Published in final edited form as:

Neuropsychol Rev. 2008 September ; 18(3): 194–213. doi:10.1007/s11065-008-9066-x.

Methodological Challenges in Causal Research on Racial and Ethnic Patterns of Cognitive Trajectories: Measurement, Selection, and Bias

M. Maria Glymour,

Department of Society, Human Development, and Health, Harvard School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA; Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA

Jennifer Weuve, and

Department of Internal Medicine, Rush Institute for Healthy Aging, Rush University Medical Center, Chicago, IL, USA; Department of Environmental Health, Harvard School of Public Health, Boston, MA, USA

Jarvis T. Chen

Department of Society, Human Development, and Health, Harvard School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA

Abstract

Research focused on understanding how and why cognitive trajectories differ across racial and ethnic groups can be compromised by several possible methodological challenges. These difficulties are especially relevant in research on racial and ethnic disparities and neuropsychological outcomes because of the particular influence of selection and measurement in these contexts. In this article, we review the counterfactual framework for thinking about causal effects versus statistical associations. We emphasize that causal inferences are key to predicting the likely consequences of possible interventions, for example in clinical settings. We summarize a number of common biases that can obscure causal relationships, including confounding, measurement ceilings/floors, baseline adjustment bias, practice or retest effects, differential measurement error, conditioning on common effects in direct and indirect effects decompositions, and differential survival. For each, we describe how to recognize when such biases may be relevant and some possible analytic or design approaches to remediating these biases.

Keywords

Causal research; Racial and ethnic disparities; Cognitive trajectory; Neuropsychological research; Counterfactuals; Directed acyclic graphs; Measurement error; Selection

Neuropsychological research encompasses efforts to improve both clinical prediction and research focusing on the etiology of disease and impairments. At the heart of the latter is the question: Why do different populations experience different rates of cognitive impairments? As reviewed in accompanying articles in this issue, differences in cognitive outcomes across racial and ethnic groups likely reflect the combination and accumulation of multiple social and environmental experiences over the life course. Etiologic research on racial and ethnic

determinants of neurologic disorders is concerned with disentangling these causal pathways. However, the observed differences in cognitive outcomes may originate, in part, from flaws in study design or analysis, introducing additional difficulties to the process of making valid causal inferences. In neuropsychological research, the potential biases implicit in much causal research are exacerbated by complex measurement and selection problems. In this paper we focus explicitly on these methodological problems relevant to causal research. Thus, much of what we say is not relevant if the goal is strictly to improve diagnoses or predict how patients will fare in the future. We focus on how to improve understanding of the factors that explain the trajectories experienced by patients, and how to predict the consequences of interventions or treatments.

In the last few decades there has been an outpouring of research on causal inference, from a range of disciplines, and this has produced solutions, or at least partial solutions to many vexing methodological problems (Winship and Morgan 1999; Pearl 2000). The results of this work are now disseminating to applied fields, frequently resulting in important advances (Mark and Robins 1993; Hernán et al. 2000; Spirtes et al. 2001; Cole et al. 2005; Hernan et al. 2005). Even when solutions are not yet available, these causal frameworks and vocabulary help to sharpen conceptualization and discussion of the problems. We begin by summarizing the counterfactual account of causation and by describing the link between statistical associations and causal relationships. This first section is crucial for much of the rest of the paper, although the other sections can be read more or less independently. The following sections address confounding and mediation tests, and insights into these familiar topics that arise from modern understanding of causation. Three sections of the paper address problems that arise specifically because of difficulties in measuring an underlying construct: ceiling/floor effects, baseline adjustment in longitudinal analysis, and screening bias. Finally, we present a counterfactual account of practice or retest effects and how to estimate these, and a discussion of differential survival. Throughout, to provide a context for discussion, we specifically focus on cognitive aging and dementia as outcomes, and efforts to understand the effects of race and education (which we call exposures) on these outcomes. The methodologic issues discussed here generally apply to any cognitive outcomes and, in particular, to studies of cognitive trajectories or change. Each of the topics discussed here has special relevance to research related to race and ethnicity: randomization of the exposures of interest is either impossible (as with race) or difficult (as with education); mediation tests are of crucial interest; the pervasive differences in cognitive test performance across racial and ethnic group exacerbate the problems introduced by imperfect measurement; and severe selection processes operate throughout life. These difficulties would operate even if the assessment tools at our disposal functioned equivalently across racial and ethnic groups, but they are much worse because many of our measurement instruments perform differently across racial and ethnic groups. We briefly discuss diagnostic accuracy of cognitive measures in the screening section, and this topic is covered in much greater detail in the Pedraza and Mungas article in this issue.

Causation Versus Association

Counterfactuals

For two random variables X and Y , we say X causes Y if, had X taken a different value than it actually did—and nothing else temporally prior to or simultaneous with X differed—then Y would have taken a different value. The idea of the counterfactual is key: the causal effect is the contrast between what actually happened given that the exposure took the value that it did, and what *would have happened* if the exposure had, counter to fact, taken a different value. Potential outcomes are a related concept referring to the range of possible values the outcome variable might have taken for any individual, under different possible treatment values. Often we consider causation to be probabilistic, rather than deterministic; we can

then modify the definition to say, had X taken a different value, this would have resulted in a different probability distribution for Y . Causation also appears probabilistic if we assume that a population is composed of both people for whom the outcome is changed by the value of the exposure, and people whose outcome would not differ under different exposures.

In general, it is not possible to observe the value that Y would take under multiple different values of X for any given individual, because X takes on only one of its possible values for any single person. For this reason, we cannot directly calculate the causal effect of X on Y for an individual. Instead, we try to estimate the causal effect for a population by comparing the distribution of Y for people with different values of X . Suppose X is a dichotomous variable indicating exposed or unexposed. To estimate the causal effect, we assume that the distribution of Y observed among those who were exposed represents the distribution of Y that would have been observed for those people who were not exposed, if (counter to fact) they had been exposed. Under this assumption, sometimes called exchangeability, the difference in the distribution of Y among the exposed and the distribution of Y among the unexposed represents the causal effect of X on Y . The assumption of exchangeability would be questionable, however, if other causal pathways could have resulted in a systematic difference between the exposed and the unexposed (or, if the distribution of Y among the observed unexposed would be different than the distribution of Y among the exposed had they been *unexposed*, counter to fact).

Causation, as defined above, is distinct from statistical associations. We say X and Y are statistically independent if knowing the value of X does not provide any information about the value of Y , but statistically associated or dependent if knowing the value of X provides some information about the likely value of Y , even if this information is very limited and amounts to a modest change in the probability distribution of Y . For example, if dementia is more prevalent among people with limited education, then knowing someone's education provides information, albeit not certainty, on the probability that the person has dementia. Thus we say education and dementia are statistically associated, which is quite different from saying that education causes dementia.

Causal Diagrams and the Link Between Statistical and Causal Relationships

Throughout this paper, we will use causal DAGs as an intuitive tool to represent causal assumptions and show how statistical associations arise from causal relationships. We do not provide a comprehensive introduction to DAGs and their applications here, but refer interested readers to any of several alternative sources (Greenland et al. 1999; Pearl 2000; Spirtes et al. 2001; Glymour 2006a, b; Glymour and Greenland 2008). DAGs are similar to the box-and-arrow diagrams that one might scribble on a napkin, but there are two formalizations: when we draw an arrow, we mean to represent a causal effect, not merely a statistical association. If there is any factor that influences two or more variables shown in the DAG, we will also show that factor in the DAG. In other words, any confounder of any two variables in the DAG will itself be depicted. DAGs differ from structural equation models because they are entirely non-parametric.

