

## Alternatives to randomized trials for estimating treatment effects

### Introduction

---

We said in Chapter 9 that randomized blinded trials are the best way of determining treatment effects because they minimize the potential for confounding, co-interventions, and bias, thus maximizing the strength of causal inference. However, sometimes observational studies can be attractive alternatives to randomized trials, because they may be more feasible, ethical, or elegant. Of course the issue of inferring causality from observational studies is a major topic in classical risk factor epidemiology. In this chapter, we focus on observational studies of treatment effects rather than risk factors, describing methods of reducing or assessing confounding that are particularly applicable to such studies.

### Confounding by indication

---

We discussed in Chapter 9 that confounding refers to the distortion of the effect of variable A on the outcome C by a third variable B, which is both associated with A and a cause of C. We focus on treatments that are supposed to be beneficial, that is, to have a  $RR < 1$  for a bad outcome. One type of confounding makes treatments appear better than they really are – for example, finding a beneficial treatment effect when, in truth, the treatment either has no effect or causes harm. In this situation, a confounder is associated with receiving the treatment and reduces the risk of a bad outcome (Fig. 10.1).

An example is use of vitamin E to prevent cardiovascular disease. Multiple observational studies suggested a protective effect, but randomized trials have found no benefit (Eidelman et al. 2004), suggesting that some other factors (e.g., better diet, exercise, or health awareness) are the true cause of the lower risk of cardiovascular disease among users of vitamin E (Fig. 10.2).

Alternatively, when a confounder that is associated with receiving the treatment increases the risk of a bad outcome, it can mask or reduce the apparent benefit of the

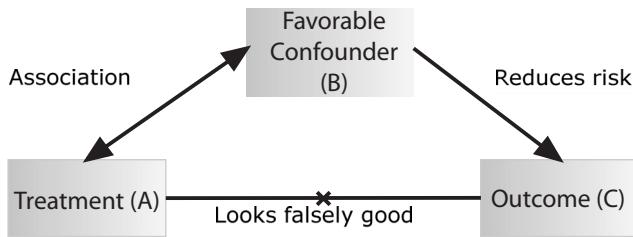


Figure 10.1 Confounder is the true cause of improved outcomes making the treatment look more effective than it really is.

treatment<sup>1</sup> (Katz 1999). For example, if only the sickest people get the treatment in question, the treatment may look harmful even when it actually helps. This effect is often called “confounding by indication,” because often those in whom the treatment is most indicated are those at highest risk of the bad outcome that the treatment is designed to prevent.

An example of confounding by indication is diuretic treatment of hypertension in diabetics. It is now clear that treating hypertensive diabetics with diuretics reduces their risk of cardiovascular mortality (Turnbull et al. 2005). However, a cohort study (Warram et al. 1991) appeared to show that diuretics increased the risk of cardiovascular mortality in hypertensive diabetics compared with leaving the hypertension untreated. The confounder was the severity of cardiovascular disease. The patients with more severe disease were more likely to be treated with diuretics, and they were also more likely to die of their cardiovascular disease (Fig. 10.3).

Confounding by indication can be controlled the same way as any other type of confounding. However, thinking about confounding by indication takes some getting used to, because in classical risk-factor epidemiology, investigators are usually looking for factors that increase, rather than decrease, risk. In epidemiologic studies, protective factors that decrease risk tend not to be selectively present in those otherwise at higher risk, as happens when treatments are given to people at higher risk of the outcome. Thus, suppression is more likely to be an issue in studies of treatments than in studies of naturally occurring risk factors.

## Instrumental variables

When we discussed the “Intention to Treat” principle in Chapter 9, we acknowledged that, in randomized controlled trials, there might be an imperfect relationship between the predictor variable of interest (e.g., actual receipt of medication) and the predictor variable analyzed (group assignment). We stressed that, to maintain the strength of causal inference provided by randomization, it is important to analyze by group assignment – that is, people assigned to take the medication should be compared with people assigned to placebo, rather than comparing people who took the medication with those who did not. However, if this analysis results in

<sup>1</sup> This type of confounding is sometimes referred to as *suppression* and the confounder is referred to as a *suppressor*. (Katz 1999)

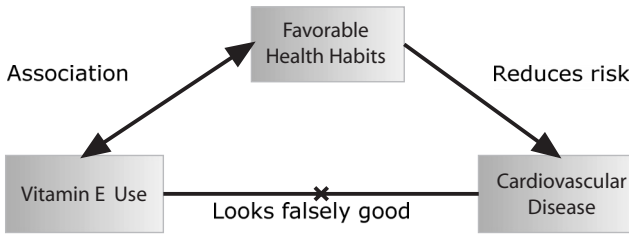


Figure 10.2 Confounding: Vitamin E seemed to reduce the risk of cardiovascular disease when presumably it is only associated with other factors that reduce risk.

significant misclassification of exposure (e.g., because some people assigned to the drug do not take it and/or some assigned to placebo obtain the active drug), power will be reduced and the estimate of effect size will be biased toward the null. Although the loss of power may be overcome by increasing the sample size, additional data and assumptions will generally be necessary to estimate how much the estimate of treatment effectiveness might have been biased toward no effect by crossover between groups.

