CHAPTER 18

EVALUATION OF PUBLIC HEALTH INTERVENTIONS

Michael A. Stoto
Leon E. Cosler

Chapter Overview

Evaluation encompasses the set of tools that are used to measure the effectiveness of public health programs by determining what works. Traditional evaluations in public health have focused on assessing the impact of specific program activities on defined outcomes. Evaluation is also a conceptual approach to the use of data—as part of a quality improvement process—in public health management. Public health organizations must continually improve upon the standards of evidence used in the evaluation of public health so that results can inform managerial and policy decision making. As public health interventions become more integrated within the community, collaboration in evaluation efforts is a growing imperative.

Evaluation concepts and methods are of growing importance to public health organizations, as well as to education and social services programs. Increasingly, public health managers are being held accountable for their actions, and managers, elected officials, and the public are asking whether programs work, for whom, and under what conditions. Public health decision makers need to know which program variants work best, whether the public is getting the best possible value for its investment, and how to increase the impact of existing programs. These evaluation questions are being asked of long-standing programs, new activities, and proposed interventions. These developments parallel today's emphasis on "evidence-based medicine" in clinical areas and suggest the growing role of "evidence-based management" within public health organizations.

In this context, *evaluation* is, first of all, a set of tools that is used to improve the effectiveness of public health programs and activities by determining which programs work, and also which program variants work most effectively. These tools derive from social science and health services research and include concepts of study design, a variety of statistical methods, and

economic evaluation tools. *Evaluation* is also a conceptual approach to the use of data—as part of a quality improvement process—in public health management.

However defined, evaluation can be useful to managers in public health who need, for example, to do the following activities:

- Judge the effectiveness of new approaches to public health service delivery systems that were developed elsewhere, and judge their potential applicability in one's own jurisdiction. For instance, do the immunization registries being tried in a number of US cities actually result in more children being immunized?
- Judge the effectiveness of new approaches to public health service delivery systems that were developed in one's own jurisdiction. For instance, does the new community-based outreach program actually result in more children being immunized? If not, why not?
- Assess how well an intervention is being implemented in one's own jurisdiction. What fraction of children is being enrolled in the community's new immunization registry at birth? Which children are left out? What can be done to improve coverage?
- Ensure accountability of contractors and other entities with a responsibility to the public health agency. Are the managed care organizations with Medicaid contracts in the community ensuring that the children enrolled in their plans are receiving all of the recommended immunizations? Are some plans doing better than others? Why?
- Demonstrate accountability of internal programs to funders or higher authorities. Are federal funds for immunization being used according to the funders' guidelines? Are they achieving the intended effect?

This chapter begins with a primer on evaluation research methods, including economic evaluation, used by public health organizations drawing on examples from immunization programs and needle exchange programs to prevent human immunodeficiency virus (HIV) infection. Special issues in the evaluation of community-based interventions are also covered, as are issues of measurement and data. The second section of the chapter deals with practical aspects of program evaluation in public health, drawing on examples from family violence and other areas, and proposes a process for evaluation in public health settings. The final section focuses on performance measurement—in organizations as well as community settings—as a form of evaluation methodology. An extended example dealing with public health preparedness illustrates the concept of performance measurement.

Evaluation Methods

Terminology

All public health programs can be characterized by their inputs, activities, outputs, and outcomes, as illustrated in terms of a childhood immunization program in Exhibit 18–1. *Inputs* are resources dedicated to or consumed by the program. Inputs can, in some settings, include organizational structures and capacities. *Activities* are what the program does with its inputs to fulfill its mission. *Outputs* are the direct product of program activities. Outputs,

Exhibit 18-1 Examples of Program Inputs, Activities, Outputs, and Outcomes

Inputs	*Activities*	*Outputs*	*Outcomes*
Resources dedicated to or consumed by the program • Money • Staff time • Facilities • Equipment • Laws • Regulations • Funders' requirements	What the program does with inputs to fulfill its mission • Educate consumers • Educate providers • Distribute vaccines to providers • Establish an immunization registry • Provide technical assistance to providers about reminder systems • Monitor immunization coverage in the population and health plans	The direct product of program activities • Number of brochures distributed • Doses of vaccines delivered • Percentage of births enrolled in registry • Number of providers who establish a reminder system • Number of providers who monitor immunization coverage • Program costs	Benefits for participants during and after program activities • Parental awareness of vaccine benefits • Provider awareness • Changed attitudes • Children immunized • Reduced burden of vaccine-preventable disease in immunized children • Reduced prevalence of vaccine-preventable disease • Cost per child immunized

sometimes called intermediate outcomes, can sometimes overlap with outcomes, depending on the stage of the intervention. *Outcomes* are benefits for participants during and after program activities.

Within this framework, a number of different types of evaluations are possible:

- Traditional evaluations in public health have focused on assessing the impact of specific program activities on defined outcomes. For instance, does the new reminder system increase the number of children immunized? Questions may also be asked concerning the impact of resources on outcomes. For example, do laws requiring complete immunization prior to school entry reduce vaccine-preventable disease?
- Economic evaluations combine program effectiveness information with economic resources (i.e., costs and benefits) in quantitative terms. They allow decision makers to prioritize public health activities in the face of finite financial resources. Which program, for example, is most effective in terms of costs per child immunized?
- Process evaluations refer to evaluations that are focused on outputs. Such an evaluation might ask, for instance, whether the change in enrollment procedures increases the number of children enrolled in a registry. In these cases, a relationship is assumed between outputs and outcomes (presumably based on research done elsewhere), and

evidence of a change in outputs is taken as indirect evidence of an impact on desired outcomes.

- Formative evaluations refer to efforts to identify the best uses of available resources, prior to a traditional program evaluation. Formative evaluation often employs qualitative methods such as focus groups or structured interviews to understand a process or system and to identify barriers and opportunities for improvement. Project Access in the San Francisco Bay area, for instance, interviewed drug users at needle exchange programs, shooting galleries, parking lots, and drug treatment centers. Researchers discovered a variety of structural barriers that prevented the users from seeking HIV counseling and testing, such as restricted hours of counseling and testing sites, lack of transportation, complications in drawing blood, and insensitive providers.[1]
- Empowerment evaluations involve an approach whereby programs take stock of their existing strengths and weaknesses, focus on key goals and program improvements, develop self-initiated strategies to achieve these goals, and determine the type of evidence that will document credible progress.[2] This approach is discussed in more detail in the following paragraphs.
- Performance measures use statistical methods and other evaluation tools on an ongoing basis to assure accountability for public health programs and to improve performance.

Efficacy Assessment

Regardless of its ultimate purposes, evaluation is essentially an applied research activity seeking to discover whether a program, in some sense, has beneficial effects for the public's health. The program could be a specific activity in one public health clinic or a comprehensive communitywide activity. The question may be retrospective—Did it work?—before the program is expanded to other venues, or current—Is it working?—to ensure accountability and improve outcomes. The issue may be comparing two or more competing interventions, or assessing whether a particular program is better than nothing. The question may also be whether the program is better in some populations, or under some particular conditions. In every case, however, the central question is one of efficacy: Is some program more effective than some alternative?

Program evaluation thus centers on questions of efficacy, but additional steps are usually necessary in order to make policy decisions about recommendations for individuals and the allocation of resources. Programs shown to be effective in controlled situations, however, may not work in settings where the conditions are different. *Effectiveness* refers to a program's ability to get results in less than optimal situations. A work site smoking cessation program developed by highly motivated and skilled health educators for university employees, for instance, may not be as effective when applied by human resources personnel assigned to a large auto manufacturer. Effective programs employ well-developed materials and training so that they can be generalized, that is, transferred from where they were developed to other settings.

The evaluation of public health interventions requires research directed at estimating the unique effects (or net effect) of the intervention, above and

beyond any change that may have occurred because of a multitude of other factors. Such research requires study designs that can distinguish the impact of the intervention within a general service population from other changes that occur only in certain groups or that result simply as a passage of time. These designs commonly involve the formation of two or more groups: one composed of individuals who participated in the intervention (the treatment group) and a second group of individuals who are comparable in character and experience to those who participated but who received no services or an intervention that was different from that under study (the control or comparison group).

To estimate the net effect of an intervention reliably, the following technical issues must be addressed:[3,4]

- The manner in which the control or comparison groups are formed influences the validity of the inference.
- The number of participants enrolled in each group (the sample size) must be sufficient to permit statistical detection of differences between the groups, if differences exist.
- There should be agreement among interested parties that a selected outcome is important to measure, that it is a valid reflection of the objective of the intervention, and that it can reflect change over time.
- Evidence is needed to show that the innovative services were actually provided as planned, and that the differences between the innovative services and usual services were large enough to generate meaningful differences in the outcome of interest.

In clinical medicine, randomization is typically regarded as essential to show that the intervention caused the effect. In public health, however, it is often not possible to randomly assign individuals or populations to interventions for ethical and practical reasons. To make judgments about causality, evaluation researchers have developed a general consensus regarding the relative strength of various study designs, with randomized control trials being the gold standard of evidence and anecdotal case reports being the weakest of the study designs.[3] In addition, a group of statistical methods known collectively as "causal modeling" have been developed to analyze data from nonrandomized experiments and other sources to infer causal relationship when possible.[5]

Statistical power is the likelihood that an evaluation will detect the effect of an intervention, if there is one. Two factors affect statistical power: sample size and effect size, a quantitative measure of the program's impact, such as a 10% improvement in immunization rates. Public health evaluation studies are often based on small samples of individuals, and thus lack sufficient statistical power to detect meaningful effects. Power can be increased by increasing the number of subjects or by increasing the number of intervention sites, as long as each site adheres to common design elements and applies uniform eligibility criteria. Evaluators planning a study must consider whether the study size and effect size are large enough to ensure a reasonable probability that the program's impact will be statistically significant. Managers considering the results of a negative study should consider whether the study has sufficient power to detect an effect.

A related criterion is the need for careful implementation of the intervention being evaluated. A careful, randomized assessment of an intervention

that is poorly implemented is likely to show, with great precision, that the intervention did not work as intended. Such a study will not distinguish between failures of the theory that is being tested and failures of implementation. Program designers need to identify critical elements of programs to explain why effects occurred and to assist others who wish to replicate the intervention model in a new setting.

Experimental Designs

The highest level of evidence occurs with *experimental designs* that include randomized controls to restrict a number of important threats to validity. In clinical settings, individual patients are assigned by some random mechanism to a treatment group or a control group, and both groups are observed to see if there is a difference in outcomes of interest. If there is a difference in outcome, it can be assumed to be due to the intervention because the randomization reduces the chance that there are no other differences between the two groups. An additional benefit is that the random allocation per se makes it possible to perform statistical inference, that is, to assess whether the observed differences can be due to chance.

Public health interventions such as immunizations are essentially personal health services, so the randomization model can be used directly. For other programs, the unit of intervention might be social units such as schools, work sites, or even whole communities. In these instances, randomization can still be carried out, but just not on an individual basis.[4,6] Although ethical and political objections are often raised, randomization can be carried out in social settings much more commonly than is currently done. As long as there is equipoise regarding the benefits and harms of the intervention, potential for participants to benefit, and a means of informed consent, randomization is ethically acceptable. Resource constraints that prevent the immediate introduction of a new program to an entire population present an opportunity to randomize which units get the intervention first and concomitantly evaluate the outcome. Waiting lists can be arranged so that every client eventually receives the service, but those who get it first are chosen at random and compared to those who receive it later.[7]

Quasi-Experimental Designs

In public health, however, random assignment commonly either is not feasible or simply is not done. In such instances, *quasi-experimental designs*, the second level of evidence in the hierarchy, can be used to assess the impact of programs. Included in this group are analyses using existing computerized databases, case-control observational studies, and series based on historical controls. Although these designs can improve inferential clarity, they cannot be relied on to yield unbiased estimates of the effects of interventions because the subjects are not assigned randomly. A before-and-after design, for example, compares some outcome in the same group before and after a program is introduced. Did traffic fatalities go down in the three months after the speed limit was lowered? Designs of this sort, however, are subject to a variety of biases or threats to validity. If fatalities decreased, can it be due to better weather after the speed limit was changed on March 15? Before-and-after de-

signs can be improved by gathering multiple data points before and after the program is introduced, and by careful examination of other factors, such as weather, that may be responsible for the apparent effect.

Another important quasi-experimental design is to have one or more comparison groups that are thought to be similar to the group receiving the intervention. If immunization coverage rates are higher in a community that has received a special program than in a neighboring community with no such intervention, a prima facie case can be made for the efficacy of the program. A slightly more complex design combines the before-and-after and control group approaches: teen birth rates are measured in two communities before and after one community attends a special school-based program. In this approach, a larger decrease in teen birth rate in the school that received the intervention than in the control school is interpreted as evidence of efficacy. The major problem with a comparison-group design is that the treatment control groups may differ in some way other than the intervention that explains the outcome. A selection bias, for instance, occurs when more advantaged population groups are more likely to choose or be chosen for a new program. In either of the examples cited, for instance, the differences may reflect a greater social advantage in the intervention group, which explains both the outcome and why they received the program.

When randomization is not possible or is not performed, statistical methods are available to reduce the effect of selection or other biases. If the experimental and comparison groups differ in some respects that may affect the outcomes, multivariate analysis can be used to "adjust" for these differences and isolate the true effect of the intervention. A work site smoking cessation program, for example, may have been tested in two workplaces that differ substantially in the proportion of male and female workers and in the proportion of blue- versus white-collar positions. Because both sex and kind of job could affect smoking cessation success, evaluators might want to adjust for these factors in their analysis. In clinical settings, where patients with more severe illness seek out academic medical centers and are also at higher risk for failure, evaluators "risk adjust" to account for these differences.

Nonexperimental Designs

The lowest level of evidence occurs with *nonexperimental designs,* which consist of case series and anecdotal reports. Although such studies can contain a wealth of useful information, they cannot support inference because they cannot control for factors such as maturation, self-selection, historical influences unrelated to the intervention, and changes in instrumentation.

When a difference is detected between treatment and control groups in a study, the first question an evaluator asks is whether the difference could be due to chance (resulting from sampling individuals to be included). Various statistical techniques, depending on the nature of the research design and data, are available to provide answers to this question. If the analysis suggests that the difference is unlikely to be due to chance, it is said to be *statistically significant.* Statistical significance is sometimes assessed through the examination of *confidence intervals* (*CI*). A 95% confidence interval is a range calculated from the data in such a way that there is a 95% chance that the range includes

the quantity being estimated. Suppose, for example, an educational program was evaluated in terms of the average difference between the scores of individuals who were in the program and a similar group of controls on a test of knowledge of HIV risk factors. If the average difference was 1.5 points on a 10-point test, and the 95% confidence interval was 0.8 to 2.2 points (0 is not in the range), the difference can be said to be statistically significant.

Economic Analyses

There are several separate methods commonly employed under the description of economic analyses. These include: cost analysis, cost-minimization analysis (CMA), cost-effectiveness analysis (CEA), cost-utility analysis (CUA), and cost-benefit analysis (CBA). Each method differs in its approach to measuring economic resources, and thus each has applicability to different situations. There are, however, common characteristics important to all of these techniques.