Causal diagrams are a tool for representing hypothetical causal structures. Inferences drawn from a diagram necessarily assume that all of the relevant causal entities and pathways have been accurately represented, based on prior causal knowledge. Because this knowledge is rarely certain, it is important for researchers to consider a range of possible causal structures, and to use the diagrams to consider how design and analytic choices may influence the validity of causal inferences under alternative causal assumptions.

There is an important link between causation and statistical associations. Statistical associations between two variables X and Y can occur for any of five reasons: chance; X

may cause Y (causation); Y may cause X (reverse causation); X and Y may both be affected by a third variable (confounding); or, the sample in which we are examining X and Y may have been selected or conditioned on a variable that both X and Y influence (collider bias). The first four explanations are familiar. Collider bias is generally not as well-recognized and can be counterintuitive, so we will introduce it briefly.

Collider Bias

Collider bias is important to understanding a number of otherwise puzzling phenomena, including the apparent co-occurrence of various conditions. For example, suppose the diagnostic criteria for a disease require that a patient fulfill at least one of two criteria, X or Y . Then among the diagnosed patients, those without X will certainly fulfill criterion Y ; and those without Y will certainly fulfill criterion X . Even if X and Y are entirely unrelated in the general population, once we restrict to patients who fulfilled one or the other criterion, the two conditions are (inversely) associated. The term “collider bias” was coined to capture the idea that the two causes X and Y collide upon the third variable. There are many familiar examples of this phenomenon, most notably selection bias. Collider bias can be induced regardless of the specific statistical technique used to condition on the common effect, i.e. via restriction, stratification, or statistical adjustment in a regression model.

The implications of the collider bias phenomenon can be quite profound and are extremely relevant to a great breadth of neuropsychological research. As discussed above, an immediate implication is that whenever diagnostic criteria for a condition can be fulfilled by *either* of two characteristics that occur with independent frequencies in the general population, then among the patient population, we expect the two characteristics to be inversely associated. These difficulties extend into etiologic research. If race influences educational attainment, then stratifying by level of education will induce a spurious association between race and any other variable that influences education through an independent process, such as IQ. As a result, if African Americans experience discriminatory processes that reduce educational access compared with whites, but IQ independently increases level of education regardless of race, then comparing blacks and whites with equivalent education will bias the association between race and IQ.

Drawing Causal Inferences From Statistical Associations

Given the sources of statistical association described above, we can make certain predictions regarding the associations that will be induced by any set of causal relationships. For example, under the causal structure shown in Fig. 1, we anticipate that V and X will be associated (because V causes X); X and Y will be associated (because V causes both X and Y); and that, conditional on X , V and W will be associated (because X is a common effect of V and W , also called a collider on the path between V and W). Because the causal structure in Fig. 1 implies certain statistical associations amongst the variables, of which the above three are only examples, we can in some cases use statistical associations (or lack of associations) to falsify this causal structure. For example, if we found that V and X were statistically independent, we would seriously doubt the causal structure in Fig. 1.

Causal research is the effort to identify a statistical association between two variables and rule out four of the five possible explanations for this association. Confidence intervals and p -values are intended to rule out the explanation that chance accounts for the association. Longitudinal designs are used to rule out one or the other causal direction, and statistical adjustment, restriction, or other types of conditioning are used to rule out the confounding explanation. In some cases, efforts to rule out confounding may introduce a spurious association due to collider bias. The motivation for distinguishing causal relationships from other explanations for statistical association is generally to inform predictions about the

consequences of interventions. For example, if the association between education and dementia arises because pre-schooling cognitive traits influence both education and dementia, then we have no reason to anticipate that increasing someone's education would reduce that person's dementia risk. On the other hand, if we can show that education influences dementia risk, this may make interventions to increase educational opportunities more appealing.

Lack of clarity about the meaning of causation and the great difficulties in providing compelling evidence about causal effects sometimes lead researchers to claim to abandon efforts at causal inference and instead focus on identifying only associations or "risk factors" without specifying a causal relationship. Nevertheless, causal effects have great substantive interest because they inform us about the consequences of interventions. The substantive interest is so great that readers frequently attempt to draw causal inferences when reading studies of association, even when such studies are accompanied by disclaimers. We argue that it is counterproductive to abandon causal research, but very necessary to acknowledge the assumptions upon which causal inferences are based.

Confounding

Causal inferences are strongest when the exposure of interest can be randomly assigned, as in the context of a trial. The strength of this design arises because randomization essentially eliminates common prior causes (confounding) as an explanation for a statistical association. Given a large enough sample size, randomization of individuals in a study results in approximately the same distribution of potential confounders in both the treatment and control groups at baseline. With randomization, the assumption of exchangeability is fulfilled and the average outcomes for the treated group represent the outcomes that the untreated group would have had, on average, if they had been treated.

Unfortunately, randomization is impossible or very difficult for a number of extremely important factors, including personal characteristics such as race or sex, or social exposures such as education. Because of the impossibility of randomizing such factors, some have even argued that race or sex should not be considered causes (Holland 1986; Kaufman and Cooper 2001). However, when we are trying to identify the causal effect of an intrinsic personal characteristic, for example, an individual's sex, the list of possible unmeasured common causes of sex and any other outcome is fairly short (Rosenfeld and Roberts 2004; Catalano et al. 2008). Thus, we can frequently think of sex as essentially randomized, and sex differences in outcomes are plausibly considered causal in many contexts, although the pathways by which sex influences outcomes are often highly contested.

Race or ethnicity is more complicated because one's race is nearly determined by the race of one's parents, thus it is impossible to determine whether racial differences in outcomes are due to one's own race or one's parents' race and the benefits or harms that may accompany parents' race. However, again, most important variables that we typically consider possible confounders do not influence race, rather race, and the interaction of race and social environment, influence these variables.

The key causal questions relate to the mediators of racial differences in cognitive outcomes. For example, if ethnic minority groups are at higher risk of dementia, there are many possible explanations; e.g., racial and ethnic minorities may have less access to school, lower quality medical care, experience an accumulation of allostatic load due to stressful discriminatory experiences, or carry a genetic characteristic that increases risk. In efforts to identify the effects of hypothesized social, environmental, or genetic factors, confounding is a tremendous challenge and extremely difficult to rule out in the absence of intervention studies. This is especially relevant to work on the relationship between education and

dementia. The *association* between education and dementia is well-established, but it is unclear if this is because education prevents or slows the pathological processes or because IQ or some unmeasured cognitive facility influences both educational attainment and dementia risk.

Controlling Confounding Through Statistical Adjustment

The conventional approach to control for confounding is to identify the possible common causes of the exposure of interest and the outcome and then condition on these common causes, for example, through stratification, restriction, or statistical adjustment in a regression model. These methods depend on the assumption that all of the common causes of the exposure and the outcome, or at least some variable along the pathway linking these confounders to either the exposure or the outcome, have been measured. In a causal diagram, we say that any sequence of lines linking the exposure and the outcome that begins with an arrow *into* the exposure is a back-door path that potentially confounds estimates of the causal effects of the exposure on the outcome. The confounding bias that flows along back-door paths is blocked if either a) one of the variables in the path is a collider, where two arrows point into that variable (and we have not conditioned on that variable), or b) we have conditioned on one of the variables in the path that is not a collider. For example, in Fig. 1, X is linked to Z via two back-door paths: $V-Y-Z$ and $W-S-T-Z$. The latter path is not biasing because S is a collider. In fact, adjusting for a collider unblocks this path, and thus adjusting for S would actually introduce bias. The former path could be blocked by conditioning on either V or Y . Such conditioning could include stratifying on V , or possibly adjusting for V in a regression model.

The assumption that all confounding paths have been blocked often seems implausible. Propensity scores are an increasingly popular approach to confounder adjustment, but they also depend on the assumption of no unmeasured common causes.