Instrumental variables use this same type of logic. The idea is that, instead of the investigator randomly assigning subjects to a treatment group, which is then (strongly) associated with treatment but not otherwise associated with outcome, the investigator identifies some other variable that is associated with the treatment of interest and thought not to be (independently) associated with the outcome. The outcome is then determined in relation to this “instrumental variable,” instead of to group assignment. For example, the instrumental variable could be related to a time or place that is associated with the receipt of the treatment, but is thought not to relate independently to the outcome. The expected bias toward the null that may occur from misclassification of exposure is overcome with a combination of a large sample size and calculations to estimate the effect of the imperfect relationship between the instrumental variable and the predictor of interest.

The concepts here are somewhat abstract, so the easiest way to explain and understand instrumental variables may be with some examples. A study one of us helped design used an instrumental variable to study delayed effects of military service during the Vietnam-era on mortality (Hearst et al. 1986). The exposure of interest in this case was military service. The outcome variable for the study was

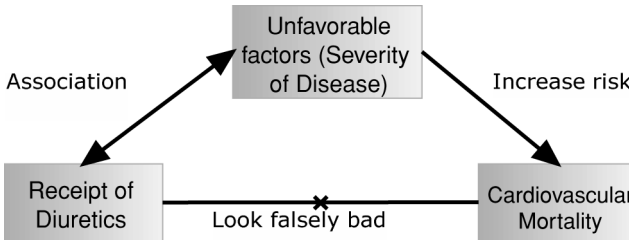


Figure 10.3 Confounding by indication. The favorable effect of diuretics on cardiovascular disease mortality is masked by the association between unfavorable factors and receipt of diuretics.

mortality, as determined from electronic death certificate registries from California and Pennsylvania that included the date of birth of the decedent and the cause of death. Although mortality rates in Vietnam veterans could easily be compared with those in nonveterans, the inference that differences in mortality after the period of service were *caused by* the military service would be weak, because men who served differed in many ways from those who did not – military service was not randomly assigned. However, for several years beginning in December 1969, annual draft lotteries were held, and 19-year-old men were randomly assigned a number from 1 to 366 based on their birth dates. Beginning in January each of the following years, men were drafted starting with the lowest lottery numbers, until a sufficient number of men had been drafted for that year. Thus, the instrumental variable chosen for this study was *eligibility* for the draft, as determined by whether a man's birth date corresponded to a draft lottery number above or below the highest number called for his year of birth (e.g., #195 for men born in 1950). Draft eligibility seemed to be a good instrumental variable, because it was associated with the predictor of interest, military service, but not independently associated with the outcome, mortality.

As it turned out, draft eligibility was a poor marker for actual military service. Only 25.6% of men whose birth dates made them eligible for the draft actually served, compared with 9.3% of men whose birth dates made them exempt. Nonetheless, the large sample size enabled the study to find statistically significantly higher mortality among draft-eligible men in the years following the period of possible military service, with observed relative risks of 1.13 for suicide ( $P = 0.005$ ), 1.08 for motor vehicle accidents ( $P = 0.03$ ), and 1.04 for all-cause mortality ( $P = 0.03$ ).

These results, analogous to an “intention to treat” analysis of a randomized trial in which there was considerable crossover between groups, provide strong evidence of causality. However, the RRs using the instrumental variable are biased toward the null if the true cause of the mortality differences was military service, not eligibility for the draft. Therefore, the investigators algebraically calculated what the RR for military service would need to be in order to produce the association found between the instrumental variable of draft eligibility and mortality. If all of the excess risk among draft-eligible men was due to military service (a hypothesis that could not be tested with this design) RRs would need to be 1.86 for suicide, 1.53 for motor vehicle accidents, and 1.25 for total mortality.<sup>2</sup>

As in the previous example, instrumental variables analyses are often done with large databases, in order to overcome the bias toward no effect caused by a loose relationship between the instrumental variable and the predictor variable of interest. Another example of this is a study by Bell and Redelmeier (2001), who tested the hypothesis that reduced staffing might increase hospital mortality in Ontario hospitals. The predictor variable of interest was staffing levels, but the investigators were concerned that staffing level could vary with (and hence be confounded by)