Cost Analysis

Cost analysis refers to any evaluation that uses the structured collection of costs without regard to evaluating health benefits or outcomes. Costs are frequently categorized into direct costs, indirect costs, and intangible costs. Direct and indirect costs can be subcategorized into medical and nonmedically related costs. Direct medical costs represent the value of the resources used specifically for the healthcare services or interventions being measured. These frequently include all medical services, diagnostic testing, and treatment including medications. Direct nonmedical costs can include necessary expenses that are not healthcare related. These costs may include patient transportation costs, child care, or the costs of advertising related to patient education. Indirect costs attempt to measure the economic value of resources that are lost as a result of contracting an illness or participating in an intervention. Commonly, indirect costs are measured as the value of lost wages or the value of lost leisure time. Intangible costs attempt to quantify costs associated with the pain or suffering associated with disease or its treatment. Intangible costs are extremely difficult to measure, and thus are rarely included in many economic analyses.[8,9]

The cost analysis technique is commonly used to conduct cost-of-illness (COI) studies. These studies attempt to quantify all costs (direct and indirect) associated with a particular illness or condition. Cost-of-illness studies assist public health decision makers in planning interventions and targeting limited research funds. Cost-minimization analysis (CMA) is another common type of cost analysis. CMA requires a thorough assessment of all relevant costs associated with two or more health interventions but further assumes that the health outcomes of each intervention are exactly the same. For example, cost-minimization analysis can be used to compare therapy with brand name versus generic drugs. This method however has been overused because of its simplicity. Frequently, the requirement of identical outcomes is not well established. When health outcomes are not equivalent, another method, such as cost-effectiveness analysis is required.[9]

Cost-Effectiveness Analysis

When comparing health outcomes that may not be the same, cost-effectiveness analysis (CEA) is a widely utilized method. CEA aggregates all appropriate costs for one or more healthcare interventions and expresses them in terms of health outcomes in their natural (nonmonetary) units. The outcomes may be a very specific single measure (e.g., dollars per life saved or dollars per case of disease prevented) or may be a composite measure that adjusts for quality of life. In situations where the denominator adjusts for quality of life, this technique has sometimes been referred to as a cost-utility analysis (CUA), which is described later.

Results of a cost-effectiveness analysis are commonly reported as two types of ratios, the average cost-effectiveness ratio and the incremental cost-effectiveness ratio (ICER). The average cost-effectiveness ratio is the appropriate summary measure when there are no comparisons between interventions, and the ratio becomes a simple description of the cost per outcome for a single intervention or treatment.[9]

More common is the situation where two or more interventions are being compared, in which case the ICER is the more useful statistic for policy makers. The ICER presents the change in cost per unit change in effect. For example, comparing two interventions (A and B) assuming intervention B is more expensive, the ICER for this comparison can be expressed as:

$$\text{ICER} = \frac{\text{Total cost(B)} - \text{Total cost(A)}}{\text{Health outcomes(B)} - \text{Health outcomes(A)}}$$

CEA results present decision makers with quantifiable trade-offs between costs and the health effects of the interventions being compared. This technique can be used for final health outcomes (e.g., patients survived or lives saved), and this technique can also be used for intermediate outcomes (e.g., number of patients who quit smoking). Intermediate outcomes are often easier and more rapidly measured. However, intermediate health outcomes should be clearly and demonstrably linked to final health outcomes to be most useful (e.g., measuring patients who quit smoking can be directly linked to lung cancer cases prevented). CEA is most appropriate when comparing health interventions with similar health outcomes. This technique cannot be used to contrast health interventions with diverse health benefits (e.g., comparing a tobacco cessation program with a prenatal care campaign).[9–12] In these situations, a specialized type of cost-effectiveness analysis utilizing a more complex health outcome is warranted.[13]

Example: Cost-Effectiveness Analysis of Smoking Cessation Programs

A recent study by Secker-Walker et al. used a cost-effectiveness analysis to assess a smoking cessation research project funded by the National Institutes of Health (NIH).[14] The project, called Breathe Easy, was conducted in four counties (two in Vermont and two in New Hampshire) and used community-based interventions targeting women aged 18–64. One county in each state, matched on salient demographic characteristics, served as a control group. Based on telephone surveys, the authors calculated decreases in the number of adult women who smoke and calculated years of life saved from standard

mortality tables. Costs were calculated from two different perspectives—intervention costs only and then from a societal perspective including evaluation costs and indirect costs. Because of the duration of the campaign, all costs were adjusted to 2002 dollars using the consumer price index (CPI). The authors computed cost-effectiveness ratios expressed as dollars per life-year saved, using multiple discount rates. Using a discount rate of 5%, the authors reported a cost-effectiveness ratio of $1922 per life-year saved (LYS) using the intervention only perspective, and $6683 per LYS when evaluation and indirect costs were included. Cost-effectiveness ratios computed with no discounting or with a 3% discount rate were not statistically significant. The authors' findings are similar in magnitude to other cost-effectiveness analyses conducted over the previous 20 years.[13]

Cost-Utility Analysis

Cost-utility analysis (CUA) is a specific type of cost-effectiveness analysis in which the health outcomes are measured in terms of some type of adjusted health utility, the most common measure being the quality-adjusted-life-year (QALY). The outcome of a QALY (and other related measures) requires some assessment of the perceived quality of life, either from a patient or a societal perspective. To use QALYs as an example—living in perceived perfect health would be valued as a health utility of 1.0 while a patient living in severe pain might value this condition as 0.5. In a CUA, an intervention that prevented someone from living for two years of life at this diminished capacity (0.5) would have the same value as one that saved one year of life in perfect health. The advantage of CUA is that different types of health interventions can be compared as long as their outcomes can be expresses in terms of QALYs. The technique combines measures of both morbidity and mortality as well as a quality-of-life assessment of a single condition or disease. As its popularity grows, this type of analysis is becoming more frequently termed a cost-effectiveness analysis in contemporary literature.

Although measuring changes in longevity has been relatively straightforward, the assessment of health-related quality-of-life (HRQOL) represents a more complex and more recent construct in these types of assessments. Quality-of-life measures (i.e., utility values) can be derived by a variety of means; frequently, these must be measured using patient surveys or a variety of role-playing scenarios based on the assumptions of game theory.[10,12,15] These surveys can be either a general index designed to measure several dimensions of health (e.g., physical health, social functioning, and mental health), or they may be disease specific. When the goals of an analysis are appropriate to a national or societal perspective, a broad-based general quality-of-life survey is recommended. Commonly used general survey instruments include the European Quality-of-Life instrument (EQ-5D), Quality of Well-Being scale (QWB), or the Medical Outcome Study Short Form (SF-6D). A recent report of the Institute of Medicine's (IOM's) Committee to Evaluate Measures of Health Benefits for Environmental, Health, and Safety Regulation provides recommendations for the use and selection of HRQOL instruments in a regulatory environment.[13] The committee suggests that (1) general health assessment tools should be used; (2) they should be sufficiently sensitive to detect differences in health status; (3) HRQOL be derived from sufficiently

large samples so as to be meaningful; (4) the instrument should be acceptable to users and the public; and (5) the instrument should be practical and inexpensive to use.[13]

Example: Cost-Utility Analysis of HIV Counseling, Testing, and Referral Programs

A contemporary application of cost-utility analysis was conducted by Paltiel et al. for the evaluation of expanded HIV counseling, testing, and referral (HIVCTR). The authors used simulation modeling to determine the effects of expanded HIVCTR in target populations with different levels of HIV prevalence. Three levels of HIV prevalence were selected:

1. 3.0% prevalence of undiagnosed HIV and 1.2% annual incidence
2. 1.0% prevalence of undiagnosed HIV and 0.12% annual incidence
3. 0.1% prevalence of undiagnosed HIV and 0.01% annual incidence

The second level represents the HIV prevalence at which the CDC had recommended routine HIV testing and the third represents the US general population. The authors use costs and quality-of-life assessments from the national AIDS Cost and Services Utilization Survey and the HIV Cost and Services Utilization Study.[16–17] The authors also use well-established guidelines for HIV testing procedures, testing results, and the effectiveness of pre- and posttest HIV counseling services. Sensitivity analyses were used to test the effects of key assumptions on the cost-effectiveness ratios (reported as dollars per quality-adjusted life-years (QALYs). In the highest risk populations, the addition of one-time screening using enzyme-linked immunosorbent assay (ELISA) for all resulted in a cost-effectiveness ratio of $36,000 per QALY. Testing every 5 years had an estimated cost-effectiveness ratio of $50,000 per QALY. In the CDC threshold prevalence population, the authors estimated a one-time screening with ELISA to result in a ratio of $38,000 per life year gained. In the US general population, one-time screening costs were estimated to be $113,000 per QALY. The authors conclude that in all but the lowest risk populations, both one-time screening as well as periodic screenings may be cost-effective based on their cost per QALY when compared to generally accepted screening interventions for other chronic conditions (e.g., cancer, diabetes, and hypertension).[18]

Cost-Benefit Analysis

Cost-benefit analysis (CBA) is an economic analysis in which both the costs and the health outcomes are expressed in dollars. The results of a CBA are thus expressed in terms of a *net benefit* representing the difference between all inputs (costs) and all outcomes (both positive and negative health consequences). The benefit/cost ratio can also be computed as the ratio of the value of benefits divided by all appropriate costs. The benefit/cost ratio can be used to rank multiple projects (with positive net benefits). This facilitates prioritization of various inteventions where resources are finite.[9,19]

Both direct and indirect costs may be included in a CBA. Because of the comprehensive nature of this method, CBAs are most often conducted from a societal perspective, particularly when evaluating a public health intervention. The expenses that are avoided because costly health problems are

prevented are also included in the valuations of program benefits. The subjective nature of some of these cost categories can sometimes make CBA findings controversial.[20]

Because all costs and outcomes are valued in dollars, CBA has the advantage of comparing vastly different health interventions and allowing decision makers to focus limited resources toward the optimal projects. However, this advantage is also the greatest limitation. CBA, by definition, requires the economic evaluation of morbidity, mortality, and quality of life. The methods used to arrive at these valuations are often debatable, and there is significant variation in the specific methods used. In addition, the evaluation of costs and benefits within a CBA design frequently aggregate data of several years (or decades) for the realization of a program's effectiveness. Thus, this method requires the adjustment of dollar values for inflation and time value (e.g., discounting). The rate selected for discounting can create large variations in a CBA's final results and therefore should be tested in sensitivity analyses.[9,19]

Example: Cost-Benefit Analysis of Folic Acid Fortification

A recent study by Grosse et al. updated the costs and benefits of folic acid fortification in the United States using both a cost-effectiveness and a cost-benefit analysis.[21] Folic acid fortification of cereal grain products has been required by the Food and Drug Administration (FDA) since 1996 in order to reduce the incidence of neural tube defects (NTDs). The authors present a thorough overview of the various economic studies that have been conducted on folic acid fortification since its implementation. For the most recent cost-benefit analysis, the authors used a cost-of-illness approach that placed a value on NTD deaths based on estimates of lost productivity in the future discounted to present day using a range of discount rates as directed by the recent Office of Management and Budget (OMB) guidelines.[21] The authors also estimate financial benefits based on the direct costs for NTDs that can be averted by folic acid supplementation. The authors estimated the lifetime costs associated with a spina bifida birth at $636,000 with $279,000 representing direct medical costs and the remaining $357,000 indirect costs. For anencephaly, the authors estimate a lifetime cost to be $1,020,000, with $1,014,000 attributed to indirect costs and $6000 the direct costs of hospitalization. For the cost-benefit analysis, the authors estimate a total economic benefit of folic acid fortification to be $422 million per year after adjusting for the cost the fortification (i.e., $3 million per year). The authors conduct a best case and worst case type of sensitivity analysis that adjusted for several of their assumptions. Under their worst-case scenarios, the authors still report a net benefit of folic acid fortification of $312 million per year.[21]

Measurement Issues Pertinent to Economic Evaluations

Because of the variation in the types of economic analyses, the assumptions and variability of their methodologies and components, researchers have sought to create some standards for economic analyses in order to improve the usefulness to decision makers. One such group was the Panel on Cost-Effectiveness in Health and Medicine, an expert-appointed panel assembled by the US Public Health Service. The panel was charged, in part, with trying to

achieve a consensus for standard components of healthcare cost-effectiveness studies and with proposing generally accepted methods for addressing the assumptions inherit in these types of analyses. A few of the most important standards for these analyses are described below.[22]

Study Perspective

Economic analyses can be conducted from a variety of perspectives or points of view, based on what entity incurs costs or acquires benefits (e.g., hospital or providers, insurers or payers, or society as a whole). The choice of perspective dictates which costs and health outcomes should be evaluated in an analysis. It is thus critical for the study perspective to be selected appropriately for the goals of an evaluation and declared prominently in any published results.[23,24]

The societal viewpoint represents the most comprehensive perspective and includes all direct medical and nonmedical costs, indirect costs, patient and family out-of-pocket costs, as well as costs or benefits that may extend beyond the intended populations. The benefits of an immunization program illustrate this—as a larger proportion of a vulnerable population becomes immunized, economic benefits are realized by everyone, even the remaining unimmunized because of the decreased risk in the entire population (e.g., effect of herd immunity). The benefits of employing the societal perspective is that, in theory, all resources are accounted for, thus controlling for costs that may only be shifted among providers and payers. Some decision makers may select a more narrow perspective for their own organization. However, Gold et al. suggests that any narrow study perspective be combined with an analysis from a societal perspective.[22] The societal perspective is thus the recommended perspective particularly when a healthcare evaluation deals with publicly funded health interventions and prevention activities.[23,24]

Time Horizon

The benefits of a health intervention often may take months or years to observe, while the costs which are incurred may be immediate. This is most critical in healthcare prevention activities, where benefits may occur far in the future. Thus it is important for economic analyses to use an appropriate time horizon that is sufficient to comprehensively capture all appropriate costs and benefits. For example, a diagnostic assay that predicts breast cancer recurrence over 10 years cannot be measured in an analysis with only a 5-year follow-up. Obviously, any gains in life expectancy would be lost with this truncated time horizon. Experts recommend that changes in life expectancy should be modeled to account for changes in survivals. When the period of potential benefits extends beyond the feasible data collection period, researchers frequently must rely on theoretical modeling.[23,25]

Discounting

Discounting is commonly used to adjust costs that may accrue over several years in order to present a common basis for comparison. Typically, costs

paid in the distant future are valued lower than present costs. It is commonly recommended that costs expended over more than one year be adjusted by using discounting. Discounting however is more controversial when applied to the accrual of health benefits. Some argue that a year of life saved should carry the same value whether it accrues in the present or in the distant future. Most economists, however, have recommended that future health benefits be discounted in the same way that costs are discounted within an economic analysis.[26]

The fundamental issue with discounting is the choice of discount rate at which costs or benefits are adjusted. The choice of discount rate can frequently affect the final results (i.e., whether an economic analysis shows cost savings or cost increases). The Panel on Cost-Effectiveness recommends that the appropriate discount rate be consistent with the contemporary cost of capital, generally between 3% and 5%, but that sensitivity analyses should be used to test rates between 0% and 7% to assess the impact of the discount rate on the final conclusions.[22,27]

Sensitivity Analyses

Because there are a number of assumptions inherent with many economic analyses, experts recommend that the impact of these assumptions on the final conclusions be tested using sensitivity analyses. This technique requires researchers to identify the plausible ranges for the values of key assumption values and recalculating the study results based on these multiple values. Several types of sensitivity analysis are commonly employed.[22,27]

The most common type of sensitivity analysis modifies one or more variables across reasonable values. Such simple sensitivity analyses can be one-way (e.g., one variable) or multiple-way (e.g., multiple variables). A threshold sensitivity analysis is similar to a one-way sensitivity analysis; however, the values of the assumptions are varied to the point at which the options being compared become equivalent or the point at which the winning strategy changes. Because often several variables need to be examined with sensitivity analysis, the possible combinations can become unwieldy. In this situation, an analysis of extremes can be employed. This type of sensitivity analysis uses a best case and worst case approach whereby all of the lowest cost and highest benefit assumptions are used (best case) and compared to a scenario using the highest cost and lowest benefit assumptions (worst case).