Propensity scores are calculated by modeling the value of the exposure. They are best developed for dichotomous exposures, and typically a logistic regression model is used to calculate the probability that each individual is exposed, using all potentially relevant covariates. This probability is used as the estimated propensity score, and individuals with similar propensity scores but different values of the actual exposure are compared to estimate the effect of the exposure on the outcome. Propensity scores are sometimes described as replicating randomized trials, but this is optimistic. Propensity scores replicate a randomized trial only if variables sufficient to block all of the back-door paths, e.g., all of the possible confounders, have been measured and included in the propensity score model. Propensity scores have an important advantage of showing whether some types of people in the population were almost *certain* to be treated, while others were almost certain not to be treated. It is often preferable to exclude such people from analyses, because they cannot be well-compared with otherwise similar individuals who received the other treatment exposure (Rubin 1997; Oakes 2004).

Avoiding Confounding by Using Natural Experiments

Another approach to estimating causal effects when randomization is not possible is to exploit natural experiments or instrumental variables. The causal structure for a natural experiment is represented in Fig. 2a. It is clearest to think of natural experiments in the same terms as analogous to randomized trials: there is some exogenous event (Z in the diagram) that changes the probability of an individual being exposed or treated (X in the diagram), but *has no other causal link* to the outcome of interest (Y). This event Z is often called an instrumental variable (not to be confused with the conventional use of the word instrument to refer to a measurement tool). In the diagram, there is also an unmeasured variable U that

confounds the association between X and Y , biasing observational associations between X and Y away from the causal effect. Valid natural experiments generally fit this causal structure, although there are some variations that can work (Pearl 2000). More formally, given a causal DAG, we say Z is a valid instrument for the effect of X on Y if (1) Z and X are statistically dependent and if (2) every unblocked path connecting Z and Y contains an arrow pointing into X (Pearl 2000).

Under this causal structure, if we find that average Y differs by treatment assignment Z , this implies that Z affects Y . If Z affects Y , this implies that X also affects Y , because there is no other possible pathway via which Z might affect Y except the one via X . This reasoning fails if there is either a direct effect from Z to Y (as in Fig. 2b) or an unmeasured common cause of Z and Y (as in Fig. 2c), which are both violations of assumption (2) above.

Glymour used this approach to test the hypothesis that education influences memory scores in the elderly, hypothesizing that compulsory schooling laws (CSLs) might operate as natural experiments for the amount of education an individual received. States in the early twentieth century increased their mandatory schooling minimums fairly frequently. If schooling benefits memory, then a child who attended school in years before such a change took effect would be expected to have worse memory performance than a child attending school after the change was implemented (Glymour et al. 2008). Results showed that children born in states with longer CSLs performed better on memory tests administered decades later.

Mendelian Randomization (MR) studies are a special case of this approach. MR designs are used to estimate the effect of a phenotype on a health outcome by exploiting an allelic variant that influences the phenotype but is thought to have no source of association with the health outcome other than pathways mediated by the phenotype (Little and Khoury 2003; Katan 2004; Smith 2004; Smith and Ebrahim 2004; Didelez and Sheehan 2007; Lawlor et al. 2008). Under these circumstances, with additional assumptions, a test of whether the incidence of the health outcome differs across levels of the genotype provides a test of whether the phenotype influences the outcome. For example, a genetic variant associated with difficulty metabolizing alcohol predicts reduced alcohol intake. This variant has been used to assess the effects of alcohol consumption on cardiovascular health (Hines et al. 2001). Connecting this to the causal diagram in Fig. 2a, Z represents the genotype, alcohol consumption would be X , and cardiovascular disease would be Y .

Data from natural experiments are often used to derive an IV effect estimate. This can be calculated as the ratio of the relation between the instrument and the outcome (the intent-to-treat effect estimate, if analyzing a randomized trial) and the relation between the instrument and the treatment (adherence). Angrist et al. proposed a specific causal interpretation of this parameter, based on the idea that some people would have been treated regardless of the value of the instrument, other people would *not* have been treated no matter what value the instrument took, while still a third group, sometimes called the cooperators, would receive the treatment if and only if assigned to receive it by the instrument. They also assumed nobody in the population was contrarian, i.e. received treatment only if assigned *not* to receive treatment, and avoided treatment only if assigned to receive it. Under these assumptions, the IV estimate provides a consistent estimate of the average effect of receiving treatment on those who received the treatment due to the value of the instrument (the cooperators (Currie 1995; Angrist et al. 1996; Greenland 2000; Angrist and Krueger 2001)).

The major limitations of IV analyses are the extreme difficulty of finding plausibly valid instruments, and the fact that the causal parameter is not the average causal effect for the

whole population, but only the effect on the cooperators. Many argue this is a substantial disadvantage: there is no way to tell if someone who was assigned to receive the treatment and did in fact receive it would have received it anyway, even if assigned not to do so. As a result, the cooperators, the subgroup to whom the IV effect estimate applies, cannot actually be identified (Robins and Greenland 1996). On the other hand, there may be contexts in which this causal effect is of greater interest than the population average effect (Glymour 2006a). The major advantage of IV analyses is that they can provide an estimate of the effect of X on Y even when there are unmeasured confounders of X and Y .

Mediation

Identifying primary mediators between a putative exposure and cognitive change is often crucial for understanding the etiology of disease and prioritizing possible intervention efforts. For example, if we believe that cognitive engagement is a key mediator between education and dementia, this suggests a different intervention strategy than if we believe the association is primarily mediated by material conditions and vascular disease such as hypertension. For many exposure/mediator combinations, it is of interest to distinguish the indirect (via the mediator) from the direct (operating directly, not via the putative mediator) effects. In Fig. 3a, the direct effect of X on Y is represented by the arrow from X to Y , and the indirect effect is represented by the sequence of arrows passing through Z : $X \rightarrow Z$ and then $Z \rightarrow Y$. The goal is generally to assess the fraction of the harm (or benefit) caused by the exposure that could be eliminated (or supplemented) by controlling the mediator and removing the effect of the exposure on the mediator. Recent decades have seen substantial progress in our understanding of direct and indirect effects, and in particular recognition of some of the limitations of conventional approaches to estimating direct effects (Robins and Greenland 1992).

Defining the Direct and Indirect Effects

The terms “direct” and “indirect” are ambiguous, because there are several alternative types of direct effects. The differences hinge on the value to which the putative mediator is held. In Fig. 3a, these all correspond to blocking the link between X and Z . A “controlled” direct effect of exposure X on outcome Y not mediated by Z is the effect of X on Y when everyone in the population is forced to receive the same level of Z (i.e. we break the link between X and Z by setting the value of Z). For example, with respect to the mediation of the education–dementia relationship via hypertension, we would say a controlled direct effect is the effect of education on dementia when everyone in the population is somehow forced to be normotensive (or, alternatively, if everyone were forced to be hypertensive). There may be a distinct value for the controlled direct effect of X on Y for every possible level of Z . In other words, the value of Z may modify the direct effect of X on Y . If X is education and Z is hypertension, then the direct effect of education on dementia if everyone were somehow forced to be *normotensive* may differ from the effect of education on dementia if everyone were somehow forced to be *hypertensive*.

It may seem absurd to imagine that everyone in the population is forced to take a single value for the mediator. Regardless of educational level, some people are likely to be hypertensive while others will be normotensive. An alternative type of direct effects, sometimes called the “pure” or “natural” direct effect, is the effect of X on Y when everyone in the population is forced to receive the level of Z they would have attained for a specific, constant, level of X (e.g. if X were 0). We break the link between X and Z , but let other factors besides X influence Z . There is a unique natural direct effect for each possible level of X .