<sup>2</sup> The calculation requires an assumption that the observed risks in eligible and exempt men are each a weighted average of risks among those who did and did not serve, and that these risks are the same in eligible and exempt men. This RR applies only to those who served *as a result* of having a low lottery number. There is no way of telling the effects of service on those who would have served regardless.

disease severity. The authors chose admission day of the week as the instrumental variable because staffing varied by day of the week, but the researchers did not believe disease severity (and/or mortality) was likely to vary by admission day for the diagnoses they studied. The assumption was that any differences in mortality by day of the week (beyond those explainable by other measured variables, such as age, race, and sex) could be attributed to staffing differences. In this design, the analysis is not of the relationship between predictor variable and outcome (patients per nurse and mortality) but rather between the instrumental variable (day of the week) and the outcome. The authors found significantly higher mortality during weekends, when staffing was lower, for 3 prospectively identified conditions they thought would be sensitive to staffing levels (ruptured abdominal aortic aneurysms, acute epiglottitis, and pulmonary embolism), but no differences for 3 control conditions (acute myocardial infarction, acute intracerebral hemorrhage, and acute hip fracture). In fact, weekend admissions were also associated with significantly higher mortality rates for 23 of the top 100 causes of death and not associated with lower mortality for any.

For some treatments, such as procedures for which good evidence about benefits, risks, or indications is lacking, there is considerable variation in the extent to which the treatments are used. This may reflect variation in the level of illness of patients being treated, as well as variable expertise with the treatments in different places. If both the treatment method and outcome are available in large electronic databases, an instrumental variable can be the proportion of patients receiving a particular treatment at the hospital at which each individual patient was seen. The hope is that the proportion of patients receiving the treatment will be associated with the treatment received but not otherwise associated with the outcome.

For example, Johnston (2000) noted marked variation among academic medical centers in the use of coiling (an endovascular procedure done by interventional neuroradiologists) and clipping (an operation done by neurosurgeons), two procedures used to treat cerebral aneurysms. A direct comparison (Johnston et al. 1999) of mortality following these two procedures suggested higher mortality from clipping (2.3% vs. 0.4%), but because patients were not randomly assigned, confounding by indication was possible (if the sicker patients were being treated with the clipping procedure). Estimating that the overall mortality risk (adjusted for age, sex, and comorbidity of the patients) should be similar across academic medical centers, Johnston used the proportion of patients with aneurysms treated by coiling as an instrumental variable. This variable was associated with the predictor but assumed not to be independently associated with mortality. Therefore, if coiling were superior to clipping, after adjusting for baseline differences, mortality should be lower in hospitals that do more coiling. This is, in fact, what he found: hospitals that did more coiling had lower adjusted mortality rates; the overall adjusted risk ratio per 10% increase in the proportion of cases treated by coiling was 0.91 for ruptured and 0.84 for unruptured aneurysms.<sup>3</sup>

<sup>3</sup> The author used a multivariate technique called Generalized Estimating Equations, which accounts for clustering of observations within hospitals, to obtain these risk ratios.

---

## Measurement of unrelated variables to estimate confounding or bias

---

Clinical trials, natural experiments, and studies using instrumental variables all are designed with a goal of minimizing or controlling confounding. An alternative approach is not to control confounding, but to make measurements that provide an indication of its importance. There are two directions from which these measurements can be made. The first is to measure another outcome that would be affected by the unmeasured confounder of concern but not by the treatment. If the treatment seems to affect this second outcome, confounding is likely to be a problem. The second is to measure another predictor variable in addition to the treatment of interest that is not felt to have a causal effect on the outcome but which should be associated with the unmeasured confounder. If confounding has an important effect on the relationship between the treatment and the outcome, it should also affect the relationship between the second predictor and the outcome. Concrete examples of these two methods should help clarify the abstract discussion above.

### Measuring another outcome

Selby et al. (1992) provide an excellent example of measuring a second outcome, subject to the same potential confounders as the outcome of interest, in order to show absence of confounding. They did a case–control study of screening sigmoidoscopy to prevent colon cancer death. The colon cancer deaths were divided into those caused by cancers that likely were and were not within reach of the sigmoidoscope. The cancers not within reach of the sigmoidoscope were the second outcome; they were presumably associated with the same confounders as those within reach of the scope, but not associated with the predictor. Although the authors used logistic regression to adjust for relevant covariables, the particularly elegant and convincing aspect of the study is their demonstration that sigmoidoscopy conferred protection against deaths from colon cancers that were within reach of the sigmoidoscope (adjusted OR = 0.41; 95% CI 0.25 to 0.69), but not from those that were beyond the reach of the sigmoidoscope (OR = 0.96; 95% CI 0.61 to 1.50). If unmeasured confounders were responsible for the apparent protective effect of sigmoidoscopy, it seems likely that they would have led to apparent protection from cancers both within and beyond the reach of the sigmoidoscope.

