Special Issues Related to Randomized Trials of Primary Prevention

Julie E. Buring

Elucidation of the biologic mechanisms of disease by basic researchers, as well as advances in diagnosis and treatment by clinicians, has historically provided, and will continue to provide, enormous benefits to affected individuals. Of equal or greater importance, however, is the evaluation of primary prevention strategies, that is, interventions that may prevent common and serious diseases in “apparently healthy” populations with no known clinical evidence of disease. Such interventions can in theory afford even more benefits to society as a whole. For example, a complete cure for leukemia would clearly be a remarkable breakthrough in medical research, eliminating approximately 20,000 deaths annually in the United States, yet even a modest reduction on the order of 20–30 percent in the development of epithelial cell cancers due to vitamin supplementation could prevent over five times that number of deaths during the same period (1). Similarly, even a 20 percent reduction in the risk of developing a first myocardial infarction would have a tremendous public health impact, given that coronary heart disease is the leading cause of mortality in the United States and other developed countries (2).

Clinical trials are powerful tools for evaluating the efficacy and safety of preventive interventions, although they are neither necessary nor desirable in all circumstances. With respect to cigarette smoking and cardiovascular disease, for example, the totality of evidence from basic research and observational epidemiologic studies was very clear in indicating the large benefits of smoking cessation, even in the absence of clinical trial data. However, the effects of most primary preventive measures are likely to be small to moderate in size—very meaningful from a clinical or public health standpoint but very difficult to detect statistically (3, 4). It is unlikely that the introduction of a new preventive measure will be accompanied by benefits as clear-cut and unequivocal as that of penicillin, which immediately reduced mortality from bronchopneumonia three- to sixfold. For the vast majority of preventive interventions, the most plausible effects will instead be on the order of a 10–30 percent difference between treatment and control groups. These differences can affect large numbers of people, especially for common and serious disease outcomes, but are difficult to demonstrate reliably. In fact, such effect sizes may easily be as large as the amount of uncontrolled confounding or bias inherent in observational epidemiologic studies.

Thus, the randomized clinical trial is the optimal strategy on which to base a judgment as to the efficacy and safety of a given preventive measure of small to moderate size. Although the design of all randomized trials involves careful consideration of issues such as adequate sample size, sufficient duration of treatment and follow-up, maximization of compliance, and design efficiency (table 1), these issues become even more crucial in prevention trials intended to detect small to moderate effects, where the ultimate goal is to provide either a definitive positive result upon which clinical practice and public policy can be based or an informative null result that could then safely permit the rechanneling of resources to other areas of promising research.

ADEQUATE SAMPLE SIZE

Although sample size must be addressed early in the planning stage of any epidemiologic study, it has particular importance in prevention trials that aim to evaluate small to moderate effect sizes. The problem with a trial of inadequate sample size is its inability either to detect an effect if one truly exists or to demonstrate clearly the lack of an effect if one does not exist. Trials of inadequate size may fail to provide definitive tests of research hypotheses simply because they are too small to rule out the play of chance as a plausible alternative explanation for the findings. In fact, such trials can actually do scientific harm if, for example, their results are misinterpreted as evidence of no effect of an intervention when the trial simply had inadequate statistical power to detect a true effect. Thus, from a strategic standpoint, even if an investigator feels confident that a new intervention will have a large benefit (i.e., a 50 percent or greater reduction in the primary endpoint), it would be preferable to design the trial to test the hypothesis of a more likely small-to-moderate benefit. If the trial is designed to find only a large effect, there will not be adequate power to detect small but nonetheless clinically important differences. However, if a large effect does in fact emerge, the trial can be stopped early by the external Data and Safety Monitoring Board.

In the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (5), for example, funded by the National Cancer Institute, the primary objectives are to determine

Received for publication January 3, 2002, and accepted for publication March 27, 2002.