Consider estimating a natural direct effect of education on dementia not mediated by hypertension, in which we hold hypertension to the value it would have taken under high-education. Even if high education reduces hypertension risk, some people with high education will nonetheless be hypertensive, while others would be normotensive if they received high levels of education but hypertensive if they received low education, and still others may be normotensive regardless of their education. For the first type of person, the natural direct effect will describe the effect of education on dementia were they hypertensive; for the second two types, the natural direct effect will be the effect of education on dementia if they were normotensive. For the whole population, the natural direct effect would be a weighted average of these values. We could repeat this calculation for the natural direct effect of education on dementia not mediated by hypertension, but holding hypertension to the value it would have taken under low education.

If the direct effect of X on Y is the same for all possible values of Z , then the natural and controlled direct effects are the same value. We will assume for the remainder of the section that this is the case, but this assumption frequently seems unlikely. If the mediator does modify the direct effect, then the sum of the direct and indirect effects can sum to more than the total effect, making decompositions of the total effect into direct and indirect effects somewhat misleading. This is similar to the problem with estimating “percent of variance explained by genetics” from twin studies, in the context of gene–environment interactions. For a discussion of alternative definitions and issues that arise when the exposure interacts with the mediator, see Robins and Greenland (1992) and Kaufman et al. (2004). We will focus for the remainder of this paper on natural direct effects.

Estimating Direct and Indirect Effects

When Z is believed to partially mediate the effect of X on Y , a common approach to quantifying the direct effect is to compare the regression coefficients for X predicting Y , with versus without simultaneous adjustment for Z (Judd and Kenny 1981; Baron and Kenny 1986). That is:

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 Z \quad (1)$$

Assuming that it is known that X affects Z , rather than that Z affects X , the coefficient β_1 is interpreted as the direct effect of X on Y . To calculate the mediated (indirect) effect, a second regression, without adjustment for Z , is estimated:

$$E(Y) = \gamma_0 + \gamma_1 X \quad (2)$$

The contrast between γ_1 and β_1 is interpreted as the portion of the effect of X on Y that is mediated by Z (MacKinnon et al. 2002). The equality of γ_1 and β_1 can be tested formally using the Hausman test (Kennedy 1998), or the more sophisticated test of Clogg et al. (1995). One limitation with this approach is that if the direct effect is incorrectly estimated, the indirect effect will also be incorrectly estimated, because the indirect estimate is simply the difference between the total effect estimate and the direct effect estimate.

This approach depends on a causal structure such as that shown in Fig. 3a. If instead the causal structure is that shown in Fig. 3b, in which Z causes X , this approach will fail to identify the direct or indirect effects. There is nothing in the regression models above that would distinguish these two structures, so it must be established from prior research or temporal order. Another case in which this approach fails is when the mediator is measured with error. In Fig. 3c, the true mediator is W , and Z is merely an imperfect marker for W .

Even a modest amount of measurement error in the mediator can substantially inflate β_1 , and thus inflate the estimated direct effect. For example, if we attempt to estimate the direct effect of race on cognitive skills, not mediated by differences in schooling, we might regress test score on race and a measure of years of education completed. However, educational attainment is only a rough marker for the true differences in schooling. Even if racial differences were entirely mediated by schooling differences, we would anticipate race would predict test score adjusted for years of schooling (Manly et al. 1999, 2002).

Another important problem with the conventional approach to mediation tests occurs when there are confounders of the mediator–outcome relationship, i.e. an unmeasured common cause of Z and Y , as in Fig. 3d. In this case, the approach described above does not generally give correct estimates of either the direct or indirect effects of X on Y , and this is due to collider bias, which we introduced earlier. These unmeasured common causes may be completely unassociated with X ; indeed this applies *even when X has been randomized*. Confounders of the Z – Y relationship will nonetheless bias the estimate of the direct effect of X on Y . This may be surprising because we are not accustomed to considering carefully whether our hypothesized mediating variables might have unidentified confounders with the outcome.

The standard approach to testing for mediation fails due to collider bias whenever the mediator–outcome association is confounded: conditioning on a common effect of two variables induces a statistical association between those variables. Thus, conditioning on Z by statistical adjustment as in Eq. 1 induces a spurious association between X and the unmeasured confounder of Z . Even if the association between X and Y would have been unbiased in model 2, it is effectively confounded (by the unmeasured confounder of Z) once we condition on Z . Whatever the causal relation between X and Y , when Z is held constant (statistically), the statistical association will reflect this causal relation plus the spurious association via U .

As an example, suppose we are interested in knowing whether the relation between education and dementia is mediated by systolic blood pressure (SBP). Unfortunately, there is an unmeasured genotype that increases both SBP and dementia risk. This genotype is unrelated to education. In Fig. 3d, X represents education, Y represents dementia, Z represents SBP, and U represents the unmeasured genotype. To estimate the effect of education on dementia not mediated by SBP, we need to compare dementia rates in people with high and low education if the value of SBP were not allowed to change in response to education. For example, if we gave someone low education but intervened to hold their SBP to the SBP they would have had, if they had high education (but changed no other characteristics of the situation), how would dementia prevalence change compared with merely giving the person low instead of high education? Of course, we cannot conduct such an intervention because we cannot actually physically set the person's SBP. The mediation analysis described above instead compares the dementia rates of people with high versus low education but who happened to have the same level of SBP. Overall, someone with high education will also tend to have lower SBP than someone with low education. A high-education person with the same SBP as a low-education person is likely to have elevated SBP for some other reason, such as the genotype mentioned above. Thus, the mediation analysis will be comparing people with high education but a high-risk genotype to people with low education and a low-risk genotype (or at least, people with low education who are not selected based on genotype). Because the genotype affects dementia, the high education people who also have high SBP will appear to be worse off than high education people overall. This will in effect underestimate the direct effect of education on dementia. Under the traditional analysis plan, if we underestimate the direct effect, we will automatically overestimate the mediated effect. This same phenomenon can be explained more formally

using counterfactual language. The implication of this problem is that possible confounders of the mediator–outcome association must be assessed just as thoroughly as potential confounders of the exposure–outcome association. Furthermore, it is very helpful to accompany mediation analyses with sensitivity analyses showing a range of plausible estimates under various assumptions about unmeasured confounders of the mediator–outcome association.

Cole and Hernán (2002) provide an accessible discussion of this problem illustrating with a numerical example. Blakely (2002), in a response to Cole and Hernán, called for careful sensitivity analyses to determine whether substantial bias is introduced under realistic assumptions about the strengths of the causal relations.

Ceiling and Floor Effects

Frequently, the outcome variable of interest is measured using an instrument that has an artificial maximum (ceiling) or minimum (floor). Throughout this section, we will focus our discussion on situations in which a measurement ceiling is present. However, measurement floors induce the same biases, albeit on the opposite end of the measurement spectrum, and managing these biases entails the same approaches (and limitations).

One way to consider the potential problems that measurement ceilings cause is to view ceilings as a source of differential measurement error. The intuition, which we will later illustrate with an example, is this: suppose the exposure X is positively related to the continuous outcome Y , but Y is measured with a ceiling. Even as the *true* values of Y increase beyond the ceiling, those *observed* values of Y remain at the ceiling, and thus their corresponding measurement errors (observed Y minus true Y) grow progressively more negative. Therefore, overall, measurement error in Y becomes a function of X .

Ceiling Bias in Cross-Sectional Studies

It is well known that when instruments with artificial ceilings are used to measure the outcome variable in a cross-sectional study, Ordinary Least Squares (OLS) regression analyses provide effect estimates that are attenuated toward zero (Kennedy 1998; Chay and Powell 2001), but in a setting with longitudinal assessments, the bias is of unknown direction.