This strategy has also been used to study factors like income and access to ambulatory care. For example, Booth and Hux (2003) studied diabetic emergencies in Ontario. They found a clear inverse association between income level (estimated from neighborhood of residence) and admissions and emergency department visits for hyper- or hypoglycemia. However, admission rates for hip fracture and appendicitis, the control conditions (second outcomes) subject to the same potential confounders but thought not likely to be sensitive to ambulatory care, did not differ. Another example (Cook and Campbell 1979, pp. 218–21) comes from a study of an intervention aimed at reducing drunk driving in Britain. A crackdown using “breathalyzers” to estimate alcohol levels in pubs was associated with an abrupt drop in motor vehicle accidents on weekend nights when pubs were open. However, there

was no change in the second outcome: accident rates occurring during hours when pubs were closed. If the decrease occurred because of a temporal trend rather than the breathalyzer intervention, the accident rate should have dropped generally, not just when the pubs were open.

### Measuring another predictor

The second approach, measuring other predictors in addition to the treatment of interest, is illustrated by a study of calcium channel blockers (CCBs) as a possible cause of myocardial infarction (MI; Psaty et al. 1995). The authors noted a progressive increase in risk of MI with increasing doses of CCBs: the adjusted OR increased from 1.13 for low dose to 1.42 for medium dose and 1.81 for high dose use. Of course one would expect that those treated with higher doses would have worse hypertension and be at higher risk, thus confounding by indication would be a major concern. However, the dose–response relationship went in the opposite direction for beta blockers – higher doses of beta blockers were associated with decreased risk of MI. Beta blockers were studied as a second predictor with a similar relationship to the confounders of interest (severity of disease) but not related to MI. The adjusted OR for beta blocker use decreased from 1.00 (reference) to 0.88 to 0.73 with increasing dose. If confounding by indication were the cause of the dose–response relationship between CCBs and MI, we might expect something similar for beta blockers, the opposite of what was observed.<sup>4</sup>

This strategy is clearly not a panacea, however. In both the Health Professionals study (Rimm et al. 1993) and the Nurses' Health Study (Stampfer et al. 1993), taking at least 400 I.U. of vitamin E daily was associated with a reduced risk of coronary heart disease, even after adjusting for all known confounders. Of course, people who take vitamin E are different from people who do not – for example, they might be more health conscious. But if that were the case, one would expect a favorable outcome in people taking a multivitamin pill or vitamin C as well, behaviors that are also associated with being health conscious. However, this was not observed. The lack of an association of the outcome with a covariable that one would expect to suffer from the same confounding as the treatment of interest suggested causality strongly enough that one of us (TBN) started to take supplemental vitamin E. Unfortunately, as mentioned above, subsequent evidence from randomized trials suggests vitamin E is of no benefit, and may even be harmful (Eidelman et al. 2004; Miller et al. 2005).

### Propensity scores

A relatively new approach to controlling confounding in observational studies of treatment efficacy is the use of propensity scores. In order for a variable to be a confounder, it has to be associated with both treatment and outcome. The usual approach to using multivariate analysis to control for confounding is to create a

<sup>4</sup> This does not prove that CCBs cause MI. If CCBs were entirely inert, but were prescribed in increasing doses for those at higher risk of MI, while beta-blockers were effective at reducing risk, we could see results like these. But one could argue that taking an inert medicine, rather than something that works, is a cause of MI.

model that includes the treatment variable and other predictors of outcome (the potential confounders). If the model fits, the coefficient for the treatment will reflect its independent contribution to the outcome.

For example, the equation for the *logistic* model can be written as:

$$\ln \left[ \frac{P(Y)}{1 - P(Y)} \right] = a + b_1X_1 + b_2X_2 + \dots + b_kX_k$$

where

“P(Y)” is the probability of the outcome Y,

“ $\ln \left[ \frac{P(Y)}{1 - P(Y)} \right]$ ” is the log-odds of the outcome,

“a” is a constant (the intercept, related to the overall probability of the outcome),

“ $X_i$ ” are the different predictor variables associated with outcome, including the predictor variable of interest as well as the potential confounders. For example, the variable of interest to you might be  $X_2$  (the treatment you are studying) and the rest would be confounders.

“ $b_i$ ” are coefficients (equal to the logarithm of the odds ratio if “ $X_i$ ” is dichotomous) associated with each predictor.

