Risk Factors, Confounding, and the Illusion of Statistical Control

Nicholas J. S. Christenfeld, PhD, Richard P. Sloan, PhD, Douglas Carroll, PhD, and Sander Greenland, DrPH

Abstract: When experimental designs are premature, impractical, or impossible, researchers must rely on statistical methods to adjust for potentially confounding effects. Such procedures, however, are quite fallible. We examine several errors that often follow the use of statistical adjustment. The first is inferring a factor is causal because it predicts an outcome even after "statistical control" for other factors. This inference is fallacious when (as usual) such control involves removing the linear contribution of imperfectly measured variables, or when some confounders remain unmeasured. The converse fallacy is inferring a factor is not causally important because its association with the outcome is attenuated or eliminated by the inclusion of covariates in the adjustment process. This attenuation may only reflect that the covariates treated as confounders are actually mediators (intermediates) and critical to the causal chain from the study factor to the study outcome. Other problems arise due to mismeasurement of the study factor or outcome, or because these study variables are only proxies for underlying constructs. Statistical adjustment serves a useful function, but it cannot transform observational studies into natural experiments, and involves far more subjective judgment than many users realize. **Key words:** confounds, risk factors, statistical control, mediators, covariates.

BP = blood pressure; **SES** = socioeconomic status; **MI** = myocardial infarction.

INTRODUCTION

In exploring risk factors for various diseases, we are often forced, by timing, economics, or ethics, to use nonexperimental designs. These designs bring with them numerous interpretational problems, including the issue of confounding. People who drink more coffee may also smoke more cigarettes and drink more alcohol (1). Determining whether coffee drinking itself increases mortality risk, and is not just a marker for some other causal factor, must be approached not by random assignment, but by statistical means. The basic technique is to include measures of potential confounders as regressors (covariates) in a regression model, or stratify the data on these confounders. People then say they have "statistically controlled" or adjusted for the potential confounders.

There are many tasks that adjustment performs well. In experimental designs, covariate adjustment can reduce the noise in outcome variation, and thus allow the manipulation effect to stand out more clearly. Statistical adjustments perform markedly less well at the epidemiologic tasks to which they are regularly put. They simply cannot convert nonexperiments to experiments because "statistical control" is fundamentally distinct from experimental control (2,3). For example, successful randomization tends to minimize confounding by unmeasured as well as measured factors, whereas statistical control addresses only confounding by what has been measured and can introduce confounding and other biases through inappropriate control (2,4–6). We shall briefly examine, with examples, unjustified conclusions that can follow adjustment for potential confounders, such as inferring that something is a causal risk factor because it predicts an outcome even after "adjustment" for possible confounders, and inferring that a factor is not causally important because its impact is markedly attenuated or eliminated by the inclusion of covariates, as can happen when one adjusts for intermediate variables, or mediators (4,7). By causal risk factor, we mean that if this factor were altered, the outcome would be altered, whereas a marker is predictive but not necessarily causal, and its manipulation need not affect the outcome variable.

Such issues are treated in detail in certain epidemiologic texts (8,9) but seem to be underappreciated in behavioral medicine research. There are other, more subtle dangers in the use of covariates that we will not discuss here but can be found treated in some detail elsewhere (2–6,9).

Statistical Control: Necessary but Not Sufficient

It is fairly easy to find risk factors for premature morbidity or mortality (10). Indeed, given a large enough study and enough measured factors and outcomes, almost any potentially interesting variable will be linked to some health outcome. Many of these associations will be chance artifacts, but some will represent replicable phenomena. Discovering such associations is useful if one's goal is simply to predict disease. Even when not directly causal, associations can help target groups for health education or screening. For example, it is probably more useful to publish information about Tay-Sachs screening in *B'nai B'rith Magazine* than to publish it in *Christianity Today*. The difficulty comes, of course, when one wants to move beyond simple prediction into health intervention, or primary prevention; this requires that we distinguish between a *marker* of a disease condition and an actual *causal risk factor*. It would be one thing to find that *B'nai B'rith Magazine* readers are more likely to be carriers of Tay-Sachs; it would be another to suggest that canceling their subscriptions would help. The problem, of course, is that magazine subscription status is associated with many antecedent factors that are related to the Tay-Sachs gene, and so is confounded by these factors.

To examine the possibility that a particular factor is not causal, but just a marker for a causal factor, a researcher would include other known or plausible risk factors as covariates and determine whether adjustment for these potential

From the Department of Psychology (N.J.S.C.), University of California, San Diego, La Jolla, California; Department of Psychiatry, Columbia University (R.P.S.), New York, New York; School of Sport & Exercise Sciences, University of Birmingham (D.C.), Birmingham, England; and the Department of Epidemiology, University of California Los Angeles (S.G.), Los Angeles, California.

