 1998). Adjusting for inflation accounts for exogenous increases in wages and other workplace-related expenditures over time that has nothing to do with the nature of the intervention. For example, increases in salary may be made as a result of collective bargaining agreements, seniority, or experience gained, and prices for supplies or equipment needed to produce the intervention may increase over time as well. Adjusting for this inflation is done by recording cost figures in constant dollars (i.e., in the dollars of a base year), by applying an index based on changes in prices over time. The index value equals 1.0 for the base year. Values for earlier years are usually lower, and values for later years are usually higher. Dollar cost estimates can then be divided by the index value for the year in which the costs were incurred, to put those estimates in terms of base-year dollars.

Analysts sometimes use the Consumer Price Index (CPI) or its Medical Care component to create the inflation index used in the base-year dollar calculation (Weinstein, et al., 1996). However, Getzen (1992) argues that neither the CPI nor its medical care component (the MCPI) is suitable for this purpose, for theoretical reasons. Neither the CPI nor the MCPI was developed for inflation adjustment, and neither adjusts well for rapid technological changes or other changes in productivity that influence purchasing power. As an alternative, Getzen argues that the Bureau of Labor Statistics' Gross Domestic Product Implicit Price Deflator is more appropriate for inflation adjustment. Newhouse (1989) makes the same argument as Getzen but recommends that the adjuster for inflation be based on the annual rate of change in per capita personal health care expenditures.[4] A prudent approach would be to avoid using the CPI or the MCPI and to conduct analyses twice, once using the GDP Implicit Price Deflator and once using Newhouse's approach, to learn how sensitive the results may be to the inflation adjustment process.

---

[4] Newhouse envisioned his adjuster to be based on national health expenditures. Grantees may also consider constructing an adjuster on changes over time in per capita health expenditures paid by the workplace that is applying the substance abuse prevention and early intervention program. The latter may be useful to adjust for local conditions, if the sample size is large enough to permit stable estimates to be generated.

Discounting. When interventions span many years, costs incurred after the first year should be discounted over and above the adjustment for inflation. Discounting means dividing the cost figures by $(1.0 + r)^t$, where r equals the discount rate and t is an exponent equal to the number of the year when the cost figure was estimated. By convention the base year is defined as year $t = 0$, the next year is defined as year $t = 1$, and so on. The base year is usually denoted as the first year in which program costs were incurred.

Discounting later-year dollars adjusts for the fact that rational consumers place a higher value on dollar costs incurred now versus the same number of dollar costs incurred later. One reason for this difference in value is that $1 spent now could otherwise have been invested to yield (for example), $1.05 next year, after adjusting for inflation. Thus, it takes more than $1 later to equal one of today's dollars (even after accounting for inflation), so today's dollars are worth more than next year's dollars.

While the need for discounting is universally accepted by economists when CBAs or CEAs are conducted, the rate at which discounting should occur has been argued for many years (Krahn and Gafni, 1993). Weinstein, et al. (1996) suggest that the discount rate "should be based on time preference, the difference in value that people assign to events occurring in the present versus the future" (page 1257). Further, they note that empirical evidence points to a difference of about 3% annually, so they recommend using a 3% discount rate. They also note that many CEAs use a 5% discount rate, so sensitivity analyses should be conducted to determine how results would change when moving from a 3% to a 5% discount rate.

Others argue that perspective matters when choosing a discount rate (Hargreaves, et al., 1998). For example, employers who pay for the substance abuse prevention and early intervention programs may wish to see a return on their investment that corresponds to the return they could have obtained by using the program dollars for other purposes. This corporate opportunity cost approach suggests that the discount rate reflect the cost of capital for the company. Often this cost of capital estimate is higher than 5%, perhaps as high as 10%, depending on the industry and the financial strength of the company. We agree that perspective is important, and that analyses should be repeated several times using different discount rates.

Component 3: Estimating and Discounting Monetary Benefits

The monetary benefits of an intervention are defined as the dollar value of the consequences of participating in it. These consequences include any revenues obtained from the program, and the dollar value of changes in health status and other outcomes associated with program participation. Examples of monetary benefits include program-related savings resulting in:

- Lower medical expenditures;
- Lower mental/behavioral health care expenditures;
- Fewer wage replacement dollars spent for days absent from work or days spent in short-term or long-term disability programs;
- Reduced property damage or fewer product defects produced by workers with substance abuse problems;
- Fewer dollars spent for workplace injury treatment;[5]
- Lower hiring and training costs associated with job terminations and subsequent personnel replacement;
- Reduced theft; and
- Fewer arbitrations associated with disciplinary actions.

The objectives of any particular substance abuse prevention or early intervention program may suggest other potential monetary benefits as well. Finally, as noted earlier, the identification of monetary benefits may also be aided by a review of program theory, other discipline-based theory, the relevant literature, interviews with participants, caregivers, or experts, focus groups, etc. As with program costs, the final list of monetary benefits to estimate will depend on the perspective taken for the CEA or CBA, the timeline for the analysis, and the budget.

Estimating program benefits can be even more difficult than estimating program costs. Given appropriate data, it may be fairly straightforward to calculate the cost of resources used to produce a program. It may be much more difficult to differentiate between benefits that are due to program participation, versus benefits that are due to other factors. The inference that benefits

[5] Miller (1997) and Lestina, Miller, and Smith (1998) describe methods for estimating the cost of workplace injuries that can be used to help estimate the benefits of a substance abuse prevention or early intervention program.

obtained are due to program participation usually requires a solid research design, a large number of participants, and sophisticated statistical testing procedures. This issue will be addressed later, because it applies to other components of the CBA or CEA as well.

<u>Negative Benefits</u>. If we use the term "benefit" to pertain to all of the consequences of participating in an intervention, some of those consequences may be negative in nature. For example, some participants may have adverse drug interactions that result from drugs taken as a part of the intervention program. Others may use more corporate services such as Employee Assistance Programs than if the intervention had not taken place. Another example of a very long-run nature is that the added health status accrued as a result of the intervention may allow participants to live longer and eventually incur costs for problems they would not otherwise have if they had not lived so long!

Weinstein, et al., (1996) argue that most negative benefits should be counted in the CBA or CEA. In a CBA this is usually done by treating negative benefits the same way that positive benefits are treated (i.e., simply add the positive and negative benefits together to obtain an estimate of overall benefits). Sometimes negative benefits are expressed on the cost-side of the CBA. As noted in the discussion of Component 5, this latter approach makes no difference if the results of the CBA are expressed as a net present value figure, but it can lead to the incorrect result if other metrics are used, such as a benefit/cost ratio. In a CEA the benefits are not expressed in monetary units, so negative benefits would be expressed on the effectiveness side of the equation in terms of the changes in utilization that lead to the negative benefits.

Weinstein, et al. (1996) cite one kind of negative benefits that may be excluded from the CBA or CEA. Their exception includes costs for diseases unrelated to the intervention that occur as a result of added years of life that are due to the intervention. They note theoretical and empirical questions which still must be answered about this issue and suggest analysts proceed as they think best. They also suggest performing analyses twice, once counting such negative benefits and once excluding them, whenever the investigator thinks that such benefits will have a large bearing on the results of the analysis.

The Weinstein et al. recommendation notwithstanding, it may be quite difficult to deal with negative benefits (or any benefits) that accrue many years in the future. In the context of the CSAP Workplace Managed Care Initiative, accounting for such benefits essentially requires moving from an evaluation framework to a forecasting framework. In an evaluation framework, talented analysts who use a solid theoretical base with a good research design and adequate data can produce good estimates of the impact of the program during a relatively short study period. Usually this involves an exercise of estimating what happened in the past, not of what is likely to happen in the future. Once the focus of the analysis is expanded to include costs or benefits that may accrue many years in the future, the random error component to the analysis becomes much more problematic. As a result, forecasts that are based on solid statistical processes can still amount to little more than speculation about program impacts. Thus, unless data about future costs and benefits are atypically available and solid, we do not recommend including extremely long-term cost and benefit figures in the CEAs or CBAs to be completed under that initiative.

Adjusting for Inflation and Discounting. Finally, as with the estimation of program costs, any benefits that are tallied should be adjusted for inflation and discounted appropriately. The same discount rates for benefits and costs should be used (Krahn and Gafni, 1993).

Component 4: Estimating and Discounting Non-Monetary Benefits

As mentioned earlier, many of the benefits of substance abuse prevention and early intervention programs are non-monetary in nature. These may include:

- Improved attitudes about substance use and risky behaviors;
- Reduced incidence or prevalence of substance abuse problems;
- Reduced likelihood of relapse into substance abusing mode;
- Lower stress;
- Lower rates of depression;
- Better quality home life;
- Better quality work environment;
- Higher job satisfaction levels;
- Better work performance and higher productivity;

- Better general health status; and
- Fewer encounters with the criminal justice system.

A variety of sources can be used to measure these benefits. Substance abuse incidence, prevalence, and relapse can be measured with survey tools, such as the National Institute on Drug Abuse National Household Survey. Incidence, prevalence, and relapse also can be measured with other customized surveys, by periodic drug testing, or by a combination of survey and clinical methods (Cook, et al., 1995). Information on general health status, stress or depression levels, quality of home life, and work environment may require survey methods as well. These might take the form of Health Risk Appraisal surveys or other survey tools such as the SF-36 (Ware and Sherbourne, 1992), the Work Locus of Control Scale (Spector, 1988), various depression scales, or others. Depression can also be measured using diagnosis codes and codes for depression medications that may be available on health care claims (Croghan, et al., 1998).

Information about some aspects of productivity and job performance may be easy to find, while others may be more difficult to obtain. For example, Miller (1998, personal communication) notes several factors related to productivity. Some of these are fairly straightforward and records may be available for them. Examples include absenteeism (sick leave), overtime pay, and liability claims. Others are more difficult to define and collect information for. These include restricted activity days, supervisor time spent re-scheduling staff and performing other activities to accommodate for substance abusers who miss time from work, the cost of distractions to co-workers, the loss of rare or unique skills that substance abusers would otherwise bring to their jobs, and poor morale. Because these factors are difficult to operationalize, good data may not exist for them and these events may not be tracked very well.[6]

With so many different types of non-monetary benefits, researchers must decide which are most important and worthy of study. Data collection costs may be quite high if survey analyses or other primary data collection techniques are used to obtain information not already included in claims files or other existing data sets. When time and budget constraints are binding, a useful

[6] Additional examples of job performance measures that are difficult to find have also been noted by WMC grantees. For example, job performance data may not be available for university faculty or senior staff, or may not be available for unionized employees. Similarly, it may be difficult to find information about disciplinary actions and associated arbitrations, and about encounters with the criminal justice system.

strategy would be to identify inherently valuable outcomes as suggested by Mohr (1992). One would then focus on those that are most important from a theoretical perspective and from the perspectives of major stakeholders.

Estimating Program Impact. Estimating the impact of the substance abuse program on non-monetary outcomes can be difficult because one must find a way to distinguish between outcomes that are due to the program and those that are due to other measurable and unmeasurable factors. A solid research design and rigorous statistical testing are often required to estimate program outcomes.

There are many research designs that have been used in evaluation studies such as those required in a good CEA or CBA. Textbooks by Cook and Campbell (1979), Hedrick, Bickman, and Rog (1993), Mohr (1992), and Rossi and Freeman (1993) describe designs that may be useful for randomized and non-randomized studies. They also describe many of the validity threats that may complicate inferences about program impact. The following lists of program designs are often encountered in the workplace literature (Heaney and Goetzel, 1997). These are sorted from most useful (but least prevalent) to least useful (but probably most prevalent):

1. *Fully randomized design, with study samples randomly selected from a larger population and subsequent randomization into treatment and control groups, along with a pre-intervention vs. post-intervention comparison procedure*. With this design, researchers usually estimate changes in average levels of outcome measures before versus after the intervention, for intervention participants (the treatment group) and non-participants (the control group). Program impact is then estimated as the change for participants minus the change for non-participants. This design is strong because, if successfully implemented with large samples of participants and non-participants, many threats to the validity of the analysis can be avoided.

2. *Randomized design without pre-post comparisons.* With this type of design, program impact is usually estimated as the post-intervention difference in mean values of outcome measures for participants versus non-participants. While this may avoid selection bias and other validity threats, the lack of a comparison of outcomes before versus after the intervention does not allow the researcher to distinguish between impacts due to the

intervention and those due to other, pre-existing trends.

3. *Quasi-experimental designs*. These designs are similar to those above, but they are executed without randomization. Rossi and Freeman (1993) note several instances in which randomization is not feasible.[7] The lack of randomization may lead to selection bias, which makes it difficult to distinguish between the impact of the program and the impact of other factors correlated with program participation. Later in this paper we describe some methods to avoid this problem in quasi-experimental research.

4. *Pre-post measures of outcome change without a comparison group*. With this design, program impact is typically estimated as the difference in average levels of the outcome measure over time, among the participant group. This design does not allow a strong inference about program impact to be made because one cannot distinguish between factors associated with participation and those not associated with participation that may lead to similar outcomes.

Later in the paper we describe some of the statistical analyses that may be required with these designs in order to estimate the impact of the intervention on non-monetary benefits. (These designs and associated statistical analyses can also be used to estimate the impact of the intervention on monetary benefits.) We will also describe major threats to the validity of the estimation process that can be avoided in well-conducted randomized and quasi-experimental CBAs and CEAs. More information on research design and validity threats can also be found in Appendix 2.

---

[7] Randomization is less feasible when: a) demand for the intervention is low, allowing too few participants for successful randomization, b) the intervention is in its early stages and must be continually refined, c) time and money are limited and resources do not permit the heavy monitoring activities required in randomized studies, and d) there is likely to be frequent crossover between participants and non-participant groups.

Discounting Non-Monetary Benefits. As with monetary costs and benefits, non-monetary benefits should be discounted (Weinstein, et al., 1996). Some people may be uncomfortable discounting non-monetary benefits, because that process implies that gains in health status that occur in the future are less valuable than those which occur today. However, the simple fact is that many of these gains are obtained in exchange for money. Therefore, a dollar spent on health care today may purchase more services than a dollar spent next year. From an economist's viewpoint, equating the health care bought in different time periods with the dollars spent on it implies that the benefits of health care should be discounted over time. Moreover, Keeler and Cretin (1983) showed that failure to discount the non-monetary benefits would lead rational purchasers to delay implementing a useful program indefinitely! This is because delays would reduce the present value of dollars needed to obtain the benefits of the program, but, without discounting, the benefits would remain the same over time. Thus, the longer one waits, the better the cost-effectiveness ratio would appear to be. This problem can be avoided by discounting non-monetary benefits.

Cropper, et al. (1992) provide an example of discounting procedures for lives saved in the future. Viscusi (1995) notes that the appropriate rate of discount for non-monetary benefits is the same as the discount rate chosen for monetary costs and benefits. Others suggest that the discount rates be different for costs vs. non-monetary benefits (Krahn and Gafni, 1993). Different rates should be used if time preference for health has already been counted when non-monetary benefits are estimated. Time preferences for health can be expressed in terms of how much someone would pay to buy health now vs. in the future. Such estimates are often made in cost-utility analyses that measure benefits in terms of quality-adjusted life years.

Component 5: Combining Discounted Costs, Benefits, and Effectiveness Measures Into a Useful Metric

The purpose of accurately estimating program costs and benefits is to combine this information into a useful metric that can be applied to judge the economic worth of the intervention, compared to its likely alternatives. Thus, costs and benefits must be estimated for the program and for each likely alternative use of program dollars.

Cost-Benefit Analysis Metrics. In a cost-benefit analysis, cost and benefit data can be combined in at least three ways:

1. *Net present value (NPV).* The NPV of the program or its alternative(s) is defined as the difference in the sum of the discounted, inflation-adjusted benefits and costs of the program over time, (i.e., as Sum $\{(B_t - C_t) / (1 + r)^t\}$. In this formula, *B* refers to program benefits, *C* denotes the program costs, *t* refers to the year in which costs and benefits are measured, *r* is the discount rate chosen, and the sum is carried out over all years in which costs and/or benefits are accrued). The alternative with the highest incremental NPV[8] is the most economically attractive, since it offers the most value measured in today's dollars (Gramlich, 1981).

2. *Internal Rate of Return (IRR).* One of the difficulties implementing the NPV is that one must choose the appropriate discount rate, *r*, a priori. Another way to cast the analysis is to solve for the discount rate that would make the decision maker indifferent toward the intervention. Since the indifference point is defined as occurring when the NPV of the project is zero, the associated discount rate (called the internal rate of return), can be found by solving for *r* in the NPV formula, when NPV is set equal to zero. If several programs are being compared, the program with the highest IRR is the most economically attractive. If only one intervention is being evaluated, the analyst should compare the IRR to the typical return on other investments made by the employer or program managers (Gramlich, 1981). The program or investment with the higher IRR would be preferable from a strict economic viewpoint.

3. *Benefit-Cost Ratio or Return on Investment (ROI) Ratio.* This measure is simply the ratio of discounted, inflation-adjusted benefits to costs (i.e., Sum $B_t/(1 + r)^t$ / Sum $C_t/(1 + r)^t$ where *B, C, t, r*, and the summation process are as defined above). This ratio specifies the estimated number of benefit dollars received per dollar spent on the program. For example, if each dollar spent on a substance abuse prevention or early intervention program of interest yields $X in reduced medical expenditures on average, X represents the benefit-cost ratio).