“k” is the number of predictor variables.

One limitation of this approach is that, if there are many potential confounders, there may not be enough outcomes in the dataset to be able to estimate their coefficients with much precision. There’s a rule-of-thumb: you would like to see at least 10 outcomes for each predictor variable. (By “outcome” we mean 10 instances of the thing you are trying to predict, e.g., deaths.) Imagine there are 1,000 patients, of whom 300 received the treatment of interest, but only 30 died. With only 30 outcomes, it will be difficult to control for confounding by more than 2 other variables aside from the treatment variable – the dataset just does not have enough outcomes to do this well.

Enter propensity scores. The idea of propensity scores is that, instead of controlling for all possible predictors of outcome, investigators instead control for predictors of the treatment. This is done by creating a model to estimate the predicted probability of treatment (or propensity to be treated). Subjects are then either matched or stratified on this propensity score, and the risks of the outcome in people who actually were or were not treated within each stratum are compared. Thus, continuing the notation above, if  $X_2$  is the treatment of interest, the model for the propensity score would look like this:

$$\ln \left[ \frac{P(X_2)}{1 - P(X_2)} \right] = a + b_1X_1 + b_3X_3 + \dots + b_jX_j$$

Note that the difference is that the probability we are trying to predict is not the probability of outcome [P(Y)], it is probability of treatment, P( $X_2$ ). The number of predictors of treatment may be different than the number of predictors of outcome, so we end up with j instead of k – 1 variables. In fact, because the model being created to estimate P( $X_2$ ) does not have to work in any dataset except the one being analyzed,

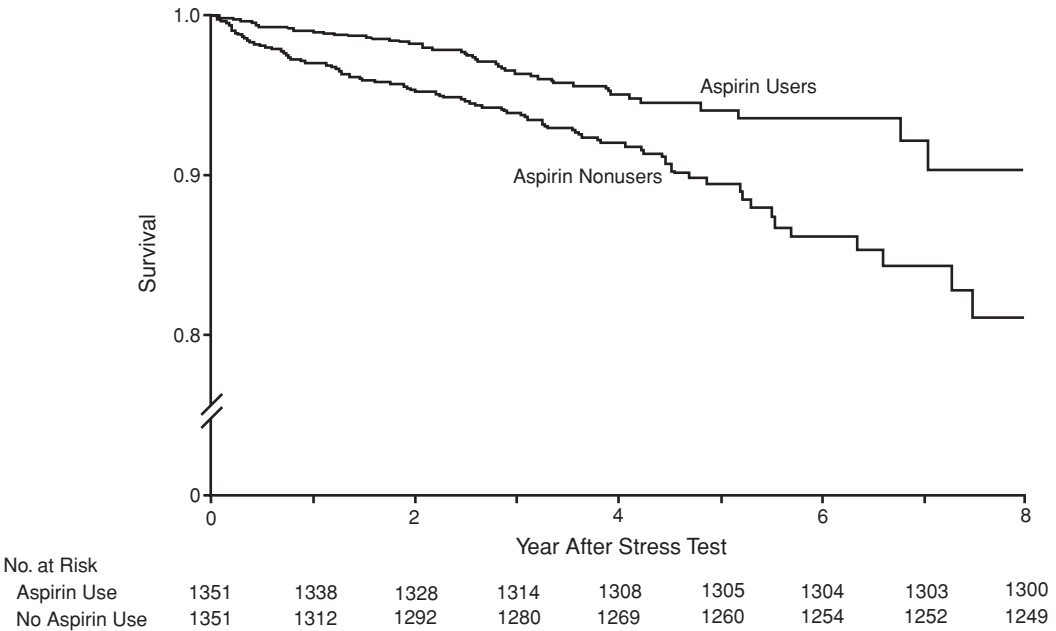


Figure 10.4 Survival of aspirin users and nonusers following stress echocardiography, matched by propensity score for aspirin use. (From Gum et al. 2001; used with permission.)

the investigator can include several predictor variables, interactions, and quadratic terms, without worrying about how the model might work on another data set.

Now the investigator can stratify on this  $P(X_2)$  variable (e.g., in quintiles of “propensity for treatment”) and compare the risk of outcome in those who actually were and were not treated within these quintiles of subjects, each of whom had approximately the same propensity to be treated. Alternatively, the investigator can match subjects who were and were not treated by their value of  $P(X_2)$  and compare outcomes between matched subjects.

For example, Gum et al. (2001) prospectively studied total mortality of 6,174 consecutive adults undergoing stress echocardiography, 2,310 of whom (37%) were taking aspirin. In unadjusted analyses, mortality did not differ between users and nonusers of aspirin – 4.5% in each group. Multivariable analysis, however, suggested a mortality benefit. This was confirmed by matching subjects by propensity scores and then comparing survival in the two groups (Fig. 10.4).