A relatively recent addition to sensitivity analyses is a Monte Carlo simulation. This technique allows researchers to prespecify multiple ranges of values and to describe the underlying mathematical distributions of the variables used in an analysis. Modeling software then randomly selects values from the specified distributions of each variable. Thousands of iterations can quickly calculate results based on these selected values, thus creating confidence intervals around the mean costs and benefits in an economic modeling situation.[22,27]

Guidelines for the Assessment of Economic Analyses

Several authors have offered specific criteria to audiences of health economic analyses to allow for their assessment of quality and objectivity. Drummond

et al. presents a list of 10 questions, the answers to which can be used for the assessment of a published economic analysis.[28] The Panel on Cost-Effectiveness in Health and Medicine designed a checklist for authors of economic analyses that should be addressed in the preparation of their work.[29] Mullins et al. provides a summary of several sets of recommendations cited among major economic analyses in the healthcare setting as a set of the following six principles:

1. An explicit statement of the study perspective should be provided.
2. A detailed description of the benefits of the program or technology should be provided.
3. Researchers should specify what types of costs were used or considered in their analysis.
4. Discounting should be used to adjust for the differential accruals of costs and benefits.
5. Sensitivity analyses should be performed to test important assumptions.
6. In the presence of multiple alternatives, summary measures should be expressed as marginal or incremental ratios.[30]

Research Synthesis

Because replication is an important part of the scientific process, a systematic review of existing studies—*research synthesis*—can provide a tool for understanding variations and similarities across studies. It can also uncover robust intervention effects. Before implementing a new program, careful public health managers check the evaluation literature to ensure that the intervention has been shown to be effective in other settings. Because such literature reviews often reveal a confusing range of different findings in evaluation studies of varying quality, techniques such as *meta-analysis* and more generally, research synthesis, are increasingly used in public health.[30,31,32,33]

Synthesis of research findings offers the potential to identify areas of agreement and to identify areas needing more research. Synthesis essentially involves a state-of-the-art literature review, presenting and analyzing the available data, and framing results so they can be translated into practice and policy. Meta-analysis is a subset of research synthesis that employs special statistical analyses of a collection of results from individual studies for the purpose of integrating the findings. This analysis can increase the statistical precision of the estimates of a program's effect.

In a research synthesis or meta-analysis, the individual study results are the raw data. Thus, in order to avoid bias, an a priori protocol for the selection of studies to be included and their analyses is needed. Search strategies should include bibliographic sources such as the National Library of Medicine (accessible through a medical library or www.nlm.nih.gov). Searching the bibliographies of review articles and studies at hand, as well as asking experts in the field for additional references, are also effective methods of research synthesis.

Once the relevant studies are identified, they can be presented through a narrative summary of each article or by an evidence table that lays out key aspects of each study, including the publication date, study population, study design and sample size, definitions of the intervention and of outcome

measures, and results. When the available studies are sufficiently similar, statistical summaries can be prepared, as illustrated in the following paragraphs.

As part of the efforts of the Community Preventive Services Task Force, for instance, researchers from the Centers for Disease Control and Prevention (CDC) identified and reviewed the effectiveness of population-based efforts to improve vaccination coverage.[34] The interventions studied included efforts to increase community demand for immunizations such as patient reminder/recall systems; programs to enhance access to immunization services by reducing out-of-pocket costs, for example; immunization mandates at school, child care, and college entry; and provider-based strategies such as provider reminder/recall systems and the assessment of immunization rates and feedback for vaccination providers. A systematic literature search yielded 126 studies of such interventions. Following a standard approach, the researchers characterized the body of evidence as strong, sufficient, or insufficient based on the numbers of available studies, the strength of their design and execution, and the size and consistency of reported effects. This analysis then formed the basis for the recommendations of the task force.[35]

Example: Efficacy and Effectiveness of Influenza Vaccines in the Elderly

In another example, researchers evaluated the efficacy and effectiveness of influenza vaccines in elderly people by identifying 5 randomized, 49 cohort, and 10 case-control studies assessing efficacy against influenza (reduction in laboratory-confirmed cases) or effectiveness against influenza-like illness (reduction in symptomatic cases).[36] Figure 18-1 summarizes the analysis of studies comparing vaccination with no vaccination for prevention of deaths caused by influenza or pneumonia in residents of long-term care facilities. Each line corresponds to one study, labeled by author. The box in the center of each line represents the study's estimate of the vaccine's effectiveness (the size of the box is proportional to sample size); the length of the line represents a 95% confidence interval for the estimate. Because the results might have differed according to the level of virus in circulation and how well the vaccine used matched the circulating viral strain, the studies were grouped according to these variables. In the lines headed "Subtotal" or "Total," the center of the diamonds represents the combined estimate of the studies above it and the width of the diamond represents the 95% confidence interval (CI) on the combined estimate. Although most of the individual studies do not show a significant reduction in risk because the 95% confidence line includes the null value of 1.0, the "Total" analysis suggests that the vaccine has a significant effect on the prevention of deaths caused by influenza or pneumonia. The overall relative risk is estimated as 0.46 with a 95% CI (0.33, 0.63), suggesting a reduction in the risk of death by more than half. Although the results in the four subgroups defined by level of viral circulation and quality of vaccine matching vary, the tests for heterogeneity do not suggest that the differences among them are significant. The results lead the authors of the review to conclude: "In long-term care facilities, where vaccination is most effective against complications, the aims of the vaccination campaign are fulfilled, at least in part. However, the usefulness of vaccines in the community is limited."[36]

In other instances, quantitative combination of results is simply not feasible because the available studies are too dissimilar. A National Research

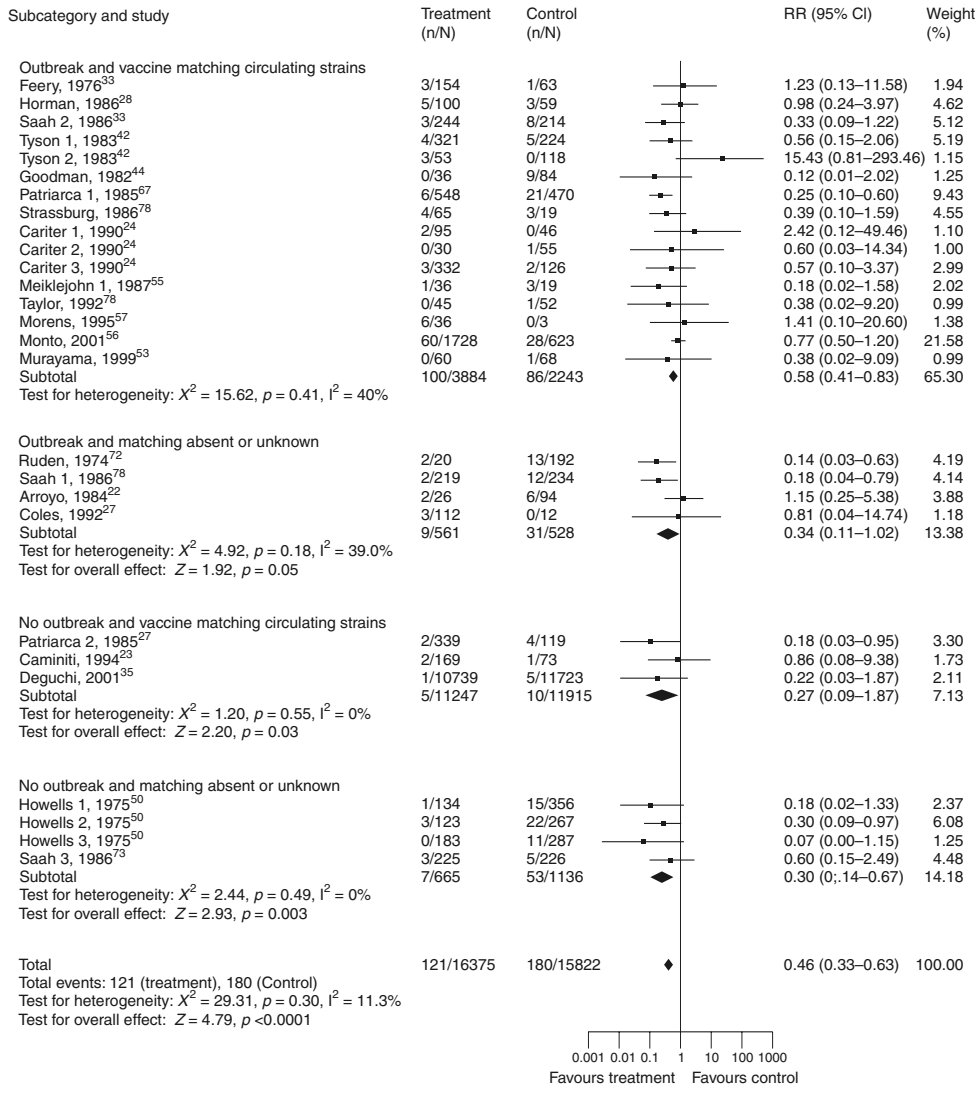

FIGURE 18-1 Influenza Vaccine Compared with No Vaccination for Prevention of Deaths Caused by Influenza or Pneumonia in Residents of Long-Term Care Facilities by Level of Viral Circulation and Quality of Vaccine Matching. Number after Names of Authors Indicates Different Databases.

Source: Reprinted from Jefferson T, Rivetti D, Rivetti A, Rudin M, DiPietrantonj C, Demicheli V. Efficacy and effectiveness of influenza vaccines in elderly people: a systematic review. *Lancet.* 2005; 366:1165–1174.

Council (NRC) evaluation of the evidence surrounding the efficacy of needle exchange programs to prevent HIV transmission provides an example.[37] The logic of this approach is clear—needles are an important vector of transmission among drug users because they are passed among users, so introducing clean needles into circulation should avert new infections. When asked by Congress to review the evidence on this issue in 1995, the NRC found dozens of evaluation studies that had been prepared. Each of these

studies, however, addressed different aspects of the problem: how needles can be distributed when state laws prohibit their possession, the costs and logistics of distributing needles and condoms, knowledge of HIV risk factors, various changes in users' needle use and other HIV risk factors, referral to drug treatment, and so on. Other studies examined the concerns of community leaders and individuals living near the needle exchange sites. Only a handful of reports examined HIV infection rates, which are difficult to study because the annual number of new infections in any study group is generally rather small. Moreover, many of the existing studies were of poor quality, which is not surprising given that many of the existing programs were not legally sanctioned.

There were, however, two groups of studies that provided useful information. In New Haven, Connecticut, researchers had been evaluating a needle exchange program run out of mobile vans.[38] One unique aspect of this program was that every needle distributed was marked with an identifying number, and each needle returned was checked using biomedical measures for exposure to HIV. These two pieces of information showed that the program reduced the infectivity of needles in circulation by approximately one third. Coupled with information from a survey of the needle exchange users, the researchers used a mathematical model to show that a needle exchange program could reduce the rate of new infections by approximately one third.

Another series of studies took advantage of the fact that one of the first needle exchange programs in the United States was organized in Tacoma, Washington, which had a preexisting enhanced hepatitis B surveillance program.[39] Hepatitis B is transmitted through blood products and sexual activity, just as HIV, but has a higher infectivity rate and shorter latency period. As a result, a group of studies in Tacoma were able to establish that needle exchange programs were effective in preventing the transmission of a blood-borne virus.

The NRC panel reviewing this evidence used a logic model to synthesize the evaluation evidence. First, a series of process studies showing that needles could be distributed efficiently ruled in the plausibility of the idea. The New Haven studies showed that needle exchange programs can increase the availability of clean equipment and reduce HIV prevalence in needles in circulation. Logically, one expects that decreasing the fraction of needles in circulation that are contaminated will lower the risk of new HIV infections, and the Connecticut model quantified this effect. The Tacoma studies confirmed the logical analysis by showing that needle exchange programs can reduce the incidence of a blood-borne disease. The NRC panel concluded, therefore, that needle exchange programs can reduce the risk of HIV infection.

Measurement

Measurement is central to evaluation. (Refer to Chapter 17 for detailed discussion of measurement.) Evaluations of program effectiveness can only assess the impact on the outcomes that have been measured, and measures of program inputs are critical for interpreting the results. The importance of measurement in performance improvement is clear from the management aphorism: what gets measured gets done. The development of measures for any evaluation involves the following four steps.

Clarify the Goals and Purposes of the Evaluation

In general, the goals and purposes of an evaluation determine the types of measures that are needed. Outcome evaluations need measures of health outcomes, whereas feasibility evaluations must focus on costs and barriers to implementation. Evaluations of programs intended to be exported to other venues must include measures of the specific intervention so that it can be replicated. Evaluations based on quasi-experimental designs require careful measures of confounding factors for adjustment purposes. Efforts to ensure accountability often require financial measures.

As illustrated in Figure 18-2, various disciplinary lenses produce different approaches to health promotion and disease prevention. At the micro-level, the *biomedical* lens focuses on biophysiological theories of disease causation and turns to biomedical interventions for solutions. The *psychosocial* lens focuses on the individual, investigating questions about individual and social behaviors such as self-efficacy and control. The *epidemiologic* lens examines disease patterns in populations and identifies differential risk factors, both biologic and environmental. The *society-and-health* lens aims to understand the way that cultural, social, economic, and political processes influence differential risks. The choice of lens underlying the intervention obviously determines the nature of the evaluation: what is measured, and so on.

An explicit "logic model" describing the logical sequence of events that connect an intervention to the desired change can be valuable in evaluating complex interventions or simple interventions in complex causal chains.

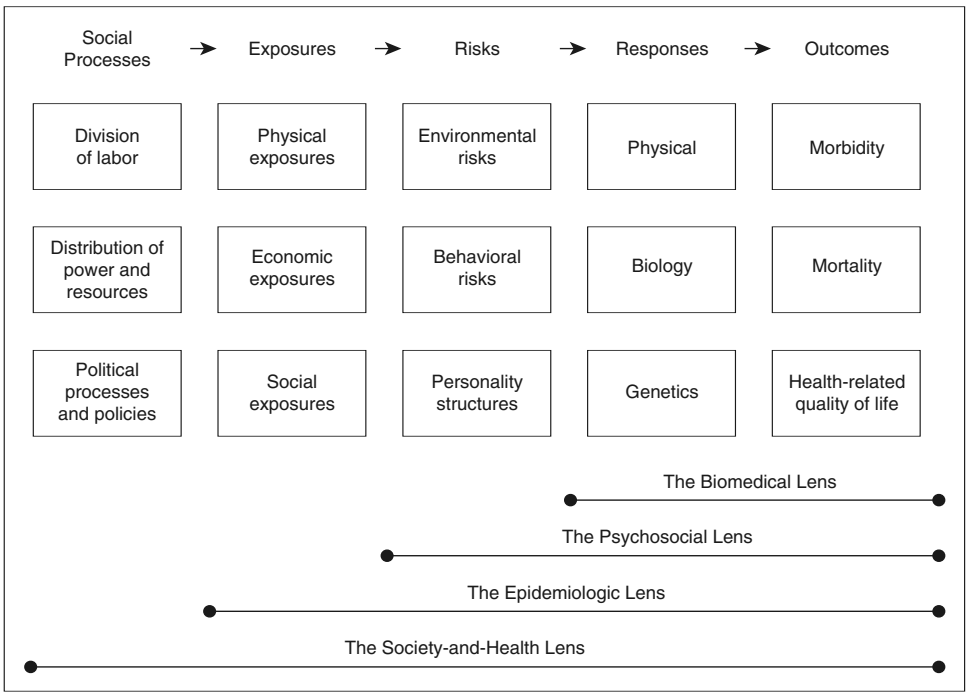

FIGURE 18-2 Alternative Disciplinary Lenses for Factors Influencing Health Outcomes
Source: Adapted with permission from *Annual Review of Public Health*, Vol 19, © 1998 by Annual Review, www.AnnualReviews.org.

Figure 18-3 illustrates a logic model for a community-based immunization program—part of a larger effort to reduce the impact of vaccine-preventable diseases. The upper component of the figure illustrates the different types of specific interventions that are possible: community-based educational programs designed to increase community demand for vaccinations, interventions to enhance access to healthcare settings, and provider-based interventions to increase the use of vaccines among those who have access to care.

Evaluations of community-based interventions require measures of awareness and access in the community. Evaluations of provider-based interventions, on the other hand, require only measures of covered populations. A beneficial effect of vaccine coverage on vaccine-preventable disease and associated morbidity and mortality is assumed, based on previous clinical studies of the vaccines themselves. Intermediate measures of vaccine coverage, and not of morbidity and mortality, are thus sufficient for evaluations on efforts to improve immunization rates. Other interventions to reduce mortality and morbidity are possible (medical treatment of individuals who contract vaccine-preventable diseases, efforts to reduce contacts between infectious and noninfectious individuals). These are not, however, usually treated in evaluations of immunization programs.