Address correspondence and reprint requests to Nicholas Christenfeld, PhD, Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109. E-mail: nicko@ucsd.edu

Received for publication December 4, 2003; revision received May 17, 2004.

DOI: 10.1097/01.psy.0000140008.70959.41

0033-3174/04/6606-0868

confounders eliminated the association between the putative causal factor and the health outcome of interest. If the association of the putative factor with disease vanished under these conditions, many researchers would conclude that it was simply a marker, and not itself part of the causal chain.

When the adjustment does not eliminate the association of the putative causal variable with disease, many researchers report that a new risk factor has been identified and argue for interventions targeted at this new risk. Before concluding that the factor identified is indeed causal, however, one needs to assess whether further adjustment for unmeasured potential confounders could eliminate the association and whether the potential confounders that were controlled were accurately measured (11,12), as well as establishing the temporal precedence of the putative causal factor (8,9,13). These needs may seem especially obvious when the study factor lacks biologic plausibility as a cause (e.g., as in the case of the creased ear lobe and elevated risk of heart disease) (14).

Because of the intuitive implausibility of reading a particular magazine being a causal risk factor for carrying Tay-Sachs, if adjusting for known risk factors (e.g., ethnic origin) did not eliminate the association, one would reasonably think that either further adjustment is needed (e.g., for geographic region), or one might question the accuracy of the included covariates (many people of Ashkenazi origins might not list Jewish as their ethnicity, preferring, say, Polish, Latvian-American, or, possibly, American). But, when causality is plausible, or when one is keen to accept a causal role, simple *caveats* often go unheeded.

Recently, it was reported that people who said that they seldom took vacations were at significantly greater risk of mortality than those who reported more frequent vacations (15). It is not hard to imagine a mechanism: vacations might mitigate stress (16), diminish anger (17), and encourage more exercise (18). Nonetheless, it is also possible that people who are healthier are more likely to go on vacations, and so vacations are not a causal factor, but only a marker of initial health status, which, naturally, will predict longevity. The authors considered this possibility and reported that vacation frequency remained a significant predictor even after adjusting for covariates such as baseline health status. One can think, however, of numerous additional potential confounders-more, in fact, than any epidemiologic study could possibly hope to have measured. Perhaps people who have more friends are more likely to take vacations to visit them, and it is having friends that is protective. Perhaps people with less demanding work settings are able to take vacation time, and it is having a low-stress work environment that promotes health. Perhaps people who travel to explore varied environments also enjoy varied food, and it is the completeness of the diet that prolongs life.

It is possible to construct a very long list of potential confounders. Of course, many would lack plausibility. Nonetheless, plausibility is a highly subjective consideration for determining whether enough potential confounders are taken into account. Identification of confounders requires *a priori* knowledge of the likely causal pathways (3,5,8,9,19–21). Unfortunately, this observation implies that the strength of any causal inference depends on the biologic plausibility of the putative factor, and the implausibility of uncontrolled potential confounders. Causal inference from observational data thus has to contain a judgmental component, which can vary across experts.

Mismeasurement and Mis-specification

While it is entertaining to come up with additional uncontrolled confounders to explain the association of vacations and health, a more subtle problem is that factors that have been statistically controlled but imperfectly measured can nonetheless still be responsible for the association, a problem referred to as residual confounding due to mismeasurement (8,9,11). To assess infrequent vacations as a risk factor, for example, it is critical to account for the possibility that sick people cannot vacation. To accomplish this, measures of initial health may be used for adjustment. However, no measure of initial health is assessed without error. Furthermore, there are a vast number of reasonable ways of measuring initial health, and clearly not all can be included. Would an initial stress test capture the critical aspect of health that is confounded with vacations? Would it be body mass index? Self-reports of vigor?

Even if the optimal measure of the confounder is chosen, but it is measured with error, then adjustment for it may not eliminate the effect of vacations. As long as vacations are a good marker of health, and the actual measure of health is crude, vacations can retain their predictive power in the model. For example, the number of cigarettes somebody buys could remain a significant predictor of lung disease, even after controlling for "smoking," if smoking is measured as the number of cigarettes the person lit during a 1-hour assessment. Even more sophisticated measures of smoking will have error, as one cannot assess the whole history of smoking frequency, the depth of inhalation, the duration before exhalation, or the amount of rebreathing of secondary smoke. In addition, without knowing the time course of cigarette damage, or the times when people might be particularly vulnerable to cigarette effects, it is not possible to assess without error the association of smoking with the study factor or the effect of smoking on disease within the study (even if much is known about these effects in previous studies).