Note that the figure is based on only 1,351 subjects in each group. This is because only 1,351 of the 2310 subjects who received aspirin had a “match,” – that is, had someone with the same propensity to receive aspirin but did not receive it – in the control group. This is not unexpected in observational studies such as this one. When the treatment is not randomized, the average propensity to receive aspirin will be higher in the group that received it than in the group that did not, which may make it difficult to match all treated subjects to untreated subjects. This loss of subjects affects both power (which was still more than adequate in this study) and generalizability. For example, the results of this study are only generalizable to patients whose propensity to receive aspirin was in a range where there was overlap between those who did and did not receive it. But this makes sense. If there are some

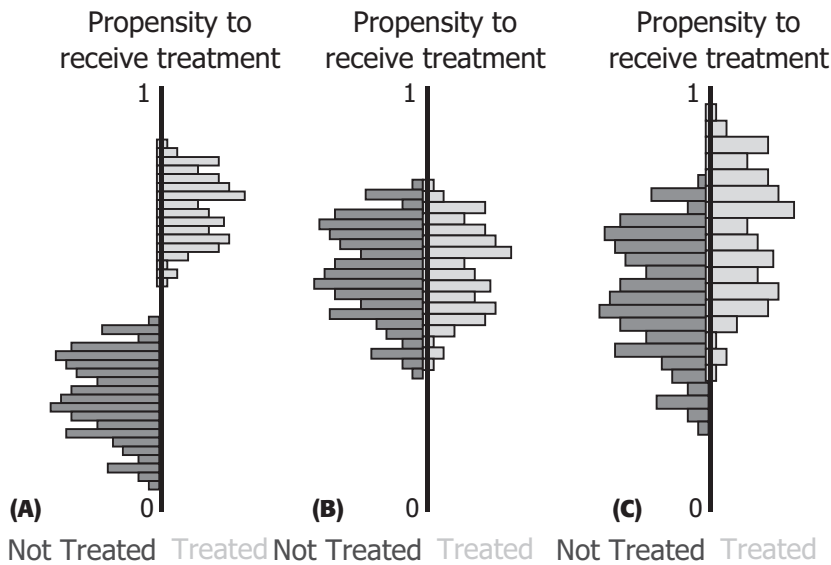


Figure 10.5 **(A)** Propensity scores do not overlap; treated and untreated groups are not comparable. A propensity analysis cannot be done and any comparison between groups is hazardous. **(B)** Propensity distributions are nearly identical. A propensity analysis is not necessary as groups are already matched or treatment was randomly assigned. **(C)** Good overlap in propensity scores; the subjects in the overlapping parts of the distribution can be studied. (Figure courtesy of Thomas Love (Case Western Reserve University Center for Health Research and Policy).)

people who absolutely should get aspirin and some who should not, their propensities will be close to 1 and 0, respectively; they will not have a match and hence will not be included in the matched results. Think of this exclusion of subjects with propensity scores near 0 or 1 as analogous to exclusion criteria in a clinical trial. If there are some persons for whom either drug or placebo is known to be contraindicated, then you neither can nor should study the difference between drug and placebo in those patients. In fact, this is another advantage of a propensity analysis: if there is little overlap between the propensity scores of those who were and were not treated, it means that those treated appear to be very different from those not treated, in terms of their indication for the treatment, and that trying to adjust for this with multivariate analysis may require questionable extrapolations beyond the data.

A propensity score analysis requires that scores overlap between a substantial portion of the treated and untreated groups. If the model predicts treatment too well, only a few subjects in the treated and untreated groups will have the same propensity score (Fig. 10.5A). For this reason, one should avoid including in a propensity score factors that are associated with receiving treatment but unlikely to cause the outcome, such as day of week or geographical location. (Note that these same factors make good instrumental variables!) On the other hand, if the propensity score distributions in the treated and untreated groups are nearly identical, there is no point in doing a propensity score analysis (Fig. 10.5B).

Propensity score analysis in an observational study of a treatment helps to separate out the effects of the treatment itself from other factors associated both with receiving

the treatment and with the outcome. However, propensity score analysis is not helpful if the goal is to identify or to quantify the effects of these other confounding factors.

