SELECTED METHODOLOGICAL ISSUES IN EVALUATING COMMUNITY-BASED HEALTH PROMOTION AND DISEASE PREVENTION PROGRAMS


Departments of Health Services, Epidemiology, and Biostatistics, University of Washington, Seattle, Washington 98195

A. R. Kristaz, C. H. Allan-Andrilla, L. J. Dey

Cancer Prevention Research Unit, Fred Hutchinson Cancer Research Center, Seattle, Washington 98104

KEY WORDS: evaluation methods

INTRODUCTION

An important recent trend in health promotion and disease prevention has been the increasing number and scope of community-based interventions. These programs are aimed at entire populations, which are usually geographically defined, and they attempt to change health behavior and disease risk.
through mass media campaigns, activation of existing community organizations, or changes in the physical or sociocultural environment. Several large programs of this kind have been mounted for cardiovascular disease prevention (30, 33, 44, 56, 71), as reviewed by Shea & Basch (79, 80), and the approach is increasingly being applied to other disease areas and populations (3, 34, 67, 89, 92). As investment in community-based programs has grown, so has the importance of evaluating their effectiveness, as evidenced in part by the recent publications of Green & Lewis (38) and Bracht (6). In this review, we focus on a selection of methodological issues that assume special importance in evaluating community-based programs, but receive little coverage in standard texts on program evaluation. These issues include:

1. Specification of the theoretical model. The design of an intervention is usually based on some theory of program action. An important early step in program evaluation is to make this theoretical model explicit: What are the key intervention components, and what are the causal mechanisms by which they are expected to work? An explicit model is needed to guide evaluation design decisions, to help identify the specific shortcomings of a program found to be ineffective, or to facilitate dissemination of an effective one. The task can be complex for community-level interventions aimed at individual-level health behavior because of the need for a multilevel conceptualization.

2. Communities as units of allocation. Because interventions aim at entire communities, an evaluation design with concurrent controls will likely involve assignment of communities en bloc to intervention and control groups. This feature has important implications for both planning study size and data analysis.

3. Allocation of a small number of communities. Cost and feasibility considerations usually limit the intervention and evaluation to a small number of communities, thus complicating the task of achieving comparable study groups.

4. Longitudinal versus repeated cross-sectional samples. Community surveys may be needed to measure change in certain key outcomes. These surveys can be conducted by either following a panel of individuals in each community over time or drawing a fresh cross-sectional sample in each community at each time point. Both approaches have unique strengths and drawbacks.

5. Validity of self-reported health characteristics. Particularly because of the highly public nature of the intervention and the inability to blind participants to their treatment group membership, the validity of self-reported data on health behavior can be a concern.

6. Measures of community environment. Assessing features of the community environment can help test the underlying causal model, detect early
program effects, and avoid excessive reliance on self-reported behavior change.

We now discuss each of these six issues in turn.

SPECIFICATION OF THE THEORETICAL MODEL

The randomized controlled trial has become a widely accepted paradigm for evaluating the effect of health interventions, against which nonexperimental methods are judged and often found wanting. The design and the size of most randomized trials are usually driven by a primary research question, which typically concerns the effect of an intervention on final outcomes. Unfortunately, this focus on final outcomes may result in overlooking the need to characterize both the intervention itself and the causal mechanisms by which it is supposed to work. Interventions then become “black boxes” whose overall effects may be detectable, but whose contents are obscure. Careful specification of the intervention and its presumed mechanism of action is an important step in designing an appropriate evaluation.

What are black box interventions? Lipsey (57) describes them as “situations for which inputs and outputs can be observed, but the connecting processes are not readily visible.” The black box then contains the causal sequence between inputs (e.g. receipt of grant funds and formation of a community coalition) and outputs (e.g. cessation of cigarette smoking). For simpler interventions, such as an immunization program, opening the black box, albeit desirable, may not be as essential to interpreting the evaluation results, replicating effective interventions, or tinkering with ineffective ones. For such interventions as community-based prevention efforts, the contents of the black box are much more complex, and their obscurity is a serious deterrent to understanding and progress.

A key reason to open black boxes is to improve interventions. With this in mind, an approach to process evaluation based on theoretical considerations has emerged in the evaluation literature (13, 14). At the heart of the approach is the notion of treatment theory, which describes how program inputs translate into outputs. An optimal treatment theory is specific enough to guide evaluation design and analysis, yet general enough to illuminate the field. The more critical need, however, is for specific applicability to the intervention under study and to the context in which it will be implemented. This need has led Lipsey (57) to label such intervention theories as “small theories of treatment.” Large theories, such as diffusion theory or exchange theory, might guide the elaboration of treatment theory, but can be too abstract and general to guide evaluation design.

A useful treatment theory provides a model to show how the program will
produce its postulated effects. At minimum, it must include key inputs (e.g. formation of a new community coalition) and outputs (e.g. avoidance of substance use by adolescents), and the sequence of events of processes connecting them. For community-based prevention programs, these events or processes must delineate a believable scenario by which the mobilization of community organizations and programs can motivate and assist individual citizens to change their behaviors. A useful small theory of treatment would describe how grant funds, program specifications, technical assistance, and other inputs translate into effective community structures that can produce and disseminate intervention components with a chance of success.

A critical aspect of useful treatment theory and process evaluation, in general, is the specification of key steps in program implementation (75). For most community-based health programs, major concerns include the functionality of the community coalition or board, the scientific quality of intervention components as actually delivered, and the exposure of community residents to those interventions.

**Treatment Theory and Evaluative Design**

A good treatment theory can greatly enhance the design, analysis, and interpretation of an evaluation (5, 57). From a study design perspective, there is almost no limit to what can be measured in a community-based program. Important events and processes may occur in the community environment or among community organizations, political leaders, health care providers, or individual members of the target population. Choosing the variables to measure requires some means of distinguishing that which is essential for determining program success or failure from the rest. Program theory provides a blueprint for measurement because, by definition, it specifies the critical steps on the path from input to output.