Model Specification Issues

A variant of the problem of poorly measured confounders is that the effect of confounding factors may not be linear. Assuming linearity on the outcome scale specified by the model, as one does by simply entering the confounder as a covariate in standard regression models, may fail to fully adjust for the confounder when the confounder effects are not linear on that scale (9,22). For example, suppose the actual risk of lung cancer is proportional to the square of the number of cigarettes one smokes. If one used only a linear term for smoking (operationalized as number of cigarettes smoked) in a logistic risk model, one would fail to control confounding attributable to the quadratic component of the smoking effect on the logit of risk (i.e., on the log odds scale).

Linearity is a strong but rarely accurate assumption. There is no reason to think that each additional cigarette smoked damages the lungs to the same extent, nor that each dollar of income adds the same increment to longevity. It is sometimes possible to take nonlinearity into account by adding nonlinear terms to a linear model, or by plotting the relationship between the variables of interest, and transforming them to create linearity. However, standard models for risks and rates use a linear model for the logit or logarithm of the outcome; as a result, they automatically imply a nonlinear relation between exposure and outcome on the original untransformed outcome scale (9). An approach that avoids such implications is to break the continuous variable into categories and allow the slope to vary from level to level, as in spline regression (9). Because of limitations of sample information and mathematical simplifications, however, one should not expect to be able to capture the true relation with much accuracy. Without enough data to determine, or enough knowledge to deduce the true, nonlinear relationship, one may have to default to an assumption of linearity on some scale, although this comes at the cost of possible bias in estimating true effects.

Confounded Errors

It has been suggested that religious involvement confers a health advantage, and a great many studies have been conducted to examine this possibility. Because researchers must rely on observational designs in which participants are self-selected for varying degrees of religious observance, these inquiries illustrate the problems described above. For example, several such studies report that people who attend church more frequently show reduced mortality (23–25). Because mobility is required to go to church, the researchers considered the possibility that people who start out healthier go to church, and would have lived longer regardless. They dealt with this problem by adjusting for measures of initial health status in regression models, and reported that church attendance remained linked to longevity. However, as with our earlier examples, there is no measure of health that captures everything important about one's state, and there are limitless other potential confounders.

The ability to attend church is not only likely to covary with initial health, but simple measures of church attendance may have confounded errors. Most studies of attendance at religious services and health outcomes collect data from interviews, during which participants are asked questions about how frequently they attend religious services. Presser and Stinson (26) contrasted the results of national surveys asking such questions with data from the Bureau of Labor Statistics on how people use their time. To collect data on time use, researchers asked subjects what they did during the last week, and Presser and Stinson examined subjects' activities on Saturdays and Sundays, reasoning that if subjects attended religious services, they would report this among their weekend activities. Religious attendance measured by time-use surveys was substantially lower than estimates derived from interview data. Presser and Stinson suggested several reasons to account for this difference, one of which was that subjects, when interviewed, may perceive questions like "how frequently do you attend religious services?" as a questions about how good a person they are, and so exaggerate their attendance.

Given that church attendance is a factor that lends itself fairly readily to quantification, it is unlikely that the problem is any smaller for other measures, such as membership in religious institutions, self-reported frequency of prayer, reading the Bible, watching or listening to religious TV or radio, and indices of comfort provided by religion. Some researchers use scales designed to summarize these behavioral dimensions in an aggregated index of religiosity. None of these scales can avoid the problem of measurement error, however, since scales are affected by errors in any of their components. The problem is not only that measures of church attendance lack accuracy, but also that they are likely influenced by other factors, some of which may be associated with health. That is, the noise in the measure is not likely to be random, and thus it might be these other factors in the measure, rather than the religiosity itself, that predicts health.

Replication

Simple replication is not a solution to the problems discussed here, for the same problems may be present in further studies. In some cases, it may be possible to find populations in which the putative risk factor is not associated with certain potential confounders, or related in the direction contrary to the one that usually obtains (27). For example, one might find a culture where physical inactivity, smoking, and high fat diets were characteristic of wealth instead of poverty. If the usual link between, say, socioeconomic status (SES) and longevity were still observed in such a population, one could now infer with greater certainty that it was not simply attributable to the poor's greater attachment to bad health behaviors. In practice, such populations may be hard to find, and the confounding will remain similar in replications.