## Summary

1. Although randomized blinded trials are the best way to establish causal relationships between treatments and outcomes, it is sometimes possible, by thinking creatively, to design observational studies that provide strong evidence of causality.
2. One approach is to identify an instrumental variable that is associated with treatment but not independently related to the outcome. Comparing outcomes between groups based on values of the instrumental variable is then similar to an intention-to-treat analysis of a randomized trial with substantial crossover between the treatment and control groups. The direction of the treatment effect is correct, but the magnitude of the effect is diminished.
3. Another approach is to test for the possibility of confounding by measuring a second outcome that is causally associated with the confounder but not the treatment. If no association exists between treatment and this second outcome, confounding is less likely to be a problem.
4. Similarly, one can identify additional predictors (other than the treatment) that are associated with the confounder but have no effect on the outcome. If these additional predictors are not found to be associated with the outcome, confounding is again less likely to be a problem.
5. Finally, one can model the propensity to receive treatment and compare outcomes of subjects with similar treatment propensities.

## References

- Bell, C. M., and D. A. Redelmeier (2001). "Mortality among patients admitted to hospitals on weekends as compared with weekdays." *N Engl J Med* **345**(9): 663–8.
- Booth, G. L., and J. E. Hux (2003). "Relationship between avoidable hospitalizations for diabetes mellitus and income level." *Arch Intern Med* **163**(1): 101–6.
- Cook, T. D., and D. T. Campbell (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Chicago, IL, Rand McNally College Pub. Co.
- Eidelman, R. S., D. Hollar, et al. (2004). "Randomized trials of vitamin E in the treatment and prevention of cardiovascular disease." *Arch Intern Med* **164**(14): 1552–6.
- Gum, P. A., M. Thamilarasan, et al. (2001). "Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: a propensity analysis." *JAMA* **286**(10): 1187–94.
- Hearst, N., T. B. Newman, et al. (1986). "Delayed effects of the military draft on mortality. A randomized natural experiment." *N Engl J Med* **314**(10): 620–4.
- Johnston, S. C. (2000). "Effect of endovascular services and hospital volume on cerebral aneurysm treatment outcomes." *Stroke* **31**(1): 111–7.
- Johnston, S. C., R. A. Dudley, et al. (1999). "Surgical and endovascular treatment of unruptured cerebral aneurysms at university hospitals." *Neurology* **52**(9): 1799–805.
- Katz, M. H. (1999). *Multivariable Analysis: A Practical Guide for Clinicians*. Cambridge, Cambridge University Press.

- Miller, E. R., 3rd, R. Pastor-Barriuso, et al. (2005). "Meta-analysis: high-dosage vitamin E supplementation may increase all-cause mortality." *Ann Intern Med* **142**(1): 37–46.
- Psaty, B. M., S. R. Heckbert, et al. (1995). "The risk of myocardial infarction associated with antihypertensive drug therapies." *JAMA* **274**(8): 620–5.
- Rimm, E. B., M. J. Stampfer, et al. (1993). "Vitamin E consumption and the risk of coronary heart disease in men." *N Engl J Med* **328**(20): 1450–6.
- Selby, J. V., G. D. Friedman, et al. (1992). "A case-control study of screening sigmoidoscopy and mortality from colorectal cancer." *N Engl J Med* **326**(10): 653–7.
- Stampfer, M. J., C. H. Hennekens, et al. (1993). "Vitamin E consumption and the risk of coronary disease in women." *N Engl J Med* **328**(20): 1444–9.
- Turnbull, F., B. Neal, et al. (2005). "Effects of different blood pressure-lowering regimens on major cardiovascular events in individuals with and without diabetes mellitus: results of prospectively designed overviews of randomized trials." *Arch Intern Med* **165**(12): 1410–9.
- Warram, J. H., L. M. Laffel, et al. (1991). "Excess mortality associated with diuretic therapy in diabetes mellitus." *Arch Intern Med* **151**(7): 1350–6.

## Chapter 10 Problems

1. Thimerosal, a mercury-containing preservative, was removed in 2001 from vaccines routinely given to infants because of concern that the total mercury dose, added up over all recommended vaccines, exceeded the U.S. Environmental Protection Agency safety limit. There are many lawsuits pending in which plaintiffs claim that autism or other neurodevelopmental problems occurred in their children because of exposure to mercury in thimerosal.

Imagine that you have access to an enormous database that includes electronic records of several million health plan members from 1990 to the present, that diagnoses of autism are recorded in the database, but that, like everywhere else, definitions may have been unstable over time. You know that Rh Immune Globulin (Rhogam<sup>®</sup>), which is given during pregnancy to most Rh-negative women who get good prenatal care, had 25 µg of mercury (as thimerosal) per dose until 2001, and you are interested in possible toxic effects of this amount of mercury on the fetus. Assume that you have plenty of subjects and accurate data on who got Rhogam and on blood types, and that some mothers in your health plan miss a lot of prenatal visits, and might not get Rhogam for that reason.