From the Division of Preventive Medicine, Brigham and Women’s Hospital, and Harvard Medical School, 900 Commonwealth Avenue East, Boston, MA 02215 (e-mail: jeburing@bics.bwh.harvard.edu). (Correspondence to Dr. Julie Buring at this address).
among men and women aged 55–74 years and free of cancer at baseline whether screening with flexible sigmoidoscopy can reduce mortality from colorectal cancer, screening with chest radiography can reduce mortality from lung cancer, screening men with digital rectal examination plus serum prostate-specific antigen can reduce mortality from prostate cancer, and screening women with cancer antigen 125 (CA125) and transvaginal ultrasound can reduce mortality from ovarian cancer. A total of 37,000 females and 37,000 males from 10 screening centers throughout the United States have been assigned at random to screening versus usual medical care and will be rescreened and followed for at least 13 years after randomization, in order to have adequate power to detect reductions in mortality of 20 percent, 10 percent, 15 percent, and 35 percent for prostate, lung, colorectal, and ovarian cancers, respectively. An external Data and Safety Monitoring Board examines the unblinded data on a yearly basis for evidence of unexpected benefit or harm that would modify or terminate the trial.

Recognition of this shortcoming of inadequate sample size to detect small to moderate effects has prompted not only the conduct of larger prevention trials but also the concomitant development of methodology to carry out such trials in a simple, streamlined, and cost-effective way (3, 6). Randomized trials, especially those testing an intervention in the treatment of a disease, traditionally have enrolled hundreds of subjects and then collected thousands of variables on each participant. Such extensive data collection, however, becomes logistically impractical in prevention trials with chronic disease outcomes such as cancer, and perhaps a decade or more before the effect becomes maximal (7). Thus changes made it far more difficult to detect declines attributable to the interventions under investigation.

In prevention trials with chronic disease outcomes such as cancer, another important consideration in determining the duration of treatment and follow-up is the proposed mechanism by which the intervention exerts its effects, as well as the latency period for the outcome of interest. Interventions such as micronutrient supplementation for the primary prevention of cancer are likely to require at least several years of treatment before any decrement in risk becomes apparent and perhaps a decade or more before the effect becomes maximal (7).

Although every effort should be made during the planning phase of the trial to choose the appropriate length of follow-up, the emergence of new evidence on mechanisms, secular declines in disease rates within the general population, and even the failure to achieve a sufficient sample size or to accrue sufficient endpoints within the trial itself may nonetheless raise the question of increasing the duration of the study beyond the planned period of follow-up. Because endpoints accumulate exponentially rather than arithmetically with continued follow-up of the study population, simply extending the duration of the study may convert an initially uninformative null result into a more conclusive finding. For example, the Multiple Risk Factor Intervention Trial initially found a statistically nonsignificant reduction in

### TABLE 1. Key design issues in primary prevention trials

- Adequate sample size
- Adequate duration of treatment and follow-up
- Achievement and maintenance of high compliance
- Run-in period
- Maximizing design efficiency
- Factorial design
- Selection of high-risk study population
coronary heart disease mortality of 7 percent, with a standard error of 12 percent, which was compatible with both the expected 22 percent decrease as well as with a null effect (8). Extending the length of follow-up led to a more definitive result’s being obtained for only a small increase in total cost. However, any such decision should be made as early in the trial as possible to maintain scientific credibility and avoid any implication that the change in design was a last-minute effort to achieve statistical significance.

ACHIEVEMENT AND MAINTENANCE OF COMPLIANCE

In addition to the total number of endpoints, another major factor that influences the power of a trial to detect a true difference between treatment groups is the difference in compliance between those groups. The effect of noncompliance in any participant, regardless of treatment assignment, is to make the intervention and treatment groups more alike, which decreases the trial’s ability to detect any true differences between the groups. High compliance in the study population is crucial to obtaining a valid result. By definition, any intervention study requires the active participation and cooperation of the study subjects. After agreeing to participate, subjects allocated to the active intervention group may deviate from the protocol for a variety of reasons, including the development of side effects, failure to take the medication or perform new behaviors, or simple withdrawal of consent after randomization. Those randomized to the comparison group may choose to obtain the active intervention on their own initiative. In the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial, for example, noncompliance or drop-out may occur in the arm randomized to the screening procedures, while contamination or drop-in may occur in the arm randomized to usual care, if these participants choose to receive nonstudy screening procedures. In addition, there may be instances in which participants cannot comply, such as when they develop a condition that requires attendance at home or clinic visit, telephone, or mail; the use of calendar packs of study medication, in which each pill is labeled with the day it is to be taken; and the use of incentives such as medical evaluations not ordinarily available from the participants’ usual source of health care. Streamlined protocols that collect only a few variables per subject will also minimize the burden on study participants and thus boost participation rates, especially among healthy individuals whose baseline tolerance for time-consuming tests and questionnaires may be lower than that of already ill patients seeking relief from disease-related symptoms. Such protocols also ensure the continued cooperation of clinical and administrative personnel, who need to balance the demands of the study with competing work-related responsibilities.