For example, Figure 1 shows the “small theory” of treatment that guides the evaluation of the Henry J. Kaiser Family Foundation’s Community Health Promotion Grants Program (89). The conceptual basis for the model (27, 39, 40), based in social learning theory (a “large theory”) (2), emphasizes modifying community norms and inducing changes in the physical, regulatory, and socioeconomic environments to make them more supportive of healthful behaviors and behavior change. To accomplish this, the model posits that projects must first activate their communities by developing a broadly based consensus among leading community organizations to address a health problem, coordinate planning, share resources, and engender broad citizen involvement. The “activated community” reaches individual citizens through high quality intervention components that change norms toward approval of healthful behaviors and disapproval of unhealthful ones (e.g. in media messages), change environments to encourage healthful behaviors and discourage
unhealthful ones (e.g. worksite smoking policies), and provide more models of individuals who have adopted healthful norms and behaviors.

Measurements in this evaluation were then selected to correspond to the major components of the treatment theory. A survey of leaders of key community organizations provides data to assess the extent to which community organizations were collaborating and generating intervention activities. Surveys of restaurants and grocery stores and reviews of legislative activity monitor environmental changes, whereas surveys of adult and adolescent residents furnish information about exposure to interventions, norms, behavioral models, as well as behaviors.

**Treatment Theory and Data Analysis**

Treatment models are also analytic models, which specify independent, dependent, and mediating variables and depict causal pathways. Judd & Kenny (47) show how modern multivariate statistical techniques can be used to test the relationships posited by a treatment theory. Lipsey (57) argues that the use of treatment theory to select appropriate, sensitive outcome measures may mitigate the common problem of insufficient statistical power in social evaluations by increasing the true effect size associated with effective treatments.

**Treatment Theory and the Interpretation of Evaluation Results**

Community-based prevention programs that address health-related behaviors may not always produce dramatic effects. Evaluation findings have often been mixed (28), controversial (70), or negative (94) and are likely to continue to be so. Treatment theory may clarify the meaning of findings by delineating the role of the treatment, or aspects of the treatment, as the cause of a positive
or negative result. Concomitantly, treatment theory may play a crucial role in disentangling bad evaluation methods from bad treatment ideas from bad treatment implementations (57).

**Treatment Theory and the Advancement of Treatment Effectiveness**

Most community-based prevention programs resemble each other, at least in general ways. Evaluations based on treatment theory should advance the state of the art by identifying the details of good ideas for replication or enhancement and bad ideas for a return trip to the drawing board.

**COMMUNITIES AS UNITS OF ALLOCATION**

Under our definition, community-based interventions are aimed at entire communities. Hence, an evaluation that uses a concurrently studied control group must generally use entire communities as controls. The unit of allocation in this design is thus the community, even though many outcome measures (including all those discussed in this section), such as smoking status or dietary fat intake, may originate with observations on individuals in those communities. Sometimes, communities are actually randomized to intervention and control groups (34, 89), as we discuss later. However, nonrandomized designs must also deal with the consequences of community-level allocation, and they require added attention to the possibility of community-level confounding factors.

Probably, the most important consequences of allocation by community are reduced statistical power and added complexity in estimating sample size requirements or statistical power. When we have person-level outcome measures, but community-level allocation and analysis, two sources of random variation must be considered and estimated: individual-level variation within a community and community-level variation within a treatment group. We must also consider two kinds of sample sizes: the number of individuals per community and the number of communities per treatment group. For a fixed total number of individuals studied, statistical power is almost always lower when allocation is by community (or cluster) rather than by individual, as shown in a short and accessible paper by Cornfield (17). At least under classical methods of analysis, part of the loss of power occurs because the number of degrees of freedom for a statistical test of treatment effect depends on the number of communities studied, not on the number of individuals studied in those communities (17, 48). When the number of communities is small, this number of degrees of freedom is also small, and the critical value that a test statistic must achieve is higher than for studies that allocate
individuals. This effect on power can grow large when the number of communities falls below about ten.

More specifically, the power to detect an effect of the intervention depends directly on the precision with which the mean level of the relevant outcome can be estimated for each treatment group. For a simple design involving randomization of \( c \) communities to an intervention group and \( c \) more to a control group, with \( n \) individuals studied per community, the expected sampling variance of the mean for each treatment group is:

\[
\frac{\sigma_C^2 + \frac{\sigma^2}{n}}{c},
\]

where \( \sigma_C^2 \) is the community-level variance component (i.e. variance in the true mean level of the outcome variable among communities) and \( \sigma^2 \) is the individual-level variance component (i.e. variance in the outcome variable among individuals within a community). As a rule, the evaluator has little control over the size of \( \sigma_C^2 \) or \( \sigma^2 \), but must estimate them both to estimate study power.

The above expression also shows that if \( \sigma_C^2 \) is at all large relative to \( \sigma^2 \), there are likely to be only modest gains from studying more individuals per community (i.e. increasing \( n \)), but potentially major gains in power from studying more communities per treatment group (i.e. increasing \( c \)). Of course, these two options for enhancing power may have quite different cost implications. In some situations, the marginal cost of each intervention site may be large, but the marginal cost of a control site may be more modest. If so, the evaluator may wish to form unequal-sized treatment groups, with more control sites than intervention sites.

An equivalent way of considering this issue (23) is to note that, under community allocation, observations on the individuals in each treatment group cannot be considered statistically independent of each other, as they can under individual allocation. Instead, observations on individuals who reside in the same community tend to be correlated. For continuous variables, the appropriate measure of correlation is the intraclass correlation, which can be expressed as

\[
\frac{\sigma_C^2}{\sigma_C^2 + \sigma^2}.
\]

The formulation based on correlated observations is thus closely linked to that based on variance components, as the intraclass correlation can be viewed as a measure of the relative sizes of the two variance components. Mickey et al
(61) discuss this issue in terms of the design effect and show how its magnitude can depend on study duration.