---

[8] The incremental net benefit estimate is defined as the difference in the inflation-adjusted, discounted, average monetary benefits and costs of two programs.

Of these three measures, most economists and government policy makers prefer the NPV, for four reasons (Warner and Luce, 1983; Nas, 1996). First, the NPV offers an answer in simple dollar terms, and it can always be calculated. Second, and in contrast to the NPV, the internal rate of return for some projects may not exist, or it may not be positive, while for others, multiple IRRs can be calculated. These situations make it difficult to interpret the results. The IRR does not exist when the benefits of the program in its first year exceed the sum of the undiscounted costs over the life of the program. The IRR is negative when no benefits accrue from the intervention. Multiple internal rates of return can result from projects in which the stream of benefits alternates from positive to negative over time (Nas, 1996). Appendix 3 illustrates some of these problems.

The third reason to prefer the NPV is that, unlike the benefit-cost or return on investment ratio, the NPV accounts for the potential difference in scale between program alternatives. Suppose, for example, that two programs are estimated to have the same benefit-cost ratio of 2.0 to 1.0. Suppose also that Program A costs $1 million and yields $2 million in benefits, while Program B costs $750,000 and yields $1.5 million in benefits. If Program B is not easily replicable and scaleable, Program A should be preferred because Program A yields more *net* benefits in today's dollars (i.e., $1 million, versus $750,000 for Program B).

The fourth reason to prefer the NPV is that, unlike the benefit-cost or return on investment ratio, the NPV is not affected by the placement of negative benefits in the equation. Suppose, for example, that two programs of equal size yield $100,000 in positive economic benefits and -$20,000 in negative economic benefits (both in inflation-adjusted and discounted terms). Suppose also that both programs cost 50,000 inflation-adjusted and discounted dollars to offer and receive. Using the NPV formula would yield identical results ($30,000), regardless of whether one puts the –$20,000 negative benefit figure on the cost side or the benefit side of the equation (i.e., ($100,000 - $20,000) - $50,000 = $100,000 – (-$20,000 - $50,000)). In contrast, the ROI ratio depends on which side the $20,000 in negative benefits are put (i.e., ($100,000 - $20,000) / $50,000 is not equal to $100,000 / ($50,000 - $20,000)).

Alternatively, there is one situation in which the NPV may be less useful for choosing a project than the other methods noted above. This occurs when more than one project can be chosen within a limited budget. Nas (1996) shows that, in such a scenario, analysts may prefer to adopt multiple projects with lower NPVs, instead of adopting the single project with the highest NPV. This strategy is useful if the sum of the NPVs from the multiple projects exceeds the NPV of the project with the highest NPV, and if the cost of adopting the multiple projects is not prohibitive.

Finally, of the methods noted above, the benefit-cost ratio is probably the most familiar to business managers. If the benefit-cost ratio is calculated, we suggest supplementing it with the net present value figure, to provide a more complete and accurate picture of the impact of the CBA. Methods for estimating the net present value are illustrated in Appendix 3.

Cost-Effectiveness Analyses Metrics. In cost-effectiveness analysis, the cost and effectiveness estimates are often combined to show the number of inflation-adjusted, discounted dollars it costs to produce one unit of benefit. Examples are the mean or median cost per relapse avoided, cost per 1% drop in substance abuse prevalence, cost per quality-adjusted life year, etc. Total costs and total numbers of the effectiveness measures should also be shown, so incremental cost-effectiveness ratios can be estimated. Generally, the program with the lowest incremental cost-effectiveness ratio is preferred.[9]

When a variety of effectiveness measures have been used in a CEA, experts or stakeholders may be consulted to judge which are most important, especially if the same alternative does not dominate the rest on all measures. Later we illustrate effective means for presenting the results of a cost-effectiveness analysis for consideration.

Component 6: Dealing With Uncertainty

Throughout the process of performing CBA or CEA the researcher will be faced with many decisions that cannot be made on the basis of empirical evidence alone. Examples include:

---

[9] The incremental cost-effectiveness ratio is defined as the difference in the inflation-adjusted, discounted, average costs of two alternatives, divided by the difference in the discounted average levels of effectiveness of those two alternatives.

- Whether to focus on charges for medical care services versus actual payments for those services;
- The perspective(s) to apply for the analysis;
- Which discount rates to use;
- Whether and how to make adjustments for differences in benefit levels across health plans or over time;
- Whether to rely on self-reported data versus secondary data such as health care claims;
- How much to value the time of program recipients and caregivers; and
- Whether to rely on wage rates and market prices as estimates of opportunity costs.

Analysts may also wish to know how robust findings would be to changes in survey questions used to assess substance abuse problems, changes in statistical testing procedures, or to various weights applied to effectiveness measures. The recommended approach for dealing with uncertainty in a CBA or CEA is to conduct sensitivity analyses.

Sensitivity analyses involve repeated re-estimation of the NPV or cost-effectiveness measures as assumptions change about important factors (Stewart, et al., 1998). For example, analysts might wish to investigate the impact of best case versus worst case scenarios for various measures of interest (Warner and Luce, 1982). Weinstein, et al., (1996) suggest that sensitivity analyses be done first by changing one assumption at a time, then by changing two or more assumptions at the same time. Mullahy and Manning (1995) concur, noting that many factors interact simultaneously to determine cost and effectiveness. They suggest that sensitivity analyses be conducted via simulations that vary several cost and effectiveness determinants at the same time, to show how the rankings of the various treatment alternatives will be affected.

Another useful sensitivity analysis technique was offered by Warner and Luce (1982). They suggest using break-even analysis to show how much an assumption would have to change to equate two interventions being compared. Finally, Hargreaves, et al., (1998) suggest using sensitivity analyses to determine whether the results of the CBA or CEA are similar for important subgroups (e.g., by gender, race, or severity of the underlying substance abuse problem).[10]

Sensitivity Analyses Versus Confidence Intervals. The cost-benefit and cost-effectiveness values obtained in CBA or CEA are point estimates that are subject to random error (Hargreaves, et al., 1998). Thus, one may wish to generate confidence intervals around those point estimates. In the past, many researchers ignored this issue and used sensitivity analyses to generate a range of values. While sensitivity analyses are indeed useful for addressing uncertainty, they cannot be used to estimate 95% confidence intervals around a mean CBA or CEA value. Hargreaves, et al. provide a formula which can be used to find a 95% confidence interval for a CBA, if net benefits can be estimated at the person level:

$$\text{95\% confidence interval is } NB \pm t_{N-1} s(NB) / N^{1/2.}$$

In this formula, NB is the sample mean of patient-level net benefits, N is the number of subjects in the analysis, t_{N-1} is the 97.5th percentile of a T distribution with N-1 degrees of freedom, and s(NB) is the sample standard deviation of NB. If the 95% confidence interval excludes zero, NB is significantly different from zero at the 5% alpha level.

Estimating a confidence interval in a cost-effectiveness analysis is more difficult because the theoretical distribution of the ratio two normally distributed random variables (e.g., of the cost and effectiveness estimates) is undefined. Hargreaves, et al. (1998) suggest the use of bootstrapping statistical techniques to address this problem, and they provide references for that technique. Bootstrapping involves repeated re-analysis of the data, each time using a random sample chosen from the original sample used in the analysis. For example, 1,000 small samples may be chosen, with replacement, from the original sample. The CEA would then be conducted

[10] One may control for factors such as age, gender, and severity of illness by doing completely separate analyses for subgroups based on these factors (e.g., for men vs. women). Alternatively, one may simply use binary indicators for these factors in a single impact analysis. The choice depends on assumptions about how these factors and others influence outcomes. For example, if women are expected to have higher health care expenditures than men regardless of their age, severity of illness, and other factors, controlling for gender with a binary indicator for female status in an expenditure regression analysis will suffice. On the contrary, if many factors affect outcomes differently for women than men, separate analyses by gender would be necessary. This issue should be considered before any impact analyses are conducted, so sensitivity analyses that compare outcomes for men vs. women can be performed accordingly.

for each small sample. The variance obtained from the results of the 1,000 analyses may then be used to construct the confidence interval of interest (Vogt, 1993).

Component 7: Presenting the Results of a CBA or CEA

Siegel, et al. (1996) present a useful set of recommendations for reporting the results of a CBA or CEA. They suggest presenting results in a standard way for easy comparison across CBAs or CEAs reported in the literature. They also note that a useful CBA or CEA report contains all of the elements of a useful research study report. These include:

1. An explanation of the framework of the analysis, including its context, goals, and objectives;
2. A detailed description of the intervention and its alternatives and their target population;
3. A justification for the perspective(s) chosen and the associated research agenda;
4. A description of the scope of the analysis and how timelines, budget, data availability, and other factors impact that scope;
5. Details about the research design to be used to estimate costs, benefits, and effectiveness;
6. Details about how costs, benefits, and effectiveness measures were obtained;
7. A clear explanation of statistical testing procedures and their limitations;
8. Total and incremental cost-benefit and cost-effectiveness estimates;
9. A detailed justification for any assumptions made, and an explanation of how these were tested in sensitivity analyses;
10. The results of the sensitivity analyses and a description of how the order of preferred alternatives vary as assumptions used to conduct the analysis change;
11. A discussion of the likely impact of choosing the program with the best incremental NPV or cost-effectiveness ratio, focusing on those who stand to lose resources if more resources are allocated to the winning program;
12. A description of study limitations, generalizability, and validity threats; and
13. A discussion of any ethical issues confronted in the analysis.

These 13 elements are often described in lengthy reports and journal articles. Reports produced for more general use should include a short (2-3 page) executive summary. Graphics (e.g., bar charts) are often easier to digest than are detailed tables, but tables should be provided in an appendix for interested reviewers.

Probably the most important point to consider when writing the CBA or CEA report or journal article is to consider the needs of the audience and stakeholders. Focus on important political, scientific, and ethical issues. Tell the audience what the timeline, budget, and data would let the analyst consider, and what could not be considered because of these binding constraints. When describing results, avoid a lengthy discussion of alternative programs that are clearly dominated by other programs, but do note reasons why programs may be dominated by others. (Dominated alternatives are those that were always found to have lower NPVs or cost-benefit ratios, regardless of the sensitivity analyses completed.)

Dealing With Many Outcomes or Sensitivity Analyses. When there are many outcome measures in a cost-effectiveness analysis or when many sensitivity analyses were conducted in a CBA or CEA, it can be a challenge to present the results for effective decision-making. Warner and Luce (1981) recommend developing a table in this case. In the context of a CEA, the rows of the table pertain to the programs being analyzed; these should be sorted by cost. The columns pertain to the cost and effectiveness figures used in the analysis. An example is shown with the following table shell:

Table 1: Example of an array of multiple effectiveness measures in CEA

		Potential Effectiveness Measures*				
Program	Total Cost (Sort from lowest to highest)	Number of SA Cases Avoided	Average Time Until Relapse	Mean Health Status Scale Score	Mean Work Locus of Control Scale	Quality Adjusted Life Years Gained (Or other effectiveness measures ...)
A						
B						
C						
D						

***The effectiveness measures noted in this table were chosen merely to illustrate the points raised in the text. In any particular CEA of interest, the relevant outcome measures may be quite different.

A second table based on incremental cost-effectiveness would be useful as well. This table would compare Program A to Program B, Program A to Program C, etc. The columns would reflect incremental costs (i.e., difference in costs between programs) and the incremental effectiveness (the difference in effectiveness measures).

The text used to describe the results shown in these tables should reference whether the order of the programs' net present values or cost-effectiveness values changes when sensitivity analyses are conducted. Alternatively, another table or set of tables could be produced, such as the following:

Table 2: Illustrating the impact of sensitivity analyses in a CBA

	Incremental Net Present Value Estimates Under Various Scenarios				
Program	First analysis	Sensitivity Analysis 1	Sensitivity Analysis 2	Sensitivity Analysis 3	Sensitivity Analysis 4
A vs. B					
B vs. C					
A vs. C					

Analysts may wish to change the presentation to graphics such as bar charts when the final presentation is made. The number of tables or graphics to include, or the number of columns to include in those tables, should correspond to the most theoretically, ethically, and politically important outcomes of interest.

The final approach taken should highlight, in text or tables, whether the results of the CBA or CEA are affected materially by the adoption of any particular assumption. For example, if a project looks financially beneficial in a CBA that focuses on the corporate bottom line but looks like a financial loser when a broader perspective is taken, this should be noted prominently in the project report and the validity of each perspective should be discussed in detail. The project report must clarify for the reader how different assumptions can lead to opposing policy recommendations. If possible, the report should also offer text on the validity of the competing assumptions and a recommendation about which assumption to place more weight upon.

Component 8: Limitations of the CBA or CEA

As mentioned earlier, time, money, and data constraints and stakeholder concerns may dictate the scope of the CBA or CEA. They may also influence the research design chosen, the nature of the statistical tests conducted, and the reliability, validity, and generalizability of the results. It is the researcher's responsibility to describe the limitations of the analysis and the implications of those limitations. This is often done by stating the context and research agenda for the CBA or CEA, noting what important research questions could not be addressed, and which distributive justice or equity concerns could not be addressed. In addition, the report should note how analysts may have been forced to deviate from optimal scientific design and analytic principles, and what that deviation implies about any conclusions drawn from the analysis. Later in this paper we describe many of the major threats to the validity of a good CBA or CEA, along with some recommendations for avoiding these threats.

CHAPTER 4

LOGISTICAL ISSUES

The previous three chapters set the stage for the SAMHSA cost research evaluation guide and defined and described the cost-benefit and cost-effectiveness analysis techniques that some CSAP Workplace Managed Care Initiative grantees may wish to apply. This chapter builds on previous material by describing many of the logistical issues that must be addressed when CBAs and CEAs are conducted. These include:

1. Deciding how data should be collected;
2. Cleaning data and dealing with incomplete data;
3. Creating person-level files for analysis;
4. Linking files and protecting the confidentiality of data;
5. Coding substance abuse problems and reliability and validity issues; and
6. Estimating financial equivalents in the absence of claims data.

These issues are discussed in turn.

Deciding How Data Should Be Collected

Hargreaves, et al., (1998) describe three major types of cost data for CBAs or CEAs. These include actual study costs, local accounting data, and aggregate data from larger regional or national surveys. Actual study costs tend to be the most accurate, because more effort is spent to identify all of the cost elements for the intervention and to collect information on the opportunity costs for those elements. Actual study costs might be based on observations of staff activities and how other resources are combined to produce the intervention (e.g., from time and motion studies or time and resource logs). Such data tend to be very difficult and expensive to collect, and many analyses therefore rely on other approaches.

Local accounting data tend to be more readily available and are most often used to estimate opportunity costs. For staff and other resources, local accounting data include wage rates, hours of service devoted to the intervention, supply prices, and prices or rental rates for other goods and services needed to produce the intervention.

Aggregate data might be used to help develop opportunity cost estimates when local data or actual study data are incomplete. For example, Hargreaves, et al. (1998) note that hospitals which participate in the Medicare program must submit annual reports detailing total costs of operation and rates of costs for services to the charges made for those services. Regional cost-to-charge ratios can be calculated and applied to a facility's charge data when its cost data are not available, in an effort to proxy the opportunity cost of providing services.

Costs Versus Charges and Payments. In many of the CSAP Workplace Managed Care Initiative interventions, benefits are expected to include reductions in medical or behavioral health care expenditures. In some interventions, direct service provision must be accounted for on the cost side of the analysis. In either case, a key question of interest is how to value the health care services used. Without detailed opportunity cost data, researchers may have to rely on charges made by providers for those services, or on payments made to providers for those services, to estimate the value of the services used.

Hargreaves, et al. (1998) suggest that payment data be used instead of charges to proxy the opportunity costs of service provision, for several reasons. First, they cite studies showing that payments are more accurate proxies for opportunity costs than are charges. Second, they note that charges often include efforts to account for bad debt on other services. Third, charges may reflect price shifting from some who are willing to pay less for a service (e.g., Medicaid programs) to those who traditionally have been willing to pay much more (e.g., private insurers). We agree that actual payment data should be used when available instead of data on charges. Those who disagree are recommended to conduct sensitivity analyses to show how results vary when payments versus charges are used.

Finally, if payments are adopted as the proxy for opportunity costs, the researcher must decide whether to include payments made by program beneficiaries directly. Beneficiary payment responsibilities include deductibles, coinsurance, or copayment amounts. Deductibles reflect the cost of initial services which are not covered by the insurance policy. An example might include the first $500 spent for inpatient care. The deductible provision states that these dollars would be paid by the policy-holder before the insurance company or health plan makes any payments for inpatient care. Coinsurance reflects a proportion of the value of services provided that are also not covered by the policy, even after the deductible is met. For example, some insurance plans pay for 80% of an approved price for inpatient care, or for outpatient care obtained outside the network of providers established by the plan. The patient must then pay the remaining 20% of the approved price for the services received. Copayments are similar in nature to coinsurance, except copayments do not usually reflect a particular percentage of the price of the service received. Rather, the copayment is usually expressed as a given dollar, up-front price for a service, such as $10 that must be paid for each physician visit. Typically, the copayment is paid at the time of the service, whereas deductibles and coinsurance may be paid later when billing occurs.

Out-of-pocket costs for deductibles, coinsurance, and copayments should be included in the CBA or CEA if the perspective taken includes that of the recipient of the intervention service. If the perspective is justifiably more limited (e.g., to reflect the bottom line amount paid by an employer for the intervention), then out-of-pocket payments would not be included in the CBA or CEA.