Identify the Concepts to be Measured

Depending on the nature of the intervention and the purpose of the evaluation, measures should be chosen to reflect the logic of the process. In the immunization example in Exhibit 18-1, for instance, measures need to be developed for each of the major inputs and activities (availability of a registry), as well as the intermediate outcomes (births enrolled in a registry) and final outcomes (in-

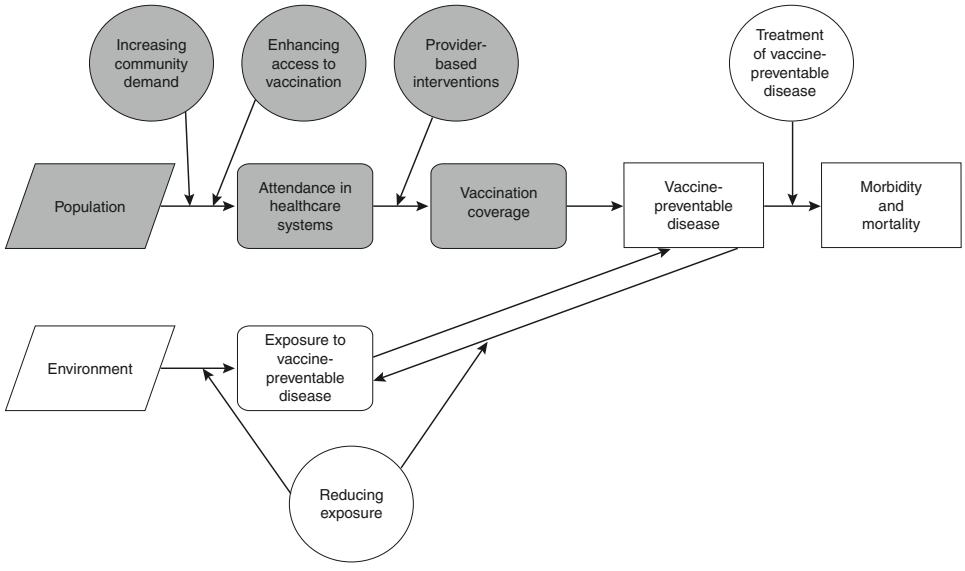

FIGURE 18-3 Evidence Model for Immunization Program Evaluations
Source: Adapted from US Centers for Disease Control and Prevention. Framework for program evaluation in public health. *MMWR.* 1999;RR-11: 1–40.

creased numbers of children immunized). In performance improvement processes (as discussed later), it is important to identify responsibility for specific activities and to choose measures that enable accountability for performance. It is also important to choose a set of measures that achieves a balance between short- and long-range goals and among levels and types of service.[40]

Complex programs require a variety of measures relating to the logic of the intervention. As a condition of funding, the Health Resources and Services Administration (HRSA) maternal and child health programs, for example, require states to report their yearly performance on a number of measures. States can choose from a variety of standard "health system capacity measures" and "health status indicators" as illustrated in Table 18-1. A group of 18 "national performance measures" covering both process and outcomes is required of all states.[41]

Develop Specific Indicators for Each Concept

General concepts such as immunization coverage and low birth weight are not sufficient for program evaluation because they can be operationalized in many ways. Careful evaluation requires that each concept be measurable by one or more specific indicators operationally defined in an unambiguous way.

Finding indicators that faithfully represent the critical concepts and that can be calculated in a timely way from available data is often a challenge.

TABLE 18-1 Examples of Specific Indicators for State Maternal and Child Health Programs

Concept	Specific Indicator
National performance measures	
Insurance coverage	The percentage of children with special health care needs age 0 to 18 years whose families have adequate private and/or public insurance to pay for the services they need
Insurance coverage	Percentage of children without health insurance
Adequate prenatal care	Percentage of infants born to pregnant women receiving prenatal care beginning in the first trimester
Immunization coverage	Percentage of 19 to 35 month olds who have received full schedule of age appropriate immunizations against measles, mumps, rubella, polio, diphtheria, tetanus, pertussis, haemophilis influenza, and hepatitis B.
Health system capacity measures	
Eligibility for publicly funded insurance	The percentage of poverty level for eligibility in the state's Medicaid and SCHIP programs for infants (0 to 1), children, and pregnant women.
Health status indicators	
Low birth weight	The percentage of live births weighing less than 2500 grams
Very low birth weight	The percentage of live births weighing less than 1500 grams

Source: Maternal and Child Health Bureau. *Maternal and Child Health Services Title V Block Grant Program: Proposed National Performance Measures, Health Systems Capacity Indicators, and Health Status Indicators.* US Department of Health and Human Services. Available at: http://mchb.hrsa.gov/grants/proposal.html.

Mortality, for instance, can be measured through general mortality rates or through disease-specific rates, which are available on a less timely basis. Healthcare costs can be measured by hospital and physician charges, but these may not accurately reflect the opportunity costs of these interventions consistent with economic theory. Quality of health care is sometimes measured through consumer satisfaction surveys, but such measures reflect only part of what policy analysts define as quality.[42] Table 18-1 illustrates the correspondence between some of the HRSA performance measure concepts and specific indicators for state maternal and child health programs.

Assess the Performance of the Proposed Indicators with Respect to Validity, Reliability, and Sensitivity to Change

Before an evaluation process goes forward, the indicators to be used must be assessed, especially in terms of validity and reliability. *Validity* is an indicator's capacity to measure the intended concept. *Reliability*, on the other hand, assesses whether the indicator consistently measures the concept. The relationship between the two is illustrated graphically in Figure 18-4.

Sensitivity to change assesses an indicator's ability to measure change that might be attributed to the intervention being evaluated. Some errors are related to chance fluctuations in epidemiologic rates. For most communities, for example, infant mortality rates fluctuate substantially from year to year simply because the numerator, the number of infant deaths, is small. Statistical measures can be used to assess the degree to which the indicator changes if and only if the concept being measured also changes. A common problem is when service records are used to assess changing disease burdens. Does a decrease in emergency department visits for asthma indicate the success of a prevention program or measures to restrict access to individuals without insurance?

Compromises must generally be made among validity, reliability, data availability, and sensitivity to change. In the area of prenatal care, for example, evaluators often use the receipt of prenatal care in the first trimester, rather than more complex measures based on official recommendations of the US Public Health Service for the frequency and content of prenatal care, because the former measure is available on birth certificates and the latter is not.[43] This is a case of trading validity for increased data availability. In many communities,

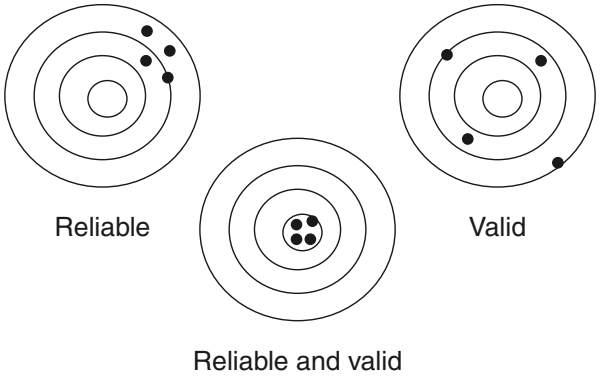

FIGURE 18-4 Illustration of the Concepts of Reliability and Validity

annual infant mortality rates are not reliable because of the small number of infant deaths. Instead of annual rates, therefore, epidemiologists commonly calculate running averages by averaging infant mortality rates over three or five years. This is a case in which reliability is gained at the expense of timeliness and responsiveness to change. Another approach that is frequently used to deal with sparse data is to use proxy measures that reflect trends and differences. The percentage of infants born at low birth weight, for example, is used rather than infant mortality because low birth weight has been shown to be strongly associated with infant mortality. This is a case of trading validity for reliability.

Data Sources

Data for evaluations in public health can be obtained from a wide variety of sources. (See Chapter 11 for an extensive description of various data sources.) The extended tobacco control example at the end of this chapter also illustrates the variety of data sources that can be accessed for evaluation efforts.

Capacity, process, and implementation measures can come from administrative records, existing reports and documents, and persons involved with the program. Administrative records provide good information on resources that are available and expended, the number of service providers assigned to a program, the number of clients served and the nature of services provided, and so on. Data of this sort can sometimes be derived from the management information systems used to run the program itself. In other cases, information can be obtained from clients, program staff, and others through surveys and focus groups. Documents such as grant proposals, newsletters, annual reports, and asset and needs assessments can also provide useful data for evaluations.

Outcome measures can also be derived from a variety of data sources, including vital statistics (birth and death records) and public health surveillance programs. Administrative and medical records from the healthcare system, because they are increasingly computerized and thus accessible, can also provide useful outcome measures. General- and special-purpose population surveys and client surveys can also provide useful information for program evaluation.

Practical Aspects of Program Evaluation

Improving the standards of evidence used in the evaluation of public health interventions is one of the most critical needs in this field. More rigorous studies are needed to better determine "what works," "for whom," "under what conditions," and "at what cost." According to a National Research Council/Institute of Medicine (IOM) report on family violence programs,[44] which can be generalized to many other areas in public health, the road to improvement requires attention to: (1) assessing the limitations of current evaluations, (2) forging functional partnerships between researchers and service providers, (3) addressing the dynamics of collaboration in those partnerships, and (4) exploring new evaluation methods to assess comprehensive community initiatives.*

*This section draws heavily on the contributions of David Cordray to the NRC report *Violence in Families*.[44]

Interventions undergo an evolutionary process that refines theories, services, and approaches over time. In the early stages, interventions generate reform efforts and evaluations that rely primarily on descriptive studies and anecdotal data. As these interventions and evaluation efforts mature, they begin to approach the standards of evidence needed to make confident judgments concerning effectiveness and cost. For instance, current policy discussions are focused on determining the effectiveness or cost-effectiveness of selected programs or strategies, conclusions that require high standards of evidence. Existing evaluation studies, however, consist mostly of nonexperimental study designs, and thus provide no firm basis for examining the impact of programs or for considering the ways in which different types of clients respond to an intervention. Nonexperimental studies, however, can reveal important information in the developmental process of research. They can illuminate the characteristics and experience of program participants, the nature of the presenting problems, and the issues associated with efforts to implement individual programs or to change systems of service within the community. Although these kinds of studies cannot provide evidence of effectiveness, they do represent important building blocks in evaluation research.

A similar developmental process exists on the programmatic side of interventions. Many family violence treatment and prevention programs, for example, have their origins in the efforts of advocates who are concerned about children, women, and the family unit. Over several decades of organized activity, these efforts have fostered the development of interventions in social service, health, and law enforcement settings that program sponsors believe will reduce violent behavior or improve the welfare of victims. Some programs are based on common sense or legal authority; others are based on broad theories of human interaction or theory borrowed from other areas. All of these interventions were preceded by research studies that identified risk factors or critical decision points in the intervention process. As programs mature and become better articulated and implemented, evaluation questions and methods become more sophisticated and complex.

It is difficult for researchers to establish good standards of evidence in service settings because they cannot exert complete control over the selection of clients and the implementation of the interventions. But several strategies have emerged that can guide the development of evaluation research as it moves from its descriptive stage into the conduct of quasi-experimental and true experimental studies. An important part of this process is the development of a "fleet of studies."[44,45] The NRC evaluation of the effectiveness of needle exchange programs discussed earlier provides one example. Although each individual study of a given project was insufficient to support a claim that the needle exchange program was effective, the collective strengths of the studies taken together provided a basis for a firm inference.

To overcome the technical challenges of conducting evaluations in public health service settings, several steps have been suggested.[44] Most importantly, research and evaluation need to be incorporated earlier into the program design and implementation process.[46] The use of innovative study designs, such as empowerment evaluation, described elsewhere in this chapter, can provide opportunities to assess the impact of programs, interventions, and strategies. Drawing on both qualitative and quantitative methods, these approaches can

help service providers and researchers share expertise and experience with service operation and implementation.

Integration of the evaluation and practice perspectives requires creative collaboration between researchers and providers who are in direct contact with the individuals who receive services and the institutions that support them. Numerous points in the research and program development processes provide opportunities for such collaboration.[47] Practitioners, for example, have extensive knowledge of the needs of clients and the nature of existing services in the community. Service providers' knowledge of details concerning client flow, rates of retention in treatment, organizational capacity, and similar details are useful in developing new interventions and a logic model that provides the framework for the evaluation. Their participation can highlight differences between new services and usual-care situations, which are often a matter of degree. Practitioners can also help to ensure that the outcomes assessed are the ones of concern. Finally, service providers have knowledge of other services and factors in the community that affect outcomes of interest.

The dynamics of collaborative relationships between evaluators and program managers require explicit attention and team-building efforts to resolve different approaches and to stimulate consensus about promising models of service delivery, program implementation, and outcomes of interest. Creative collaboration requires attention to the following issues:[44]

- Setting up equal partnerships—Tensions between service providers and researchers may reflect differences in ideology and theory about the issues being addressed or mutual misunderstandings about the purpose and conduct of evaluation research. Frontline service providers may resent the time and resources that research takes from the provision of services. True collaborative partnerships require a valuing and respect for the work of all sides. Both sides need to spend time observing each other's domains in order to better their constraints and risks.

  Recent collaborations in the evaluation of family violence interventions illustrate opportunities to address these concerns. Community agencies are beginning to realize that well-documented and soundly evaluated successes will help ensure their financial viability and even attract additional financial resources to support promising programs. Researchers are starting to recognize the accumulated expertise of agency personnel and how important they can be in planning as well as conducting their studies. Both parties are recognizing that, even if research fails to confirm the success of a program, the evaluation results can be used to improve the program.
- The impact of ethnicity and culture on the research process—Ethnicity and cultural competence influence all aspects of the research process and require careful consideration at various stages: formulation of hypotheses taking into account known cultural or ethnic differences, large enough sampling sizes to have enough power to determine differential impact for different ethnic groups, and analytic strategies that account for ethnic differences and other measures of culture. In evaluating family violence interventions, for example, there is a need for researchers who are knowledgeable about cultural practices such as

parenting and caregiving, child supervision, spousal relationships, and sexual behaviors in specific ethnic groups.

- Exit issues—The ideal relationship between a research team and the service agency is long term and sustained between large formal evaluations. Such informal collaborations can help researchers, for example, in establishing the publications needed for large-scale funding. Dissemination of findings in local publications is also helpful to the agency. Successful collaboration requires that all partners decide on the authorship and format of publications ahead of time. Thoughtful discussions are also needed before launching an evaluation about what will be released in terms of negative findings and how the findings will be used to improve services.

  Another concern is the continuation of services when research resources are no longer available. Models of reimbursement and subsidy plans are needed to foster positive partnerships that can sustain services that seem to be useful to a community after the research evaluation has been completed.

Evaluation of Comprehensive Community-Based Interventions

In recent years, public health researchers have developed a series of comprehensive community-based interventions that reflect the growing appreciation of the social determinants of health and health-related behavior.[48,49] (See Chapter 19 for a more detailed discussion of community-based prevention.) These interventions take place in schools, work sites, and even whole communities. They typically address smoking, diet, exercise, and other behavioral risk factors for cancer and cardiovascular disease.[4] Because of their complexity and scale, however, such interventions present special challenges to evaluators.

Moreover, public health interventions increasingly involve multiple services and the coordinated actions of multiple agencies in a community. The increasing prevalence of coexisting problems such as substance abuse and family violence or child abuse and domestic violence, for instance, has encouraged the use of comprehensive services to address multiple risk factors associated with a variety of social problems. In tobacco control, as illustrated below, the range of entities and activities involved is so great that in some senses, the community itself is the proper unit of analysis for evaluations.

As public health programs become a more integrated part of the community, the challenges for evaluation become increasingly complex.[43]

- Because participants receive numerous services, it is nearly impossible to determine which service, if any, contributed to improvement in their well-being.
- If the sequencing of program activities depends on the particular needs of participants, it is difficult to tease apart the effects of selectivity bias and program effects.
- As intervention activities increasingly involve organizations throughout the community, there is a growing chance that everyone in need will receive some form of service (reducing the chance of constituting an appropriate comparison group).