One compliance-maximizing strategy with wide applicability in prevention trials is the implementation of a “run-in” period prior to actual randomization. In this approach, all eligible participants experience a practice period for a number of weeks or months prior to formal randomization into a treatment group. This permits those who have difficulty adhering to the intervention program or those perceiving adverse effects to withdraw before randomization without detracting from the validity of the study. The actual format of the run-in period for a particular trial will depend on the hypothesis being tested. In studies of behavioral interventions, an appropriate test of willingness and ability to comply with the study protocol might consist of attendance at multiple screening visits, completion of forms similar to those that would be used in the actual trial, or willingness to undergo any laboratory procedures that might be required. In trials of pill-taking regimens, all the participants could take pills on the same schedule as they would in the trial, but with the exact formulation of the pills taken during the run-in (i.e., active agent vs. placebo) dependent on the postulated mechanisms of action and the state of knowledge of potential side effects.

The chief reason for using a run-in period is that the majority of subjects in a long-term trial who eventually become noncompliant tend to do so in the first months after initiation of the intervention. For example, in the Physicians’ Health Study, a completed randomized, double-blind, placebo-controlled trial of aspirin and beta-carotene
in the primary prevention of cardiovascular disease and cancer conducted among 22,071 US male physicians (9), 2-year pilot data indicated that approximately a third of participants stopped complying with the assigned treatment within 4 months, primarily because of difficulties in remembering to take a daily pill or because of the development of side effects due to aspirin. Following this initial drop in compliance, however, the loss in compliance over the remaining 20 months was small. Although the total sample size of randomized subjects will be lower if a run-in is implemented, there will be corresponding increases in power and efficiency, as well as decreased costs in tracking a smaller number of more cooperative people, which make the run-in a particularly attractive option for primary prevention trials, where it is not necessary to initiate an intervention during or immediately after an acute event. In the Physicians’ Health Study, it was demonstrated that, if 33,000 physicians were enrolled and 33 percent dropped out before being randomized, the remaining 22,000 would represent a group of proven excellent compliers who would be much more valuable to a long-term trial than the original 33,000 who could have been randomized immediately upon enrollment (10). Thus, implementation of a run-in period could substantially increase the power of the trial by yielding a group of committed compliers for long-term follow-up.

One possible limitation that has been raised with regard to the use of a run-in is that restricting a trial to a group of proven good compliers may result in a subject pool that differs from the general population with respect to outcomes. To the extent that the noncompliers who were eliminated can be followed, this question can be evaluated directly. However, it is important to remember that this issue of external validity, or generalizability, of trial findings will always be a matter of judgment (11), as those willing and eligible to be randomized into a trial are a select subgroup who may not be representative of the general population. The issue of generalizability does not, however, impact on the internal validity of the results, which may in fact be increased by the inclusion of proven good compliers. In fact, any procedure that maximizes compliance, thereby increasing the chances of obtaining a valid result, will indirectly assist in the assessment of generalizability, as the primary requirement of a generalizable study is that it be internally valid.

ISSUES IN DESIGN EFFICIENCY

Because prevention trials are often large in size and long in duration relative to other randomized trials, design strategies to increase the efficiency for a given sample size and to decrease costs need to be considered. Two approaches are the use of a factorial design and the analysis of “higher risk” subgroups.

Use of a factorial design

One approach that can improve the cost efficiency of a trial is to test two or more hypotheses simultaneously using a factorial design. In a $2 \times 2$ factorial design, for example, subjects are first randomized to treatment A or B to address one hypothesis; within each of these treatment groups, there is then further randomization to treatment alpha or beta to evaluate a second research question. Thus, participants in the trial are allocated to one of four possible regimens: each of the two active agents alone, both active agents, or both placebos.