<u>The Level of Data Collection.</u> The data used in CBAs and CEAs of substance abuse prevention and early intervention programs are likely to exist at several levels. For example, health care claims data are usually available at the service level. Personnel data are usually available at the person level or at the business unit level. Survey data are typically available at the person level, although for confidentiality reasons survey data may be aggregated to a higher level. Most analysts prefer to collect data at the most detailed level possible, such as the service level or person level. This allows the greatest flexibility for subsequent analyses, but it entails greater costs.

The following recommendations may be useful for deciding how much data to collect and at what level to collect them:

1. Collect data at the most detailed level that the budget and other constraints will allow.
2. Collect as many data elements as possible for those elements that influence the likelihood of participating in the intervention. If randomization is used to define participation status and treatment group and control group membership, collect these data elements anyway. This will allow the researcher to check whether the randomization was successful and adjust analyses if randomization did not work to equate the study groups.
3. Sort the data elements in terms of those which are most important from theoretical and stakeholder perspectives. Then collect data on as many elements as can be afforded. This will assure that subsequent controls or adjustments can be made to account for

important factors that influence the outcomes of interest.

4. Collect payment data[11] if available, and collect data on deductibles, coinsurance, and other out-of-pocket payments in addition to health insurance payments. Data on out-of-pocket expenditures will enhance the ability to conduct analyses from the participant's perspective.
5. Collect one or more unique identifiers that allow data to be merged from several sources. These might include social security numbers or combinations of other demographic variables such as name, address, gender, zip code, etc. Scramble these data for security and limit access to them to the extent possible to avoid a breach of confidentiality.
6. Establish a liaison with an expert located where the data were originally assembled or collected. This liaison can be very helpful when learning about the nuances of data and making formal requests to receive data.

Many other helpful suggestions about data collection have been offered by staff at the Workplace Managed Care Coordinating Center. For details, see the interview guides they prepared for dealing with managed care organization data systems, human resources departments, and employee assistance programs (Bray, 1998; Bray and Zarkin, 1998a, b).

Cleaning Data and Dealing With Incomplete Data

Once data are collected their suitability for analysis must be investigated. This investigation involves efforts to "clean" the data, to identify and purge incorrect values, to replace incorrect values or missing values with correct values (if possible), and to learn the nuances of the data that may influence the ability to perform subsequent analyses. There are several useful steps that should be applied in the data cleansing process. These include:

---

[11] Generally, payment data are collected for all services incurred during the time period of interest. This is easy if the time period of interest occurs well after all claims have been paid. If the time period of interest occurs very soon after claims are incurred, many claims may not be processed and paid when the payment data are collected, resulting in missing payments. In such cases, the claims for services used should still be collected, and the analyst may wish to impute expected payments for those claims that have not been processed and paid yet. Expected payments can be taken from the fee schedule agreed upon by the health plan and its providers, or by using the imputation strategy noted later in this report.

1. Fully documenting each data source by generating complete data file layouts and dictionaries that describe the location and content of each variable;
2. Developing standard definitions and associated coding processes for variables that appear in multiple files;
3. Developing detailed data quality reports to check on the timeliness, availability, reliability, and validity of the data; and
4. Generating algorithms to impute missing data or to replace erroneous data.

Data File Layouts and Dictionaries. Many data preparation and analysis programs (e.g., SAS, SPSS) produce data file layouts describing the location, size, and format for variables included in the file. However, many programs allow insufficient room for detailed descriptions of the content of each variable. Analysts often develop more descriptive tables or text files to avoid this problem. These tables and text files often include details about survey question wording and response categories, how missing data were coded or otherwise handled, how imputed observations were identified and handled, etc. The time spent producing more detailed dictionaries is often well spent, because questions or concerns about the nature of the data may arise that would not otherwise become known. Raising these questions with the liaisons from the original sources of the data will usually result in a more thorough knowledge of the strengths and limitations of the data.

Standard Definitions and Coding Processes. In the context of the CSAP Workplace Managed Care Initiative, the need for standard definitions and coding processes is twofold. First, within any particular grantee project, researchers may have to merge data from multiple sources for the CBA or CEA. These sources might include health care claims data, separate pharmacy records or behavioral health care carve-out records, absenteeism files, health plan enrollment files, survey data files, EAP files, and other sources. These files will vary in content and quality, yet there may be some overlap that allows linkages to be made across files.

Linking fields might include name and address fields, social security numbers or other numerical identifiers, demographics such as age and gender, and other location information. When these fields exist in multiple data sources they should be checked for consistency and data availability. Conversations with the data liaisons may allow “gold standards” to be developed which dictate which source of information is the best for any given field. The final analysis file that is built from a combination of input files should include the gold standard variables and note their sources. Any data that are replaced in the final analytic file should be retained in the original source files, however. In addition, any data that are imputed in the final analytic files should be flagged and additional variables should be added to denote how the imputation was carried out.

The second requirement for standard definitions and coding processes within the CSAP Workplace Managed Care Initiative arises from the need to perform a cross-site evaluation using standard CBA or CEA techniques. With nine different funded projects, many different definitions may be used for key variables that influence or measure cost and effectiveness. A substantial amount of time has already been spent in Steering Committee meetings to reduce this variation. Some recommendations have also been put forward in this document. Individual grantees may benefit from reviewing those recommendations with their data liaisons and then conferring with the cross-site evaluators to assure consistency across the nine projects.

Data Quality Reports. A variety of sources can be used to check the availability, consistency, reliability, and validity of data elements. Examples include health care claims files from multiple providers, billing records, health plan enrollment files, medical records, and conversations with data liaisons.

The degree to which these sources are consulted often depends on the budget for the CBA or CEA. For example, questionable data on specific service items or charges in the claims data can be verified by reviewing medical records or provider bills, but this can be a very time consuming and expensive process. While medical record data are often viewed as a gold standard for comparison to other sources, the reality is often that medical records are filed improperly, that portions are missing, or that recording practices are not optimal. For seriously ill persons the medical record will be quite voluminous. If it is not indexed well, it may be hard to find corroborating materials. For these reasons, tracking and reviewing medical records can be difficult and expensive, often ranging from $75-125 per record reviewed. If the project budget will not support this activity, or if the analyses require data not found in medical records, a variety of other methods can be used.

At a minimum, data cleaning efforts should include analyses of data availability. For example, a simple table can be created for every field in the database to show how often data are missing, how often available data have valid entries, the range of those entries, and basic statistics such as mean, median, and standard deviation values. Scanning this table may raise questions about missing data, outlier values, and coding conventions that the data liaison can answer.

A better data cleansing approach includes comparative information from the multiple providers of data. For example, if more than one health care provider or health plan will be submitting data, the above mentioned table can be structured with columns for each provider or plan. This will allow easy comparison of data attributes across plans and may raise further questions about data quality that can be investigated with the data liaison.

In addition to data availability and coding convention investigations, other logical studies of data quality should be done. For example, a comparison of maternity services by gender should note no such services for males, and total payments for all services should be positive.[12] Next, utilization and expenditure reports should be generated that compare rates and dollars for inpatient, outpatient, pharmacy, behavioral health care, and other utilization measures across providers or plans.

Throughout the data cleansing process, questionable data values should be flagged and then sorted by the original source (e.g., by doctor, health plan, location, etc.). The odds of incurring questionable values should be noted by source, to learn whether one or more sources are particularly problematic and need help to produce more reliable or valid data.

Throughout the data cleaning process, the analyst may notice payment fields with negative values or multiple claims submitted for services received on the same day. Over a lengthier time window, one may also see claims from several dependents associated with the major policy holder. Some of these occurrences may be reasonable and some may not, and it is the analyst's job to find the reasons why these phenomena occur. Some relatively straightforward checks are noted next.

There are at least two reasons for observing negative payment amounts. Sometimes the negative amount represents a correction because payment was inadvertently made for services not covered under the insurance policy. This can be checked by comparing the service with the negative payment to a list of covered services. In other cases, payment fields with negative values represent efforts to recoup overpayments for a covered service incurred by the beneficiary. This may be checked by searching for an earlier claim filed for service(s) received on the same day as

[12] Negative payment amounts that reflect post-hoc adjustments are common, but the total payments over a given period such as a year should still be positive.

the service noted on the negative payment claim. Once the earlier service claim(s) are found, a review of the payment history for those claims should be conducted, to verify that the payment offered was indeed higher than contracted amounts. If this is the case, total expenditures for the time period of interest should include the negative payment amount. If no previous claims can be found, or if the documented payment history is incomplete, the analyst should note this in the data quality report and check with the data liaison before deciding whether to include or exclude the claim with the negative payment value. Alternatively, the analyst can sum the charges from the two claims (i.e., the one with the positive payment amount and the one with the negative payment amount) and compare the result to the range of payments made for the same service for other beneficiaries. If the summed amount is typical of payments for the same service incurred by other people, the negative claim is likely to be a valid one.

A similar strategy should be taken to address multiple claims found for services received on the same day. It is common to find multiple claims, because the "one-stop shopping" characteristic of many health facilities lets patients receive laboratory tests, x-rays, or other services on the same day. Many times ancillary services are used within a few days of a physician visit or a scheduled hospital admission. Analysts who find multiple claims for service on the same day can therefore search for another claim that includes the physician visit or the hospital admission within a reasonable time period, such as a week or two from the date of the multiple claims. Finding a claim for the doctor's visit or the hospital stay would provide evidence that the multiple claims for ancillary services are okay. Generally, multiple claims are questionable only if two or more claims can be found for exactly the same service on the same day, with the same payment amount and provider identification number. Such claim records should be put aside, sorted by source, and then discussed with the data liaison to determine whether the duplicates should be excluded. The prevalence of questionable multiple claims instances should be noted on the data quality report.

Finally, during the data cleaning process, the analyst should sort claims according to the identification number of the major policy holder. Such an identification field is typically found on the claim record, so that enrollment can be verified before payment is made. Once the sort is made, the analyst should check the claims to see if one or more dependents of the major beneficiary are noted in the claim stream. This can be seen by noting different *patient* identification numbers on claims with the same major policy holder identification number. For those with multiple patient identification numbers, the analyst should check records in the enrollment data base to verify that coverage exists for multiple persons under the major policy holder's account. Major policy holders sometimes change the number of dependents they cover due to major life changes (e.g., marriage, divorce, death), and the enrollment data base should include information about these changes. If verification of coverage cannot be found, the analyst should check with the data liaison to determine whether the claims for dependents should be dropped. The prevalence of questionable covered lives should be noted on the data quality report.

A guiding premise for data cleaning is that analysts should take the time to think about as many different types of data problems that would compromise analyses. It is also worth conferring with data experts to develop a list of expectations about data availability, content, coding practices, ranges, and utilization and expenditure variables. Once these problems and expectations are documented, programmers can test the data to look for problems and try to verify expectations. Questions arising from this process can then be referred to the data liaisons and remedial action can be taken and flags can be created to warn analysts of problematic observations.

Imputing and Replacing Data. Sometimes missing or erroneous data may be imputed logically from other sources. Simple examples include imputing age from birth dates in enrollment files, or replacing missing diagnosis codes in disability files with diagnosis codes found in medical claims data. When very simple logical imputations are not possible, some researchers use hot deck or Bayesian approaches to impute data (Little and Rubin, 1987). These techniques involve selecting replacements for missing data based on data from variables that are correlated with the variable with the missing observations. The statistical properties of imputed data are noted in Little and Rubin (1987), and the techniques they describe may be quite useful. However, the use of imputed data may require multiple iterations of data analysis, each differing by the imputed value(s) chosen. This may complicate the analysis and add to the cost of conducting the CBA or CEA.

Regardless of the technique chosen to impute data, we recommend that researchers create flags to denote each imputed observation and how the imputation was performed. This will allow for sensitivity analyses later to illustrate whether the imputation strategy affected the results of the CBA or CEA.

Finally, we do not recommend imputing values for key cost or benefit outcome measures of interest, unless these can be done with logical imputations. Very often, missing outcome data occur non-randomly. For example, some program participants may decline to answer questions about substance abuse because they do not wish to divulge their alcohol or drug use patterns. When outcome data are missing in a non-random fashion, a selection bias may occur. Methods for dealing with this bias are noted later in this paper, and in Little and Rubin (1987).

Creating Person-Level Files for Analysis

When data from multiple sources will be used for the CBA or CEA, input datasets will often be in different formats from each other and from the format of the analytic file which will be the result of the merge. For example, a file from the human resources department containing demographic information usually contains one observation per person, while a health care claims file usually contains one observation per health care service. An absenteeism file usually contains one line per person per time period, while a disability program file often contains one line per disability case. The standard analytic file is most useful if it contains one line per person. This section offers some tips for creating person-level files for analysis.

In the scenario just noted, it is easy to handle the demographic information; one simply copies the demographic fields of interest from the human resources source data set to the final analytic file.[13] Figuring out how to summarize the health care claims, the absenteeism data, and the disability program data requires a little more thought. There is no single correct way to summarize the information. Depending on the analysis desired, the methods will change. Also, a different method will most likely be used for each type of data.

The building blocks of the CBA are the cost and benefit data, which often come from health care expenditures or expenditures for other phenomena (e.g., for sick days, disability program days, etc.). Expenditure data are usually summed over the study period of interest, or over a meaningful subperiod, such as a year. The summation may be performed using a few different methods, even for the same analysis. For example, a common expenditure variable reflects total expenditures for the period of interest, which adds all health care service (inpatient, outpatient, mental health, and drug) expenditures within the period. In addition, each of the types of services can be summed separately, producing measures of inpatient expenditures, outpatient expenditures, mental health expenditures, and drug expenditures. In one analysis, one may wish to determine whether subjects' expenditures differed from one time period to the next, so different expenditure variables must be created for each time period. In yet another analysis, one may wish to analyze mental health costs separately, so all expenditures associated with diagnoses that fall within a specified ICD-9 code range or major diagnostic category should be summed separately.

Summing values is an easy way of combining information from several different observations over a time period. However, some variables must be aggregated in a different way. For example, diagnosis and procedure codes cannot be combined easily when forming a person-level file. For certain analyses, one may know which diagnosis codes are of interest when building the analytic file. However, at other times one may have no prior knowledge of which diagnoses will be analyzed later.

The most common way to deal with this issue is to create arrays of diagnosis and procedure codes. The number of array elements depends on the analysis and the length of the study period. The longer the study period, the more diagnoses and procedure codes a patient is likely to have. To reduce the size of the analytic file, one may limit the maximum number of

---

[13] If conflicting information about demographics is noted across files, a gold-standard should be chosen or created from those conflicting data. The gold standard can be retained in the final analytic file, but the conflicting source data should be retained in the original files if needed later.

diagnoses and procedure codes that are used. For example, one may save up to 12 (or some other meaningful number of) codes.

To make sure that the codes saved are the most relevant, codes associated with the most costly services should be kept first. To allow enough space to keep the codes that reflect the whole spectrum of care a patient receives, each specific diagnosis or procedure code should only be saved once. Finally, to allow for a more complete analysis of the casemix of the population of interest, one may save all of the unique diagnosis and procedure codes.

In our experience, and as noted by Miller (1998, personal communication), the vast majority of subjects will have fewer than a dozen unique diagnosis codes in any given year, and most subjects have less than six or seven. With chronically ill persons, however, the number may be much larger.[14] In this instance, saving all the unique diagnoses may lead to space problems. Researchers should consider the tradeoff between saving space and using less than the complete number of unique diagnoses. Saving space is important for large computer files, since computing power would be maximized and processing time can be minimized. However, retaining fewer diagnoses than are available may compromise the ability to describe the casemix of the sample. It may also limit the ability to control for comorbidities in statistical analyses. This may lead to bias if comorbidity patterns differ for substance abuse program participants and non-participants.

Absenteeism and Disability. Information on absence and disability variables are often found in the data sets as per-person or per case variables for a given time period. For example, in many human resources data sets there are up to 12 absence day variables, which represent the number of days absent for each month during a calendar year for a person. It is easy to sum these variables to create a total absence days variable if the study period is defined by calendar time. However, many research studies are based on a "trigger" event. A trigger event is an occurrence that signals the start of a study period. For example, it may be the date that a subject takes a health risk assessment or the date a subject joins a substance abuse program. The study period might be defined as the year beginning with the trigger date or as six months before to six months after the trigger date, for example.

[14] For another project focusing on program beneficiaries with disabling physical and mental health problems, MEDSTAT researchers found 99% of the clients had 20 or fewer unique ICD-9 diagnosis codes, but many had more than 12. The first 20 unique codes were therefore saved for the analytic file.

For the first and last months of the study period, it may be impossible to determine which part of the absence days fell within the study period and which part fell outside the period. For example, if the study period based on a trigger date is 3/20/95 to 3/19/96 and the data indicate that the subject had five absence days during March of 1995, it is impossible to determine whether all or some of them occurred after March 20th. In these cases, the number of absence days falling within the study period may need to be imputed. In the example, 12 days of March of 1995 fall within the study period out of 31 days in the month. The number of absence days in March can be imputed by multiplying the five absence days by 12/31, which yields 1.94. Since the analysis will consider group means or totals, it does not matter that 1.94 is not an integer. For each of the other months besides the first and last in the study period, the absence and disability days would be summed and added to the imputed first and last month totals to obtain a study period total.