Specific Methods for Estimating Sample Size and Power

Statistical tools useful for study planning have been developed for a variety of study designs that involve allocation by community. Donner and colleagues (22, 23) and Hsieh (43) provide guidelines for studies of simple, two-group comparisons that involve continuous or dichotomous outcome measures. Shipley et al (81) describe and illustrate methods for designs involving randomization of matched pairs of communities. Hsieh (43) discusses an approach for power calculations when communities are to be randomized within two or more strata, and when treatment effects are to be measured in terms of a pretest/posttest comparison over time. Koepsell et al (48) suggest an approach that can be used when the time path of a program effect is of central interest, as may be true for an evaluation developed around a specific intervention model. They also discuss different approaches for longitudinal versus repeated cross-sectional samples of individuals studied over time. Earlier work by Gillum et al (35) also considers the problem of allowing for dropouts over time.

Obtaining Estimates of Community-Level Variance

One of the greatest challenges in estimating power and sample-size requirements in community-based studies is providing estimates of the community-level variance component, \( \sigma_c^2 \). [For a design involving comparison of changes over time, the evaluator would instead supply an estimate of \( \sigma_{CT}^2 \), the community-by-time interaction variance, against which treatment-by-time interactions would be tested. See Koepsell et al (48).] Depending on the outcome variable of interest, suitable estimates can sometimes be derived from public data sources, such as the Centers for Disease Control Behavioral Risk Factor Survey, or from previous studies.

Several statistical methods for estimation of variance components have been proposed (77). Particularly when the number of individuals studied varies across communities, these methods can yield different estimates. As a practical matter, both the BMDP and SAS computer packages have procedures for computing variance components. In BMDP, procedure P3V provides both maximum likelihood or restricted maximum likelihood (REML) estimation methods. In SAS, the corresponding procedure is PROC VARCOMP.

For illustration, Table 1 presents estimates of \( \sigma_c^2 \) and of \( \sigma^2 \) for current smoking status, as derived from three studies that involved data collection across several communities: the evaluation of the Kaiser Family Foundation Community Health Promotion Grant Program (89), the RAND Health Insurance Experiment (64), and a survey of cancer-related risk behaviors con-
ducted in Washington State by the Cancer Prevention Research Program at the Fred Hutchinson Cancer Research Center (1990, unpublished data). In each of those studies, cities or counties were the communities of interest, and smoking status was coded as 0=nonsmoker, 100=smoker to yield estimates in a convenient numerical range. Within each study, estimates of \( \sigma_C^2 \) obtained by the three statistical methods are generally similar, e.g. they range from 8.8 to 10.7 for the Kaiser data. However, the point estimates are quite different across studies; REML estimates range from 5.4 for the RAND data to 30.3 for the Washington State data. Despite the relatively large number of individuals studied, estimates of \( \sigma_C^2 \) from these data sets are based on small numbers of communities and thus have rather wide confidence limits. When such data are used in study planning, it may be wise to use several estimates of \( \sigma_C^2 \), which vary through a plausible range and yield “optimistic” and “pessimistic” estimates of sample size or statistical power.

Particularly for large data sets, the task of computing variance component estimates can be time-consuming and costly, and an investigator may lack the resources to do so. Sometimes, only published, community-level means or prevalences may be available. In these situations, an investigator can obtain a crude point estimate of \( \sigma_C^2 \) by simply computing the variance of the set of community-level means or prevalences. On average, such an estimate tends to

### Table 1

Examples of individual- and community-level variance components for current smoking status

<table>
<thead>
<tr>
<th></th>
<th>Kaiser Community Health Promotion Grants Program</th>
<th>RAND Health Insurance Experiment (at entry)</th>
<th>Washington State Cancer-Related Behavioral Risk Factor Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. communities</td>
<td>15</td>
<td>6</td>
<td>35</td>
</tr>
<tr>
<td>Total no. individuals</td>
<td>8726</td>
<td>5094</td>
<td>1642</td>
</tr>
<tr>
<td>Prevalence of smoking</td>
<td>24%</td>
<td>37%</td>
<td>26%</td>
</tr>
<tr>
<td>Individual-level variance (( \sigma^2 ))</td>
<td>1800.7</td>
<td>2342.3</td>
<td>1990.7</td>
</tr>
<tr>
<td>Point estimate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>95% conf. limits</td>
<td>(1746.2, 1855.1)</td>
<td>(2249.5, 2435.3)</td>
<td>(1850.9, 2130.7)</td>
</tr>
<tr>
<td>Community-level variance (( \sigma_C^2 ))</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REML Method</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point estimate</td>
<td>10.7</td>
<td>5.4</td>
<td>30.3</td>
</tr>
<tr>
<td>95% conf. limits</td>
<td>(0, 21.7)</td>
<td>(0, 16.5)</td>
<td>(0, 79.9)</td>
</tr>
<tr>
<td>ML Method</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point estimate</td>
<td>8.8</td>
<td>3.8</td>
<td>25.1</td>
</tr>
<tr>
<td>95% conf. limits</td>
<td>(0, 17.6)</td>
<td>(0, 12.1)</td>
<td>(0, 67.2)</td>
</tr>
<tr>
<td>Method of Moments</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point estimate</td>
<td>10.0</td>
<td>5.0</td>
<td>20.1</td>
</tr>
</tbody>
</table>
be conservative (i.e. too large) and probably still has wide confidence limits if based on a small number of communities. But, at least the estimate gives an investigator an idea of $\sigma^2$ for use in study planning.