The principal advantage of the factorial design is its ability to answer two or more questions in a single trial without loss of sensitivity and for only a marginal increase in cost (12). This is a particularly attractive option when one hypothesis is not at the time as scientifically mature as the other but is still promising and worthy of investigation. For example, when the Physicians’ Health Study was begun, the hypothesis that aspirin could be effective in the primary prevention of cardiovascular disease was widely accepted, given the extensive evidence from the secondary prevention or treatment trials of those who had already experienced a cardiovascular event (13). In contrast, the evidence for beta-carotene in the prevention of cancer was more preliminary, based on observational studies of dietary intake (14). Thus, although the beta-carotene hypothesis was promising, it was unlikely at that time to warrant funding of a trial to evaluate that question alone.

Ideally, additional treatments in a factorial design should not complicate trial operations, materially affect eligibility requirements, or cause side effects that could lead to poor compliance or losses to follow-up. In addition, the possibility of an interaction between treatment regimens must be considered. In the Physicians’ Health Study, because there was no reason to believe that aspirin and beta-carotene would interact with each other in any material way as their postulated effects worked via different mechanisms, a factorial design was an effective way to study both while compromising the effect of neither.

Fortunately, when interactions do exist, they tend to affect only the magnitude of observed treatment effects rather than their direction from one of benefit to harm or vice versa. Moreover, although the possibility of interactions is sometimes viewed as a potential limitation of a factorial trial, this design is in fact the only way to uncover their existence and evaluate, for example, whether two drugs in combination differ with respect to efficacy or side effects, than either drug alone (15). This has been particularly relevant in the evaluation of the chemoprevention potential of micronutrients, which are often made available to the general public in combination formulas.

Selection of a high-risk study population

As discussed earlier, accumulating a sufficient number of endpoints requires the specification of an adequate duration of follow-up. An additional strategy is to identify individuals who may be particularly likely to develop an endpoint of interest; that is, select a higher-risk group. One might consider factors such as age, sex, or personal or family history to identify such individuals. Alternatively, one might measure baseline biochemical parameters that are considered to be risk factors or that might modify the effect of the inter-
vention and then do a stratified analysis to assess the relative benefit among different subgroups. This will increase the sensitivity of the trial to identify which particular subgroup of participants, if any, stands to benefit most from the intervention. In the designing of the Physicians’ Health Study, for example, it was considered that if the true reduction in epithelial cancer incidence due to beta-carotene were at least 30 percent, the power to detect that difference using the postulated sample size of 22,000 would be adequate. On the other hand, if the risk reduction were only 10 percent overall, it would not be possible to detect such an effect with any assurance among this number of participants. However, this 10 percent overall reduction might result from the combination of a much larger effect confined exclusively to that subgroup of doctors who had low carotene, retinol, or retinol-binding protein levels at baseline with a much smaller or null effect in doctors lacking this biochemical profile. It was reasoned that this important public health finding could easily be detected if participants were stratified by baseline levels of these parameters, and that future public health intervention could then be aimed at that particular subgroup. The availability of these values would enable the investigators to identify which particular subgroup of doctors, if any, stood to benefit most from beta-carotene supplementation. Thus, prerandomization blood specimens were collected by mail from almost 15,000 of the participants. At the end of the trial, it was found that 12 years of supplementation with beta-carotene produced no evidence of benefit in terms of the incidence of malignant neoplasms (16). However, men in the lowest quartile of plasma beta-carotene at baseline assigned to beta-carotene supplementation had a significant reduction in risk of prostate carcinoma (17), supporting longer follow-up of participants in all randomized trials of beta-carotene supplementation for this endpoint.

Many questions of primary prevention have first been evaluated in trials of secondary prevention, that is, among those with a previous history of the disease. Because such individuals are at higher risk of a subsequent event than is the general population, trials of secondary prevention can in general be smaller in size and of shorter duration and still accumulate an adequate number of events. However, the demonstration of a benefit of an agent or intervention—or the finding of a lack of benefit—among those diagnosed with a disease, while suggestive, does not provide direct evidence on its role in primary prevention.

CONCLUSION

The ultimate goal of each methodological choice in any randomized trial is to allow clear refutation or confirmation of the hypotheses under investigation. Randomized trials are crucial to advance knowledge concerning possible interventions for the prevention of common and serious diseases in the general population. Careful consideration of design and conduct issues particularly relevant to prevention trials, which, due to the most likely postulated small- to moderate-sized effects, are usually larger in size and longer in duration than treatment trials, will help to ensure that either a definitive positive result or a truly informative null result is obtained.

REFERENCES