Utilization Measures. Other variables that are often summarized for analysis reflect health care utilization. Examples of these include inpatient admissions, outpatient services, and drug prescriptions. Other utilization variables include program participation values, such as number of times visiting a counseling center. In general, these utilization measures are summed in a similar manner to the expenditure variables.

A Few Final Thoughts on Data Set Creation. In some cases, especially with very large study populations and data sets, the creation of the analytic data set may take a long time and require a lot of disk space. This makes it problematic to go back and add new variables to the data set. The best way to deal with this issue is to anticipate variables that may be potentially useful in an analysis. Consulting program-based or discipline-based theory, talking to stakeholders, and reviewing the literature thoroughly will usually produce a fairly complete list of variables needed for analysis.

Next, it is best to formulate variables in a manner that increases flexibility later. For example, the analysis plan may call for a variable representing the number of unique major diagnostic categories (MDCs) that a patient may be classified into. It may be better to create 25 binary variables that represent whether the patient had a diagnosis falling into each of the 25 MDCs.[15] One may use these 25 variables in the analysis, or one may create one variable representing the unique number of MDCs each person had. Similar approaches can be taken with

---

[15] See Fetter, et al. (1991) for a definition and description of the 25 MDCs. MDCs 19 and 20 refer to mental health and substance abuse problems, respectively.

other variables. For example, if total expenditures are called for in the analysis plan, inpatient, outpatient, and pharmacy expenditure variables can be used to provide that information while increasing analytic flexibility.

Finally, as in any business task, it is important to document and communicate the methods for variable creation. Most of all, this helps to keep all researchers involved on the same page, eliminating surprises later and reducing the likelihood of having to recreate the analytic file. Also, documentation should note the decisions made about important variables, such as the formulation of outcome variables and the categories created for key explanatory variables. If the decisions are not agreed upon by everyone involved, it is much easier to change the documentation beforehand than to recreate the analytic file later.

Linking Data From Multiple Files and Guarding Confidentiality

The key consideration when merging two or more data files is finding a good linkage, or 'key', field. A good key field must have three qualities: it should uniquely identify a person, it must be found in all of the data sets to be merged, and it must be complete in all data sets.

One of the most commonly used key fields is the Social Security Number (SSN). Because of its unique nature, this field is most often recorded by health plans and providers in their data as a contract identifier. In addition, the completeness of the SSN field is not usually problematic in health care claims data, because almost every health care record requires a contract identifier.

However, the SSN, like every other key field, is not perfect. First, many Americans have health care coverage provided by their employer, making the employee the point of reference for the coverage. Therefore, on most health care records where a dependent is the patient, the employee's SSN is reported instead of the actual patient's SSN. Dependents and spouses are generally given a unique suffix, and this suffix or other characteristics (relation to employee, age/gender) can be used as additional key fields to the SSN when dependents and spouses are to be analyzed.

Second, not all health care plans use the SSN as their contract identifier. Similarly, employee assistance programs and other voluntary programs may not collect social security numbers, to reinforce the confidentiality of their services. If the SSN is used as the contract identifier in some data sets but not others, a conversion map linking SSN to the other identifying field is required before merging can take place.

Third, SSN, like any other key field, is susceptible to incorrect or missing coding of its values. For example, the SSN on a particular record may have been typed incorrectly or it may not have been entered at all. These situations may occur on few records, but will lead to an imperfect merge.

Fourth, the SSN is more likely than any other field to have confidentiality issues associated with it. Since SSNs are assigned by the government and are used as person identifiers by all types of organizations, including financial institutions, schools, and courts, a wealth of information can be found about a person just by knowing their SSN. In the last few years, with the Internet and other technologies emerging that provide quick access to information that was previously almost impossible to obtain, the confidentiality issue has become much more of a concern.

A field that has all of the benefits of the SSN field, but doesn't have the same confidentiality concerns is a modified SSN. The modified SSN usually is the result of a scrambling algorithm applied to the SSN. An example of a simple algorithm is an algorithm that switches the 1^{st} and 8^{th} digits, the 2^{nd} and 5^{th} digits, the 3^{rd} and 9^{th} digits, and the 6^{th} and 7^{th} digits. Under this algorithm, the SSN 123-45-6789 would result in a modified SSN of 859-42-7613. Many other, more complicated, scrambling algorithms exist as well. For example, the Oregon Medicaid program uses a bits permutation approach to scramble social security numbers in a way that makes it extremely difficult, if not impossible, to guess the original SSN.

The modified SSN works well as long as the following two requirements are met. First, the modified SSN must be included in every data set to be merged. If some data sets contain SSN while others contain a modified SSN, or if the data sets contain multiple variations of the modified SSN, large problems will occur with the merge. This situation is worse than having two different types of key fields in the data sets to be merged because the SSN and the modified SSN may look alike to the naked eye, especially if the data set is large. Thus, merges may occur between records that represent different individuals.

Second, the scrambling algorithm should be known only to those who have access to the original SSNs. If the algorithm becomes known to those who should not have access to SSNs, confidentiality concerns may become problematic.

Other fields are sometimes used as the key field in merges. For example, a synthetic family identifier may be created, which uniquely determines a family represented by an employee. In addition, a within-family member identifier may be created, which uniquely determines each member within the family. For example, a family of five will share one family identifier but have five different member identifiers. Together, the family and member identifiers uniquely represent each person in the data.

When analyses are performed which utilize insurance claims, health plan enrollment data, absenteeism data, and other health productivity and management databases, the family and member identifier fields often cannot be used as the key fields, because they usually do not exist in all of the databases. It may often require the written permission of the employer and the employee to make the linkages between key fields and develop a conversion map.

Finally, when identifying numbers are not available but names, addresses, and other descriptive information are available, a phonetic-based matching system can be used to link files. One example of such a system is the National Center for Health Statistic's matching system, which is often used by NCHS staff to merge outside sources of data with vital statistics such as birth or death certificates (NCHS, 1990). This system may be used for other types of merges as well. One benefit of the NCHS approach is that the analyst can readily identify duplicate records by checking record counts and sorting by the variables used for the merge process. These can be viewed by the analyst and rules can be devised to reduce the number of duplicates or to omit questionable records. In some instances, NCHS staff may perform the merge for a fee.

Other Confidentiality Concerns. One of the challenges of dealing with multiple types of data files is protecting their confidentiality, and the confidentiality of any files derived from them. The Federal government requires an assurance of compliance with appropriate methods for preserving confidentiality, and it is best to confer with data security personnel when dealing with this matter. Often, data security policies require identifying a data custodian who is responsible for the security of all confidential data. This custodian will develop data confidentiality training procedures, rules dictating who may have access to data and for how long access will be granted, and whether or not data may be shared by multiple parties. A data security log may also be developed to monitor the use of data for the research project. Rules for the use of data on mainframe and personal computers may also be developed, as may rules for protecting computer accounts. Data storage and disposal procedures may be required, and regulations for display of data on computer screens may also be made. The protection of physical facilities that house data will also be of concern, along with plans for dealing with potential breaches of security due to catastrophes such as fires or other problems.

Many researchers are used to dealing with a small number of data sources, and these policies and procedures can be adhered too easily in that context. However, when many more data sources are added to the mix, confidentiality requirements may differ by source and negotiations may be required to sort out the confidentiality procedures in a way that maximizes the security of data.

Coding Substance Abuse Problems and Reliability and Validity Issues

Given the variety of programs being implemented and evaluated by the nine CSAP Workplace Managed Care Initiative grantees, there may be a variety of methods used to identify persons with substance abuse problems. These include survey techniques, medical record reviews, and identification based upon diagnosis codes in claims databases. Fowles, et al., (1998) show that the sensitivity and specificity of the survey-based and claims-based techniques vary, using medical records review as a gold standard. Sensitivity refers to the ability to identify a disorder via claims or surveys, in clients who really have that disorder. Specificity refers to the ability to identify those who do not have the disorder in claims or surveys, from those who really do not have it.

In the Fowles, et al. study, the sensitivity of the claims-based technique was very low for substance abuse problems (0.20), while the sensitivity of the survey-based technique was higher (0.60). Specificity was high for both techniques (0.99). These specificity and sensitivity estimates may not be generalizeable because of the small sample size used in the Fowles, et al., study, but their results suggest that reliance on medical records or self-report may be advisable. If budget constraints do not allow a complete review of medical records, reviewing a sample of records for coding accuracy is still worthwhile. If even that is not affordable, focusing first on self-reported data and then on claims is probably advisable.

<u>Methods of Coding for Substance Abuse Problems</u>. There are several methods used across the country to identify persons with substance abuse problems in claims data. These include the Diagnostic and Statistical Manual of Mental Disorders, Third Edition (DSM-III), the revised version of DSM-III (DSM-IIIR), the Fourth Edition of DSM (DSM-IV), the International Classification of Diseases, Ninth Edition, Clinical Modification (ICD-9-CM), and the Tenth Edition of the ICD (ICD-10). Other methods may exist as well. Unfortunately, crosswalks do not exist to make one-to-one correspondences between the codes used in each of these methods.

Moreover, the reliability of coding within a method varies according to the type of substance abuse, and reliability varies between methods as well (Cotter et al, 1997). Over time, however, coding reliability and validity have improved (Wray, et al., 1997). Improvements will probably continue as long as valid codes are required by external entities for reimbursement or quality assurance purposes.

For purposes of the CSAP Workplace Managed Care Initiative cross-site evaluation, consistency of coding methods across grantees is desirable, but probably not enforceable. That is, grantees probably cannot make changes in coding systems used by managed care organizations or other providers. If multiple coding systems can be used within a given grantee site, it would be useful to perform sensitivity analyses to determine if results of the CBA or CEA differ by coding methods.

The Availability of Diagnosis Codes. One of the most problematic features of any claims-based coding mechanism is the unavailability primary or secondary diagnosis codes. Diagnosis codes for substance abuse or chemical dependency problems may not always be recorded, to avoid stigmatizing the patient. This has been noted for other mental health problems as well, such as depression (Croghan, et al., 1998). If survey data are not available to help find persons with substance abuse problems, researchers sometimes address this problem by searching for claims for medications that are only used for the problem of interest. Those who were prescribed the drug of interest (e.g., antabuse for alcohol problems) would then be added to the pool of candidates who were identified on the basis of diagnosis codes alone. While this method is imperfect, it is still preferable to relying solely on the diagnosis codes noted in the claims files.

Secondary diagnosis codes for substance abuse or chemical dependency problems are also missing in some cases. Wray, et al., (1997) showed that secondary codes were often missing in electronic claims files because of a lack of space in those files or on the paper forms used to create those electronic files. They showed that many more diagnosis codes could be found in medical records than on claims forms. This problem appears to be more severe in outpatient claims than in inpatient claims, and the problem has lessened over time. Nevertheless, researchers are cautioned that it may be difficult to adjust well for complicating or comorbid conditions when using claims data. Some improvement is likely when a longer time window is used to find diagnoses, however. Searching for secondary codes over a year-long interval, for example, is likely to yield a more complete set of diagnosis codes for chronically ill persons than a search over a shorter period would yield.

Estimating Financial Equivalents in the Absence of Claims Data

Because HMOs often pay providers on a capitated basis, they may not have complete claims data that would otherwise be used to estimate expenditures for medical care. In addition, many HMOs do not have accounting systems that are set up to estimate patient-level costs, and financial estimates of the cost of producing care are therefore non-existent in many HMO settings. Thus, estimates of the cost of treating HMO clients are often the result of tallying the services provided and multiplying the value of those services by a dollar conversion factor. The following text describes how this process may work, first for outpatient services and then for inpatient care.

<u>Outpatient Services.</u> Outpatient services can be organized by site, according to whether those services were provided in the doctor's office or in a facility setting, such as a hospital outpatient department. Services provided in a doctor's office can be disaggregated further into services provided by doctors themselves versus technical services provided by other personnel. Physicians' services can further be disaggregated into the work provided by the doctor, the associated administrative expense, and those expenditures related to medical malpractice insurance. The majority of the dollars spent in the treatment process for physicians' services relate to the doctor's time. The monetary value of this time can be estimated by denoting the doctors' services used to treat clients using CPT procedure codes, multiplying those services by their associated relative value units (RVUs), and then multiplying the total number of RVUs by a dollar-per-RVU conversion factor. CPT procedure codes are described in detail by the American

Medical Association (AMA, 1997).

Relative value units were developed for the Health Care Financing Administration (HCFA) by researchers at Harvard University (Hsiao, et al., 1988). These researchers estimated the amount of work involved in producing physicians' services so the Federal government could fairly reimburse doctors for those services under the Medicare program. An RVU valued at 1.0 is an indicator of the amount of physician work needed to produce a reference service. Most other services also have RVU values, with values less than 1.0 denoting less work and values exceeding 1.0 denoting more work than is required for the reference service. For a given patient, the sum of RVUs from all of the services applied may be taken to estimate the total amount of doctors' work needed to treat that patient in a time period of interest. The dollar-per-RVU conversion factor established by HCFA (which is currently $36.69) may then be applied to estimate the dollar value of physicians' services. The resulting dollar value may then be adjusted slightly for differences in the cost of producing physicians services in different geographic areas, using HCFA's geographic practice cost index (HCFA, 1989).

HCFA uses a similar process to estimate the value of maintaining physicians offices and to estimate the value of malpractice insurance for doctors' services. RVUs for these services exist and can be monetized in a similar manner. The dollars estimated for office services and for malpractice insurance can then be added to the imputed value of physicians' work, to complete the imputation of the value of physicians' services.

Physicians' services are not the only important contributor to the cost of outpatient care, however. Many outpatient services are provided by technical staff (e.g., laboratory or radiology staff). HCFA maintains separate fee schedules for these services, and these fee schedules can be used to impute the cost of technical services that are provided in office settings.

The cost of outpatient services that are provided outside of the doctor's office (e.g., in hospital outpatient departments) is more difficult to estimate. No fee schedules are used by HCFA to pay for these services. Many of these services are reimbursed at cost under the Medicare program. Thus, if the participating HMOs do not have data on the actual cost of these services, and neither does HCFA, it will be more difficult to impute the value of those services. If CPT codes can be assigned to these services, the RVU estimation process described above can be used to monetize them, but our conversations with staff at the Medicare Payment Assessment Commission lead us to believe that these costs will be understated by an unknown factor. If clients participating in the substance abuse program and those in any comparison groups are equally likely to use outpatient services in a facility setting, no bias should result when cost comparisons are made between these groups. If clients and comparison group members use facility-based outpatient services at different rates, some bias may result from the imputation process. Researchers are advised to compare utilization rates by site of service for clients and comparison group members, to assess the likelihood for bias in the imputed estimates of facility-based costs.

Once costs have been estimated for the doctor's work, for administrative services in doctors' offices, for malpractice insurance, for services provided by non-physician personnel, and for outpatient services provided in a facility setting, the total value of those outpatient services can be estimated by adding the value of each service type for each study participant.

Inpatient Services. The value of inpatient resources for subjects in the CBA or CEA cannot be estimated in the same fashion as the value of outpatient services, because RVUs do not exist for individual inpatient services and resources. However, two alternatives may exist to estimate the value of inpatient services provided to protocol clients. First, many HMOs do not capitate payments for inpatient services; these are often paid for directly on a negotiated fee-schedule basis. Thus, actual dollar expenditures may be available from at least some HMOs for inpatient services. If these expenditures are available, they may be used, either directly or indirectly, in the cost analyses performed for the protocol.

If actual expenditure data are not available, inpatient hospital expenditures can be imputed using the diagnosis-related group (DRG) payment amount for the DRG in which the hospitalization can be assigned, multiplied by the HCFA area-wage index to account for geographic differences in the cost of producing inpatient care. For those hospitalizations with a large (outlier) number of inpatient days, additional costs for the number of days beyond the outlier threshold established for the DRG can be estimated on a cost-per-day basis, using cost-per-day figures which may be obtained from HCFA. These figures would also be adjusted to control for geographic variations in the cost of producing inpatient care.

Expenditures for professional fees associated with inpatient visits must be added to the DRG-based imputed amount. Information about professional fees must be obtained from doctors' billing records or the health plan's fee schedule. The DRG payment rate does not reflect professional fees for inpatient services.

Once each hospitalization has been assigned a financial value, these values may be added for each patient to impute his/her inpatient treatment costs for the time period of interest. This process should be completed for each patient, regardless of whether actual financial data are available for inpatient services. To assess the validity of the imputation process, imputed values may then be compared to actual inpatient expenditures in those settings where the actual inpatient data are available. If the correlation between these two sets of inpatient figures is high, the imputed figures can be used in subsequent analyses. If the correlation is low, an additional conversion factor can be applied to the imputed figures to make them closer in value to the actual dollar figures. This conversion factor could, for example, be defined as the overall ratio of actual dollars to imputed dollars for inpatient services, using data from all of the HMOs, which have actual inpatient dollar figures.

Total Cost of Providing Treatment. After financial values for outpatient and inpatient services have been estimated, they can be added to impute the total dollar value of services provided to protocol clients in the study period.[16] (As noted earlier, cost figures from latter years must be discounted and cast in terms of constant, base-year dollars.) If actual dollar expenditure data for inpatient services are available for a large proportion of program participants, cost analyses can be conducted in two ways. One set of analyses can be conducted using imputed inpatient data for some HMOs and actual data for others, and another set of analyses can be conducted using imputed data for all data contributors. The results can be compared to illustrate the sensitivity of the CBA or CEA results to the cost-estimation process.