**Analysis Strategies**

Oft-quoted advice by Cornfield (17) is: "Randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception ... and should be discouraged." Whiting-O'Keefe & Simborg (91) have also commented on the all-too-common practice of ignoring the proper unit of analysis in studies that involve assignment of aggregates to treatment conditions.

Space limitations permit only brief mention of several suitable analysis techniques. Randomization tests (65) provide a valid method to test for program effects with minimal statistical assumptions. These tests are more feasible to implement with small numbers of study units, especially in the present era of cheap computing power. However, they have the decided disadvantage of never being able to reject the null hypothesis if the number of possible assortments of study communities into treatment groups is very small. Traditional analysis-of-variance methods for hierarchial (nested) study designs may be used and are readily implemented when the number of observations per community is relatively constant across communities (25). Analysis using community means as though they were elementary observations can also be a straightforward and valid approach for such "balanced" designs. The above-mentioned BMDP procedure P3V can accommodate designs with unequal sample sizes (20). The analysis of variance is most applicable for continuous outcome measures, but it may also be suitable for dichotomous outcomes if the number of observations per community is reasonably large and if community prevalences are not too close to 0 or 1.

Donald & Donner (21) have suggested a method that accounts for randomization by cluster when combining $2 \times 2$ contingency tables across communities. Donner & Donald (24) have proposed analytic methods when randomization by cluster has been carried out within strata. Zeger et al (95) have described powerful and flexible analysis methods for correlated dichotomous outcomes by using generalized estimating equations in the context of longitudinal studies. Software to implement their approach is not yet widely available, however.

**ALLOCATION OF A SMALL NUMBER OF COMMUNITIES**

Often, funding agencies or communities themselves decide whether a program is to be mounted in a particular community or set of communities, and
an evaluator may have little say in the matter. On other occasions, a multi-community program may be set up as a planned social experiment, thus allowing evaluation considerations to affect the process by which communities are designated as intervention or nonintervention sites. But, even when an evaluator has the luxury of allocating communities to treatment groups, it may be far from clear how best to do so. Here, we consider two aspects of the decision.

**Should Communities Be Randomized?**

When only a few study communities are available to be allocated randomly to an intervention and a control group, there is an increased risk of a major imbalance between groups on important confounding factors, whether these factors are known or unknown. One can argue that some possible outcomes of simple randomization would be unacceptable, such as those that put intervention and control communities into the same media market and lead to cross-contamination. For that reason, for example, and to minimize investigator travel, communities close to study headquarters are sometimes chosen as intervention sites, thus leaving communities farther away as controls.

Nonetheless, even when only a few communities are available for study, a random allocation process has much to recommend it (51), especially when processes other than simple random allocation are considered. The difficulty of creating acceptably balanced treatment groups results chiefly from the limited number of communities available for assignment, and that difficulty remains whether randomization is used or not. Other methods for achieving balance, such as matching or stratification, can be used in conjunction with randomization. In the COMMIT project, for example, 11 pairs of communities were formed, and one member of each pair was chosen at random to be the intervention site (34); in the Kaiser Health Promotion Evaluation Project, a form of restricted randomization was used after study communities were arranged into strata (89). Restricted randomization can also be sued to deal with the problem of shared media markets by ruling out certain unacceptable study group configurations in advance and selected one of the remaining acceptable configurations at random, as long as each community ultimately has an equal chance of becoming an intervention or a control site. (This may be a particularly suitable context in which to use a randomization test for statistical inference.) In brief, although a carefully designed random allocation process may not prevent problems of treatment group comparability as neatly as it does with larger samples, it need not complicate them either. And, randomization offers other advantages: namely, a firm basis for formal hypothesis testing and a public perception of even-handedness in forming the comparison groups that is hard to achieve any other way.
Should Communities Be Matched?

As noted above, matching can be used with or without randomization to achieve some degree of comparability between intervention and control groups or to enhance power. Theoretically, the best factor on which to match is one that is highly correlated with change in the outcome variable; in practice, there may be limited knowledge about which community characteristics qualify as good matching factors on this basis. Freedman et al (34) showed that a matching scheme that incorporated geographic proximity and community size appeared to perform well in forming matched pairs that were similar with regard to the prevalence of the target behavior at baseline. However, Martin et al (59) suggest that when the number of study communities is small, matching should be used only in the presence of a very good matching factor, chiefly because the loss of degrees of freedom that results from using the community pair (rather than the individual community) as the unit of analysis can seriously compromise power and, in fact, weaken the comparison.

LONGITUDINAL VERSUS REPEATED CROSS-SECTIONAL SAMPLES

A central goal of most community-based health promotion programs is to reduce risky health behaviors in study communities. Surveys of community residents at two or more points in time are often required to obtain direct evidence on whether this goal is met. These surveys may use either longitudinal samples, which consist of a panel of individuals in each community who are surveyed repeatedly, or repeated cross-sectional samples, which consist of a fresh sample of individuals from each community on each survey occasion (usually with only a small probability of repeated selection of the same individual). Although this discussion is in terms of samples of individuals, similar comments apply to other possible subunits within a community, such as restaurants or schools.

Several writers (1, 29, 36, 73) have commented on the relative merits of the longitudinal and repeated cross-sectional sampling approaches. The choice between the two depends on the correspondence between sample type and program objectives, on relative susceptibility to biases, on statistical power trade-offs, and on cost. Table 2 summarizes factors to be considered in this section.