Several approaches have been suggested for analyzing outcomes measured with ceilings, including: ignoring the ceiling, deleting everyone with a score equal to the ceiling, using Tobit regression models, using median-based regression, and, in the case of longitudinal analyses, adjusting for the baseline value of the outcome variable.

To make this discussion explicit, we will focus on the case when immediate recall (Recall) or attention is the outcome of interest, and it is measured with a Digit Span Forward test (Digits), when the test was stopped at a maximum of eight items. The same concerns apply for any continuous latent construct (e.g., executive function, visuospatial performance), when measured with an instrument with a low ceiling. We assume a simple function g relates recall to digits score i.e., $\text{Digits} = g(\text{Recall}) + \epsilon$, and we further assume that this function is linear for most of the observable range:

$$\text{Digits} = \alpha_1 + \alpha_2 \text{ Recall} + \epsilon \quad (3)$$

This function is not entirely adequate, however, because although the maximum possible digits score is 8, we assume that recall has no maximum. Therefore, we modify the formula relating digits to the unobserved construct recall as follows:

$$\text{Digits} = \min(8, \alpha_1 + \alpha_2 \text{ Recall} + \varepsilon) \quad (4)$$

To estimate the effect of an exposure X on recall, we would regress recall on the exposure of interest:

$$E(\text{Recall}) = \beta_0 + \beta_1 X \quad (5)$$

Unfortunately, this is not possible because we have no direct measure of recall, so instead we regress digits score on exposure:

$$E(\text{Digits}) = \gamma_0 + \gamma_1 X \quad (6)$$

With a ceiling, unless β_1 is equal to 0, γ_1 does not equal β_1 . It is easy to see this graphically (Fig. 4). Any high values of recall tend to be reclassified downwards to the ceiling. If there is a positive relation between the exposure and recall, the negative error in measuring recall will be more common for high values of the exposure. Thus, the increment in the outcome expected for each increment in the exposure will be underestimated, and the coefficient will be attenuated towards 0. The consequence of a measurement floor is identical: the coefficient will be attenuated towards 0 compared with the coefficient that would have been observed without the measurement floor.

Ceiling Bias in Longitudinal Studies

All of the issues discussed above apply quite generally to repeated-measures contexts, regardless of the number of assessment waves. For simplicity, we will consider a situation in which we have only two measurements of recall: baseline and follow-up. We are interested in whether the exposure affects the amount of change in recall experienced by an individual over the follow-up period, as in the equation below, in which j indexes time of observation:

$$E(\text{Recall}_j) = \beta_0 + \beta_1 X + \beta_2 \text{Time}_j + \beta_3 X * \text{Time}_j \quad (7)$$

When recall is measured with a ceiling using the digits test, β_1 , the estimated cross-sectional relation of X to recall at baseline, will be attenuated as discussed above. We interpret β_2 as the change in recall over time expected among subjects with $X=0$, and β_3 as the effect of X on change in recall. If average recall changes over time, then β_2 and β_3 will be biased, but in directions that depend on the study, and are often unknown. This is because the ceiling restricts how much change in the outcome we get to see, and the change could be in either direction. Although it is tempting to assume in the case of recall in old age that the direction of change is likely to be downwards, practice effects (described in “Practice or Retest Effects”) could easily alter this. Consider four types of trajectories: (1) person’s recall corresponds to a ceiling score or higher at baseline and follow-up; (2) person’s recall is below the ceiling at baseline but corresponds to a score equal to the ceiling or higher at follow-up; or (3) person’s recall is at the ceiling at baseline but declines by follow-up, or (4) person’s recall is below the ceiling at baseline and follow-up. Unless our study includes only type 4 people, we will generally observe biased estimates of the effect of exposure on change.

Solutions and Non-solutions

In some cases, it may be tempting to delete observations that hit the ceiling, with the reasoning that these are the clearly mismeasured observations, and we can calculate an unbiased effect estimate using the other observations. This approach can be disastrous. Deleting measurements with the maximum score is an example of conditioning on the

dependent variable. One way to see the blunder in this approach is to, first, recall that nearly all phenomena—especially cognitive skills—are labile (naturally vacillate around a central point) and are usually further measured with a degree of error (e.g., Eq. 3 above), irrespective of any ceilings that a measurement instrument might impose; and second, note what would happen to a high-recall person, a person who, without measurement error, was right on the cusp of the ceiling. If this person had a negative measurement error (i.e., $\epsilon < 0$, and thus $\text{Digits} = a_1 + a_2 \text{ Recall} + \epsilon < \text{ceiling}$), he might stay in the data set. On the other hand, if he had a positive error, he would be deleted. Thus, deleting ceiling values is tantamount to deleting all the high-functioning people with positive measurement error and keeping the high-functioning people with negative measurement errors. This approach induces serious bias in both the cross-sectional and longitudinal cases. One equally problematic variation on this approach is to redefine a continuous outcome variable, such as cognitive function, to exclude people with a particularly poor score, in essence, imposing an artificial floor on observed recall. For example, considering determinants of cognitive function among those without cognitive impairment introduces this sort of bias, when impairment is defined as a cognitive test score below a pre-defined threshold.

Another option sometimes adopted to handle ceilings in the longitudinal case is adjustment for baseline value of the outcome. The reasoning is that, if someone starts high, they cannot “go up” much, while someone who starts low has more potential to increase at the follow-up measurement, and therefore a fair analysis should evaluate the effect of X on change within “level fields” of baseline performance and improvement potential. This is the intuition that often motivates analyses estimating change within levels of baseline score. Unfortunately, as we describe in “Cognitive Change and Baseline Adjustment” below, this approach can lead to serious bias if there is even modest unreliability or instability in the outcome measurement (as hypothesized above; Yanez et al. 1998, 2002; Glymour et al. 2005).

Tobit regression was developed specifically to address the problem of “censored” observations in linear regression. The assumption of Tobit is that the latent (unobserved) outcome variable is normally distributed, and under this assumption Tobit provides an unbiased effect estimate. The effect estimate is based on the Inverse Mills Ratio, which estimates how much the mean of a distribution shifts as its tails are censored. However, unlike OLS, Tobit regression is quite sensitive to violations of the homoscedasticity and normality assumptions (Greene 2000).

Another approach to avoiding bias from ceilings is to use median- (or any quantile-) based methods, such as Least Absolute Deviations (LAD) regression. Instead of minimizing the sum of the squared residuals as in OLS, LAD regression chooses the beta coefficients in order to minimize the sum of the absolute values of the residuals:

$$\sum_{i=1}^n |y_i - x_i\beta| \quad (8)$$

In OLS models, regression coefficients describe the expected difference in mean outcome that corresponds to a one-unit change in the predictor. In contrast, regression coefficients in LAD models describe the expected change in the *median* outcome given a one-unit change in the predictors. This approach is equivalent to quantile regression for the median.

To provide a sense for why minimizing the sum of the absolute deviations (as opposed to the squares of the deviations, as in OLS) gives the median, imagine a data set with only three observations on Y : -20, 0, and 2. Now choose the constant c that minimizes the sum of $|Y - c|$. The best choice for c is 0, the median of the outcome values. If c is 0, the sum of the

absolute deviations is 22. A value of c one unit above the median is one unit closer to the observations above the median, but one unit farther from all the observations below the median *and* one unit farther from the median observation. (In this example, choosing $c=1$ gives a sum of 23.) The same reasoning applies for choosing a c below the median.

LAD regression may be preferable to OLS regression when the dependent variable is subject to substantial measurement error, because LAD, unlike OLS, is robust to outlying values of the dependent variable (Rousseeuw and Leroy 1987). In particular, the potential appeal of LAD regression in the context of measurement ceilings is that median levels of an outcome are unlikely to be altered in the presence of a ceiling. In fact, LAD regression is unbiased if the median is below the ceiling in *every* stratum of the covariates (e.g., for all values of the exposure). However, in strata where the observed median equals the ceiling (equivalently, when the true, unobserved median equals or exceeds the ceiling), the LAD regression coefficients are biased.