[16] An alternative to the RVU / DRG approach noted above is to estimate mean values for outpatient and inpatient services from other sources. Examples of these sources include Medicaid data files and files of inpatient, outpatient, and pharmaceutical services maintained in The MEDSTAT Group's MarketScan Family of Data Bases. These sources may be preferable if bias is problematic due to differential use of outpatient facility-based services by intervention clients and comparison group members. Another alternative, which applied only to inpatient care, was described by Barnett (1997). He showed how to generate person-level cost estimates from data on patient-level utilization and facility-level costs.

CHAPTER 5
STATISTICAL ISSUES

This chapter describes some important statistical issues that must be addressed to accurately estimate the impact of the program or intervention being evaluated in the CBA or CEA. These include:

- Choosing appropriate analytic techniques;
- Avoiding threats to validity;
- Using two-part econometric models to estimate program impacts;
- Performing other econometric analyses and adjustments; and
- Dealing with grouped data.

The first four of these issues are common to all CBAs and CEAs. The last, dealing with grouped data, is noted because grouped data are common for some of the CSAP Workplace Managed Care Initiative grantees.

Choosing Analytic Techniques

As with any type of research, the analytic techniques used for cost-benefit or cost-effectiveness analyses can heavily influence the findings. This problem was one of the motivational factors behind a series of articles on CEA in the Journal of the American Medical Association in 1996 (Russell, et al., 1996). Those articles described some of the techniques that were noted in Chapter 3. However, even if one adheres to all of the major requirements of a good CBA or CEA, analytic techniques can still influence findings and legitimate differences of opinions may arise about analytic techniques that some researchers prefer over others.

This section describes some analytic issues to keep in mind when conducting CBAs and CEAs. These refer to:

1. The success of the patient or client recruitment process;
2. The success of the randomization process;
3. Dealing with non-randomized designs; and

4. The appropriate use of an intent-to-treat design.

These issues are addressed in turn below.

The Patient Recruitment Process. Entry into many randomized research studies is voluntary, and often those clients who agree to participate are the only ones randomly assigned to the intervention and comparison groups. Clients who refuse to participate are often not subjects of the clinical outcome and cost analyses.

Refusing to participate in a new intervention tends to occur in a non-random fashion. Non-participants may differ from participants in terms of their demographics, locations, health status, aversion to risk, and willingness to submit to unknown treatment processes provided by unknown doctors. If these differences exist and are uncontrolled, estimates of the impact of program may lack external validity, having unknown or known but limited generalizability.

If data on the differences between participants and non-participants are available, those data can be used to adjust statistical comparisons to enhance the generalizability of the results. There are several ways to use such data. One appealing method common to the survey literature is to use data on determinants of participation in the study to estimate the predicted probability that each person participates. These predicted probabilities can be obtained from a logistic regression. The dependent variable for the regression would be a binary indicator of whether each person participated in the study. Independent variables would measure various predictors of participation status. Predicted values from the logistic regression can then be used to calculate the predicted probability of participating in the intervention, for each sample member. The inverse of these predicted probabilities would be used as weights in subsequent statistical analyses. This approach increases the weights of those who participate in the intervention, so they represent the non-participants as well (Kalton and Kasprzyk, 1986). Kalton and Kasprzyk describe the usefulness of this and other weighting techniques, and DuMouchel and Duncan (1983) and offer a statistical test for whether such techniques are necessary.

The Randomization Process. If successful, randomization to intervention and control groups will avoid many threats to internal validity. This occurs because randomization minimizes the likelihood that members of these two groups differ with regard to factors that influence treatment outcomes and costs. In addition, if successful, randomization greatly simplifies the analytic process, because descriptive statistics such as means and medians can be compared to make inferences about the impact of group membership on treatment outcomes and costs. In a situation where simple random sampling has been successfully applied, multivariate analytic techniques such as linear regression need not be applied to disaggregate the impact of group membership from other factors that influence outcomes or costs, because those other factors are equally likely to affect each group of interest. Thus, if randomization is successful, analyses may be limited to bivariate techniques.

Before CEAs or CBAs are conducted, one should note the success of the randomization process by comparing the characteristics of the intervention and control groups. This comparison might focus on demographics, baseline clinical measures, the existence of comorbid conditions at baseline, and the clients' use of services in a period before entry into the protocol. Statistical tests of proportions, means, and medians may be used for the assessment of the randomization process. If the results of these tests suggest that significant differences exist between the intervention and comparison groups, multivariate estimation techniques should be used for subsequent cost comparisons to adjust post-hoc for these differences. These multivariate techniques may include linear, logistic, or proportional hazard analyses, depending on the utilization or cost measure of interest. The major independent variable in these analyses would be a binary indicator to denote whether participants were in the intervention or comparison group. Additional variables would reflect those factors that are found in the randomization assessment process to further differentiate these groups.

Finally, a breakdown of the randomization process may not result in just measurable differences between clients in the intervention and comparison groups. Unfortunately, such a breakdown may also result in differences in unmeasurable factors that influence group membership and the outcome or cost measures of interest. In such a scenario, selection bias may result, threatening the internal validity of the study. Heckman (1976) and others (e.g., Olsen, 1980), showed how to overcome this problem. More information about selection bias adjustments is provided below.

Using An Intent-to-Treat Design. For many impact analyses an intent-to-treat design is used. The major feature of an intent-to-treat design is to conduct statistical analyses between groups whose membership is defined on the basis of their status when the initial randomization or baseline event occurred (Gibaldi and Sullivan (1997); Sclar, et al., (1998)). In this scenario, clients would be defined as members of intervention or comparison groups regardless of what happened after that baseline event. For example, clients who crossover from the intervention to the comparison group would always be counted in their first group, even though they were not always subject to that group's defining features (i.e., they did not always receive the treatment originally specified). If crossovers occur frequently enough, the result of using an intent-to-treat design in this manner will be an underestimate of the true impact of the intervention. We recommend that researchers follow participants to determine whether and when crossover occurs. If more than a few (e.g., more than five percent) of the participants change groups, one should account for this in the statistical analyses.

If the number of crossover participants is small, there may be little worry in applying the intent-to-treat logic as specified above, although the result will be a conservative estimate of the impact of the intervention. If the number of crossover participants is large, a separate group of crossovers should be identified and used as a third group of interest in the analysis.[17] Depending on whether clients were randomly assigned to treatment or control groups to begin with, bivariate or multivariate techniques can then be used to estimate the impact of the intervention and crossover status.

Including a separate group of crossovers in the statistical analyses will lead to more homogenous patient groupings. This in turn will lead to less bias in the statistical analyses of differences between those in the intervention and comparison groups. If crossover status is due to characteristics that must be associated with the intervention (e.g., due to compliance issues associated with drug side effects or cost), the report of the analysis should state this to provide a complete picture of the impact of that therapy. If crossover status is due to factors not related to the intervention, it would not be fair to allow these factors to influence the estimated impact of the intervention on clinical outcomes or cost. Clients who crossover from one group to another should therefore be queried to find out why crossover occurred.

---

[17] An alternative might be to compare outcomes according to the proportion of time spent in the intervention group versus the comparison group (e.g., less than 25% of the time, from 26-50% of the time, etc.).

Avoiding Threats to Validity

Using Cook and Campbell's (1979) text as a guide, Hargreaves, et al. (1998) describe many threats to the validity of mental health research studies. These include threats to internal validity such as:

1. Regression to the mean: the tendency of extreme scores or values of cost or effectiveness measures to regress (return) to less extreme scores upon subsequent measurement;
2. Testing and instrumentation effects: the impact of a pretest process to influence posttest values, or the impact of changes in measurement strategies on observed outcomes;
3. History and maturation: naturally-occurring changes shared by subjects in pretest and posttest studies;
4. Loss of subjects from follow-up;
5. Loss of subjects from assigned service (i.e., the crossover problem); and
6. Selection bias (i.e., non-random differences between treatment and comparison group members).

These internal validity threats can often be avoided with a high quality design (e.g., proper randomization to treatment and comparison groups, with a pre-post feature to the study). The existence of any of these problems can lead to confounding the effects of the intervention with other factors.

<u>Selection Bias.</u> Selection bias is, arguably, the most important threat to the internal validity of a study. This bias often occurs in the absence of a successfully randomized method for allocating subjects to treatment and control groups. Randomization, if successful, would work to minimize the observable and unobservable differences between the treatment and control groups. In this scenario, observed differences in medical expenditures, absenteeism, or other outcomes between program participants and control group members could then be ascribed to the program of interest.

In the absence of a randomized design (which is clearly not possible for most corporate programs or their evaluations), other methods must be used to minimize the impact of observable and unobservable differences between program participants and non-participants, when estimating the impact of program participation on the outcomes of interest. The failure to apply such methods could result in selection bias, which means that one cannot be sure whether differences in the outcomes of interest are due to program participation or to those other observable and unobservable factors.

In the absence of a randomized design, researchers often use multivariate regression analyses to remove the impact of observable differences in age, gender, race, and other factors, to provide a more accurate estimate of the impact of the program. However, by themselves, multivariate analyses may not eliminate selection bias. Thus, without further modification, the resulting estimates of the impact of program participation may still be biased to an unknown degree.

To minimize selection bias, one must consider the possibility that those who participated in the program of interest were different in *unobserved* ways from those who did not use program services. If these unobserved characteristics are also correlated with the major outcome variables, biased estimates of program impacts may be obtained.

There are at least three well-known methods that can be used to avoid selection bias in a non-randomized study. The easiest one to implement is known as the propensity score approach. A second method is referred to as the Heckman-type adjustment; the third technique is known as an instrumental variables approach. These are described below. One of these methods should also be used in randomized studies, if the randomization did not work to minimize differences between treatment and control group members because of differences in compliance or drop out rates or other factors. This fact is often not considered in poorly executed randomized studies.

Propensity Scores. The propensity score technique has been described in detail by Rosenbaum and Rubin (1984), Robins, Mark, and Newey (1992), and Drake and Fisher (1995). Basically, the method of propensity scores involves estimating the probability of participation in a substance abuse prevention or early intervention program, based upon observable characteristics like demographics, business unit, management/non-management status, etc., or based on survey information about motivation to take care of oneself and other measures. The predicted probabilities are estimated for the entire sample, both participants and non-participants, using probit or logistic regression. The participant and non-participant subjects are then sorted by their predicted probabilities. Next, the sorted probabilities are used to match groups of participants and non-participants. For example, participants with predicted probabilities in the range of 0.0 – 0.20 would be compared to non-participants within the same range, etc. Analyses would then be carried out within each of these groups using the matched observations. This method is an attempt to compare "apples to apples," since all of the comparisons are based on participants and non-participants with similar predicted probabilities of participating in the intervention.

Connors, et al. (1996), illustrates the propensity scoring approach in a study of critically ill heart clients. While their study does not address substance abuse problems, it is a good example of how the propensity scoring approach can be used in studies requiring input from physicians and other clinical staff who are most familiar with randomized studies and less familiar with non-randomized alternatives.

The Heckman Approach. Heckman (1976, 1979) proposed a two-step approach to test for sample selection bias and to correct it if present. The first step involves the estimation of a probit regression model of program participation. The estimated probabilities of participation that can be obtained from the probit regression model would then be used to predict the risk (or "hazard") of not participating, given that the individual had the option to participate in the program. In the subsequent outcome equations to be estimated in the second step, the hazard rate would be entered as an additional explanatory variable. Heckman (1976) showed that the hazard rate represents the effects of unobserved variables on the outcome variables of ultimate interest. If the regression model used to create the hazard rate accounts for all of the major determinants of program participation, including the hazard rate variable in the outcome analyses will remove the selection bias.

Instrumental Variables Analysis. A third technique for dealing with selection bias is known as instrumental variables analysis. This approach is similar to Heckman's, in that a first stage regression may be used to create a new variable, known as the instrument, which is highly correlated with treatment status but not correlated with the error term in the outcome regression of ultimate interest. The regression used to produce the instrument may be the same probit or logistic regression used in the Heckman or propensity score techniques. The instrument may then be created as the predicted probability of participating in the intervention program, based on demographics, health status, or survey-based measures such as those noted above.

Standard econometric software such as LIMDEP can be used to estimate the regressions of ultimate interest with the instrument, to find consistent (but not necessarily unbiased) estimates of program impact (Greene, 1995). Consistent estimates are those that are unbiased only in very large samples, so small-sample studies using the instrumental variables approach may not yield unbiased estimates. Pindyck and Rubinfeld (1991) provide a basic description of the instrumental variables approach. The approach is also mentioned in Hargreaves, et al. (1998), but their characterization of the second-stage estimation process using the instrumental variable is incorrect. In contrast to their suggestion, it is not appropriate to merely substitute the instrument for the binary measure of participation status in the regression of ultimate interest. The appropriate methods for using the instrument may be found in textbooks by Kmenta (1986) and Greene (1990).

Choosing Among the Approaches. None of the approaches for selection bias adjustment has been identified as universally superior to the others, and the choice of method may be dictated by the analyst's preference, by sample size, or by econometric theory. When sample size is sufficient and theory suggests that any approach will do, one may find the propensity score approach to be preferable. The propensity score approach is the easiest to use, understand, and explain to a lay audience. Moreover, the propensity score approach is the only one that may satisfy theoretical concerns when the major outcome variable of interest is binary, or when many subjects incur no expenditures in a study period. (See the text below on two-part econometric models for a discussion of this latter phenomenon).

Another reason to choose the propensity score approach is that, in practice, the major alternatives may fail. For example, the Heckman approach may not lead to stable estimates of the impact of an intervention unless at least one factor that influences the decision to participate in the intervention can be identified that does not also influence the outcomes of ultimate interest. An example might include a variable describing the location of the worksite, if that location influences access to corporate-based program facilities or services. Another example might include variables describing employees' motivation to take care of themselves or their aversion to the risks of unknown treatments. If variables such as these are not included in the regression model used to generate the predicted probabilities of participation in the intervention, those predicted probabilities may not adjust well for unobserved factors that influence participation status and the outcomes of ultimate interest. Moreover, the predicted probabilities may be so highly correlated with other variables in the regressions of ultimate interest that reliable estimates of the influence of those variables may not be obtained. In practice, this is problematic less often with the propensity score approach.

Criticisms of Selection Bias Models. Some researchers have questioned the ability of post-hoc adjustments to control for selection bias. In an influential article, LaLonde (1986) argued that such adjustments to evaluate job-training programs have failed to reach the same policy conclusions as social experiments with randomized designs. Heckman and Smith (1995), however, have argued convincingly that the reason for LaLonde's conclusion was the lack of a good model for predicting whether sample members participated in the intervention. When Heckman and Smith repeated LaLonde's analysis using an alternative model with better data, they obtained the same results as the fully randomized social experimental design--the social science analogue to randomized clinical trials.

Conclusion. The message in Heckman and Smith (1995) is clear: post-hoc adjustments can work, and they work best when good predictors of program participation are available and when sample sizes are large. Efforts to control for unmeasurable differences between participants and comparison group members should focus first on the propensity score technique. The propensity score technique is easier to implement than the Heckman-type adjustment if most of the variables that predict program participation are also good predictors of the outcomes of ultimate interest. The propensity score approach is also easier to explain to those not familiar with it. The Heckman approach and the instrumental variables approach are more difficult to follow, and the instrumental variables approach is less useful in small-sample studies. Small sample size may also be a problem for the propensity score approach, because that approach requires dividing the sample into several subsamples for separate analyses. All three models may have limited value if the major predictors of program participation cannot be measured. Researchers should therefore confer with program staff and stakeholders about this issue before estimating program effects with these techniques.

<u>Threats to External and Statistical Conclusion Validity</u>. Hargreaves, et al. (1998) list the following threats to the generalizability of cost-outcome studies in mental health:

1. Systematic exclusion: the failure to represent all eligible members of the population in the CBA or CEA. Usually this can be avoided by successful random sampling from the population of interest, before allocating subjects to treatment and control group status.
2. Having a non-random pattern of consent refusals. Sometimes subjects refuse to participate in the study and the pattern of refusals is not random.
3. Applying treatments in atypical settings or under atypical conditions; and

4. Failure to "blind" providers or data analysts to the method of treatment.

These threats to the generalizeability of the study can often be avoided by: a) careful screening of the target population, b) sampling randomly from that population, c) using more typical settings or treatment conditions, and d) carefully blinding all clients, providers, and data analysts to the interventions provided.

Finally, Hargreaves, et al., (1998) note low statistical power as a threat to the conclusions drawn from statistical analyses in mental health studies. Low power refers to the low likelihood of correctly detecting a difference in outcomes or costs between those in the intervention and comparison groups.

Low statistical power can result from several factors, but most often results from having too few subjects in the treatment and comparison groups. The smaller the number of subjects in each group, the larger the within-group variances tend to be. The larger variances contribute to larger standard errors estimated in statistical tests, and larger standard errors may result in a lower likelihood of finding statistically significant differences between groups. Low power can be avoided by including large samples in the analysis. It can also be avoided by increasing the magnitude of the difference in cost or outcome that is deemed important, or by increasing the Type 1 (alpha) error[18] rate in the analysis. Low power also can be overcome in many instances by changing the testing scenario to include "randomization" tests.