Correspondence with Program Objectives

An important question is whether the intervention seeks primarily to change the health behavior of individuals, or to change the prevalence of risky behaviors in the community. These two kinds of changes are not the same, as
Table 2  Factors influencing a choice between longitudinal and repeated cross-sectional samples

<table>
<thead>
<tr>
<th>Factor</th>
<th>Longitudinal</th>
<th>Cross-sectional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program objective</td>
<td>Directly measures change in individual health characteristics</td>
<td>Directly measures change in community prevalence of health characteristics</td>
</tr>
<tr>
<td>Selection bias at recruit</td>
<td>May be worse because participation is not anonymous</td>
<td>Participation may be anonymous</td>
</tr>
<tr>
<td>Attrition</td>
<td>Losses to follow-up may be related to behavior being evaluated</td>
<td>Not a problem</td>
</tr>
<tr>
<td>Testing</td>
<td>Repeated questioning may be a co-intervention</td>
<td>Not a problem</td>
</tr>
<tr>
<td>Maturation</td>
<td>Panel gets older, while community at large may not</td>
<td>Not a problem</td>
</tr>
<tr>
<td>History</td>
<td>Panel consists of more long-term community residents with exposure to “local history”</td>
<td>Less a problem</td>
</tr>
<tr>
<td>Cross-contamination</td>
<td>Not a problem</td>
<td>Movement between intervention and control communities may dilute intervention effect</td>
</tr>
<tr>
<td>Statistical power</td>
<td>Higher for fixed sample size and intervention effect</td>
<td>Lower</td>
</tr>
</tbody>
</table>

communities are dynamic populations whose membership can change over time because of births, deaths, and in- and out-migration. A decline in the community prevalence of a behavior over time may occur, even in the absence of any individual-level behavior change, if individuals who join or leave the community differ systematically from other community residents in terms of their health behavior. Community-based programs usually do seek individual-level behavior change. But, sometimes, they may also change the social environment, deliberately or otherwise, through recruitment of persons with healthy behavior and out-migration of those with risky behavior, e.g. a worksite health promotion program may succeed in institutionalizing a preference for non-smokers in hiring decisions, and it may make workplace smoking policies uncomfortable for smokers so that they seek jobs elsewhere. Other factors being equal, longitudinal samples are theoretically better suited to isolating program effects on individual behavior change, whereas repeated cross-sectional samples are better suited to measuring program effects on community-wide prevalence.

If the survey involves a large fraction of the community, and if population turnover is low, one may, in fact, generate a longitudinal subsample within...
the cross-sectional samples by repeated selection of the same individuals. In other situations, there may be ways to circumvent limitations of a specific sampling approach by altering other aspects of the survey methodology. For example, respondents in a follow-up cross-sectional survey can be asked about their length of residence in the community and about any changes in their health behavior that occurred during the study period. It may also be possible to supplement a longitudinal sample or to replace those lost to follow-up with newcomers during the study to render its composition more representative of the community at each time point, even though this option complicates data analysis.

Susceptibility to Bias

Table 2 also highlights certain sources of bias that can affect longitudinal and repeated cross-sectional samples differently; thus, "bias" means any systematic difference between measured characteristics of the sample and the corresponding true characteristics of the population supposedly represented by the sample.

Self-selection at recruitment can occur under either sampling approach because of nonresponse. Active refusal to participate is an important component of nonresponse (31, 42, 88), and concerns about privacy account for many refusals in some surveys (19). Although respondents can participate anonymously as part of a single cross-sectional sample, members of a longitudinal sample must reveal their identities and consent to be recontacted. These additional demands may further jeopardize willingness to participate.

Attrition affects longitudinal, but not repeated cross-sectional, samples. Attrition can be large: In the Stanford Five-City Project, 39% of the baseline cohort completed three follow-up surveys over a five-year period (28). Several longitudinal studies have found that individuals who smoke at baseline are more likely to drop out than those who do not (41, 45, 46). Other studies have found that subjects who are harder to follow are more likely to have worse exercise habits (54) or higher levels of substance abuse at follow-up (41, 53, 69). These findings suggest that losses from a cohort often occur preferentially among those with worse health habits.

Testing effects occur when changes in reported behavior are caused (or inhibited) by the act of repeated questioning. They affect longitudinal samples only. Although the possibility of such effects has long been known by psychologists (8, 16) and shown for nonhealth behaviors, such as voting (49), little evidence is available concerning testing effects on reported health characteristics. In the MRFIT STUDY (63), a larger discrepancy between self-reported and thiocyanate-adjusted quit rates for smoking in the intervention group compared with the control group at follow-up suggested the possibility of testing-treatment interaction. A study by Bridge et al (7) sug-
gests that repeated questioning resulted in shifting attitudes about cancer. Murray et al (62) found greater declines in smoking in a repeatedly questioned cohort of adolescents compared with a single, comparably aged cross-sectional sample. They inferred that the surveys themselves may have accounted for part of the difference.

Maturation occurs in a longitudinal sample, which ages over time, whereas the age distribution in the community and in repeated cross-sectional samples may change very little. Any age-related phenomenon may thus appear to change in a longitudinal sample over time, even if the change had no relation to a community intervention.

History may also preferentially affect longitudinal samples, which necessarily consist of longer-term community residents. Stable members of the community may have more exposure to local, nonprogram-related events that cause behavior change.

Cross-contamination of treatment groups is at least a theoretical possibility, if mobility among study communities is high. With repeated cross-sectional samples, a follow-up survey participant may have recently moved from a control community to a study community, or vice versa, thus rendering the subject's exposure status unclear. This kind of bias can be more of a concern if "community" is broadly defined to include such settings as workplaces or schools.

Although these sources of bias can interfere with the degree to which the sample reflects the community at a given time point, they do not necessarily result in a biased estimate of program effect. If attrition affects longitudinal samples similarly in intervention and control communities, for example, this source of error would "cancel out" in a comparison between study groups. Likewise, bias that remains stable over time could still allow accurate estimation of a change in the prevalence of a characteristic over time. The strongest evaluation designs used to date have assessed program effect by comparing changes over time between intervention and control groups. Under such a design, the estimate of program effect would be biased only if there is interaction among size of bias, treatment group, and time, e.g. if repeated surveying renders a person more susceptible to an intervention effect, or if attrition of persons with unhealthy behavior occurs differently in the intervention group versus the control group. Unfortunately, little empirical evidence is available to judge how serious such potential threats to validity are in practice.