Censored Least Absolute Deviations (CLAD) regression is a modification of LAD that avoids this problem. CLAD regression iteratively recalculates the LAD coefficients, omitting observations whose *predicted* outcome values exceed the ceiling at each iteration until a LAD equation exists such that no observations meet this omission criterion. CLAD estimates are centered around the true effect estimates (for the median) under a broad range of distributions of the residuals (Powell 1984). The statistical efficiency of LAD relative to OLS depends on the distribution of the residuals: if the regression residuals are homoscedastic and normally distributed, OLS has more statistical power than LAD, although the coefficients are identical. The CLAD process of dropping observations whose predicted values lie outside the observable range will (generally) reduce the statistical power of the method.

Another option, which may also be of substantive interest, is to estimate the regression for a quantile other than the median. For example, even if the median is at the ceiling in many covariate strata, the 40th percentile might not be. The CLAD approach can be used with any quantile, conceptually addressing the question of how that quantile differs across levels of the exposure variable. Although rarely examined, it may frequently be of interest to examine how various percentiles of an outcome correspond to an exposure. For example, by comparing the CLAD coefficients that correspond to different quantiles of cognitive function score, we could ask whether age induces declines in high functioning individuals at the same rate as it induces declines among the low functioning.

One limitation of CLAD is that there are not analytical solutions to the standard errors. However, it is straightforward to bootstrap the standard errors. If the data comprise repeated measures on an individual, the observations are likely to be autocorrelated, and if this autocorrelation is not taken into account, the standard errors will be underestimated. To avoid this problem, the bootstrap can draw a resample of individuals, instead of resampling observations within individuals (Rust and Rao 1996; Carpenter and Bithell 2000). CLAD can also be difficult and time-consuming to implement. Because there are not analytic solutions to the coefficients, they must be identified by iteration. Models composed of numerous categorical independent variables may pose particular difficulty. Finally, in fitting a CLAD model to our data, we assume that the relation of the exposure and covariates to the outcome is the same above the ceiling (in the unobserved region) as below it.

The analytical biases induced by the presence of outcome measurement ceilings have rightfully been a source of concern and have spurred the development of innovative corrective methods. However, it is worth considering that the magnitude of any ceiling-generated bias is directly proportional to the fraction of observations that hit the ceiling: if

only a small percent are affected, the bias will be minimal. Moreover, for many cognitive outcomes, the assumption of a normally distributed latent variable appears plausible, and Tobit models may be adequate.

Cognitive Change and Baseline Adjustment

In research on the determinants of change in cognition, a crucial analytic decision is whether to adjust for baseline cognitive status. For example, if we were interested in the effect of education on change in cognitive function between time 1 and time 2, we must choose between the following two models:

$$\text{score}_{i2} - \text{score}_{i1} = \gamma_0 + \gamma_1 \text{education}_i + \gamma_2 \text{score}_{i1} + \varepsilon_{it} \quad (9)$$

$$\text{score}_{i2} - \text{score}_{i1} = \beta_0 + \beta_1 \text{education}_i + \delta_{it} \quad (10)$$

where $\text{score}_{i,t}$ represents our measure of cognitive function for person i and time t . The baseline adjusted change score model in Eq. 10 produces coefficients for education that are identical to the following model in which only time 2, not the change score, is used as the dependent variable:

$$\text{score}_{i2} = \gamma_0 + \gamma_1 \text{education}_i + \gamma_2' \text{score}_{i1} + \varepsilon_{it}' \quad (11)$$

There is a common intuition that it is preferable to adjust for baseline scores (Eq. 9) if baseline scores predict follow-up scores (as they nearly always do) and are also correlated with the exposure of interest (as is also typical). Contrary to this intuition, in many very common situations, adjusting for the baseline score will introduce a bias into estimates of the effect of an exposure on cognitive change, when the results without baseline adjustment would be unbiased. Any of several causal structures could account for a discrepancy between baseline adjusted and baseline unadjusted models. In general, if exposure predicts baseline level of the outcome, conditioning on this baseline measure induces a spurious correlation between the exposure and change score in either of two common situations:

1. Measures of the outcome fluctuate due to imperfect measurement reliability or latent variable instability; or
2. Change has already occurred prior to the baseline measurement; the rate of change experienced in the past predicts the future rate of change, and exposure is unaffected by baseline function (rather, exposure directly or indirectly affects baseline function).

Whenever either of these criteria is met, exposure is likely to be a statistically significant predictor in baseline-adjusted change score regression models even when there is no causal effect of exposure on change. Similarly, if there is a causal effect, baseline-adjusted models will provide biased effect estimates.

In both cases, the bias can be anticipated by noting that baseline adjustment compares exposed individuals and unexposed individuals who have the same observed baseline score. If exposure predicts a higher baseline true value of the outcome, then there must be some explanation for why an exposed person would have the same score as an unexposed individual. For example, suppose that two apparently healthy 65 year olds, one with no schooling and the other with a college education, each required 86 s to complete the Trail making Test B (TMT). The low education person is probably at his or her group median or better. The college educated individual is performing well below group norms for the TMT

(Tombaugh 2004). One possible explanation is that the college-educated person is a true outlier from his or her population mean. However, it is also possible that the college-educated person's test was inaccurate, or the person experienced a transient impairment on the day of the exam, such as depression or sleep-deprivation. A third possibility is that the college-educated person has experienced some form of ongoing cognitive deterioration, which is likely to continue. If the score is due either to measurement error or the result of an ongoing process of deterioration, baseline adjustment will bias the estimated effect of exposure on true cognitive change.

The bias in the context of measurement error occurs because of regression to the mean. We anticipate the individual from the high-performing population who received a low score due to a transient condition or measurement error is likely to "regress" to his or her high-performing mean. The person from the low-performing population, despite having an identical baseline observed score, is likely to regress back to his or her low-performing population mean. Regression to the mean, which would occur in the absence of any causal effect of exposure on change in true score, will make it appear that those in the more educated, high-performing group, have an upward change score, compared with those in the less educated, low-performing, group (Yanez et al. 1998, 2002; Glymour et al. 2005).

In this first situation in which measurement error in the outcome is present, the bias resulting from baseline adjustment typically exaggerates the observed exposure–change association in the direction of the exposure–baseline association (e.g., if the exposure is related to high levels in baseline cognitive score, then baseline-adjusted analyses of change will produce estimates for the exposure–change relation that are too high). In the second situation in which baseline adjustment can introduce a bias, even in the absence of measurement error, the bias typically runs in the opposite direction. Bias introduced by baseline adjustment when the baseline measurement is taken in the middle of the course of an ongoing change process is essentially due to selection, but sometimes more colorfully termed "horse-racing bias." (Peto 1981). The term horse-racing is based on the point that if you begin watching horses compete in the middle of the race, the one who is ahead already is probably the faster horse. If you begin observing cognitive trajectories in the middle of a change process, the person with the lower score probably has the worse illness. Whatever features of the person or illness that resulted in large declines *before* you measured his or her cognitive status are likely to continue to operate during the follow-up period. In the absence of measurement error, a low performing person from a high-performing population likely has some unobserved problem that has resulted in low performance, such as an aggressive form of illness. In contrast, a high performing person from a low-performing population probably has some unobserved asset that has resulted in high performance, despite his or her group mean. If you compare these two people, despite having the same baseline score, you expect that the "decliner" will continue to decline. Any characteristic that is associated with being a "decliner" in the baseline group will predict future decline. If two individuals have the same baseline score but one is from a generally high-performing population, it is likely this person is in fact a "decliner." Thus, in this baseline adjustment bias, any characteristic that typically indicates membership in a high-performing population will spuriously predict decline. For example, if a college educated 65 year old takes 86 s to complete the TMT, we might suspect the person is suffering from incipient AD (Alzheimer disease). In contrast, there is no reason to think that a 65-year old with no schooling who requires 86 s to complete the TMT is suffering from AD. Comparing the college-educated person with a score of 86 to the person with no education with a score of 86 is tantamount to comparing someone with undiagnosed AD to a healthy individual. This comparison will make it appear that education predicts cognitive decline.