Design a study to take advantage of these facts in order to assess the impact of this dose of mercury on the risk of autism. What groups would you compare to have the greatest strength of causal inference?

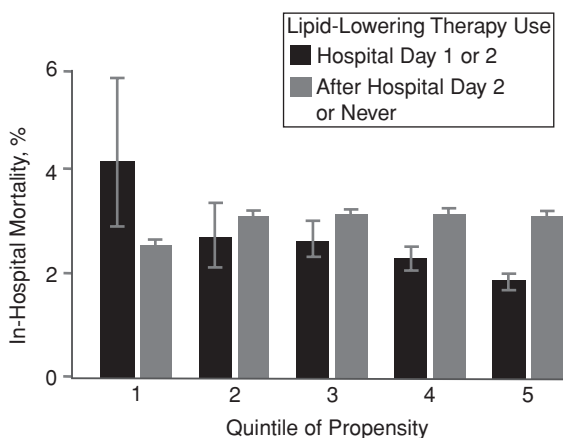
2. You plan to use a large clinical database to perform a retrospective cohort study to investigate whether screening digital rectal examinations (DRE) by primary care providers prevent deaths from prostate cancer.
  - a) What would be your main *predictor variable*?
  - b) Your outcome variable will be prostate cancer death rates in different groups, so the numerators will be men who died of prostate cancer. What should the denominators of those rates be?
  - c) What could you do to address the problem of confounding (or selection bias)?
3. There is some evidence that regular use of nonsteroidal antiinflammatory drugs (NSAIDs) for arthritis may reduce the risk of colon cancer. You wish to design a

study to see whether occasional ibuprofen use reduces colon cancer risk. Explain why you might want to ask about acetaminophen use and how you would use the answers to such questions.

4. Consider a cohort study (Nichol et al. 2003) that estimated the effect of flu vaccination on mortality during the 1998–1999 and 1999–2000 flu seasons. The study population included almost 300,000 subjects at least 65 years old, of whom about 58% were vaccinated. Among vaccinated and unvaccinated subjects, 1.2% and 2.0%, respectively, died during the flu season.
  - a) What confounder(s) could make flu vaccine appear less effective than it really is?
  - b) What confounder or bias might have biased the results in the opposite direction?
  - c) Using multivariable analysis to adjust for potential confounder(s), the authors estimated an RR of about 0.5, an ARR  $\approx$  1.0% and an NNT  $\approx$  100 for all-cause mortality.

Since it is hard to believe that half of deaths could be prevented by the flu vaccine, you are concerned that there are potential unmeasured and unmeasurable confounders. Without doing a whole other study, how could you test the question of whether the flu vaccination's apparent effect in reducing mortality during the flu season was actually due to unmeasured factors that were associated with flu vaccination?

5. Lindenauer et al. (2004) reported that perioperative use of lipid-lowering agents may decrease mortality following cardiac surgery by about 30% to 40%. They controlled for confounding by creating a propensity score.
  - a) Describe in words what the propensity score for this study was.
  - b) Figure 1 from that paper (reprinted below) shows that mortality was lower among users of lipid-lowering drugs in all but the first quintile of propensity.



In-hospital mortality associated with lipid-lowering therapy in propensity-based quintiles. Error bars indicate 95% confidence intervals. Seventeen patients (0.002%) were excluded from multivariable analysis due to missing data; therefore, among 780,574 patients, mean lipid-lowering therapy use per quintile of propensity was 0.5% (quintile 1,  $n = 156,114$ ), 1.9% (quintile 2,  $n = 156,115$ ), 9.8% (quintile 3,  $n = 156,115$ ), 10.9% (quintile 4,  $n = 156,115$ ), and 31.3% (quintile 5,  $n = 156,115$ ). From Lindenauer et al. (2004). Used with permission.

- i) Why are the error bars for the mortality estimate for the left-most column of the graph so much longer than those for the other columns?
- ii) It appears that, for subjects in the lowest propensity quintile, use of lipid lowering drugs on hospital day 1 or 2 appeared to be harmful rather than beneficial. Assume for this question that there is no random error and no confounding – that is, that the results in the figure are accurate and causal. What implication does this have for promoting increased use of such drugs to reduce perioperative mortality after noncardiac surgery?

### References for problem set

---

- Lindenauer, P. K., P. Pekow, et al. (2004). “Lipid-lowering therapy and in-hospital mortality following major noncardiac surgery.” *JAMA* **291**(17): 2092–9.
- Nichol, K. L., J. Nordin, et al. (2003). “Influenza vaccination and reduction in hospitalizations for cardiac disease and stroke among the elderly.” *N Engl J Med* **348**(14): 1322–32.