The difference between studying a population in which the potential confounder does not exist and a population in which it must be statistically controlled is illustrated in work on depression and health following a coronary event. Depression is often found to be a significant predictor of short-term cardiac mortality following myocardial infarction (MI), and this relationship sometimes persists after statistical adjustment for potentially confounding measures of initial disease severity (28). The often imprecise assessment of baseline cardiac disease severity (29), however, makes inferring a causal role for depression difficult. Studies of the predictive utility of depression in which it was not associated with usual measures of cardiac disease severity on entry to the study have found little relationship between depression and mortality following MI (30–32). Consistent with such reports are results of the ENRICHD study, a randomized controlled trial aimed at reducing symptoms of major and minor depression in MI patients in order to promote cardiac health. Although the trial was moderately successful in alleviating symptoms of depression, it does not appear to have reduced all-cause mortality or the recurrence of non-fatal MI over an average follow up period of 41 months (33,34). While such null findings are certainly not definitive, they remind us that depression may simply be a marker of disease severity, not an underlying cause of death. Statistical adjustment using the linear component of imperfect measures of disease severity is not sufficient to discount such a possibility.

A related point is made by results of recently published research on the association between perceived stress and cardiovascular disease, in which the population studied is one in which high stress levels are not linked, as they frequently are, to socioeconomic disadvantage (35,36). In a population of more than 5000 Scottish men, followed over 21 years, it was men in higher occupational class groups that had higher perceived stress scores at entry to the study. This result may reflect historical variations in discourse patterns as much as, if not more than, variations in stress exposure. In the early 1970s, when the perceived stress data were collected, it is likely than stress was more comfortably part of the vocabulary of higher as opposed to lower SES groups. In terms of 21-year health outcomes, the prevalence and incidence of angina, diagnosis of which relied largely on symptom reporting, increased with increasing perceived stress. However, no such positive associations emerged for ischemia, mortality from cardiovascular and coronary heart diseases, or all-cause mortality. The usual inverse association between mortality and SES obtained.

It is possible that the positive association between higher perceived stress and angina reflected reporting bias, namely, a tendency for participants reporting higher stress to also report more symptoms of chest pain. The absence of an analogous association with more objective measures of heart disease suggests that the reported symptoms may not have reflected cardiac disease. These analyses illustrate how a putative cause (perceived stress) might emerge in other studies as a predictor of health outcomes: through reporting bias for subjective outcomes like angina and through confounding (of stress by SES) for objective outcomes like death. The results of this study do not conflict with previously reported associations between measures of stress and health outcomes; rather, they demonstrate how such associations could emerge spuriously, and so conflict with certain conclusions drawn from previous studies.

Constructs Versus Operationalizations

Many of the foregoing problems arise because there is usually a difference between a particular operationalization and the underlying construct. In assessing the effect of religiosity on health, we might like to control for initial health status, and we may even report that we have done so. The best we can do, however, is adjust for the effects associated with a particular measure of health, or limited set of measures, using a particular statistical method. Similarly, if we want to demonstrate the benefits of religiosity for longevity, we might measure church attendance. However, church attendance is not the same as religiosity, and self-reports of attendance are certainly not (26). Even if there is a benefit, it could be due to the weekly walk, the cookies after the service, genuflecting, or any number of other things. After all, a measure that one person calls religiosity another might call social support.

The use of conceptual constructs for the variables of interest is an essential shorthand for research. We cannot hope to describe all of the differences between people who check, on a questionnaire, that they attend religious services weekly and those who check the box suggesting they do not. We may call the whole package religiosity, but this does not mean we have defined (let alone measured) religiosity in an accurate way. Similarly, health is not the same as a score for activities of daily living. The solution is not simple replication, for using the same operationalization would not tell us what the operationalization misses. One proposed solution is conceptual replication, using operationalizations that share nothing but the construct of interest with the original ones. Preparing people for medical procedures seems to have benefits, whether that preparation is done by videotaped message, physician, or hospital roommate (37). Relative poverty is associated with reduced life expectancy, even when tombstone size, rather than a more traditional measure, is used as the operationalization of SES (38). In pregnant women, the effects of stress on gestational length seem to hold when that stress is an earthquake, rather than indexed by a high score on self-report inventories (39). Discrepancies among the findings that emerge with different operationalizations can provide clues as to which aspects of the construct are critical.

The gap between the theoretical construct and the particular operationalization is similar to the problem of "reification," an issue that Gould (40) illustrates with intelligence testing. A score on an IQ test is not the same as intelligence, just as a high score on a scale for activities of daily living is not health. Creating an IQ test and giving it a name can obscure the conceptual leap and can suggest that there is a real underlying entity that can be completely captured by scores on a unidimensional scale, when there is no such entity.