Statistical Power

A major attraction of the longitudinal-sample approach is its greater statistical power to detect change. This gain in power results from, and is quantitatively dependent on, intertemporal correlation in health characteristics at the individual level: The more stable the characteristic, the greater the advantage of
a longitudinal sample for detecting a hypothesized change of a given size. Schlesselman (76) and Cook & Ware (15) discuss the statistical principles that underlie this conclusion. Koepsell et al (48) discuss performance of sample-size calculations for both sampling approaches.

Given that the longitudinal-sample approach may be more susceptible to a variety of biases, as discussed above, Martin et al (58) derived a simple inequality that shows how large the added bias must be to outweigh the power advantages of a longitudinal sample, at least for a simple design situation. Specifically, consider a design in which \( r \) = the correlation between individual’s baseline and follow-up health behavior status, \( n \) = the number of individuals surveyed per occasion, \( b_L \) = the amount of bias in the estimate of mean change from baseline to follow-up based on a longitudinal sample, \( b_X \) = the corresponding bias for a cross-sectional sample, and \( s^2 \) = the overall variance in behavior. Martin showed that when \( r < n(b_L^2 - b_X^2)/2s^2 \), then a cross-sectional sample approach yields a lower expected mean-squared error than a longitudinal-sample approach.

Unfortunately, a confident choice between sampling approaches depends on having good advance estimates of the likely extent of several kinds of bias and of the expected intertemporal correlation in the characteristics being measured. Moreover, all of these factors can be expected to vary from one behavior to another, so that the superior sampling approach for studying one behavior may be inferior for studying another. Perhaps for these reasons, several evaluations of large-scale community interventions have used both longitudinal and cross-sectional samples; these evaluations usually let the baseline survey sample serve both as the longitudinal sample and as the first cross-sectional sample (29, 44, 89). Building on this practical stratagem of safety through redundancy, Thomquist & Anderson (86) have recently proposed what they nickname a “belt and suspenders” method for combined analysis of data from longitudinal and cross-sectional samples, which uses generalized estimating equations.

VALIDITY OF SELF-REPORTED HEALTH CHARACTERISTICS

In community-based health promotion and disease prevention studies, information about health behavior is often gathered directly from individuals through interviews or self-administered questionnaires. There is a widespread belief that people are inclined to overreport desirable health behaviors and underreport undesirable health behaviors. As more attention is paid to health behaviors in the media, in public places, in worksites, and in clinical practice, individuals, families, and different social groups may become sensitized to socially desirable forms of behavior. Therefore, methodologies to investigate and improve the validity of self-reports are important to develop and apply.
One major approach is to search for "objective" measures of behavior, on the assumption that they are free of subjective bias. Biochemical validation tests, such as those used in smoking research, are prized for their criterion validity. These "gold standard" measures, however, may be too costly, as well as vulnerable to between-individual variation in absorption, metabolism, and excretion (37). One investigative team even concludes that "... questionnaire response appears to be the standard against which physiologic test of smoking must be judged, not vice versa" (68). Self-reports often become the only feasible method for collecting data on health behaviors. We summarize here published evidence for the validity of self-reports for two forms of health behavior that have been common targets of community-based interventions: cigarette smoking and dietary behavior. We also discuss the major methodologies for evaluating and improving these reports.

**Cigarette Smoking**

A recent review and meta-analysis of studies, which uses biochemical validation of smoking behavior, suggests that self-reports of cigarette smoking obtained by in-person interviews have fairly high sensitivity and specificity among adult respondents who participate in community studies, when examined in relation to a biochemical measure of smoking status (66). Similar validation studies, which have been carried out among students, suggest that self-reports among adolescents involved in smoking cessation interventions are less accurate. Biochemical validation remains desirable in evaluations of smoking cessation interventions.

Biochemical validation cannot determine, however, the accuracy of reports regarding smoking consumption, i.e. the number of cigarettes smoked (85). Nor can biochemical tests be used to validate smoking histories that yield estimates of risk in terms of pack-years. Lifetime smoking consumption is likely underreported, given the difficulties of long-term recall.

Several methodological techniques have been used to evaluate and improve self-reports of smoking behavior. Studies of surrogate reports of behavior, usually next-of-kin and particularly spouses, indicate that self-reports of cigarette smoking correlate highly with surrogate reports (60).

Other studies have suggested that informing subjects that a biochemical measure of cigarette smoking, such as salivary cotinine or expired carbon monoxide, is to be obtained improves the validity of self-reports (4, 26). In some instances, bogus measurement procedures are used, or biochemical samples are obtained but never analyzed. This approach has been called the "bogus pipeline." When genuine objective measures were used in research with adolescents, Bauman & Dent (4) found that adolescents who had recently smoked reported significantly greater amounts of smoking if they were informed about the biochemical measure before completing the questionnaire.
Unfortunately, published studies evaluating self-reports of cigarette smoking seldom contain the actual questions used to classify smokers and nonsmokers. Thus, the form and content of the questions themselves are difficult to evaluate. The actual wording of questions can influence the responses given and, hence, the categorization of respondents as smokers (87). Therefore, studies asking about smoking should report or reference the questions used, so that this potential source of invalidity can be examined and controlled. They should also report whether subjects were told before answering questions that they would later be asked to provide a specimen for biochemical validation.

Dietary Behavior

A problem with assessing dietary behavior through self-report is that eating is a mundane, frequent behavior that a person does with relatively little attention. At least three methods have been used in community-based studies to assess dietary change: nutrient intake (diet records, 24-hour recall, and food-frequency questionnaires); biochemical measures (primarily serum cholesterol); and approaches aimed at the specific targets of the intervention (measures of individual behavior, such as “Yesterday, did you eat a vegetable with dinner?”, or environmental measures as discussed below, such as percent of supermarket milk shelf space devoted to lowfat milk) (50).