- As program activities saturate the community, it is necessary to view the community as the unit of analysis in the evaluation. Outcomes, however, are typically measured at the individual level. At a minimum, appropriate statistical models are needed to take the different levels into account.[6]
- The tremendous variation in individual communities and diversity in organizational approaches impede analyses of the implementation stages of interventions.
- An emphasis on community process factors (ones that facilitate or impede the adoption of comprehensive service systems), as opposed to program components, suggests that evaluation measures require a general taxonomy that can be adapted to particular local conditions.

Conventional notions of what constituted a rigorous evaluation design are not easily adapted to meet these challenges. Some authors have concluded that randomization is simply not feasible, and that conventional alternatives to randomization are technically insufficient.[50,51] Weiss proposed an alternative evaluation model based on clarifying the "theories of change" that explores how and why an intervention is supposed to work.[51] The evaluation should start with the explicit and implicit assumptions underlying the theory guiding the intervention efforts; this theory is generally based on a series of small steps that involve assumptions about linkages to other activities or surrounding conditions. By creating a network of assumptions, it is possible to gather data to test whether the progression of actions leads to the intended end point.

Other researchers note that the theory of change perspective provides some basic principles to guide collaborative evaluations.[52] First, the theory of change should draw on the available scientific information, and it should be judged plausible by all the stakeholders. Second, the theory of change should be doable—that is, the activities defined in the theory should be able to be implemented. Third, the theory should be testable, which means that the specification of outcomes should follow logically from the theory.

Community interventions are characterized by small relative effects. Strong interventions typically yield a 2% to 5% reduction in the prevalence of risk factors such as smoking or lack of exercise, or a similar percentage reduction in average serum cholesterol or blood pressure. As Geoffrey Rose observed, changes of this magnitude in entire populations are likely to have large effects on disease risk and the burden of morbidity and mortality.[53,54] Thus, from a public health perspective, the impact of an intervention is a product of both its efficacy in changing individual behavior and its reach, meaning the proportion of the population reached either through their direct participation or indirectly through diffusion of intervention messages throughout the community, work site, or school.[55] Anna Tosteson and colleagues, for instance, estimated that population-wide strategies to reduce serum cholesterol are cost-effective if cholesterol is reduced by as little as 2%.[56] Thus, although the effects of community interventions may appear small by standards of clinical research, these interventions can have a substantial public health impact.

Small effect sizes, however, create significant statistical difficulties in the evaluation of community interventions. Although the number of individuals

involved in community trials is often large, the number of units of allocation that are randomized (schools, work sites, or whole communities) is typically very small. Furthermore, the power and precision of statistical tests depends more on the number of units of allocation than on the number of individuals.[6] Taken together with small effect sizes, these features of community interventions make it very difficult to achieve statistical significance in conventional terms. In other words, a true 2% reduction in a risk factor, which has great public health significance, might not achieve statistical significance in a study with thousands of individuals in a small number of communities.

A number of statistical approaches may help resolve this problem. First, more efficient statistical methods are needed to improve investigators' ability to detect small differences. This can come through increasing the number of units of allocation in studies or making better statistical use of the existing information through, say, the use of appropriate hierarchical statistical models.[57] Alternatively, where separate interventions have used parallel methods in similar populations, meta-analysis can be useful in increasing statistical power. Alternatively, future interventions might be planned with enough parallelism that meta-analysis is appropriate after the individual results are available.

Empowerment Evaluation

Empowerment evaluation represents a new use of evaluation concepts, techniques, and findings to foster improvement and self-determination. It has its roots in education and social services but has many applications to public health. This form of evaluation draws on empowerment processes, in which attempts to gain control, obtain needed resources, and critically understand one's social environment are fundamental. It is designed to help people help themselves and improve their programs using a form of self-evaluation and reflection. Empowerment evaluation is necessarily a collaborative group activity, not an individual pursuit.[2]

As illustrated in Figure 18-5, empowerment evaluation involves six steps. These six steps are described below using the example of coalitions in three Kansas communities for the prevention of adolescent pregnancy and substance abuse.[58]

1. Take stock. Determine where the program stands, including strengths and weaknesses, and identify community concerns and resources. A series of listening sessions—informal public meetings in which individuals identified problems, barriers to addressing the problem, resources for change, and potential solutions—were held to engage key leaders, people affected by the problem, and people who could contribute to addressing the problem. The groups included religious leaders, youth, parents, teachers, health officials, and representatives from informal neighborhood groups and community organizations.
2. Focus on setting missions and establishing goals. Determine where you want to go in the future with an explicit emphasis on program improvement. In Kansas, the initial mission and goals were based on initiatives that had shown some success in reducing adolescent preg-

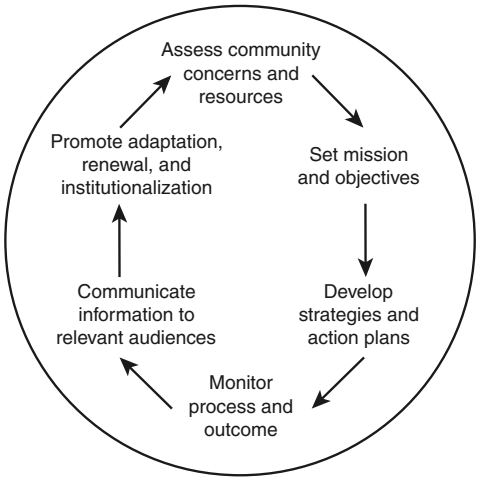

FIGURE 18-5 Process of Empowerment Evaluation
Source: Adapted with permission. Fawcett SB, Paine-Andrews A, Francisco VT, et al. Empowering community health initiatives through evaluation. In Fetterman DM, Kaftarian SJ, Wandersman A, eds. *Empowerment Evaluation: Knowledge and Tools for Self-Assessment and Accountability.* Thousand Oaks, Calif: Sage; 1996:170.

nancy and substance abuse in other Kansas communities. Following a shooting in one community, the coalition modified its objectives and action plan to reflect community concerns about youth violence associated with substance use.

3. Develop strategies and action plans to accomplish goals and objectives. Each community developed its own action plans, based on the model developed in other communities, consisting of proposed changes in programs, policies, and practices in a variety of sectors. Schools, for example, were to implement a "comprehensive K–12 age-appropriate sexuality education curriculum."
4. Monitor process and outcome measures to document progress toward goals. In Kansas, measurements were based on a monitoring system that was based on logs and administrative records to assess process and intermediate outcomes, constituent surveys of process and outcome, school-based behavioral surveys, community-level indicators such as the pregnancy rate, and interviews with key participants.
5. Communicate information to relevant audiences. Regularly sharing accomplishments and keeping constituents informed of progress are important to maintaining community support, obtaining additional resources, and ensuring accountability. In Kansas, data were shared with the coalition leadership, the community at large, and the Kansas Health Foundation, the primary sponsor.
6. Promote adaptation, renewal, and institutionalization. The monitoring data helped the Kansas coalition recognize accomplishments and redirect energies when necessary. In one community, for example, high levels of substance abuse service provision and low levels of community action indicated that the coalition was becoming a service agency rather than a catalyst for community change.

The Evaluation Process

A working group from the CDC recently published a set of standards and a framework for effective program evaluation in public health.[59] This framework derives from both the practical experience of the CDC and other public health practitioners and from the published standards and recommendations of practitioners in public health, social services, and education.

To be effective, evaluation efforts must meet the following four standards:

1. Utility—Evaluation must serve the information needs of intended users.
2. Feasibility—Evaluation efforts must be realistic, prudent, diplomatic, and frugal.
3. Propriety—Evaluators must behave legally, ethically, and with regard for the welfare of those involved and those affected.
4. Accuracy—Evaluation must reveal and convey technically accurate information.

The CDC's framework describes evaluation efforts as a cycle consisting of the following six steps (Figure 18-6). Although the steps are logically ordered, all are interrelated, and most actual evaluation efforts require iteration and feedback among the steps.

Identify and Engage Stakeholders. A number of different parties have an interest in the outcome of any evaluation; they include persons involved in and affected by the program, as well as the primary users of the evaluation, and especially those who will use it to make decisions about resources or policy. Some stakeholders are obvious: program participants, service providers, and so on. Other stakeholders, however, are less direct. Employers, for instance, may have an interest in a school-based drug prevention program if it increases the productivity of graduates hired by the company.

Engaging these stakeholders means fostering input, participation, and power sharing in the planning and conduct of the evaluation and in the interpretation and dissemination of the findings. This engagement can improve

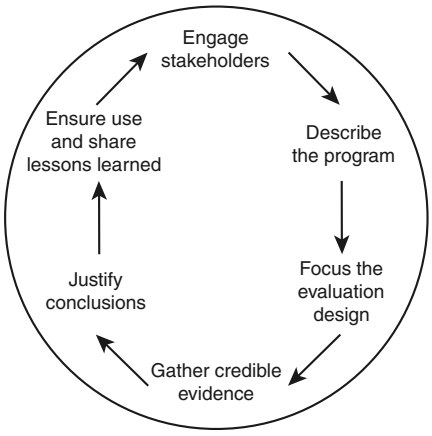

FIGURE 18-6 CDC Framework for Program Evaluation in Public Health
Source: Adapted from Centers for Disease Control and Prevention. Framework for program evaluation in Public Health. *MMWR.* 1999;48(AR-11):1–40.

credibility, clarify roles and responsibilities, ensure cultural relevance and understanding, help protect human experimental subjects, and avoid real or perceived conflicts of interest. Ultimately, it helps to increase the chance that the evaluation will be useful and have a positive impact.

Describe the Program. Before a program can be evaluated, it must be described in terms of the needs it is to address and its purpose, activities, resources, and expected effects. Placing a program in a larger context and clarifying why program activities are believed to lead to the expected changes improves the evaluation's fairness and accuracy. It also permits a balanced assessment of strengths and weaknesses and helps stakeholders understand how program features fit together and relate to a larger context. A clear description of the program as actually implemented is especially important for new activities, which may be implemented in very different ways in different localities. Such descriptions are essential in assessing why interventions work in some settings and not in others.

Focus the Evaluation Design. If an evaluation is to be useful, it is important to clarify at the outset its purpose and uses, as well as the potential users of the results. Assessing the effectiveness of a new program may require different research designs than ensuring the accountability of contractors. Different users are interested in different research questions, depending on their interests, authority, and responsibility. Consider, for example, a community-wide campaign to educate people about the need to call 911 immediately after chest pains begin, and specifically to educate them not to worry about being embarrassed if the pain turns out not to be a heart attack. The local chapter of the American Heart Association, which may have been responsible for getting out the word through the media, would want to know how many people saw the ads and whether they understood and recalled the message. Emergency room staff would focus on how individuals were treated after they arrived, and look specifically at those individuals without heart attacks. Managed care organizations might look at the pretreatment approval process. Specific research methods for evaluation are discussed above.

Gather Credible Evidence. To ensure accuracy and meet the needs of users, evaluation efforts must be based on credible quantitative and qualitative indicators. Indicators are specific measures of program attributes or outcomes that pertain to an evaluation's focus or questions. An evaluation of an educational program to prevent teen violence would require, for example, information on the intensity of the intervention as well as the outcome. Specific indicators of program intensity might be the number of hours of student contact time in the ninth grade or the number of minutes of airtime on a specific group of radio stations that teens favor. Outcome indicators could be the percentage of students who recall the basic message of the program and the number of violent incidents in the targeted schools in the year following program implementation.

Data that are timely must be identified or developed, and steps must be taken to ensure the credibility of the data to the intended users of the evaluation. If data are provided by an agency with a stake in the outcome of the evaluation, auditing or other steps to ensure accuracy may be necessary. Often, compromises must be made between validity, reliability, timeliness, and credibility.

Justify Conclusions. The conclusions of an evaluation are justified when they are based on evidence that is credible to the stakeholders and that

addresses their values and concerns. Standards are explicit statements of stakeholders' values that are operationalized in a way that allows evaluators to judge an intervention's success. A program to reduce the consequences of sexually transmitted disease (STD), for example, might be evaluated in terms of its feasibility in public health clinics and managed care organizations, sensitivity to public values about sexuality, and reduction of disease burden in the population.

The conclusions and recommendations of an evaluation are the product of statistical analysis appropriate to the design of the evaluation and the synthesis of all of the available data, comparing indicators to appropriate standards. An STD evaluation might include an examination of the number of gonorrhea cases reported to the public health department before and after the intervention. This analysis, however, must be interpreted in context. Does a drop in reported cases reflect a decrease in disease, or in individuals getting treatment in private settings or not at all? Is a shift of cases from public clinics to private clinics a desired outcome? Analyses of this sort generally require a substantial amount of professional judgment by evaluation specialists.

Ensure Use and Share Lessons Learned. Evaluation results do not translate into informed decision making and appropriate action without deliberate efforts to ensure that the findings are appropriately disseminated and called to policy makers' attention. Evaluations must first be carefully designed with the concerns and interests of the stakeholders in mind, as discussed above. Dissemination efforts must be planned to ensure that the evaluation results are brought to the attention of those persons or organizations that are in a position to use them in a form that is understandable and useful to that audience. Evaluations must follow up with stakeholders to ensure that the results are understood and not misused. Finally, opportunities for feedback can be useful to evaluators in creating an atmosphere of trust among stakeholders and in refocusing future evaluation efforts, if necessary.

Some managers in public health say, as a point of pride, that they always evaluate every program that they implement. Indeed, some federal agencies and private foundations require evaluation of all funded projects. How does this square with clinical medicine, where physicians do not "evaluate" every procedure that they perform? Having once shown that a procedure works, there is no need to evaluate it every time. The answer is that there are many forms of evaluation, and there is usually one appropriate for any situation in public health. If an intervention has been shown to work in another community, for instance, health officials might want to check that it is being properly implemented and works under the conditions in their communities. Evaluation techniques can also be useful to ensure that a program continues to be properly implemented, an approach known as *performance measurement* as discussed in Chapter 17.

Performance Measurement and Improvement Processes

In recent years, public health and healthcare policy makers have come to realize the importance of population-based data on health status and the determinants of health for effective policy determination, especially for improving

the accountability of managed care organizations, public health agencies, and other entities that can contribute to the public's health. Part of this realization is due to the nature of public health and the emerging importance of preventive medicine: their impact can only be seen in statistical terms such as declining rates of lung cancer attributed to smoking reductions many years earlier. There are no grateful patients or families who know that they have been "saved" by the intervention of a particular physician or hospital. Managed care organizations and the purchasers of their services, moreover, have come to realize that performance measures based on data from the covered populations can be used to hold plans accountable for providing quality services. Similar approaches are beginning to be applied in public health settings as well.[60,61]

In response, a wide range of health data systems and approaches have been developed at the national level. *Healthy People, Healthy People 2000,* and *Healthy People 2010* have clarified the importance of specific, quantitative, population-based health measures for setting public health policy.[62–65] Other examples include the model standards developed by the American Public Health Association and others; Mobilizing for Action through Planning and Partnerships (MAPP), developed by the National Association of County and City Health Officials; the measures used in a Planned Approach to Community Health (PATCH); and the measures proposed by the National Civic League's Healthy Cities/Communities project.[66–70] Taking this approach further, and consistent with the Government Performance and Results Act (GPRA), the US Department of Health and Human Services (HHS) proposed a series of Performance Partnership Grants—to include specific performance measures for states receiving the funds—to replace a number of current block grants.[70] Although this specific approach has not been implemented, performance measurement has become increasingly common in the public health sector.[70] The HRSA's Maternal and Child Health Services block grants, for instance, now require annual performance measures at the state level, as discussed earlier.[41] After September 11, 2001, new federal funding for state and local efforts in relation to bioterrorism, and more recently pandemic influenza, has also come with requirements for performance measures, as discussed below.