To be sure, the leap from construct to operationalization is sometimes a small one. For example, biologic sex is fairly easy to capture, reliably and validly, if not quite perfectly. A recent report on excess heart disease among New York City residents and visitors (41) concluded that the elevations in mortality were not due to an unusual sex ratio. We trust this result because we believe that the measure captures a very real and profound difference among individuals (the presence or absence of a Y chromosome) with little error. More problematically, however, the same report concludes that the differences are not due to SES or ethnicity. In reality, the appropriate claim is that differences in mortality are not explained by adjustments for rough measures of particular operationalizations of these constructs. The number of years of education as recorded on a death certificate is not SES, even though it may be about as good as any other simple measure. Similarly, the division into white, black, Hispanic and other is not "race," a complex social and personal construct. Given the limitations of death certificates and other available data, the researcher often has little choice but to adopt such approximations. Nonetheless, with the exception of a few concepts, such as sex and all-cause mortality, inexactness of measurement and the multidimensional nature of the most constructs make it unlikely that any simple measure will capture the construct.

Mediators Versus Confounders

The final problem we shall discuss does not arise from limited reliability or validity. It instead arises from the dilemma that several different causal explanations are possible when adjustment does reduce or eliminate the predictive power of the study exposure. One possible explanation is confounding, ie, the study exposure is a marker of some causal factors but is not itself directly involved in the causal chain from exposure to disease. Thus, for example, one might dismiss reading the *B'nai B'rith Magazine* as a causal factor in carrying Tay-Sachs if one eliminates its predictive power by including Jewish religious affiliation as a covariate. In another context, however, the study exposure might have a causal effect that is mediated by the variables used for adjustment. This is a problem of overadjustment (9). In such a case, one might falsely conclude that the study factor is not causally important, and never test possibly efficacious interventions based on that factor. Including a mediator in the model will usually reduce or eliminate the predictive power of the original factor (9,20,42), but eliminating its predictive power does not refute causality any more than establishing its predictive power demonstrates causality.

Consider the hypothesis that excessive blood pressure (BP) reactions to stress lead to hypertension. We could test this idea by measuring BP reactivity and resting BP levels in a large group of people. We would then follow these people, ideally for decades, and, by measuring resting levels again, determine which people showed a significant rise in levels, or became hypertensive. Given our hypothesis, we ought to find that excessive reactivity is a risk factor for later hypertension, but we might be concerned that reactivity is just a marker for elevated BP resting levels, and is not important per se. Consistent with this concern, suppose those with somewhat elevated resting BP at the initial testing were the ones with large BP reactivity scores. To control for this apparent confounder, we would adjust for initial resting BP levels in regression analyses. This would tell us whether BP reactivity contributes any predictive information beyond simply knowing initial resting BP level. Such an analysis may reveal that reactivity was no longer very predictive, with most of the variation in follow-up blood pressure levels being accounted for by initial resting levels.

If our goal were to predict who was most likely to develop hypertension, it would be reasonable to dismiss reactivity, given the modest amount of additional variation it explained and the extra cost involved in its measure, and focus instead on initial resting levels. However, reactivity still could be causally related to future BP status, e.g., if heightened reactivity preceded initial elevated resting BP level, it could be responsible, in part, for the initial elevated resting BP level. Thus, an intervention that reduced reactivity might still decrease the likelihood of developing hypertension. In this example, controlling for BP levels is the equivalent of failing to predict the outcome of a foot race from genuine causal factors, like fitness, because we have controlled for the runners' position halfway into the race.

Other examples show that we may face situations in which a single variable may have both confounding and mediating roles. Suppose that people who take more vacations are less likely to die over some 5-year study period, but including initial health status in the regression model eliminates this association. If people in poor health take fewer vacations, this elimination might in part reflect removal of confounding by health status. However, if people's vacationing tendencies are fairly constant over the years, then health status on entry to the study will reflect the cumulated health impact of a lifetime's vacation habits, and health status will in part be a mediator of vacationing effects. Similarly, if regularly attending church provides health benefits, then initial health status will also be a mediator of the effect of church attendance.

The confusion between mediators and confounders will be less of an issue if the risk factor is not stable over time. If a person only becomes a high BP reactor shortly before the first testing session, then this will not yet be reflected in resting BP levels, and reactivity will then have an opportunity to predict later hypertension even controlling for resting levels. Similarly, if a person has only just started to take vacations, or attend church, then these behaviors will not be reflected in initial health status and will have the opportunity to predict subsequent health, with initial health status included as a covariate. If these changes are out of the control of the subject, it can create at least a quasi-experimental design. That is, if people started taking vacations because of a change in company policy (rather than because they suddenly felt vigorous, made friends, or had free time), and other people stopped for the same reason, then it would be possible to assess the effect of vacation independently of initial health status. This is the situation exploited by research on post-coronary artery bypass graft recovery and the influence of social support (43). In this work, since the social support of interest is that provided by hospital roommates, it is separate from the support available before the patients needed bypass surgery. Thus, its status as a predictor of recovery can be determined largely independently of any effects social support may or may not have had on initial health status. Looking at changes in the risk factor is similar to predicting stock prices. Knowing how well a company is run will not help a securities trader, since that information is already fully incorporated, or discounted, in the price of shares (44). Knowing about changes in management, however, will help predict price fluctuations.