The lack of a criterion measure of dietary intake in free-living persons is the major problem in evaluating the validity of these measures. Assessing convergent validity (concurrence among different measures) is a common alternative. In general, correlations among various nutrient intake measures are rarely above 0.6 and, depending upon the nutrient, are frequently as low as 0.3 (55). Even for food frequency questionnaires, which are designed to minimize intra-individual variability, test-retest correlations are rarely above 0.65 and may be as low as 0.2 (84).

A special threat to validity arises from the nonblinded nature of most community dietary intervention studies. If the intervention program has an effective public education component, residents of intervention communities understand the relationship between food and health better and pay greater attention to food and food choices. These intervention effects could influence measurements in the absence of behavior change, thus confounding any interpretation of contrasts between intervention and control communities. The act of retesting a cohort may also produce biases in reported behaviors, as discussed earlier. Unfortunately, there are few data with which to substantiate or estimate the magnitude of these potential biases.

In practice, 24-hour recalls and food records have usually been deemed too expensive, time-consuming, and difficult to administer for use in large-scale
community studies. Food frequency questionnaires and abbreviated questionnaires of behavior specifically targeted by the intervention, methods that rely upon retrospective reports of dietary habits, are most often the only practical means for assessing dietary change. The cognitive processes that underlie responding to these methods are complex. For example, food frequency judgments require individuals to assign "typical frequency" and portion size judgments for what is often a long list of food items. An inferential process, by which a frequency judgment is derived at the time the question is asked, must occur. Little research has been done to investigate these cognitive processes and their potential for biasing reported dietary behaviors (82, 83).

Epidemiologic studies of dietary behavior (93) have found that respondents' reports of current dietary behaviors, or recall of previous behaviors, depend on whether the foods are perceived as socially desirable or personally relevant. Comparisons with daily food records have indicated overestimates of up to 50% on food frequency judgments for "healthy" foods and underestimates of up to 30% for "unhealthy" foods (74).

Various approaches need to be investigated to both assess and minimize these biases in dietary recall. For example, social desirability and food salience scales may be included in evaluation schemes (18). Less direct approaches include making the dietary intake assessment an adjunct to some other task not so closely related to health habits (e.g. embedding it in a longer series of questions about consumer buying behavior). Another approach might be to include bogus foods (e.g. lowfat olive oil) in food frequency questionnaires to estimate the overreporting of "healthy" foods.

Both laboratory-based and community-level studies are needed to advance our understanding of how individuals evaluate and report health behaviors, and whether any biases we find differ for persons in community intervention and control communities. Over the last few decades, the accumulated research suggests that self-reports of smoking require biochemical validation in intervention studies, particularly with adolescents in school-based cessation programs. The lack of such biochemical measures for self-reported dietary behavior adds considerable complexity to the assessment of an inherently complicated and multifaceted behavior.

MEASURES OF COMMUNITY ENVIRONMENT

Several complicating factors that arise in assessing the outcome of community-level interventions enhance the attractiveness of a class of measures that we call "environmental" indicators. This section briefly describes the complicating factors, defines environmental indicators and places them in the context of other community-level measures, and gives some examples. Cheadle et al (12) provide a more complete discussion of this class of measures.
As noted above, two difficulties arise in assessing community-level interventions: the impossibility of blinding individual subjects to the presence of the intervention, which threatens the validity of self-reported attitudes, behaviors, and outcomes; and the complexity of the mechanism of action by which community programs change individual behaviors, with many intermediate steps in the behavior change process. As illustrated by the causal model shown Figure 1, these intermediate steps often involve modifying that which can be labeled the "community environment," defined broadly to include the legal, social, and economic, as well as the physical, environment. Components of the health-related community environment include institutions (stores, worksites, political institutions), geography (air, water quality), media messages (TV, radio, print), laws, and regulations (smoking ordinances).

Environmental indicators thus serve two functions in an evaluation. First, they provide an indicator of shared attitudes and/or collective behavior that does not rely on self-reports. Second, they capture features of the environmental link in the chain that connects health-promotion programs to changes in health-related behavior.

Environmental indicators are derived from observations of the community context in which people live. To clarify this notion, it is useful to relate them to other "community-level measures," i.e. approaches to characterizing the community as a whole as opposed to individuals or subgroups within it. Community-level measures can be divided into three sub-categories: "individual-disaggregated"—information originally obtained on individuals for whom individual-level covariate data (e.g. demographic characteristics) are available that can be considered in analyzing and interpreting community-level summary statistics; "individual-aggregated"—measures derived from individual-level information, but available only in aggregated form; and "environmental indicators"—measures based on observations of the community environment.

Most community-level measures that have been used in evaluating health promotion programs fall into the first category, i.e. are based on individual-level measures (e.g. interview surveys, physiologic measures) through which additional information on each respondent is available. These data are most frequently gathered by investigators who are evaluating a particular health-promotion program, but could easily include other public-use data available at the individual level (birth and death tapes, hospital discharge abstracts). Community-level measures formed by aggregating individual-level data, devoid of identifiers, to the community level include data collected by agencies other than the program evaluators. Examples of aggregate measures include census data, mortality rates, traffic-accident statistics, and most economic data (e.g. sales information).
Environmental indicators, the third class of community-level measures, are derived from observations of aspects of the community environment that, like other community-level measures, are then aggregated to the level of the community. For example, the number, type, and visibility of nonsmoking signs in a workplace (which can be regarded as a small community) are an environmental indicator of the attitudes of the workers and management in that workplace toward smoking. Greater degrees of militancy toward smoking among employees and management will probably be associated with more and better-advertised no-smoking areas. In addition, the number and character of workplace no-smoking signs are indicators of the environmental influences acting on employees.