At the local level, many communities currently prepare community report cards for health, based in part on one or more of these efforts, but generally uncoordinated with their neighbors.[40,71] The availability of appropriate data is one of the common weaknesses of these approaches. The basic demographic and epidemiologic data available in communities, on which many of these report cards draw, often do not reflect the full spectrum of the dimensions of health or its determinants. Difficult technical problems with the existing measures and lack of data availability (especially for small geographic areas) have further limited the applications of population-based health assessment measures in public health practice.[72]

Performance Improvement in Managed Care

Managed care organizations and other organized healthcare delivery systems are increasingly using performance measures or report cards based on their defined populations to hold themselves accountable to members and purchasers. In recent years, the federal Health Care Financing Administration (HCFA), the Joint Commission on Accreditation of Healthcare Organizations

(JCAHO), the National Committee for Quality Assurance (NCQA), the Institute of Medicine, and other groups have developed a variety of performance measures for hospitals, providers, health plans, and managed care organizations.[73–76] Because it is responsible for delivering care to a defined group of enrollees, managed care makes possible, for the first time, accountability in terms of quality of care for populations, including access to care and health outcomes.[77] Going beyond current practices, David Kindig has proposed that population-based health outcome measures should be the driving force in the market-based management of health plans, and that the health care for entire populations eventually should be managed with these measures.[78]

This trend presents an important opportunity for public health organizations as guarantors of the public's interest in the accessibility, content, and quality of health services. Rather than provide childhood immunizations directly through their own clinics, for example, public health organizations can work with branches of government responsible for the oversight of Medicaid and the regulation of health insurance to ensure that managed care immunization rates are audited and available to purchasers and the public.

The NCQA's *Health Plan Employer Data and Information Set* (HEDIS) is a prominent set of performance measures that deserves some attention, in particular because of its increasing use in Medicaid and other publicly funded managed care.[74] (See Chapter 11 for a more detailed discussion on HEDIS.) HEDIS is a set of 22 standardized performance measures designed to ensure that purchasers and consumers have the information they need to compare the performance of managed healthcare plans reliably. The performance measures in HEDIS are related to many significant public health issues such as cancer, heart disease, smoking, asthma, and diabetes. The NCQA finds that managed care plans that consistently monitor and report on quality are showing significant improvements in quality, resulting in a substantial positive effect on the health of the American public.[74] Some of the measures most relevant to public health are shown in Table 18-2.

The partial list of HEDIS indicators in Table 18-2 illustrates two important points regarding performance measurement. First, there is a substantial overlap between the HEDIS measures and other public health performance measures. Childhood immunization and prenatal care measures, for example, are also included in the HRSA Maternal and Child Health block grant performance measures discussed above. The specific form of the indicators, however, can vary by application. The HEDIS but not the HRSA measures, for instance, calls for one dose of chickenpox vaccine. The HRSA measure applies to all children in the state aged 19–35 months; the HEDIS measure applies only to 2-year-old children enrolled in a health plan. Lack of immunization associated with lack of access to health care, therefore, appears in the HRSA measure but not in the HEDIS measure.

Second, the breast cancer screening and cholesterol management measures illustrate two different approaches to incorporating clinical practice guidelines into performance measures. The US Clinical Preventive Services Task Force currently recommends biannual mammography for women over age 40. HEDIS focuses its indicator on "women between the ages of 52 and 69 who have had at least one mammogram during the past two years."[72] The cholesterol management measure, on the other hand, takes a tertiary prevention approach, focusing on people who have already had a cardiovascular event.[72]

TABLE 18-2 Examples of Specific Performance Measures in HEDIS 3.0

Concept	Specific Indicator
Timeliness of prenatal care	Percentage of women beginning their prenatal care during the first trimester or within 42 days of enrollment if already pregnant at the time of enrollment
Childhood immunizations	Percentage of children who turned 2 years old during the measurement year and received the following vaccinations: 4 doses of diphtheria-tetanus-pertussis, 3 doses of polio, measles-mumps-rubella, 3 doses of *Haemophilus Infuenza* type b (Hib), 3 doses of hepatitis B, and 1 dose of chickenpox
Advising smokers to quit	Percentage of smokers or recent quitters age 18 and older who received advice to quit smoking from a health professional
Breast cancer screening	Percentage of women between the ages of 52–69 who have had at least one mammogram during the past 2 years
Cholesterol management	Percentage of health plan members 18–75 years of age who had evidence of an acute cardiovascular event and whose LDL-C was screened and controlled to less than 130 mg/dL or less than 100 mg/dL in the year following the event

Source: National Committee for Quality Assurance. *The State of Managed Care Quality 2005.* Washington, DC: National Committee for Quality Assurance; 2005. Available at: http://www.ncqa.org/Docs/SOHCQ_2005.pdf.

Community Health Improvement Processes

The IOM has proposed a community health improvement process drawing on the existing use of evaluation tools in a community health setting.[40] Other authors describe similar processes using somewhat different terms, but the basic ideas and issues are typically the same: ownership by communities, a broad definition of health, a cross-disciplinary approach to intervention, and sharing of responsibility among stakeholders for both decision making and accountability. The IOM's model also can be thought of as an example of empowerment evaluation.*

The rationale for the community health improvement process (CHIP) model is that because a wide array of factors influence a community's health, many entities in the community share the responsibility of maintaining and improving its health. Responsibility shared among many entities, however, can easily become responsibility that is ignored or abandoned. At the level of actions that can be taken to protect and improve health, however, specific entities can and should be held accountable, with assignments made through a collaborative process. Because resources and concerns of communities differ, each will have to determine its own specific allocation of responsibility and accountability. Once accountability is assigned, communities can use performance monitoring to hold community entities accountable for actions for which they have accepted responsibility.

*This section draws heavily on the IOM report, *Improving Health in the Community: A Role for Performance Monitoring.*[40]

Growing out of this perspective, a CHIP that includes performance monitoring can be an effective tool for developing a shared vision and for supporting a planned and integrated approach to improve community health. A CHIP offers a way for a community to address collective responsibility and marshal resources of its individuals and families, the medical care and public health systems, and community organizations to improve the health of its members. A CHIP should include two principal interacting cycles based on analysis, action, and measurement (Figure 18-7). The overall process differs from standard models primarily because of its emphasis on measurement to link performance and accountability on a community-wide basis.

The health assessment activities that are part of a CHIP's problem identification and prioritization cycle should include production of a community health profile that can provide basic information to a community regarding its demographic and socioeconomic characteristics and its health status and health risks. This profile would provide background information that could help a community interpret other health data and identify issues that need more focused attention.

For example, the set of indicators for a community health profile might include the following:

- Sociodemographic characteristics, such as the high school graduation rate and median household income
- Health risk factors, such as child immunization coverage, adult smoking rate, and obesity
- Healthcare resource consumption, such as per capita healthcare spending
- Health status, such as the infant mortality rate by race/ethnicity, numbers of deaths due to preventable causes, and confirmed child abuse and neglect cases
- Functional status, such as the proportion of adults in good to excellent health
- Quality of life, such as the proportion of adults who are satisfied with the health care system in the community

Within the CHIP framework, performance monitoring takes place in the analysis and implementation cycle. A community may have a portfolio of health improvement activities, each progressing through this cycle at its own pace. A prototype performance indicator set for vaccine-preventable diseases is shown in Exhibit 18-2. Measures such as these can be further articulated to clarify the accountability of individuals and families, the medical care and public health systems, and community organizations.

To make operational the concept of shared responsibility and individual accountability for community health, stakeholders need to know, jointly and as clearly as possible, how the actions of each potentially accountable entity can contribute to the community's health. Thus, a CHIP should include the development of a set of specific, quantitative performance measures that link accountable entities to the performance of specific activities expected to lead to the production of desired health outcomes in the community. Selecting these indicators will require careful consideration of how to gain insight into progress achieved in the health improvement process. A set of indicators should balance population-based measures of risk factors and health outcomes and health systems-based measures of services performed. Process measures

(such as availability of insurance coverage for immunizations) might be included, but only to the extent that there is evidence that links them to health outcomes. To encourage full participation in the health improvement process, the selected performance measures should also be balanced across the interests and contributions of the various accountable entities in the community, including those whose primary mission is not health specific.

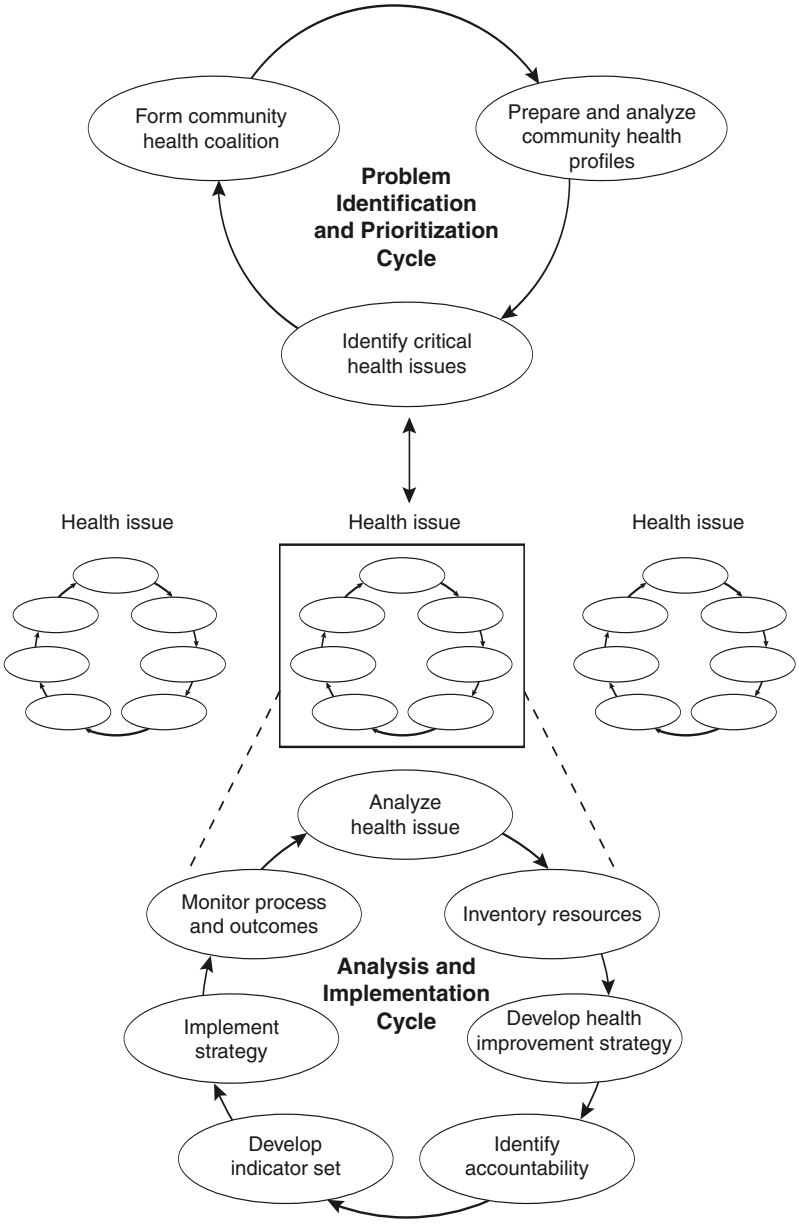

FIGURE 18-7 A Community Health Improvement Process
Source: Reprinted with permission. Institute of Medicine. *Improving Health in the Community: A Role for Performance Monitoring*. Washington, DC: National Academy Press; 1997.

Exhibit 18-2 Sample Prototype Indicator Set: Vaccine-Preventable Diseases

- Immunization rate for all children at 24 months of age.
- Immunization rate at 24 months of age for children currently enrolled in managed care organizations.
- Immunization rate at 24 months of age for children currently enrolled in Medicaid.
- Existence in the community of a computerized immunization registry that provides automated appointment reminders; if available, the percentage of children in the community included in the registry.
- Among children with commercial health insurance coverage, percentage with full coverage for childhood immunizations.
- Percentage of Medicare enrollees who received an influenza immunization during the previous calendar year; percentage who have ever received a pneumococcal pneumonia immunization.
- Pneumonia and influenza death rates for persons age 65 and older.
- Existence in the community of an active childhood immunization coalition involving health service providers, the local health department, parents, and interested parties.

Source: Adapted with permission. Institute of Medicine. *Improving Health in the Community: A Role for Performance Monitoring.* Washington, DC: National Academy Press; 1997.

Example: Indicators of Public Health Preparedness

The anthrax attacks in 2001 and the threat of bioterrorism, as well as emerging infectious diseases such West Nile virus and SARS, have raised concerns about the public health system's ability to respond to emergencies. (See Chapter 23.) Since then, the federal government has distributed almost $5 billion to strengthen state and local public health, as well as hospital preparedness.[79] An additional $350 million was invested in FY 2006 to help these departments prepare for pandemic influenza.[80] Investments of this magnitude demand accountability measures. In addition, many state and local health departments want measures to guide quality improvement efforts, whether they are internally or externally initiated.

Measuring the preparedness of public health systems faces a number of challenges. First, serious public health emergencies are rare, so outcomes (no matter how construed) can not be assessed by direct observation. The infrequency of such emergencies also makes it difficult to learn from experience about what activities work best to increase preparedness. Second, an effective public health emergency response is complex and multifaceted. In any given situation it is difficult to say what an optimal response is, and certainly to capture it in objectives measures. Third, public health systems are fragmented. There are city, county, regional, and state health departments and offices as well as federal agencies, and these structures vary from state to state and often within states. Public health systems also include partner agencies such as hospitals and physicians, emergency medical services (EMS) agencies, agricultural and environmental protection agencies, police, and others. Many of these do not think of themselves as public health agencies, and they are certainly not

under the control of a local or state health official. As a result, accountability for preparedness is diffused. Finally, it should be noted that unlike agencies such as fire departments whose primary purpose it to respond to emergencies, very few people in public health have preparedness as a full-time job.

In response to the need for accountability, four basic types of performance measures have emerged. Perhaps the simplest approach is *informal assessment*: health departments or other knowledgeable parties simply judge how well prepared they are for various public health emergencies. Another approach is embodied in detailed *standards-based assessments* asking whether health departments or other community agencies have undertaken various preparedness activities or meet specified functional standards. Systematic reviews of public health functions during *proxy events* such as major disease outbreaks can also serve to measure aspects of preparedness. Finally, *drills and exercises* of various types are frequently used to raise awareness and for planning and training and less commonly to assess preparedness; they also represent a way to assess preparedness, as discussed below. For the remainder of this chapter we will focus on measures appropriate for measuring functional capacities through drills and exercises.

To develop meaningful and useful measures of public health preparedness for any of these approaches, a logic model that specifies the critical goals and objectives of public health preparedness, as well as how various functions, processes, and resources contribute to meeting them, is needed. One such logic model (Figure 18-8) specifies the goals and objectives of public health preparedness, as well as the functional capabilities and capacity-building activities intended to assure those goals and objectives.[81] In particular, this model helps analysts to distinguish between (1) what the public health system needs to be able to do (the capabilities it needs) during an emergency, and (2) what must be done before an emergency occurs to build capacities. Clearly these two are related, but the tools and approaches needed to measures each are quite different. As discussed below in more detail, inventories and checklists are commonly used to measure what a community has done to build capacity (i.e., has it done what is recommended). On the other hand, proxy events, drills, and exercises can be used to assess how well it did respond or might respond to future emergencies.

The logic model suggests that the overall *goal* of public health preparedness is to mitigate the morbidity and mortality as well as the psychological, social, and economic consequences of a biological attack, a naturally occurring disease outbreak, or other similar disaster (right hand side of model). This goal assumes that effective actions can greatly reduce the health and social consequences of a disease outbreak (especially if a contagious agent is involved), but that 100% prevention is not possible. As such, this goal leaves out other public health activities, such as childhood immunization, that deal with ongoing health problems. Leaving this goal out of this logic model does not imply that these public health activities are any less important than preparedness. Indeed, reductions in performance in these areas can be seen as a cost of public health preparedness efforts.