Randomization tests were developed for use with small samples and are useful for making statistically-based comparisons between those samples when sample sizes cannot be increased or when one is not comfortable artificially increasing Type 1 error or effect sizes simply to increase power. To complete a randomization test, one first computes the observed test statistic of interest (e.g., a difference in mean values). Then one randomly re-assigns clients to treatment and comparison groups and computes a new test statistic. This process is repeated many times, such as for all permutations of treatment-control group membership. Then one records the probability that the observed test statistic was exceeded in these permutations. This probability is referred to as the p-value for the randomization test. Randomization tests and statistical power are described

[18] Type I error is defined as concluding that costs or outcomes are different when in fact they are not.

in more detail in O'Brien, et al., (1994), O'Brien and Drummond (1994), and Eddington (1995).

Estimating Program Impacts With Two-Part Models

In many analyses of health care expenditures, absenteeism, disability, and other health and productivity data, analysts will observe many clients or clients with zero values in a time period of interest. For example, it is not unusual to observe 15-30% of the covered lives in a health plan who use no medical care services in any given year. Similarly, many employees take no sick leave, and many more do not use short-term or long-term disability programs in any given year. Thus, these data sets may include a large number of zero values, followed by other positive values for those who did use medical care or other services. In addition, among the group of persons who incurred some service use, there may be a small proportion of extremely large (i.e., outlier) values of medical expenditures or absent days for those who are extremely ill. These two phenomena cause estimation problems that are often dealt with by using two-part regression models.[19]

Two-part estimation techniques are useful for at least three reasons (Duan, et al, 1983; Manning, 1998). First, the variables that influence whether or not any services are used may be different from the variables that determine how many services are used once the decision to seek care has been made. Second, even if the same variables influence the decision to seek care and the amount of care used, the importance of each variable may differ in each analysis. Third, in a scenario in which outlier values exist, the error term of the regression analysis used to estimate the impact of a substance abuse program may not be normally distributed. This may result in estimates of the impact of the program that are neither precise nor robust (Manning, 1998). Analysts often address this problem by re-defining the dependent variable of interest by taking its natural log value. Taking the log value often normalizes the regression error terms, but log values do not exist for those without any service use. Hence, as described below, a two-part model can be used to analyze separately those who did incur some service use.

Two-part models are estimated as follows. First, a probit or logistic regression model is estimated to address the probability that each person uses any services of interest. For example,

[19] Remember that regression models may also be required to estimate substance abuse program impact when randomization into the program or its comparison groups did not occur, or when randomization failed to equate the characteristics of each group. In this section we expand the discussion of regression models to account for other characteristics of the observed data.

the regression may address whether these probabilities vary, on average, for those who participated in a substance abuse prevention or early intervention program and those who did not, controlling for confounding variables such as demographics, location, etc. The predicted values obtained from the first-stage probit or logistic regression can be manipulated to predict the probability of incurring any service use for each observation in the data set.

The second part of the two-part model usually involves an ordinary least squares regression of the magnitude of service use for the subsample of observations that incurred any use. This part typically uses the log-transformed dependent variable, although other transformations can be used as well (Manning, 1998). The purpose of this piece of the two-part model may be to investigate whether the magnitude of service use differs, on average, for substance abuse program participants versus non-participants, controlling for confounding factors. If transformed properly, the predicted values obtained from this part of the model reflect the estimated service use for average members of the participant and comparison groups, controlling for other factors. The transformation process is described in detail below; it is needed to transform predictions of log values of the outcome variable into predictions of actual values.

The results obtained from each piece of the two-part model can then be combined to estimate the overall impact of program participation on service use. This may be done by multiplying the predicted probability of incurring any services (obtained from the first part of the model) by the estimated magnitude of service use (obtained from the second part of the model). This multiplication is usually performed for the subset of persons who incurred any service use.

<u>The Transformation Process: Using the Smearing Estimate</u>. Because the second part of the two-part model is typically estimated in log terms, analysts must transform the predicted outcome values into non-log terms in order to facilitate interpretation of the results. As Manning (1998) notes, if the analysis of interest pertains to monetary expenditures, this transformation may be necessary because “Congress does not appropriate log dollars,” and “First Bank will not cash a check for log dollars” (page 285). In other words, no one cares about log values of the outcome variables; people care instead about real values.

There are many options for transforming log values back into real values; these are described in Duan, et al. (1983). The option most often used involves the “smearing estimate,” which Duan developed. The smearing estimate corrects for the underestimate of non-logged values that would arise by simply using the exponentiated value of the logged predicted values to estimate the effect of the substance abuse program. Without applying a smearing estimate, underestimates of program effect would arise because using the logged value of the dependent variable in the two-part model changes the variance of that dependent variable as well as its magnitude. The smearing estimate accounts for this change in variance.

The smearing estimate is usually applied as a constant for a subgroup of interest. The constant is multiplied by the exponentiated, predicted values obtained in the second part of the two-part regression model. The value of the constant is obtained by averaging the exponentiated residuals from the regression used in the second part of the two-part model. To avoid problems that arise when the exponentiated residuals are correlated with the independent variables used in the regression, Mullahy (1998) notes that separate smearing estimates can be estimated for subgroups of interest. For example, separate estimates can be generated for substance abuse program participants and non-participants. In practice, the value of the smearing estimate is usually larger than, but close to, 1.0.

<u>Applying the Two-Part Model</u>. In practice, the following steps are used to estimate a two-part model.

1. Estimate a probit or logistic regression to address the probability of incurring any service use or expenditures. The dependent variable for this analysis should be coded as 1 for those who incurred any services or expenditures. For others the dependent variable should be coded as zero. Independent variables should include measures of participation in the substance abuse program and confounding factors such as demographics, etc.
2. Generate predicted values from the probit or logistic regression and transform them into probabilities. Generate these predicted probabilities twice, once assuming that all observations in the data set reflect participants in the substance abuse program, and once assuming that all observations reflect non-participants. This may be accomplished by setting the participation indicator equal to 1 in a first analysis and by setting that indicator equal to 0 in a second analysis. Set all of the values of the other independent variables equal to their respective means. Then multiply the regression coefficients by the associated mean values and solve for the predicted values twice, once when the program participation indicator is set equal to 1.0, and once when the participation indicator is set equal to 0.

3. Estimate the ordinary least squares (OLS) regression of the magnitude of service use or expenditures. This regression is estimated only for those with non-zero values for the dependent variable expenditure or utilization measure.
4. Generate predicted values from the OLS regression twice, once for program participants and once for non-participants, using the methods noted in Step 2 above.
5. Generate separate smearing estimates for program participants and non-participants, by averaging their respective exponentiated residual values from the OLS regression.
6. Exponentiate the predicted values of the OLS regression obtained in Step 4, and multiply the results for participants and non-participants by their respective smearing estimates. This will produce estimates of expected outcomes for participants and non-participants that are cast in actual (i.e., non-log) terms.
7. Multiply the predicted probabilities obtained in Step 2 and the predicted outcome values obtained in Step 6. The result will be person-level estimates of the outcome variables of interest that account for the fact that many observations have zero utilization or expenditure levels while adjusting for differences in covariate factors.

8. Estimate mean values from the data obtained in Step 7. Do this twice, once assuming that all persons are substance abuse program participants, and once assuming that all persons are non-participants.
9. Perform a t-test on the difference in the two means obtained in Step 8. The standard error for this t-test can be obtained by using the Delta Method noted by Manning (1998), or by bootstrapping techniques. The result obtained from this test will show whether substance abuse program participants and nonparticipants varied significantly in terms of outcome measures which are often valued at zero and which may be non-normally distributed.

<u>Applying the Two-Part Model With Adjustments for Selection Bias</u>. In a context in which selection bias is likely to be problematic, the application of the two-part model is complicated substantially. In this case, we recommend using the propensity score technique to adjust for the selection bias. If the propensity score technique is used, the two-part model must be estimated several times, for each subgroup of participants and non-participants who have similar predicted probabilities of participating in the substance abuse program. It would not be appropriate to apply the Heckman method or the instrumental variables method for selection bias adjustment in the context of a two-part model. This is because the underlying theory behind the Heckman model and the instrumental variables model does not apply when a dependent variable of interest is binary, as it is in the first part of the two-part model. For example, Heckman's inverse Mills ratio is meaningless when trying to adjust for selection bias in an analysis of whether substance abuse program participants and non-participants differ in the likelihood of incurring any health care expenditures in a study period. The instrumental variables technique is also meaningless in that context, leaving the propensity score technique as the method of choice among these three.

<u>Other Models</u>. A variety of other models can be used to estimate the impact of a substance abuse program in non-randomized settings or when randomization fails to equate the participant and non-participant groups. These models are mentioned briefly by Mullahy (1998). One model may be of particular interest, because it avoids the use of the smearing estimate. This model is labeled by Mullahy as a "modified two-part model." This model uses a non-linear least squares approach for the regression estimated on those who have non-zero values of service use or expenditures. The popularity of this model is expected to increase, in part because models that use the smearing estimates sometimes perform poorly. Poor performance of models that use

smearing estimates may be due to unknown or uncorrected correlations between the values of the independent variables in the regression model and the exponentiated regression residuals. Details about this problem may be found in Mullahy's article.

Other Econometric Analyses and Adjustments

To maximize the likelihood of obtaining unbiased, consistent, and efficient estimates of the impacts of treatment components and other factors on costs and effectiveness, researchers may adjust analyses for other problems. These problems include measurement errors, the interdependence between cost and effectiveness measures, statistical outliers, collinear predictors, and censored cost or effectiveness data. Each of these problems is described below.

Mullahy and Manning (1995) note that biased and inconsistent estimates of treatment impacts may result when errors exist in the measurement of cost or effectiveness variables. Biased and inconsistent estimates may also be found when errors occur measuring the interventions being studied or when measuring other factors that are correlated with those interventions. In many studies the extent of measurement error is unknown. To detect and help avoid measurement error, reliability and validity analyses should be performed for the cost and effectiveness measures of interest. Data on service use should be obtained and triangulated from multiple sources, in attempts to reduce missing data and measurement errors.

Mullahy and Manning (1995) and Siegel, et al., (in press) note that, in many situations, estimates of the incremental costs and effectiveness of treatment approaches are not independent. Because of this, it is possible to find that incremental cost-effectiveness ratios are not significantly different from zero, even when the numerators and denominators of those ratios are both found to be statistically significant. Mullahy and Manning note that as the correlation between the numerator and the denominator of the cost-effectiveness ratio becomes more negative, the confidence interval for the ratio increases.

Siegel, et al., argue that the interdependence between costs and effectiveness should be considered explicitly in the cost-effectiveness analyses. This can be done by constructing patient-level differences between cost and effectiveness measures, as opposed to comparing mean values from large groups. Siegel, et al., provide formulas for several cost-effectiveness measures and for the associated confidence intervals. Interestingly, they did not provide such formulas for incremental cost-effectiveness measures. Alternatively, multivariate analyses of costs and effectiveness may be conducted using "seemingly unrelated regression" techniques (Pindyck and Rubinfeld, 1991). These techniques would account for the interdependence between costs and effectiveness measures by adjusting for the non-zero correlations between the error terms from the cost and effectiveness regressions.

In cost-effectiveness analyses it is not uncommon to find a wide range of cost values. In some analyses, the distribution of costs may be highly skewed. When skewness is due to questionable cost data for a subset of observations, the estimates of treatment impacts may be biased. To adjust for this, extremely high or low cost values are identified and then sorted by treatment type, source of data, or other factors, to learn what the reasons may be for outlier status. If these analyses suggest that some of the influential observations are of questionable validity, those observations should be excluded from the analyses.

Next, it is not uncommon in CEAs or CBAs to find that some factors which influence the cost of the intervention or its outcomes are highly correlated. Failure to adjust for these correlations will lead to unstable statistical test results. Belsley, Kuh, and Welsch (1980) describe methods that can be used to identify linear combinations of predictor variables that lead to collinearity problems. Generally, researchers simply drop one or more of the highly correlated variables to fix this problem. If all of the highly correlated variables are required to test a program theory, however, one may combine them statistically using principal components analysis. Subsequent analyses can then be used to estimate the impact of each of the correlated variables. This process is described briefly by Kennedy (1992).

Finally, much of the data to be used in a CBA or CEA may be censored (i.e., observed for only partially as long as desired). For example, data before and after an arbitrarily designated start or end time for the study may not be available. Alternatively, employees may switch or lose jobs, changing or losing health insurance coverage in the process. Data on subsequent health care use therefore may not be available. Some analysts control for the amount of time that data are observed by annualizing the cost or effectiveness measures (e.g., Goetzel, et al., 1998). Others use methods based on survival analysis techniques that have been specially designed for use with censored data. An explanation of these methods can be found in Gardiner, et al. (1995), who constructed cost-effectiveness ratios when the effectiveness measures were censored and the cost data were not.

Dealing With Grouped and Categorical Data

Some of the grantees under the CSAP Workplace Managed Care Initiative will not have access to individual level cost or effectiveness data. Rather, most variables will be measured as group means or proportions, with the groups corresponding to business units, plants, or other employee locations. Other variables may be categorized, even if they exist at the individual level. For example, medical expenditures may be known only in terms of ranges of dollars spent, if survey data are required to measure those expenditures.

When data are grouped or categorized, a major question of interest is whether one can still estimate the cost or effectiveness of a substance abuse program efficiently and without bias. Kmenta (1986) shows that the degree of bias or efficiency depends on the variance of the true individual-level data within each group, or on the distribution of underlying values of those variables that are measured categorically.

<u>Grouped Data</u>. Suppose one wants to estimate the impact of a substance abuse program on person-level behavioral health care expenditures, but data are only available at the level of the plant or other employee location. No individual-level expenditure data are available, and participation in the program is measured as the proportion of eligible or referred employees within a group who actually participated in the program. Other covariates such as demographics and other factors are available only as mean values (e.g., mean age) or proportions (e.g., percent male).

In this scenario, the data allow us to directly estimate relationships between participation in the program and average behavioral health care expenditures, controlling for mean values of other variables. In contrast, one may wish to know what the impact of the program would be if measured with individual-level data. Under these conditions, Kmenta shows that if ordinary least squares regression techniques are used with grouped data to estimate the program's impact on expenditures, the impact estimate will be unbiased. However, the standard error of the program impact estimates will be biased and inefficient (i.e., other standard error estimators would be more accurate and have smaller variance). Without further manipulation, conclusions about the statistical significance of the program impact estimate may therefore be incorrect. Fortunately, Kmenta goes on to show how to correct the least squares regression estimates to avoid the biased and inefficient standard error estimates (see pages 366-373). These corrections require knowledge of the number of individuals in each group, along with the group means or proportions.

As a final note, Kmenta shows that the r-squared measure for the least squares regression with grouped data is also biased, towards 1.0. This means that the estimate of how much variability in expenditures can be explained with the grouped data is higher than it would be if the analysis was based on individual level data. However, the degree of the bias is usually small if there are large numbers of individuals in most groups.

Categorical Data. Suppose that the data are categorized in a way that provides the analyst only with information about ranges of behavioral health care expenditures. For example, data may show whether individuals had expenditures equal to $0, $1 - $500, $501 - $1000, and other increments. When the dependent variable of interest is categorized, some analysts may choose the midpoint of the interval as the dependent variable value for the regression used to estimate program impact. Under these circumstances, Kmenta shows that estimates of the impact of program participation on medical expenditures will be biased. The bias will be small, however, if the values of the independent variables are not highly correlated with the errors associated with using the midpoint of the expenditure category as the value of the dependent variable in the regression.[20]

Similarly, if the values of the expenditure variable are known exactly but the independent variables are categorized, some bias will result if analysts use the midpoint of the category ranges for the values of the independent variables in the regression analysis. The bias is small only if the true, individual-level values of the independent variables are uniformly distributed over their range, from lowest to highest.[21]

Conclusion. Based on the information in Kmenta's text (1986), grouped data appear to be less problematic to deal with than categorized data. Program impact estimates based on group data are likely to be unbiased. Any bias in the standard errors of the program impact estimates may be handled fairly easily, using formulas that Kmenta provides. In contrast, there are no easy

[20] An alternative estimation process would use multinomial logistic regression to estimate the odds of having expenditures in any given category (compared to a reference category), based upon program participation and other factors. No bias would result from this process if other regression assumptions are met.

[21] As an alternative, analysts may create binary variables to denote which category each independent variable observation falls into. For example, binary indicators for age group or income may be created. Analyses based on these binary indicators will not be biased, but the results will not provide an overall estimate of a one-unit change in age or income on the dependent variable of interest.

corrections for estimates based on categorical data. The project report should either note that some bias may exist in the program impact estimates, or other analyses (see footnotes) should be conducted that do not allow estimation of the impact of one-unit changes in independent variables on changes in dependent variables.

CHAPTER 6

CONCLUDING REMARKS

The length and breadth of this report suggest that well conducted cost-benefit and cost-effectiveness analyses are difficult to do. A wide variety of conceptual, statistical, and logistical issues must be considered. These issues may require coordination across several disciplines. Many types of staff must confer with the analyst. These may include economists, epidemiologists, physicians, clinical personnel, social scientists, human resources personnel, other workplace staff, computer programmers, and data set developers. The wider the variety of outcome measures to be addressed, the more data sets to be analyzed, the wider the variety of issues to be considered, and the greater the number of interactions and consultations there will be. Thus, CBAs and CEA are truly multidisciplinary efforts that require a high degree of organization, analytical sophistication, and skill.