Although baseline adjustment often biases models of change, there are situations in which baseline adjustment is useful, or when neither a baseline-adjusted nor an unadjusted model could identify the causal effect of interest. To determine the best option, it is often useful to draw a causal diagram, including measurement error and cognitive change, and determine adjustments that would be needed to block the backdoor paths between the exposure of interest and cognitive change.

Practice or Retest Effects

Differences in cognitive test scores observed across repeated assessments reflect actual changes in cognitive function, period effects due to test administration or other contextual factors, and the effects of practice (McCaffrey and Westervelt 1995; Salthouse et al. 2004). Frequently, it is only the actual changes in cognitive function that are of substantive interest, and practice and period effects are considered nuisances. Practice effects are especially troubling because prior research indicates that the magnitude of practice effects depend on characteristics of the test itself, test administration, and the test-taker (Horton 1992; Wesnes and Pincock 2002). For example, people with low education may have more to gain from repeated encounters with a cognitive assessment, whereas people with high education have already taken a great number and variety of cognitive assessments, and have little to gain from any single new encounter. In this case, we may see larger practice effects among those with low education. If these practice effects are not accurately identified and separated from cognitive aging, our estimates of the effect of education on cognitive aging will similarly be biased. The same concerns hold for any exposures that may modify the magnitude of practice effects. In epidemiological studies of determinants of cognitive change, it is therefore important to carefully characterize the practice effects for each population subgroup for each test. We begin by presenting a counterfactual definition of practice effects, describing why practice effects cannot be directly estimated for individuals and considering alternative approaches to estimating such practice effects.

A Counterfactual Definition of Practice Effects

Let $Y_{i,j,k}$ represent person i 's test score when taking a test in year j for the k th occasion. The practice effect P at occasion k is then defined as the difference between person i 's actual score on the test and the score s/he would have achieved if s/he had never before taken the test:

$$P_{i,j,k} = Y_{i,j,k} - Y_{i,j,1} \quad (12)$$

Unfortunately, it is never possible to observe both of these values for the same individual. No person can both take the test for the first time at year j and take the test for the k th time at year j . It is tempting to estimate the practice effect by using change scores between successive assessments for example, if assessments are taken in year $j-1$ and year j :

$$P_{i,j,k} = Y_{i,j,k} - Y_{i,j-1,1} \quad (13)$$

The difficulty with this approach is that change score reflects not only practice effects, but also true change in the construct being measured and period effects. To identify the practice effect with this approach, we would need to know the magnitude of the period and aging effects so we could back these out.

Approaches to Estimating Practice Effects

Despite the impossibility of identifying individual practice effects, it is possible to estimate the average practice effect for a population if some individuals are randomly assigned take the test earlier than others. The early-starters take the test for the first time in year $j-1$ and

repeat the test at time j . The late starters take the test for the first time in year j . Under randomization, the average test scores of the late-start group represent the average test score individuals in the early start *would have had* if they had taken the test for the first time in year j (option 1 in Fig. 5). The average practice effect can be estimated by comparing the averages between the randomized groups:

$$\bar{P}_{j,k} = \bar{Y}_{j,k} - \bar{Y}_{j,1} = \bar{Y}_{\text{early},j,k} - \bar{Y}_{\text{late},j,1} \quad (14)$$

Because both groups are assessed in year j , period effects do not introduce any bias into this estimate of the average practice effect. The groups were randomly divided, so average age should be equivalent in the two groups. Practice effects could then be calculated for any subgroup of interest, e.g., within strata of age, education, race, or comorbidity. This approach is exploited by Thorvaldsson et al. (2006)

True randomization of the timing of first cognitive assessments is rare in epidemiologic studies of the elderly. In the absence of randomization, various approaches are used to estimate practice effects. One approach is to make assumptions about the magnitude of cognitive aging and period effects, and subtract these from the total observed change to estimate the practice effect (option 2 in Fig. 5). For example, if we assume no “true” cognitive change and no period effects, then the difference between test score at occasion 1 and test score at occasion k for any individual is exactly equal to that individual’s practice effect. Many studies use only brief delays between first and subsequent tests and depend on this assumption of zero aging and zero period effects to estimate the magnitude of the retest effect. It is unclear to what extent such short term estimates generalize to longer delays.

If the assumption of no true cognitive change seems implausible, it is also possible to estimate the rate of cognitive change using, for example, the cross-sectional age differences and then back these out of change scores to estimate the practice effect. The assumptions for this approach are very strong: if the age effect is estimated incorrectly, whether based on prior knowledge or cross-sectional contrasts, the practice effects will also be estimated incorrectly. For example, if cross-sectional estimates of age differences inflate the rate of cognitive aging because successive cohorts are better equipped for neuropsychological tests, the practice effect will be proportionally underestimated. Furthermore, we have to assume that period effects are zero.

Because of the well-documented difficulties with estimating rate of cognitive aging using cross-sectional age comparisons, an alternative is to estimate the rate of cognitive aging using differences in the follow-up times for individual participants. Large studies generally stage follow-ups over months or even years, because it is impossible to complete thousands of interviews in a very short time period. This results in a time difference between the first and last interviews conducted in a given wave that is sometimes nearly as large as the intended time gap between waves. In such a design, the difference in performance between the first and last interviewees in a wave can be used to estimate the effect of cognitive aging and backed out of the practice effects (McArdle and Woodcock 1997). A similar approach would exploit selective refusals, study participants who agree at time 1, refuse to be interviewed at time 2, but rejoin the study at time 3. This is valid only under several assumptions. The reasons that some people were interviewed early and others late must have nothing to do with the cognitive outcome of interest. We must assume that no factor that influences the timing of the interview also independently predicts cognitive test performance. If late contacts or selective refusers tend to be people with heavy burdens of disability or illness, this is likely to introduce bias. Second, we must assume that the practice effect does not decay with longer follow-up intervals. For example, if the earliest re-interview occurs after 12 months delay and the latest re-interview occurs after 24 months

delay, we can calculate the average rate of cognitive aging and thus the magnitude of the practice effect, but only if we assume that the practice effects are equal for a 12 month or a 24 month delay. Finally, we must assume that there are no period effects.

With more than two assessments, we might identify a practice effect if we imposed an assumption regarding the functional form of practice effects, such as that practice effects do not occur after the second time an individual takes a test (Collie et al. 2003). This assumption is also used to justify the dual baseline approach, in which the two tests are given in quick succession and the second assessment is used as the baseline score, on the assumption (McCaffrey and Westervelt 1995; Collie et al. 2003).

A more appealing approach would be to exploit pseudorandomized designs that simulate option 1 in Fig. 5 and as described earlier in this section. Staggered assessment schedules such as those delivered by a randomized design may be mimicked by exploiting variability in the timing of assessments introduced by study design. These may be seemingly arbitrary features of the history of the study, such as when the study received additional funding, or when certain eligible groups were “released” to interviewers. Ideally, such staggering would be formally built into study designs. To do so would be relatively cheap, because it entails delaying cognitive assessments for a randomly chosen subsample, and the time thus freed could be used for other assessments.