There are several strands in the existing literature relevant to environmental indicators. Since the mid-1960s, a substantial literature on social indicators and social indicator models has accumulated in sociology (9, 52, 72). For example, Carley (9) presents indicators derived from the Social and Economic Accounts System (SEAS) developed by Fitzsimmons & Lavey (32), which organizes 477 community indicators into 15 sectors. Health sector social indicators in the SEAS include individual-aggregated measures (e.g. number of deaths per 1000 live births), as well as measures that could be classified as environmental indicators: number of full-time equivalent physicians, hospitals, and hospital beds.

Another close relative to environmental indicators in the existing literature are the "unobtrusive" or "nonreactive" measures first collected and categorized by Webb et al (90). A measure is unobtrusive if the object of interest is unaware of being observed. Nonreactive measures do not suffer from the problem of reactivity bias, i.e. the "true" response is not altered by the process of measurement. All unobtrusive measures are nonreactive, but some nonreactive measures may be highly obtrusive (e.g. blood tests). In many cases, these unobtrusive measures would be classified as individual-level measures under our scheme, as the observations are made on individuals and then aggregated to get an estimated mean or proportion for the group of interest. However, several other measures reported in the literature are based on characteristics of the community environment [e.g. graffiti (78)] and can, therefore, be classified as environmental indicators.

Table 3 provides examples of community-level measures related to tobacco use. These measures are categorized along two dimensions: the measurement category (individual-disaggregated, individual-aggregated, and environmental) and the obstrusiveness and reactivity bias likely to be associated with the measure. The environmental measures are further subdivided according to the component of the environment being measured (e.g. workplace, restaurant).

The examples in Table 3 may help clarify the earlier discussion of terminology. The newspaper poll of attitudes could, in principle, be shifted to the
<table>
<thead>
<tr>
<th>Measurement Category*</th>
<th>Reactive</th>
<th>Nonreactive</th>
<th>Unobtrusive</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Individual-Disaggregated</strong></td>
<td>Phone survey of smoking status, attitudes toward smoking in public places</td>
<td>Cotinine, other biochemical measures of smoking status</td>
<td>Sample of household garbage to count cigarette butts</td>
</tr>
<tr>
<td><strong>Individual-Aggregated</strong></td>
<td>Published results of newspaper poll of attitudes toward smoking in public places</td>
<td>Classroom-level cotinine measures</td>
<td>Cigarette sales; Election results from a vote over a nonsmoking ordinance</td>
</tr>
<tr>
<td><strong>Environmental</strong></td>
<td>Interview with company CEO; his/her views on smoking policy</td>
<td>Survey of company smoking policies</td>
<td>Prevalence of designated nonsmoking areas</td>
</tr>
<tr>
<td><strong>Worksite</strong></td>
<td>Staff response to customers who smoke in nonsmoking areas</td>
<td>Proportion of restaurants with some nonsmoking seating</td>
<td>Ashtray prevalence; visibility of nonsmoking signs</td>
</tr>
<tr>
<td><strong>Restaurants and other public spaces</strong></td>
<td>Interviews with key informants: community attitudes</td>
<td>Interviews with key informants: legislative history of ordinances</td>
<td>Does community have a nonsmoking ordinance?</td>
</tr>
<tr>
<td><strong>Community as a whole</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*See text for discussion of measurement categories.*
individual-disaggregated category, if the newspaper collected demographic information on the respondents and made the individual-level information available to outside investigators. The worksite environmental indicators cover aspects of company smoking policy. The interview with the company president is likely to be colored by concern about public relations, and thus subject to a considerable amount of reactivity bias. The company will also be aware that a survey is being conducted of its smoking policy, but because the assessment could focus on written policy statements, there is less chance of an untruthful response. The observation of the prevalence of no-smoking areas could be made unobtrusively, if admittance were gained for some reason other than to conduct such a survey (e.g. the observations could be made by an employee).

The advantages of environmental indicators have already been noted: They are frequently unobtrusive and, therefore, not subject to response bias. And, they are measurements of important intermediate factors in health-promotion interventions. The drawbacks of environmental indicators are the same ones that have held back the development of unobtrusive measures in social psychology: lack of persistent and credible efforts to assess and improve the validity and reliability of candidate measures (78). An effort to overcome this lack of evidence for environmental indicators has begun, however. For example, the reliability of a grocery store instrument designed as an environmental indicator of dietary habits has been assessed as part of the evaluation of the Kaiser Family Foundation Community Health Promotion Grants Program (11, 89). The validity of the grocery store instrument has also been assessed, by comparing the results of the survey with a phone survey of individuals in the same communities (10). Only through such a process of accumulating information about validity, reliability, and responsiveness to change can a fair test of these measures be conducted.

CONCLUSIONS

At present, the community-based approach to health promotion appears to be in an expansion phase, spurred in part by the apparent success of several large-scale, community-wide programs aimed at prevention of cardiovascular disease. New programs are now being developed for a wider array of health conditions, the definition of “community” is being broadened to include both larger and smaller social units, and the range of target populations is being widened demographically and socioeconomically. Many newer community-based programs are being mounted with fewer resources and a different mix of intervention modalities than their predecessors. All of these factors emphasize the importance of rigorous evaluation to determine when, where, how, and for whom the community-based approach succeeds. We hope that the
above discussion helps sensitize evaluators to the special challenges they face in attempting to answer those important questions and kindles the interest of methodologists to develop new and better evaluative tools.

ACKNOWLEDGMENT

This work was supported by Grant No. CA 34847 from the National Cancer Institute.

Literature Cited

with the cluster as unit of randomization. 


75. Scheirer, M. A. 1987. Program theory and implementation theory: implications