In these examples, some of the effect of the study exposure is already reflected in the mediator, and so adjustment for the mediator masks the total impact of the exposure. On the other hand, the mediator may still confound exposure effects. The only solution to such problems is to employ methods that appropriately account for the time-varying nature of the mediator and the study factor; these methods require accurate repeated measures of both variables (7,45).

There are numerous such factors that can differentially affect etiology and prognosis, and this makes statistical control for health status at some intermediate point problematic. Living near a trauma center should enhance recovery, without lowering accident risk. Being sober on the highways offers considerable protection from crashing, but once they have crashed, it is not necessarily the case that sober people will recover faster. Vacations might reduce the risk serious illness more than they speed recovery. There are many other increasingly complicated and unwieldy possibilities that we are happy not to enumerate. However, the general point is that indiscriminate adjustment for covariates can lead to peculiar and erroneous conclusions.

The effects of most sociodemographic variables are mediated by other factors. Having a small income and an unfulfilling job, no spouse or friends, no vacations, and being a man or an atheist do not directly cause illness. There will be numerous intermediate variables, ranging from general ones like a sense of self-determination to the biochemical, such as release of stress hormones. Including accurate measures of any of these variables can reduce or even eliminate the predictive power of sociodemographic factors, and thus make them appear unimportant. Nonetheless, eliminating predictive power by controlling mediators does not in any way imply that a sociodemographic variable is fruitless for study or for meaningful interventions, or that it is not causally important. It may even be a more effective variable for intervention than intermediate biologic variables.

In experimental work, the putative causal factor is manipulated, and mediators are easily identified by their occurrence after the intervention. One can conclude, within statistical error, that differences between groups were produced by treatment assignments (intent-to-treat), rather than confounding (although noncompliance may still lead to confounding of *received* treatment). With observational designs, however, distinguishing a mediator from a confounder must often rely on common sense, intuition, background knowledge, or one's theoretical or ideological persuasion. If one found that the longevity differences between doctors and nurses changed on sex adjustments, one would be safe assuming that sex was a confounder, rather than a mediator (although other confounders would still need to be considered). However, looking at high-stress jobs and health, should one regard smoking and drinking habits as confounders, or mediators? Without randomization of jobs, one's conclusions are necessarily limited and are highly dependent on background information and educated guesses. One person's mediator may be another person's confounder, even when only one of the two can be right, as certain medical controversies have illustrated (20).

The other implication of the multiply-mediated nature of risk factors is that it should be possible to find unlimited numbers of genuine causal risk factors for just about any health outcome. If, for example, eating French fries increases heart-disease risk, then any factor that increases French fry consumption will be causally linked to health. Those factors could include poverty, having a high-stress job, residing near a fast-food restaurant, social isolation, and distaste for leafy-green vegetables. All these factors could all increase the number of fries in the diet, thus causing heart disease. In other words, they could all be independent causal risk factors, although all are mediated by a common pathway, the French fry.

An additional problem is that controlling a mediator can create confounding, and this new confounding may inflate or reduce the observed association; it may even create a completely spurious association when no effect is present. We do not discuss these problems, but illustrations can be found (4,5,7).

In conclusion, despite the inherent limitations of nonexperimental studies, they serve important, even essential, functions in behavioral and health research, for randomized trials are often impractical or unethical (46). Even when they can be done, randomized trials may fail to answer the original question when they employ a questionable operationalization of the construct of interest. One need look no further than discovery of the dangers of tobacco, asbestos, hormone replacement, and various industrial chemicals, to find cases in which observational studies have greatly benefited public health, although those studies were often supplemented by laboratory experiments and pathology or toxicology. As weaker and weaker associations have been explored, however, failures have occurred, for example, overly broad if not completely erroneous claims for the benefits of beta-carotene supplementation in preventing cancer, or in hormone replacement therapy in preventing heart disease, each refuted by later large-scale randomized trials. Many other proposed interventions remain to be subjected to such tests, and hence suitable caution is needed. The failures of observational research underscore the importance of understanding potential sources of error, such as inadequate control of confounders, controlling only for poor assessments of confounders, discounting a construct as an explanation when only one particular operationalization can be discounted, and confusing confounders and mediators.