For a community to meet this goal, in the event of an emergency, it must meet the following objectives: (1) identify and characterize the nature of the outbreak or attack as quickly as possible; (2) mount an early and effective response including providing health care to those affected, taking action to

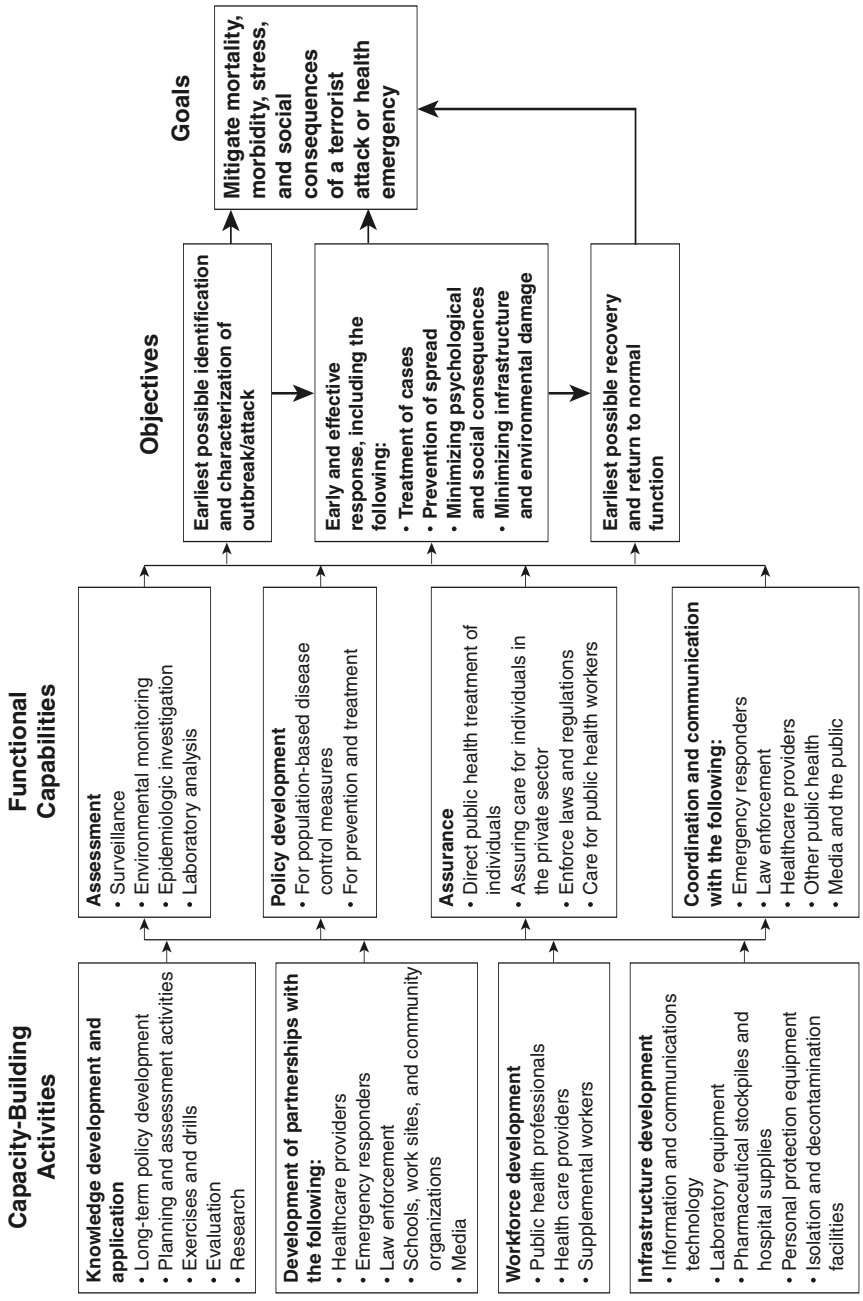

FIGURE 18-8 Public Health Preparedness Logic Model

prevent spread, and minimizing the psychological and social consequences; and (3) enable the earliest possible recovery and return to normal function. Of course, depending on the emergency, some objectives will be more important than others. The evidence for the connection between these objectives and the goals is based on logic and general experience with public health emergencies and other natural disasters, especially infectious disease outbreaks in the past.

If a community is to meet the objectives above during an attack or emergency, it must be capable of carrying out certain functions. Some of these *functional capabilities*, such as surveillance, must operate on a continuous or ongoing basis, and others must be available or enhanced in an emergency. To the extent that these capabilities also relate to other ongoing public health responsibilities, they are more likely to be ready and effective in an emergency because those responsible have experience and the systems have been tested. The functional capabilities in the logic model—the building blocks of a public health response—are categorized according to the Institute of Medicine's three core public health functions—assessment, policy development, and assurance.[82] In addition, coordination and communication have been added because of their prominence in emergencies. *Assessment* includes disease surveillance, environmental monitoring, epidemiological investigation, and laboratory analysis. In this context, *policy development* refers to the ability of public health systems to develop and implement policies during a public health emergency covering population-based disease control activities such as quarantine, isolation, mass prophylaxis, vector control, as well as develop and communicate clinical policies relating to infection control, prevention, and treatment. *Assurance* covers care provided by official health departments to affected individuals (mass immunization clinics, for instance) as well as the enforcement of laws and regulations in support of population-based disease control activities such as isolation and quarantine. Assurance also includes health department involvement in assuring private-sector care for affected individuals through the activation of surge capacity and the Strategic National Stockpile (SNS) and special care for public health or other healthcare workers if needed. *Coordination and communication* within public health and with a variety of other organizations are in support of the first three functional capabilities, but are important enough to merit special consideration.

Many of the *capacity building activities* that states and local health departments are currently undertaking with federal funding support are intended to build the functional capabilities identified in the previous section. *Knowledge development and application* includes the development of policies and protocols in advance of an emergency, planning and assessment activities, exercises and drills, evaluation, and research. Because an effective public health response requires more than the health department per se, the second category of capacity-building activities is focused on the *development of partnerships to support emergency operations*. The partners could include hospitals, physicians, and other healthcare providers, including mental healthcare providers; emergency responders; law enforcement agencies; schools, worksites, and other community organizations; and the media. *Workforce development* activities include the recruitment, training, and preparation of public health professionals, including the designation and prevaccination of smallpox response teams, and of healthcare providers including mental health care providers in the private sector who might be called upon in a public

health emergency. This category also includes the identification and training of a supplemental workforce, such as nurses and volunteers willing to participate in mass immunization efforts. Other *infrastructure development* activities focus on information and communications technology, laboratory equipment, pharmaceutical stockpiles and hospital supplies, personal protection and decontamination equipment, and isolation and decontamination facilities.

Using Tabletop Exercises to Measure Preparedness

In a public health tabletop exercise, public health officials and others who would be involved in responding to a public health emergency are assembled in a room (around a table) and are asked by a facilitator to respond to a fictional scenario as they themselves or their organizations would respond if the conditions were real. This is followed by a "hot wash" during which participants are asked to evaluate the response and to suggest changes that are needed to respond more effectively in the future. An after-action report focusing on observed strengths and weaknesses is prepared by the facilitators.

The emergency response community has long used tabletop exercises to motivate and train, and to develop, test, and evaluate emergency response plans. More recently, public health agencies and healthcare providers have adopted this approach for the same purposes. Going beyond these uses, tabletop exercises can also be used to assess the level of a public health's preparedness. In one such approach, individuals from the local health department and other governmental (e.g., police) and private-sector entities (e.g., hospitals, Red Cross) were assembled, presented with information suggesting a serious disease outbreak such as smallpox, and in one or more stages were asked how they would respond.[83]

Another application used a possible smallpox attack. Seven counties in California exhibited a wide variation in their readiness to detect and respond to this attack. Common sources of variation included: initial steps in beginning an epidemiologic investigation, ability to communicate with most of the doctors and hospitals in the community to initiate active surveillance, practices regarding isolation and quarantine, communication with the public, procedures for collection and testing of biological specimens, beliefs about smallpox vaccination, and understanding of legal authority. Commonly identified gaps included: lack of information systems, significant training needs for public health workforce, inadequate numbers of public health professionals, lack of knowledge of potentially vulnerable or difficult-to-reach population subgroups, and lack of community involvement in planning.[84]

To illustrate this approach in more detail, consider the avian influenza tabletop exercises developed by the Harvard School of Public Health.[85] Should a pandemic influenza virus emerge and reach the United States, its impact on health and well-being, as well as the social and economic disruption it causes, will depend on how well health departments, other government agencies, and the private sector are prepared to respond effectively to a public health emergency. Addressing this concern, federal, state, and local pandemic flu plans are currently being developed, and exercises are being employed to evaluate communities' level of preparedness. To effectively judge public health system performance in an exercise, it helps to begin with a logic model such as the one described above. This model can help clarify which functional capabilities im-

plemented by a variety of public and private community organizations can contribute to overall public health preparedness. More importantly, the model helps evaluators distinguish between (1) what the public health system needs to be able to do during an emergency (the functional capabilities that will be assessed during the exercise), and (2) what must be done before an emergency occurs to build capacities.

In the form of a checklist, Table 18-3 illustrates measures of preparedness that exercise evaluators can use to judge the performance of avian influenza

TABLE 18-3 Preparedness Domains and Indicators for a Pandemic Influenza Tabletop Exercise

Surveillance and Epidemiology

- Receive and respond to urgent case reports.
- Investigate and track reported cases.
- Track information (i.e., newly hospitalized cases, newly quarantined cases) for policy makers.
- Build laboratory capacity (i.e., rapid identification of unusual influenza strains), including ability to ship specimens to state or CDC lab.
- Link with and share data among different surveillance systems (e.g., state DOH, CDC, other communities and states, local hospitals, etc.).
- Step up surveillance capacity in time to initiate containment protocols.

Disease Control and Prevention

- Determine the legal authorities regarding isolation and quarantine.
- Make available procedures to manage isolation and quarantine.
- Have the capacity to support people in quarantine (e.g., preidentified sites, support for home quarantine).
- Develop infection control policies and disseminate them to hospitals and healthcare providers.
- Implement community interventions such as school closings.
- Conduct mass screening.
- Distribute limited medical supplies (including vaccines) to priority groups.
- Control population movement in and out of the community.

Mass Care

- Ensure continuity of healthcare operations.
- Assure health care for all cases that meets relevant standards of care.
- Protect healthcare workers with personal protective equipment (PPE) and infection control practices.
- Increase hospital capacity (including ventilators and other equipment) when necessary.
- Activate and use the Strategic National Stockpile (SNS).
- Prioritize the use of limited medical supplies.
- Provide security within healthcare facilities and at mass point-of-dispensing (POD) sites.
- Coordinate medical reserve.
- Ensure the provisions of mortuary services.

Communication Within Public Health (broadly defined)

- Provide current information (i.e., newly hospitalized cases, newly quarantined cases) to decision makers.
- Disseminate infection control policies to hospitals and healthcare providers.
- Communicate with public health agencies in neighboring communities and the state.
- Communicate within the local public health system (including other government agencies).
- Communicate with hospitals and healthcare providers.

(continues)

TABLE 18-3 (continued)

Communication with the Public

- Communicate with the general public up-to-date outbreak information, disease control requirements, individual risk reduction, and when and where to seek medical care.
- Communicate with the public to minimize fear.
- Communicate with marginalized population groups through trusted sources.

Leadership and Management

- Initiate and use the incident command system.
- Identify activities that will be performed at a state, local, or coordinated level.
- Interact with local, state, and federal officials with regard to the delegation of legal and law enforcement responsibilities.
- Identify the authority for declaring a public health emergency.
- Gather resources in support of implementing action.
- Assess and manage local resources.
- Address and respond to cross-jurisdictional needs.
- Assist special needs populations.
- Respond flexibly, in proportion to the magnitude and severity of the scenario and available resources.
- Anticipate psychosocial needs and activate appropriate services.
- Integrate community-based organizations in the response.

tabletop exercise participants. The checklist is organized into six preparedness domains, and each domain has 3 to 11 specific indicators. Numerical scores could be assigned, say using a 5 point scale ranging from 1 (response not sufficient) to 5 (response exceeded expectations) for each domain and a three point scale for each indicator as follows: 1: not sufficient; 2: sufficient; 3: exceeds expectations. Numerical scores of this sort are of course arbitrary, but knowledgeable and sufficiently trained assessors can use them effectively to evaluate the community's response to the scenario and hence provide a measure of preparedness.[86]

Based on feedback from the participants, we found that tabletop exercises of this sort were generally effective in helping a community assess the level of preparedness of its public health systems. The scenarios presented were seen as realistic and elicited critical areas of public health system response. Observations that the participants made in the workshop and the after-action reports helped health departments to identify areas of strength and weakness, and many indicated that they planned to take action to address weaknesses. Some of the critical elements of this approach seemed to be (1) assembling the appropriate participants from the local public health department and representing partner agencies, (2) a realistic and challenging outbreak scenario, tailored to the local situation (i.e., including local institutions such as hospitals and reflecting existing capabilities and responsibilities), (3) simulation of actual decision making through built-in stops during which participants are asked what they would do, and (4) facilitation and evaluation by knowledgeable individuals from outside the community being assessed, using a structured framework such as Table 18-3.

• • •

A large and growing set of evaluation methods now exists to assist public health organizations in measuring and improving their programs and op-

erations. These tools can be used to monitor the quality, outcomes, and efficiency of public health activities carried out by single and multiple institutions. These tools can also be used to examine the effects that interventions outside the domain of public health have on community health. Public health organizations must continually work to improve the standards of evidence used in evaluating public health interventions, so that results can inform managerial and policy decision making. Moreover, as an increasing share of public health interventions are carried out through multi-institutional partnerships, public health organizations must meet the growing imperative for collaboration in evaluation efforts. By doing so these organizations can move closer to the goals of evidence-based management and the gains in population health that it promises.

Chapter Review

1. Evaluation addresses the following questions:
   - Which programs work, for whom, and under what conditions?
   - Which program variants work best?
   - Is the public getting the best possible value for its investment?
   - How can the impact of existing programs be increased?
2. All public health programs can be thought of in terms of *inputs, activities, outputs,* and *outcomes*.
3. Economic evaluations include costs and benefits in quantitative terms—for example, which program is most effective in terms of dollars per child immunized?
4. Formative evaluation refers to efforts to identify the best use of available resources, prior to a traditional program evaluation. Formative evaluation often employs qualitative methods such as focus groups or structured interviews.
5. *Statistical power* is the likelihood that an evaluation will detect the effect of an intervention, if there is one. Two factors affect statistical power: sample size and effect size, which is a quantitative measure of the program's impact.
6. Research synthesis—systematic reviews of existing studies, including meta-analysis, is increasingly used in public health to uncover robust effects.
7. The goals of an evaluation determine the types of measures that are needed. Outcome evaluations need measures of health outcomes, whereas feasibility evaluations focus on costs and barriers to implementation.
8. A CDC framework for evaluation consists of a six-step cycle:
   - Identify and engage stakeholders.
   - Describe the program.
   - Focus the evaluation design.
   - Gather credible evidence.
   - Justify conclusions.
   - Disseminate evaluation results to improve the program.