The purpose of this guide has been to describe how to conduct good CEAs and CBA of substance abuse prevention and early intervention programs. While many conceptual issues were described, we also tried to illustrate questions, problems, and issues that must be addressed in a real-life, applied setting outside the research laboratory. Such settings are typically non-randomized ones, lacking the same degree of control in experimental studies. Validity threats such as selection bias and low statistical power are often problematic in these settings, and we offered some suggestions for dealing with these problems. Despite the lack of rigor that may be used in well-conducted randomized trials, sound, rigorous, quasi-experimental methods are still available for the analyst to use. Hopefully this research guide has described these techniques in a manner that will facilitate high-quality evaluations of the impact of substance abuse prevention and early intervention programs.

Finally, even a lengthy research guide cannot address every question and issue that is pertinent to a good CBA or CEA. Many important questions remain to be addressed by the CSAP Workplace Managed Care Initiative grantees and subcommittees. Other questions and issues can be investigated in more detail by consulting the appendices and references listed at the end of this guide.

GLOSSARY

The purpose of this glossary is twofold. First we define terms which have appeared earlier in the text, for quick reference. Second, we offer some useful definitions of terms such as absence rate, turnover rate, and disability program expenditures, to help guide data collection strategies across the nine CSAP Workplace Managed Care Initiative grantee sites. The final data collection strategy will be decided by the sites after more detailed review and discussion by the Steering Committee.

Absence rate - unscheduled absence as a percent of scheduled workdays. This rate does not include: long-term absences after the first four days; vacations, holidays, or other scheduled leave; or absence of less than a full day. Monthly absence rates are calculated by employers and collected as part of the Bureau of National Affair's (BNAs) Quarterly Employment Survey. BNA then calculates the monthly median rates and the average of monthly median rates for the year. In the BNA report, rates are calculated as [Number of worker-days lost through unscheduled absence during month / (Average number of employees) * (No. of scheduled workdays)] * 100. (Source: Bureau of National Affair's definition, 1995).

Absenteeism - time spent away from work. May be classified separately due to employee sick leave, personal days, mental health days, jury duty, vacation, holidays, family illness or bereavement, family medical leave act, workers compensation program days, short-term disability, or long-term disability. Substance abuse program theory should be used to determine which of these types of absenteeism are appropriate for analysis of the impact of a substance abuse prevention or early intervention program.

Array - a group of items saved in a SAS or other analysis file. For example, groups of diagnoses or procedure codes may be kept for subsequent analysis. The SAS Language and Procedures Guide defines an array as: (1) a method of grouping variables of the same type for processing under a single name and (2) a method of defining an area of memory as a unit of information.

Benefit - the positive or negative monetary consequences of participating in a substance abuse program. Usually differentiated from "effectiveness," which pertains to the non-monetary consequences of participation.

Benefit-cost ratio *(also known as return on investment ratio)* - the inflation-adjusted, discounted benefits of a program or intervention divided by the inflation-adjusted, discounted costs of providing and consuming the program. Values above 1.0 generally denote economically attractive programs that provide more than one dollar in benefits for each dollar spent on the program.

Break-even analysis - an analysis designed to determine the number of dollars of costs or the value of benefits that would have to be assigned to make two alternative programs equally attractive (Warner and Luce, 1982).

Bootstrapping - a process of repeated subsampling, with replacement, from a larger sample, followed by analysis of each repeated subsample. Analyses with the subsample are used to estimate variances or standard errors of variables of interest (Vogt, 1993).

Censored data - data about an event or phenomenon of interest that are unavailable for periods of time or groups of people. For example, medical expenditures may be unavailable for persons who switch health plans, or for time periods before or after employment or some other event of interest.

Charges - the prices of health care services or other goods and services imposed by suppliers of those services. Charges typically exceed the costs of producing those services and sometimes reflect additional moneys to recoup bad debt or to offset losses or lower payments from some customers.

Coinsurance – the portion of the covered health care cost for which the person insured has the responsibility to pay, usually based on a fixed percentage; a percentage of cost to be paid by the insured, having already paid the maximum deductible for the year. (Source: Rognehaugh R, The Managed Care Dictionary)

Copayment - the portion of the covered health care cost for which the person insured has the responsibility to pay, usually as a fixed fee for a specific service type (e.g., $10 per doctor visit).

Cost - the monetary value of resources used to produce or consume a program or intervention.

Cost-benefit analysis (CBA) - A systematic method for valuing over time the monetary costs and consequences of producing and consuming substance abuse program services. Results from a CBA are often provided in terms of a net present value figure, which shows the difference in inflation-adjusted, discounted costs and benefits of the program in today's dollars or in the dollars of a base year of interest. Results may also be shown in terms of an internal rate of return or a benefit cost ratio (see definitions for these terms).

Cost-effectiveness analysis (CEA) - A systematic method for valuing over time the monetary costs and non-monetary consequences of producing and consuming substance abuse program services. Results from a CEA are often shown in terms of total costs and total levels of effectiveness (e.g., total quality adjusted life years saved or total numbers of substance abuse cases avoided), or in terms of cost per unit of effectiveness.

CPT-4 – Current Procedural Terminology, 4th Edition. Unique sets of 5-digit codes developed by the American Medical Association that apply to the medical service or procedure performed by providers and used as a standard in the industry; used for billing. (Source: Rognehaugh R, The Managed Care Dictionary)

Deductible – the minimum threshold payment which must be made by the enrollee each year before the plan begins to make payments on a shared or total basis. (Source: Rognehaugh R, The Managed Care Dictionary)

Discount rate - the rate at which future dollars or future units of effectiveness are devalued, relative to current dollars or units of effectiveness.

Discounting - the process of devaluing future dollars or units of effectiveness to reflect preferences for dollars or goods or services now, versus in the future.

Effectiveness - the non-monetary consequences of participating in a program or intervention.

Gross Domestic Product Implicit Price Deflator - an index developed by the Bureau of Economic Affairs which may be used to adjust for inflation. Details can be found at the BEA web site: http:\\bea.doc.gov/bea/dn/dpga.pdf)

Imputation - the process of replacing missing data. May be done logically (based on other existing data) or with statistical techniques based on variables that are correlated with the variable with the missing data. See Little and Rubin (1987) for examples and details.

Inherently valued outcome - defined by Mohr (1992) as an outcome which is valued “for its own sake rather than for the sake of achieving something further” (page 15).

Incremental net benefit value – the difference in the inflation-adjusted, discounted, average benefits and costs of two alternative programs.

Incremental cost-effectiveness ratio – the difference in the inflation-adjusted, discounted, average costs of two programs, divided by the difference in discounted average levels of effectiveness of the two programs.

Intent-to-treat design - an evaluation design in which analyses are conducted upon the basis of a treatment or comparison group assigned or chosen at baseline, regardless of how long observations remained in that group.

Internal rate of return - the discount rate associated with a net present value figure of $0. Programs with higher internal rates of return are more economically attractive.

Internal validity - refers to the ability to make statements about causal relationships between variables (Cook and Campbell, 1979). Internal validity threats may diminish the truthfulness of those statements.

Long-term disability expenditures - includes salary continuation payments for those covered by insured, self-administered, or trust plans. (Source: U.S. Chamber of Commerce definition, 1995).

Lost workday cases - cases that involve days away from work or days of restricted duties at work, or both. (Source: U.S. Dept. of Labor, Bureau of Labor Statistics, 1995).

Net present value - the inflation-adjusted, discounted benefits of a program or intervention, minus the inflation-adjusted, discounted costs of producing and consuming it, expressed in today's dollars or the dollars of a base year of interest.

Opportunity cost - the value of resources used to produce or consume goods or services in their next best alternative use.

Outlier data - extremely high or low values of a variable of interest.

Practical significance - a result or value of sufficient magnitude that it is important to program providers, clients, policy makers, or other stakeholders.

Productivity - defined generally by economists as the amount of output of a good or service produced per unit of input needed to produce it. May be measured more easily in manufacturing processes in terms of goods or units produced per staff member or machine. More difficult to measure for services, because the boundaries that define services may be less well understood or the quality of services produced may be more difficult to measure**.**

Productivity correlates - factors related to productivity such as various forms of absenteeism, restricted activity days, employee morale, production delays, job tenure, etc.

Propensity score - in the context of performing adjustments for selection bias, the propensity score is the predicted probability that each client participates in a substance abuse program.

Quadratic Equation – an equation with the general format of $ax^2 + bx + c = 0$, where *a*, *b*, and *c* are constants and *x* is a variable.

Quadratic Formula – a formula which can be used to solve a quadratic equation for *x*. The quadratic formula is: $x = \{-b \pm (b^2 - 4ac)^{0.5}\} / 2a$.

Quality-adjusted life year - a measure that combines gains from reduced morbidity (quality gains) and reduced mortality (quantity gains) into a single measure. The combination is based on the relative desirability of these different outcomes, with preferences for each health state accounted for in the analysis. See Drummond, et al. (1997) for details.

Randomization test - a process of repeated testing used to estimate p-values for statistical tests with small samples. See Eddington (1995) for details.

Relative Value Unit - a measure of the relative amount of work it takes to complete a medical procedure, compared to a reference procedure. See Hsiao, et al. (1988) for details.

Selection bias - a bias in the estimate of a program effect that arises from the inability to separate the impact of the program on an outcome of interest from the impact of other factors that are correlated with program participation and outcome measures. Such bias often occurs in non-randomized or poorly randomized settings, resulting in treatment and comparison groups that differ on measurable and unmeasureable factors. For example, self-referral to (or self-selection into) a substance abuse program may result in substantial differences between substance abusers who participate and those who do not participate in the program. These differences, along with participation status, may influence observed outcomes.

Sensitivity - in the context of the accuracy of diagnosis coding, sensitivity refers to the ability to identify persons with a particular disorder using claims data or survey data, among persons who really have that disorder.

Sensitivity analysis - a process of repeating the CBA or CEA several times, varying one or more assumptions necessary to carry out the analysis each time, to see how robust the results are to these changing assumptions.

Specificity - in the context of the accuracy of diagnosis coding, specificity refers to the ability to identify those who do not have a disorder of interest using claims data or survey data, among those who really do not have that disorder.

Short-term disability expenditures - includes company payments for sickness and accident benefits beyond any sick leave or other days not included in the short-term disability program. For example, many companies do not pay for the first five consecutive absence days under a short-term disability program.

Stakeholders - persons or groups who have strong opinions about the design, function, or outcomes of a program or intervention.

Statistical power - The ability to detect a relationship between an intervention and outcomes of interest when that relationship really exists.

Turnover rate - includes all permanent separations, whether voluntary or involuntary. Monthly turnover rates are calculated by employers and collected as part of the Bureau of National Affair’s Quarterly Employment Survey. BNA then calculates the monthly median rates and the average of monthly median rates for the year. Monthly rates are calculated as (Number of separations during month / Average number of employees on payroll during the month) * 100. (Source: Bureau of National Affair’s definition, 1995). SAMHSA grantees may wish to calculate separate turnover rates for voluntary and involuntary separations if their programs are more likely to affect one type of turnover than another.

Workers Compensation payments - Includes actual disbursements for injuries and illnesses covered under Workers Compensation program rules.

Workplace injuries and illnesses - Nonfatal occupational illnesses or injuries which involve one or more of the following: loss of consciousness, restriction of work or motion, lost worktime, transfer to another job, or medical treatment (other than first aid). (Source: U.S. Dept. of Labor, Bureau of Labor Statistics, 1995)

REFERENCES

American Medical Association. *Physicians' Current Procedural Terminology – CPT'98*. Chicago, IL: The American Medical Association, 1998.

Barnett PG. 1997. "Research without billing data: Econometric estimation of patient-specific costs." *Medical Care* 35: 553-563.

Belsley DA, Kuh E, Welsch RE. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York, NY: John Wiley & Sons, Inc., 1980.

Black TR. *Evaluating Social Science Research: An Introduction*. Newbury Park, CA: Sage Publications, Inc., 1993.

Bray JW. *An Interview Guide for Human Resources Data Systems*. Research Triangle Park, NC: Research Triangle Institute, 1998.

Bray JW, Zarkin GA. *An Interview Guide for Managed Care Organization Data Systems*. Research Triangle Park, NC: Research Triangle Institute, 1998a.

Bray JW, Zarkin GA. *An Interview Guide for Employee Assistance Program Data Systems*. Research Triangle Park, NC: Research Triangle Institute, 1998b.

Cochrane WG. *Sampling Techniques, 3rd Edition*. New York, NY: John Wiley & Sons, Inc., 1978 (pp. 127-130).

Connors AF, Jr., Speroff T, Dawson NV, Harrell TC, Jr., Wagner D, Desbiens N, Goldman L, Wu AW, Califf RM, Fulkerson WJ, Jr., Vidaillet H, Broste S, Bellamy P, Lynn J, Knaus WA. The effectiveness of right heart catheterization in the initial care of critically ill clients. *Journal of the American Medical Association* 276(11): 889-897, 1996.

Cook RF, Bernstein AD, Arrington TL, Andrews CM, Marshall BS. Methods for assessing drug use prevalence in the workplace: A comparison of self-report, urinalysis, and hair analysis. *The International Journal of Addictions* 30: 403-425, 1995.

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

Cottler LB, Robins LN, Helzer JE. 1989. "The reliability of the CIDI-SAM: A comprehensive substance abuse interview." *British Journal of Addiction* 84: 801-814.

Croghan TW, Obenchain RL, Crown WE. What does treatment of depression really cost? *Health Affairs* 17: 198-208, 1998.

Cropper ML, Aydede SK, Portney RP. "Rates of time preference for saving lives." *American Economic Review* 82 (2): 469-472, 1992.

Donaldson C. The (near) equivalence of cost-effectiveness and cost-benefit analyses: Fact or fallacy? *Pharmacoeconomics* 13: 389-396, 1998.

Drake C, Fisher L. Prognostic Models and the Propensity Score. *International Journal of Epidemiology*, 24: 183-187, 1995.

Drummond MF, O'Brien BJ, Stoddart GL, Torrance GW. *Methods for the Economic Evaluation of Health Care Programs, 2nd Edition*. New York, NY: Oxford University Press, 1997.

Duan N. "Smearing estimate: A nonparametric retransformation method." *Journal of the American Statistical Association* 78: 605-611, 1983.

Duan N, Manning WG, Jr., Morris CN, Newhouse JP. "A comparison of alternative models for the demand for medical care." *Journal of Business and Economic Statistics* 1: 115 -

126, 1983.

DuMouchel WH, Duncan GJ. “Using sample survey weights in multiple regression analyses of stratified samples.” *Journal of the American Statistical Association* 78: 535-543, 1983.

Eddington ES. *Randomization Tests*. New York, NY: M. Dekker, 1995.

Fetter RB, Brand DA, Gamache D. *DRGs: Their Design and Development*. Ann Arbor, MI: Health Administration Press, 1991.

Fowles JB, Fowler EJ, Craft C. 1998. “Validation of claims diagnoses and self-reported conditions compared with medical records for selected chronic conditions.” *Journal of Ambulatory Care Management* 21: 24-34.

Gardiner J, Hogan A, Holmes-Rovner M, Rovner D, Griffith L, Kupersmith J. Confidence intervals for cost-effectiveness ratios. *Medical Decision Making* 15: 254-263, 1995.

Getzen TE. Medical care price indexes: Theory, construction, and empirical analysis of the U.S. series 1927-1990. *Advances in Health Economics and Health Services Research* 13: 83-128, 1992.

Gibaldi M, Sullivan S. Intention-to-treat analysis in randomized trials: who gets counted? *Journal of Clinical Pharmacology* 37: 667-672, 1997.

Goetzel RZ. Program evaluation. In O’Donnell M and Harris J (eds.): *Health Promotion in the Workplace, 2nd Edition*. Albany, NY: Delmar Publishers, Inc., 1994.

Goetzel RZ, Dunn RL, Ozminkowski RJ, Satin K, Whitehead D, Cahill K. Differences between descriptive and multivariate estimates of the impact of Chevron Corporation’s Health Quest Program on medical expenditures. *Journal of Occupational and Environmental Medicine*

40: 538-545, 1998.

Gramlich EM. *Benefit-Cost Analysis of Government Programs*. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1981.

Greene WH. *Econometric Analysis*. New York, NY: Macmillan Publishing Company, 1990.

Greene WH. *LIMDEP, Version 7.0*. Castle Hill, New South Wales, Australia: Econometric Software, Inc., 1995.

Hargreaves WA, Shumway M, Hu T-W, Cuffel B. *Cost-Outcome Methods for Mental Health*. San Diego, CA: Academic Press, 1998.

Health Care Financing Administration (HCFA). Report to Congress: Medicare Physician Payment. HCFA Publication No. 03287, 1989.

Heaney CA, Goetzel RZ. A review of health-related outcomes of multi-component worksite health promotion programs. *American Journal of Health Promotion* 11: 290-308, 1997.

Heckman JJ. The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a simple estimator for such models. *Annals of Economic and Social Measurement* 5: 475-492, 1976.

Heckman JJ. Sample selection as a specification error. *Econometrica*, 47(1):153-161, 1979.

Heckman JJ, Smith J. Assessing the case for social experiments. *Journal of Economic Perspectives*, 9(2):85-110, 1995.

Hedrick TE, Bickman L, Rog DL. *Applied Research Design: A Practical Guide*. Newbury Park, CA: Sage Publications, Inc., 1993.

Hsiao WC, Braun P, Dunn D, Becker ER. Resource-based relative values. *Journal of the American Medical Association* 260: 1988.

Kalton G, Kasprzyk D. "The treatment of missing survey data." *Survey Methodology* 12(1): 1-16, 1986.