Recognizing the hazards of the statistical control of confounding should lead to special cautions in the conclusions offered in published findings. Referring to these statistical adjustments as reducing the plausibility of alternative explanations, rather than as eliminating them, seems more in keeping with their potential. The measurement of covariates should be accorded as much attention in the design as in the assessment of the critical predictor and outcome variables. If one wants to measure the effect of religiosity, say, on health outcomes following an MI, one should design the initial health status covariate measure at least as carefully as the outcome measure. The payoff for a finely crafted exposure or outcome measure may be obvious. Perhaps because of the secondary role played by confounders, there seems to be less incentive to invest in confounder measurement, and perhaps even some temptation to get by instead with simple, convenient measures.

Examples such as we have discussed illustrate why treating confounders lightly can be counterproductive. Just as studies often include multiple measures of the critical outcome, they can include multiple measures of a potential confounder. For example, in work on depression and mortality following an MI, it has been found that in the critical assessment of initial health status, length of initial hospital stay, Killip Class, and Peel index are still usefully supplemented by a measure of discharge medication, which appears to be capturing some additional, unique information about the attending cardiologists' assessment of illness (31,47). Although all of these measures together may not adequately capture initial health, omitting any one of them (e.g., discharge medications) will almost certainly lead to uncontrolled confounding.

Short of randomization, the only way to avoid grappling with the above issues for a particular confounder is to find a population in which one can be certain that the confounder varies little. For example, confounding by initial health status would unlikely be strong in a cohort of new military recruits, simply because these recruits would have been selected for good health. But even then, confounding by other factors (such as race) will still have to be controlled.

There are times at which research is concerned simply with prediction rather than intervention (as happens sometimes in health-services research), in which the confounding issues we have discussed are not of as great concern. But more often intervention is the ultimate goal, in which case causal inference is essential. In those settings, it is essential to remember that "statistical control" is nothing more than a highly fallible process filled with judgment calls that often go unnoticed in practice.

REFERENCES

1. Talcott GW, Poston WS 2nd, Haddock CK. Co-occurrent use of cigarettes, alcohol, and caffeine in a retired military population. Mil Med 1998;163:133–8.
2. Greenland S. Quantifying biases in causal models: classical confounding vs. collider-stratification bias. Epidemiology 2003;14:300–6.
3. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999;10:37–48.
4. Cole SR, Hernan MA. Fallibility is estimating direct effects. Int J Epidemiol 2002;31:163–5.
5. Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol 2002;155:176–84.
6. Greenland S, Brumback BA. An overview of relations among causal modelling methods. Int J Epidemiol 2002;31:1030–7.
7. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology 1992;3:143–55.
8. Szklo M, Nieto FJ. Epidemiology: beyond the basics. Gaithersburg, MD: Aspen; 2000.
9. Rothman KJ, Greenland S. Modern epidemiology. Philadelphia: Lippincott-Raven; 1998.
10. Hopkins PN, Williams RR. A survey of 246 suggested coronary risk factors. Atherosclerosis 1981;40:1–52.
11. Greenland S. The effect of misclassification in the presence of covariates. Am J Epidemiol 1980;112:564–9.
12. Davey Smith G, Phillips AN, Neaton JD. Smoking as 'independent' risk factor for suicide: illustration of an artifact from observational epidemiology. Lancet 1992;340:709–12.
13. Kraemer HC, Stice E, Kazdin A, Offord D, Kupfer D. How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. Am J Psychiatry 2001;158:848–56.
14. Elliott WJ, Powell LH. Diagonal earlobe creases and prognosis in patients with suspected coronary artery disease. Am J Med 1996;100:205–11.
15. Gump BB, Matthews KA. Are vacations good for your health? The 9-year mortality experience after the multiple risk factor intervention trial. Psychosom Med 2000;62:608–12.
16. Pickering T. Cardiovascular pathways: socioeconomic status and stress effects on hypertension and cardiovascular function. Ann NY Acad Sci 1999;896:262–77.
17. Siegman AW, Smith T, editors. Anger, hostility, and the heart. Hillsdale, NJ: Lawrence Erlbaum; 1994.
18. Tikkanen HO, Hamalainen E, Sarna S, Adlercreutz H, Harkonen M. Associations between skeletal muscle properties, physical fitness, physical activity and coronary heart disease risk factors in men. Atherosclerosis 1998;137:377–89.
19. Greenland S. Randomization, statistics, and causal inference. Epidemiology 1990;1:421–9.
20. Greenland S, Neutra RR. Control of confounding in the assessment of medical technology. Int J Epidemiol 1980;9:361–7.
21. Robins JM. Data, design, and background knowledge in etiologic inference. Epidemiology 2001;12:313–20.
22. Royston P. A useful monotonic non-linear model with applications in medicine and epidemiology. Stat Med 2000;19:2053–66.
23. Oman D, Reed D. Religion and mortality among the community-dwelling elderly. Am J Public Health 1998;88:1469–75.
24. Hummer RA, Rogers RG, Nam CB, Ellison CG. Religious involvement and U.S. adult mortality. Demography 1999;36:273–85.
25. Koenig HG, Hays JC, Larson DB, George LK, Cohen HJ, McCullough ME, Meador KG, Blazer DG. Does religious attendance prolong survival? A six-year follow-up study of 3,968 older adults. J Gerontol A Biol Sci Med Sci 1999;54:M370–6.
26. Presser S, Stinson L. Data collection mode and social desirability bias in self-reported religious attendance. Am Sociol Rev 1998;63:137–45.
27. Phillips AN, Davey Smith G. Bias in relative adds estimation owing to imprecise measurement of correlated exposures. Stat Med 1991;11:953–61.
28. Frasure-Smith N, Lespérance F, Talajic M. Depression and 18-month prognosis after myocardial infarction. Circulation 1995;91:999–1005.
29. Carroll D, Lane D. Depression and mortality following myocardial infarction: the issue of disease severity. Epidemiol Psichiatr Soc 2002;11:65–8.
30. Lane D, Carroll D, Ring C, Beevers DG, Lip GYH. Mortality and quality of life 12 months after myocardial infarction: effects of depression and anxiety. Psychosom Med 2001;63:221–30.
31. Lane D, Carroll D, Ring C, Beevers DG, Lip GYH. In-hospital symptoms of depression do not predict mortality three-years after myocardial infarction. Int J Epidemiol 2002;31:1179–82.
32. Mayou RA, Gill D, Thompson DR, Day A, Hicks N, Volmink J, Neil A. Depression and anxiety as predictors of outcome after myocardial infarction. Psychosom Med 2000;62:212–8.
33. Berkman LF, Blumenthal J, Burg M, et al. effects of treating depression and low perceived social support on clinical events after myocardial infarction: the Enhancing Recovery in Coronary Heart Disease (ENRICHD) randomized trial. JAMA 2003;289:3106–16.