References

1. DeCarlo P. Project Access: Research That Feeds Back into the Community. *CAPS Exchange*. San Francisco, Calif: Center for AIDS Prevention Studies; 1999.
2. Fetterman DM. Empowerment evaluation: an introduction to theory and practice. In: Fetterman DM, Kaftarian SJ, Wandersman A, eds. *Empowerment Evaluation: Knowledge and Tools for Self-Assessment & Accountability*. Thousand Oaks, Calif: Sage; 1996:3–46.
3. Green SB, Byar DP. Using observational data from registries to compare treatments: the fallacy of omnimetrics. *Stat Med*. 1984;3:361–370.
4. Sorensen G, Emmons K, Hunt MK, Johnston D. Implications of the results of community intervention trials. *Annu Rev Public Health*. 1998;19:379–416.
5. Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu Rev Public Health*. 2000;21:121–145.
6. Murray DM. *Design and Analysis of Group-Randomized Trials*. New York, NY: Oxford University Press; 1998.
7. Hoaglin DC, Light RL, McPeek B, Mosteller F, Stoto MA. *Data for Decisions: Information Strategies for Policymakers*. Cambridge, Mass: Abt Books; 1982.
8. Sloan FA, ed. *Valuing Health Care: Costs, Benefits, and Effectiveness of Pharmaceuticals and Other Medical Technologies*. Cambridge, Mass: Cambridge University Press; 1995.
9. Understanding and using the economic evidence. In: Zaza S, Briss PA, Harris KW, eds. *The Guide to Community Preventive Services: What Works to Promote Health?* Oxford, UK: Oxford University Press; 2005:449–463.
10. Haddix AC, Teutsch SM, Corso PH, eds. *Prevention Effectiveness: A Guide to Decision Analysis and Economic Evaluation*. 2nd ed. Oxford, UK: Oxford University Press; 2003.
11. Muennig P. Introduction to cost-effectiveness. In: Muennig P, Khan K, ed. *Designing and Conducting Cost-Effectiveness Analyses in Medicine and Health Care*. San Francisco, Calif: Jossey-Bass; 2002:1–32.
12. Russell LB, Siegel JE, Daniels N, et al., Cost-effectiveness analysis as a guide to resource allocation in health. In: Gold MR, Siegel JE, Russell LB, Weinstein MC, eds. *Cost-Effectiveness in Health and Medicine*. Oxford, UK: Oxford University Press; 1996:3–24.
13. Miller W, Robinson LA, Lawrence RS. *Valuing Health for Regulatory Cost-Effectiveness Analysis*. Washington, DC: National Academies Press; 2006.
14. Secker-Walker RH, Holland RR, Lloyd CM, Pelkey D, Flynn BS. Cost effectiveness of a community based research project to help women quit smoking. *Tob Control*. 2005;14:37–42.
15. Torrance GW, Feeny D. Utilities and quality-adjusted life years. *Int J Technol Assess Health Care*. 1989;10:559–575.
16. AIDS Cost and Services Utilization Survey (ACSUS). Public Use Tapes 4 and 5 [Database]. Springfield, Va: National technical Information Service; 1992.
17. Bozzette SA, Berry SH, Duan N, et al. The care of HIV-infected adults in the United States. *N Engl J Med*. 1998;339:1897–1904.
18. Paltiel AD, Weinstein MC, Kimmel AD, et al. Expanded screening for HIV in the United States—an analysis of cost-effectiveness. *N Engl J Med*. 2005;352:586–595.
19. Parasuraman S, Salvador C, Frick KD. Measuring economic outcomes. In: Pizzi LT, Lofland J, eds. *Economic Evaluation in US Health Care: Principles and Applications*. Sudbury, Mass: Jones and Bartlett; 2006:15–40.

20. Nichol KL, Lind A, Margolis KL, et al. The effectiveness of vaccination against influenza in healthy, working adults. *N Engl J Med*. 1995;33: 889–893.
21. Grosse SD, Waitzman NJ, Romano PS, Mulinare J. Reevaluating the benefits of folic acid fortification in the United States: economic analysis, regulation, and public health. *Am J Public Health*. 2005;95(11):1917–1922.
22. Gold MR, Siegel JE, Russell LB, Weinstein MC, eds. *Cost-Effectiveness in Health and Medicine*. Oxford, UK: Oxford University Press; 1996.
23. Farnham PG, Haddix AC. Study design. In: Haddix AC, Teutsch SM, Corso PS, eds. *Prevention Effectiveness: A Guide to Decision Analysis and Economic Evaluation*. 2nd ed. Oxford, UK: Oxford University Press; 2003:11–27.
24. Luce BR, Manning WG, Siegel JE, Lipscomb J. Estimating costs in cost-effectiveness analysis. In: Gold MR, Siegel JE, Russell LB, Weinstein MC, eds. *Cost-Effectiveness in Health and Medicine*. Oxford, UK: Oxford University Press; 1996:176–213.
25. Torrance GW, Siege JE, Luce BR. Framing and designing the cost-effectiveness analysis. In: Gold MR, Siegel JE, Russell LB, Weinstein MC, eds. *Cost-Effectiveness in Health and Medicine*. Oxford, UK: Oxford University Press; 1996:54–81.
26. Garber AM, Weinstein MC, Torrance GW, Kamlet MS. Theoretical foundations of cost-effectiveness analysis. In: Gold MR, Siegel JE, Russell LB, Weinstein MC, eds. *Cost-Effectiveness in Health and Medicine*. Oxford, UK: Oxford University Press; 1996:25–53.
27. Armstrong EP. Sensitivity analysis. In: Grauer DW, Lee J, Odom TD, eds. *Pharmacoeconomics and Outcomes: Applications for Patient Care*. 2nd ed. Kansas City, Kan: American College of Clinical Pharmacy; 2003:231–245.
28. Drummond MF, O'Brien B, Stoddart GL, Torrance GW. *Methods for the Economic Evaluation of Health Care Programmes*. 2nd ed. Oxford, UK: Oxford University Press; 1997.
29. Siegel JE, Weinstein MC, Torrance GW. Reporting cost-effectiveness studies and results. In: Gold MR, Siegel JE, Russell LB, Weinstein MC, eds. *Cost-Effectiveness in Health and Medicine*. Oxford, UK: Oxford University Press; 1996:276–303.
30. Mullins CD, Flowers LR. Evaluating economic outcomes literature. In: Grauer DW, Lee J, Odom TD, eds. *Pharmacoeconomics and Outcomes: Applications for Patient Care*. 2nd ed. Kansas City, Kan: American College of Clinical Pharmacy; 2003:246–273.
31. Louis TA, Fineberg HV, Mosteller F. Findings for public health from meta-analysis. *Annu Rev Public Health*. 1985;16:1–20.
32. Egger M, Davey Smith G, Altman D, eds. *Systematic Reviews in Health Care: Meta-analysis in Context*. 2nd ed. London: BMJ Books; 2001.
33. Mosteller F, Colditz GA. Understanding research synthesis (meta-analysis). *Annu Rev Public Health*. 1996;17:1–23.
34. Shefer A, Briss P, Rodewald L, et al. Improving immunization coverage rates: an evidence-based review of the literature. *Epidemiol Rev*. 1999; 21:96–142.
35. Task Force on Community Preventive Services. Interventions to improve vaccination coverage in children, adolescents, and adults. *Am J Preventive Medicine*. 2000;18(1S):92–96.
36. Jefferson T, Rivetti D, Rivetti A, Rudin M, DiPietrantonj C, Demicheli V. Efficacy and effectiveness of influenza vaccines in elderly people: a systematic review. *Lancet*. 2005;366:1165–1174.
37. Institute of Medicine, National Research Council, Normand J, Vlahov D, Moses LE. *Preventing HIV Transmission: The Role of Sterile Needles and Bleach*. Washington, DC: National Academy Press; 1995.

38. Kaplan EH, Heimer R. HIV incidence needle exchange participants: estimated from syringe tracking and testing data. *J Acquired Immune Deficiency Syndromes*. 1994;7:182–189.
39. Hagan H, Des Jarlais DC, Friedman SR, Purchase D, Alter MJ. Risk for human immunodeficiency virus and hepatitis B virus in users of the Tacoma syringe exchange program. In: *Proceedings, Workshop on Needle Exchange and Bleach Distribution Programs*. Washington, DC: National Academy Press; 1994:24–34.
40. Institute of Medicine. *Improving Health in the Community: A Role for Performance Monitoring*. Washington, DC: National Academy Press; 1997.
41. Maternal and Child Health Bureau. *Maternal and Child Health Services Title V Block Grant Program: Proposed National Performance Measures, Health Systems Capacity Indicators, and Health Status Indicators*. Rockville, MD: Health Resources and Services Administration; undated.
42. Donaldson MS. *Measuring the Quality of Health Care*. Washington, DC: National Academy Press; 1999.
43. US Public Health Service Expert Panel on the Content of Prenatal Care. *Caring for Our Future: The Content of Prenatal Care*. Washington, DC: US Department of Health and Human Services; 1989.
44. Chalk R, King PA. *Violence in Families: Assessing Prevention and Treatment Programs*. Washington, DC: National Academy Press; 1998.
45. Cronbach LJ, Snow RE. *Aptitudes and Instructional Methods: A Handbook for Research on Interactions*. New York, NY: Irvington Publishers; 1981.
46. Reiss AJ Jr, Roth JA, National Research Council. *Understanding and Preventing Violence*. Washington, DC: National Academy Press; 1993.
47. Cordray DS, Pion GM. Psycho-social rehabilitation assessment: a broader perspective. In: Glueckauf R, Sechrest LB. *Improving Assessment in Rehabilitation and Health*. Newbury Park, Calif: Sage; 1993:215–240.
48. Evans RG, Stoddart GL. Producing health, consuming health care. In: Evans RG, Barer ML, Marmor TR. *Why Are Some People Healthy and Others Not? The Determinants of Health of Populations*. New York, NY: Aldine De Gruyter; 1994.
49. Patrick DL, Wickizer TM. Community and health. In: Amick BC, Levine S, Tarlov AR, Walsh DC, eds. *Society and Health*. New York, NY: Oxford University Press; 1995:46–92.
50. Hollister RG, Hill J. Problems in the evaluation of community-wide initiatives. In: Connell JP, Kubisch AC, Schorr LB, Weiss CH, eds. *New Approaches to Evaluating Community Initiatives*. New York, NY: Aspen Institute; 1995:127–172.
51. Weiss CH. Nothing as practical as good theory: exploring theory-based evaluation for comprehensive community initiatives for children and families. In: Connell JP, Kubisch AC, Schorr LB, Weiss CH, eds. *New Approaches to Evaluating Community Initiatives*. New York, NY: Aspen Institute; 1995:65–92.
52. Connell JP, Kubisch AC. *Applying Theories of Change Approach to the Evaluation of Comprehensive Community Initiatives: Progress, Prospects, and Problems*. New York, NY: Aspen Institute; 1996.
53. Rose G. Sick individuals and sick populations. *Int J Epidemiol*. 1985;14: 32–38.
54. Rose G. *The Strategy of Preventive Medicine*. New York, NY: Oxford University Press; 1992.
55. Abrams D. Conceptual models to integrate individual and public health interventions: the example of the work-place. In: Henderson M, ed. *Proceedings of the International Conference on Promoting Diet Change in*

Communities. Seattle, Wash: Fred Hutchinson Cancer Research Center; 1991:173–194.
56. Tosteson ANA, Weinstein MC, Hunink M, et al. Cost-effectiveness of population-wide educational approaches to reduce serum cholesterol levels. *Circulation*. 1998;95(1):24–30.
57. Murray DM. *Design and Analysis of Group-Randomized Trials*. New York, NY: Oxford University Press; 1998.
58. Fawcett SB, Paine-Andrews A, Francisco VT, et al. Empowering community health initiatives through evaluation. In: Fetterman DM, Kaftarian SJ, Wandersman A, eds. *Empowerment Evaluation: Knowledge and Tools for Self-Assessment & Accountability*. Thousand Oaks, Calif: Sage; 1996: 161–187.
59. US Centers for Disease Control and Prevention. Framework for program evaluation in public health. *MMWR*. 1999;RR-11:1–40.
60. Perrin EB, Koshel JJ. *Assessment of Performance Measures for Public Health, Substance Abuse, and Mental Health*. Washington, DC: National Academy Press; 1997.
61. Perrin EB, Durch J, Skillman SM. *Health Performance Measurement in the Public Health Sector: Principles and Policies for Implementing an Information Network*. Washington, DC: National Academy Press; 1999.
62. US Department of Health, Education, and Welfare. *Healthy People: The Surgeon General's Report on Health Promotion and Disease Prevention*. Washington, DC: US Government Printing Office; 1979.
63. US Department of Health and Human Services. *Promoting Health/ Preventing Disease: Objectives for the Nation*. Washington, DC: US Government Printing Office; 1980.
64. US Department of Health and Human Services. *Healthy People 2000: National Health Promotion and Disease Prevention Objectives*. Washington, DC: Office of the Assistant Secretary for Health; 1991.
65. US Department of Health and Human Services. *Healthy People 2010: Understanding and Improving Health*. Washington, DC: US Government Printing Office; 2000.
66. American Public Health Association, CDC, et al. *Healthy Communities 2000: Model Standards*. Washington, DC: American Public Health Association; 1991.
67. National Association of County and City Health Officials. *Achieving Healthier Communities through MAPP: A User's Handbook*. Washington, DC: NACCHO; 2004.
68. Kreuter MW. PATCH: its origin, basic concepts, and links to contemporary public health policy. *J Health Educ*. 1992;23:135–139.
69. US Centers for Disease Control and Prevention. *Planned Approach to Community Health: Guide for the Local Coordinator*. Atlanta, Ga: 1995.
70. National Civic League. *The Healthy Communities Handbook*. Denver, CO: National Civic League; 1993.
71. Fielding JE, Halfon N, Sutherland C. Characteristics of community report cards—United States, 1996. *MMWR*. 1997;46:647–649.
72. Stoto MA. Public health assessment in the 1990s. *Annu Rev Public Health*. 1992;13:59–78.
73. Joint Commission on Accreditation of Healthcare Organizations. Evolution of *Performance Measurement at the Joint Commission 1986–2010*. Oakbrook Terrace, Ill: Joint Commission on Accreditation of Healthcare Organizations; undated. Available at: http://www.jointcommission.org/NR/rdonlyres/333A4688-7E50-41CF-B63D-EE0278D0C653/0/EvolutionofPM.pdf. Accessed August 22, 2006.

74. National Committee for Quality Assurance. *The State of Managed Care Quality 2005*. Washington, DC: National Committee for Quality Assurance; 2005.
75. Institute of Medicine. *Performance Measurement: Accelerating Improvement*. Washington, DC: National Academy Press; 2006.
76. Darby M. *Health Care Quality: From Data to Accountability*. Washington, DC: National Health Policy Forum, George Washington University; 1998.
77. Stoto MA, Abel C, Dievler A. *Healthy Communities: New Partnerships for the Future of Public Health*. Washington, DC: National Academy Press; 1996.
78. Kindig DA. *Purchasing Population Health: Paying for Results*. Ann Arbor, MI: University of Michigan Press; 1997.
79. Schuler A. Billions for biodefense: federal agency biodefense budgeting, FY2005–FY2006. *Biosecurity Bioterrorism: Biodefense Strat, Pract, Sci*. 2005;3:94–101.
80. US Department of Health and Human Services. HHS announces additional $225 million for state and local pandemic influenza preparedness efforts [Press release]. July 11, 2006. Available at: http://www.hhs.gov/news/press/2006pres/20060711.html. Accessed July 19, 2006.
81. Stoto MA, Dausey D, Davis L, et al. Learning from experience: the public health response to West Nile virus, SARS, monkeypox, and hepatitis A outbreaks in the United States. *Rand Corporation Technical Report*. 2005.
82. Institute of Medicine. *The Future of Public Health*. Washington, DC: National Academy Press; 1988.
83. Dausey DJ, Diamond A, Meade B, et al. Preparedness training and assessment exercises for local health departments. *RAND Health*. 2005. TR-261-DHHS.
84. Lurie N, Wasserman J, Stoto MA, et al. Local variation in public health preparedness: lessons from California. *Health Aff*. 2004; Suppl Web Exclusives: W4-341-53.
85. Harvard School of Public Health, Center for Public Health Preparedness. Toolkit to assist public health in conducting preparedness exercises. Available at: http://www.hsph.harvard.edu/hcphp/products/exercises/HSPHCPHP%20Avian%20&t%20Pandemic%20Influenza%20Tabletop.pdf. Accessed August 22, 2006.
86. Stoto MA, Biddinger P, Cadigan R, Savoia E. Using tabletop exercises to evaluate a community's preparedness for pandemic influenza. Presented at the annual meeting of the American Public Health Association, Boston, Mass, November 2006.