Keeler EB, Cretin S. "Discounting of life-saving and other non-monetary benefits." *Management Science* 29: 300-310, 1983.

Kennedy P. *A Guide to Econometrics, 3rd Edition*. Cambridge, Massachusetts: The MIT Press, 1992.

Kmenta J. *Elements of Econometrics, 2nd Edition*. New York, NY: Macmillan Publishing Company, 1986.

Krahn M, Gafni A. Discounting in the economic evaluation of health care interventions. Medical Care 31: 403-418, 1993.

LaLonde, R. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* 76(4): 604-620, 1986.

Lestina S, Miller TR, Smith GR. Creating injury episodes using medical claims data. Prepared for the 3rd Steering Committee Meeting of the CSAP Workplace Managed Care Initiative, Feb. 19-20, 1998.

Little RJA, Rubin DB. *Statistical Analysis With Missing Data*. New York, NY: John Wiley & Sons, Inc., 1987.

Luft HR. Health services research as a scientific process: The metamorphosis of an empirical research project from grant proposal to final report. In: Defriese GH, Ricketts TC, and Stein JC (eds.), *Methodological Advances in Health Services Research*. Ann Arbor, MI: Health Administration Press, 1989.

Manning WG, Jr. "The logged dependent variable, heteroscedasticity, and the retransformation problem." *Journal of Health Economics* 17: 283-295, 1998.

Miller TR. "Estimating the costs of injury to US employers." *Journal of Safety Research* 28(1): 1-13, 1997.

Mohr L. *Impact Analysis for Program Evaluation*. Newbury Park, CA: Sage Publications, Inc., 1992.

Mullahy J. "Much ado about two: Reconsidering retransformation and the two-part model in health econometrics." *Journal of Health Economics* 17: 247-281, 1998.

Mullahy J, Manning WG, Jr. "Statistical issues in cost-effectiveness analyses." Chapter 8 in Sloan F (ed.), *Valuing Health Care: Costs, Benefits, and Effectiveness of Pharmaceuticals and Other Medical Technologies*. New York, NY: Cambridge University Press, 1995.

Nas T. *Cost-Benefit Analysis: Theory and Application*. Thousand Oaks, CA: Sage Publications, 1996.

National Center for Health Statistics. *National Death Index User's Manual*. Hyattsville, MD: U.S. Dept. of Health and Human Services, Public Health Service, Centers for Disease Control, National Center for Health Statistics, DHHS Publication No. (PHS) 90-1148, Sept. 1990.

Newhouse JP. Measuring medical prices and understanding their effects. *Journal of Health Administration Education* 7: 19-26, 1989.

O'Brien BJ, Drummond MF, LaBelle RJ, Willan A. In search of power and significance: issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Medical Care* 32: 150-163, 1994.

O'Brien BJ, Drummond MF. Statistical versus quantitative significance in the socioeconomic evaluation of medicines. *PharmacoEconomics* 5: 389-398, 1994.

Olsen RJ. A least squares correction for sensitivity bias. *Econometrica* 48: 1815-1820, 1980.

Ozminkowski RJ, Branch LG. On the economic analysis of interventions for aged populations. In: Hickey T, Speers MA, and Prochaska TR (eds.): *Public Health and Aging*, Baltimore, MD: The Johns Hopkins University Press, 1997.

Pindyck RS, Rubinfeld DL. *Econometric Models and Economic Forecasts, 3rd Edition.* New York, NY: McGraw-Hill, Inc., 1991.

Revicki DA. Health care technology assessment and health-related quality of life. In: Banta HD and Luce BR (eds.), *Health Care Technology and Its Assessment*, New York, NY: Oxford University Press, 1993.

Robins J, Mark S, Newey W. "Estimating exposure effects by modeling the expectation of exposure conditional on confounders." *Biometrics*, 48: 479-495, 1992.

Rognehaugh R. *The Managed Care Dictionary*. Gaithersburg, MD: Aspen Publishers, Inc., 1996.

Rosenbaum P, Rubin D. Reducing bias in observational studies using subclassification on the propensity score." *Journal of the American Statistical Association* 79(387): 516-524, 1984.

Rossi PH, Freeman HL. Evaluation: A Systematic Approach (5th Edition). Newbury

Park, CA: Sage Publications, Inc., 1993.

Russell LB, Gold MR, Siegel JE, Daniels N, Weinstein MC. The role of cost-effectiveness analysis in health and medicine. *Journal of the American Medical Association* 276: 1172-1180, 1996.

Sclar DA, Skaer TL, Robison LM, Galin RS, Legg RF, Nemec NL. Economic consequences with antidepressant pharmacotherapy: A retrospective intent-to-treat analysis. *Journal of Clinical Psychiatry* 59, Suppl 2: 13-17, 1998.

Siegel JE, Weinstein MC, Russell LB, Gold MR. Recommendations for reporting cost-effectiveness analyses. Journal of the American Medical Association 276: 1339-1341, 1996.

Siegel C, Laska E, Meisner M. “Statistical methods for cost-effectiveness analyses.” *Controlled Clinical Trials* (in press).

Spector PE. Development of the work locus of control scale. *Journal of Occupational Psychology* 61: 335-340, 1988.

Stewart KG, Levy D, Rosenbach ML, Bartosch WJ. *Estimating Costs and Outcomes of Substance Abuse Prevention Strategies: Technical Report*. Rockville, MD: Center for Substance Abuse Prevention, Substance Abuse and Mental Health Services Administration, Dept of Health and Human Services, 1998. DHHS Publication No. 98-3235.

Viscusi WK. “Discounting health effects for medical decisions.” Chapter 6 in Sloan F (ed.), *Valuing Health Care: Costs, Benefits, and Effectiveness of Pharmaceuticals and Other Medical Technologies*. New York, NY: Cambridge University Press, 1995.

Viscusi WK, Moore MJ. “Rates of time preference and valuations of the duration of life.” *Journal of Public Economics* 8: 397-417, 1989.

Vogt WP. *Dictionary of Statistics and Methodology*. Newbury Park, CA: Sage Publications, 1993.

Ware JE, Sherbourne CD. The MOS 36-item short form health survey (SF-36). I: conceptual framework and item selection. *Medical Care* 30: 473-483, 1991.

Warner KE, Luce BR. Cost-benefit and Cost-effectiveness Analysis in Health Care: Principles, Practice, and Potential. Ann Arbor, MI: Health Administration Press, 1982.

Weinstein MC. “From cost-effectiveness ratios to resource allocation: Where to draw the line?” Chapter 5 in Sloan F (ed.), *Valuing Health Care: Costs, Benefits, and Effectiveness of Pharmaceuticals and Other Medical Technologies*. New York, NY: Cambridge University Press, 1995.

Weinstein MC, Siegel JE, Gold MR, Kamlet MS, Russell LB. Recommendations of the Panel on cost-effectiveness in health and medicine. *Journal of the American Medical Association* 276: 1253-1258, 1996.

Wray NP, Hollingsworth JC, Petersen NJ, Ashton CM. 1997. “Case-mix adjustment using administrative databases: A paradigm to guide future research.” *Medical Care Research and Review* 54: 326-356.

APPENDIX 1

ESTIMATING THE COST OF ALTERNATIVE INTERVENTIONS

Stewart, et al. (1998) present a cost-allocation matrix which can be used to illustrate how to estimate the cost of alternative substance abuse prevention strategies. We modified their matrix a bit below, by reducing the number of interventions to simplify the subsequent explanation. We also added the “total” rows at the bottom and the monetary cost figures to each cell. We then operate under the assumption that each alternative project runs for three years (beginning in 1998), and we show how costs vary by year. We first show each year’s costs in nominal (unadjusted) dollars. Then the inflation adjustment is applied to the totals, assuming an annual inflation rate of 3.5%. Inflation-adjusted dollars are cast in terms of their 1998 values. A 5% discount rate is applied in years 2 and 3 as well, to adjust for the changing value of money over time.

We these assumptions in mind, the three cost allocation matrices are shown below, for two fictitious alternatives. The first alternative involves no change in the way a managed care organization and employer identify substance abusers and refer them to treatment. The second alternative involves better methods to find substance abusers (e.g., more frequent drug testing and enhanced training for managers and workers to spot potential problems), more rapid referral to the Employee Assistance Program, better communication between the EAP and the managed care organization, more rapid treatment, and closer monitoring of progress.

Costs for Year 1 of Alternative Interventions to Reduce the Prevalence of Substance Abuse Among Employees of Company A

Resources	**Undiscounted Intervention Cost – Year 1 (1998 A.D.)**	
	Method 1 – Status Quo	Method 2 – Enhanced case finding, faster referral, better coordination of services
People's time	500,000	600,000
Space	150,000	150,000
Transportation	50,000	75,000
Equipment	300,000	350,000
Supplies	100,000	200,000
Telecommunications	45,000	100,000
Information services	100,000	200,000
Financing / insurance	75,000	75,000
Other resources	200,000	200,000
Total	**1,520,000**	**1,950,000**

Costs for Year 2 of Alternative Interventions to Reduce the Prevalence of Substance Abuse Among Employees of Company A

Resources	**Undiscounted Intervention Cost – Year 2 (1999 A.D.)**	
	Method 1 – Status Quo	Method 2 – Enhanced case finding, faster referral, better coordination of services
People's time	525,000	630,000
Space	157,500	157,500
Transportation	52,500	80,000
Equipment	315,000	375,000
Supplies	105,000	210,000
Telecommunications	50,000	112,500
Information services	110,000	220,000
Financing / insurance	70,000	70,000
Other resources	210,000	210,000
Unadjusted Total in 1999 dollars	**1,595,000**	**2,065,000**
Total Adjusted for Inflation (in 1998 dollars, with 3.5% inflation rate)	**1,541,063**	**1,995,169**
Total Adjusted for Inflation and Discounted (in 1998 dollars, with 5% discount rate)	**1,467,679**	**1,900,161**

Costs for Year 3 of Alternative Interventions to Reduce the Prevalence of Substance Abuse Among Employees of Company A

Resources	Undiscounted Intervention Cost – Year 3 (2000 A.D.)	
	Method 1 – Status Quo	Method 2 – Enhanced case finding, faster referral, better coordination of services
People's time	575,000	660,000
Space	175,500	175,500
Transportation	55,500	80,000
Equipment	335,000	390,000
Supplies	125,000	225,000
Telecommunications	55,000	120,000
Information services	125,000	250,000
Financing / insurance	67,000	67,000
Other resources	230,000	230,000
Unadjusted Total in Year 2000 dollars	**1,742,500**	**2,197,500**
Total Adjusted for Inflation (in 1998 dollars, with 3.5% annual inflation rate)	**1,626,642**	**2,051,390**
Total Adjusted for Inflation and Discounted (in 1998 dollars, with 5% annual discount rate)	**1,475,412**	**1,860,671**

It is important to note that more detailed budgets may be derived prior to completing each cost matrix. Such budgets would be necessary to itemize each category, showing, for example, the amount of time that each paid employee and unpaid volunteer contributed to the project, multiplied by their paid or market wage rates. Space costs might be itemized to reflect mortgage or lease rates in various locations where people work. Similarly, costs for equipment, supplies, telecommunications, information services, travel, financing and insurance, and other resources may be location specific.

Estimating the cost of each item in each category is where the real work lies in the costing process. As noted in Chapter 3 and in Hargreaves, et al. (1998), accounting methods or cost study methods may be used to derive these costs, but the focus should be on estimating the opportunity cost of each resource. Nothing is free, even if it is donated to the project, because even donated resources could have been used for other valuable purposes. The highest value of those alternative uses represents the opportunity cost of the resource.

Next, one should note that costs for all of the items in the matrix may change at different rates over time. In the example noted above, staffing costs from year 2 to year 3 increased by about 9.5% (in unadjusted dollars), whereas travel costs increased by only 4.8%. This might happen if some former staff quit and new staff are added, for example.

Once the cost allocation matrices have been developed, it is simple to estimate the total, discounted, inflation-adjusted cost of each alternative. In the example above, this would be done by adding the values in the last row of each table. In discounted, inflation-adjusted 1998 dollars, the Status Quo alternative would cost $4,463,091 and the cost of expanded case finding, faster referral to treatment, and better coordination of services would be $5,710,832.

The fact that the cost of expanded case finding, faster referral to treatment, and better coordination of services is estimated to be higher than the cost of the Status Quo alternative is meaningless at this point, for two reasons. First, we concocted these numbers, just for the sake of illustration; these figures were not derived from any real project. We could easily have made up numbers that illustrated lower costs for expanded case finding, etc., but that may not seem realistic. Second, and more important, when choosing among alternatives cost alone should *not* be the deciding factor. In fact, choosing the less costly alternative may result in a waste of resources (i.e., it may be penny-wise but pound foolish, as the saying goes). From an economist's perspective, the choice of which alternative to adopt should be based upon the *relationship* between the benefits of each alternative and its costs. It is possible that the benefits of expanded case finding, etc. outweigh their additional costs. It would be the analysts's duty to investigate this issue in a well-conducted cost-benefit or cost-effectiveness analysis Methods for relating benefits to costs are described in Chapter 3, and in Appendix 3 below.

APPENDIX 2

Section on "Program Design," taken from Goetzel RZ, Program Evaluation, in *O'Donnell MP and Harris JS (eds.) Health Promotion in the Workplace, 2nd Edition,* Albany, NY: Delmar Publishers, Inc., 1994. (Used with permission).

APPENDIX 3

EXAMPLES OF NET PRESENT VALUE AND INTERNAL RATE OF RETURN

In this appendix we illustrate the net present value and internal rate of return calculations for fictional projects.

Net Present Value. Suppose the fictional project being evaluated lasts for three years, having the following stream of undiscounted, inflation-adjusted cost and benefit dollars.

Year	Benefit	Cost
0 (base year)	0	1,000
1	500	500
2	2,000	500

As noted earlier in Chapter 3, the net present value formula is the sum of the discounted difference between benefits and costs. For this fictional project, assume a 5% discount rate, *r*, is chosen. The NPV calculation therefore is as follows:

$$NPV = \{(0 - 1000) / (1 + 0)^0 \quad + \quad (500 - 500) / (1 + 0.05)^1 \quad + \quad (2000 - 500) / (1 + 0.05)^2\}.$$

Since costs and benefits are not discounted in the base year, the denominator of the first term in the NPV equation equals 1.0. If we carry out the math in the NPV equation, we obtain:

$$\begin{aligned} NPV &= -1000 + 0 + 1500 / 1.1025 \\ &= -1000 + 1360.54 \\ &= \ 360.54. \end{aligned}$$

Thus, the net present value of this fictional project equals $360.54. Investing now in this three-year project would be equivalent to receiving $360.54 today! If multiple projects were being evaluated, the project with the highest NPV would be preferable (with the one rare exception noted by Nas (1996), earlier in Chapter 3).

Internal Rate of Return. The internal rate of return is defined as the discount rate which would lead investors to be indifferent toward the project. Mathematically, this is defined as the rate that must occur when the NPV equals zero. For the fictional project noted above, we therefore solve the NPV equation for the discount rate, *r*. Thus, the formula for the IRR is as follows:

$$0 = (0 - 1000) / (1 + 0)^0 + (500 - 500) / (1 + r)^1 + (2000 - 500) / (1 + r)^2$$

Carrying out the math in this equation, we obtain:

$$0 = -1000 + 0 + 1500 / (1 + r)^2$$
$$\text{or } 1000 = 1500 / (1 + r)^2$$
$$\text{or } 1000 (1 + r)^2 = 1500$$
$$\text{or } (1 + r)^2 = 1.5$$
$$\text{or } (1 + r)(1+r) = 1.5$$
$$\text{or } r^2 + 2r + 1 = 1.5$$
$$\text{or } r^2 + 2r - 0.5 = 0.$$

This last equation is a quadratic equation, and we can use the quadratic formula to solve for *r*. The quadratic formula for this equation is:

$$r = [-2 \pm \{4 - 4(1)(-0.5)\}^{0.5}] / 2$$

If we solve this quadratic we obtain $r = 0.2247$ or $r = -2.2247$. If we insert either one of these values into the NPV formula noted above, the NPV calculation will yield zero. Thus, the IRR is either 22.47% (a great return on investment!) or –222.47% (an awful one!). We noted earlier in Chapter 3 that the IRR can sometimes be either positive or negative. This less-thansatisfying result is one reason to choose the NPV as the better method for estimating the value of this fictional project.

A Non-existent Internal Rate of Return. Finally, we noted in Chapter 3 that the IRR may not even exist. An example of this case follows. Suppose that another fictional project has the following stream of undiscounted, inflation-adjusted benefits and costs.

Year	Benefit	Cost
0 (base year)	2,500	1,000
1	500	500
2	2,000	500

If we solve for the internal rate of return, we will eventually obtain another quadratic equation. (We are leaving out the intermediate steps for brevity, but the reader can easily check the results by trying to solve for *r* in the NPV formula for this project, when NPV is set equal to zero.) The quadratic obtained is as follows:

$$r^2 + 2r + 2 = 0$$

If we apply the quadratic formula to solve this equation, we obtain:

$$r = [- 2 \pm \{4 - 4(1)(2)\}^{0.5}] / 2$$

The problem with this quadratic formula is that we must take the square root of a negative number (i.e., $4 - 4(1)(2) = -4$). The square root of this number ($-4^{0.5}$) is not a real number. Thus, the IRR does not exist. For this project, one would be better off relying on the net present value as an estimate of its economic worth.