34. Louis AA, Manousos IR, Coletta AP, Clark AL, Cleland JGF. Clinical trials update: The Heart Protection Study, IONA, CARISA, ENRICHD, ALIVE, MADIT II and REMATCH. Eur J Heart Fail 2002;4:111–6.
35. Macleod J, Davey Smith G, Heslop P, Metcalf C, Carroll D, Hart C. Are the effects of psychosocial exposures attributable to confounding? Evidence from a prospective observational study on psychological stress and mortality. J Epidemiol Commun Health 2001;55:878–84.
36. Macleod J, Davey Smith G, Heslop P, Metcalf C, Carroll D, Hart C. Psychological stress and cardiovascular disease: empirical demonstration of bias in a prospective observational study of Scottish men. Br Med J 2002;324:1247–53.
37. Suls J, Wan C K. Effects of sensory and procedural information on coping with stressful medical procedures and pain: A meta-analysis. J Consult Clin Psychol 1989;57:372–9.
38. Smith GD, Carroll D, Rankin S, Rowan D. Socioeconomic differences in mortality: evidence from Glasgow graveyards. Br Med J 1992;305:1554–7.
39. Glynn LM, Wadhwa PD, Dunkel-Schetter C, Chacz-Demet A, Sandman CA. When stress happens matters: effects of earthquakes on stress responsivity in pregnancy. Am J Obstet Gynecol 2001;184:637–42.
40. Gould SJ. The mismeasure of man. New York: Norton; 1996.
41. Christenfeld N, Glynn LM, Phillips DP, Shrira I. New York City as a risk factor for heart attack mortality. Psychosom Med 1999;61:740–3.
42. Judd CM, Kenny DA. Process analysis: estimating mediation in treatment evaluations. Eval Rev 1981;5:602–19.
43. Kulik JA, Mahler HIM, Moore PJ. Social comparison and affiliation under threat: Effects on recovery from major surgery. J Pers Soc Psychol 1996;71:967–79.
44. Fama EF. Efficient capital markets: a review of theory and empirical work. J Finance 1970;25:383–433.
45. Robins JM. Control of confounding by intermediate variables. Stat Med 1989;8:679–701.
46. Smith GGS, Pell, JPP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. Br Med J 2003;327:1459–61.
47. Lane D, Lip GYH, Carroll D. Is depression following acute myocardial infarction an independent risk for mortality? Am J Cardiol 2004;93: 1333–4.