Differential Measurement Error and Screening

Diagnostic accuracy is a difficulty not only in clinical settings, but also for etiologic research. In many cases, we are interested in measuring a latent construct (such as “cognitive decline,” “depression,” or “anxiety”) that cannot be measured directly. We must rely instead on instruments or expert observers to make the diagnosis. Thus, in both cases, diagnosis may be subject to some degree of error. Additionally, in many settings, diagnosis as “having” or “not having” a particular condition is based on meeting some arbitrary threshold on a continuously-measured instrument, which may further introduce error. These topics are discussed in the Pedraza and Mungas paper in this issue, regarding psychometrics of cognitive diagnostic instruments. We focus in our discussion on a general framework for causal inference given measurement error, and note some selected issues of screening that may arise in the analysis of large survey data.

The diagram in Fig. 6 illustrates the phenomenon of differential measurement error. In this diagram, the exposure X is a direct cause of outcome Y , but the true value of Y is unobserved, and the measured value of Y depends on both the true value of Y and some error. Performance of the measurement instrument for Y is differential by X , and this is reflected in the path from X to Y_{measured} through error γ . Thus, it will be impossible to estimate the effect of X on Y_{true} given the observed data, since any analysis using X and Y_{measured} will estimate the effect of X on Y based on the combination of the indirect path from X to Y_{measured} through Y_{true} and the measurement path from X to Y_{measured} through γ . Any factor that influences the error will appear to predict Y_{measured} , if that factor is also associated with X , then the effect estimates for X will be biased.

This problem applies even in two-stage screening designs, if the preliminary screening tests operate differently by subgroup. For example, a screener may have 70% sensitivity in highly educated subpopulations but 90% sensitivity in populations with less education. Even if the screen is followed by a comprehensive clinical evaluation, and the clinical evaluation has *perfect* sensitivity and specificity, the screening bias is likely to introduce a bias into effect estimates for education. In the worst case, the screening items may operate so poorly in some subgroups that they do not even identify the worst performers. One response to the concern about differential sensitivity in screening instruments is to adjust the criteria to

reflect population norms for specific demographic subgroups, such as racial or ethnic groups, or educational levels. This could be conceptualized as redefining the criteria so that there is no association between X and Y_{measured} in Fig. 6. Obviously, this causes a serious problem for etiologic research into the effects of race/ethnicity or education on the outcome. The association between X and Y_{measured} reflected both differential measurement error and a true causal process. Once the diagnostic criteria are modified in this way, it becomes almost impossible to investigate whether these characteristics influence true disease risk (Berkman 1986).

Diagnostic inaccuracy also compromises research into the predictors of the trajectory after diagnosis. Many cognitive disorders are diagnosed in part on the basis of performance below a threshold. Because of imperfect reliability and instability in cognitive assessments, after selecting the subgroup with very low scores (e.g., the impaired), the cognitive score is subject to regression to the mean. After diagnosis, a regression to the mean phenomenon is very likely to occur so that patients appear to perform slightly better at successive interviews.

Differential Survival

Prospective, longitudinal studies are an important source of epidemiologic information on the natural history of cognitive trajectories, and the effect of social environmental variables on health outcomes over the life course. Prospective studies permit researchers to establish temporal ordering of exposures and outcomes, and can avoid some of the well-known difficulties with case-control selection in retrospective studies (Schaie 1992; Schaie and Hofer 2001). However, studies of aging cohorts over long stretches of time inevitably run into problems with loss to follow up. Mortality is a particularly thorny cause of loss to follow up, since it is an “absorbing state”, and no amount of aggressive follow up can permit researchers to observe the events that would have occurred in those who died if they had not died at their observed death times. When loss to follow up among the exposed and unexposed is differentially related to other variables, this can induce bias in the estimation of causal effects.

Consider a concrete example. At most ages, mortality rates for African-Americans are substantially higher than mortality rates for whites. The result of this phenomena is that from any given birth cohort, a much smaller fraction of African Americans survive to age 70 compared to the fraction of whites who survive to age 70. It is frequently thought that the few survivors in the African American cohort are likely to on average be “hardier” in some respect than the larger group of white survivors. Hardiness presumably arises because of some unobserved genetic, social, or behavioral characteristics of those people. Indeed, the black-white mortality gap narrows at older ages, and this selective survival phenomenon is frequently cited as the reason for this gap, although other explanations have also been posited. The diminishing mortality gap ultimately leads to a “crossover” in life expectancy among the oldest old. The age at which this crossover occurs has been migrating up in successive cohorts, presumably as the mortality differentials earlier in life diminish with successive cohorts. This phenomenon is best documented in black-white comparisons, but can be seen in almost any high/low mortality group, including comparisons between whites and Asians (in which whites are the high-mortality group), and, in some cases whites and Hispanics (in which Hispanics are the high-mortality group; Markides and Machalek 1984; Corti et al. 1999; Mohtashemi and Levins 2002; Thornton 2004). Furthermore, although we focus here on mortality, the same selection phenomena apply to any outcome that can only occur once, such as onset of a chronic disease. For example, similar concerns apply to research on the effect of race or ethnicity on the incidence of dementia diagnosis. The key is that once you are diagnosed with dementia, you are no longer at risk of developing dementia

in the future: previously diagnosed cases are removed from the risk set. Comparable mortality crossovers among black Americans have also been documented for coronary heart disease (Anonymous 1998; CDC 1998; Corti et al. 1999; Williams et al. 1999) and stroke survival in the UK (Wolfe et al. 2005).

These mortality crossovers, and similar patterns of changing racial/ethnic differences in health outcomes across the age range, likely reflect a mix of substantive and artefactual mechanisms. For example, these patterns of mortality have been attributed to differential errors in age reporting by race (Elo and Preston 1994), race-specific birth cohort effects, and heterogeneity among blacks in terms of birthplace and migration (Corti et al. 1999), or selection bias due to differential survival. Substantive causal interpretations of attenuated coefficients among older cohorts are plausible, because early-life exposures may have larger effects in mid-life than in later old age. Thus, etiologic research must attempt to disentangle the substantive from the artefactual. We focus here on the possibility that selective survival introduces this bias.

Selective survival can bias estimates of the effect of race on mortality risk if any other unmeasured factor (including those *initially unrelated to race*, such as genetic make-up) affects survival to old age. This can result in exaggerations—or even reversals—of the black–white mortality ratio among older age groups (Mohtashemi and Levins 2002; Glymour and Greenland 2008; Glymour 2006b). Survival bias is related to collider bias we discussed in the first section of this paper, in which “survival” is the common effect of race and an unobserved characteristic.

Consider an extremely simplified hypothetical population composed of 50% whites and 50% blacks. Suppose that both race and genetic endowment influence risk of mortality, but at birth these variables are independent: half of blacks and half of whites have “healthy” genes. Suppose a white person with healthy genes has a 1% chance of dying in each year (thus a 99% chance of surviving); a white person with unhealthy genes has a 2% chance of dying; a black person with healthy genes also has a 2% chance of dying; and a black person with unhealthy genes has a 4% annual mortality rate. Although at birth half of the white population carried the healthy gene, and half of the black population carried the healthy gene, prevalence of the healthy gene among survivors will increase each year (Fig. 7). This relative increase will be faster among blacks, because blacks are dying more quickly. After 80 years at these rates, 69% of surviving whites would be expected to carry the healthy gene, while 84% of blacks would be expected to carry the healthy gene. Because mortality rates are so much higher for blacks than whites, we expect that elderly blacks are more likely to carry some beneficial characteristic. This will bias estimates of the *causal* effect of being black on mortality risk unless we have measured that other beneficial characteristic, e.g., the healthy genotype. Even though the 90-year-old black woman might have a longer life expectancy than a white woman of the same age, the black woman would have lived *even longer* if she had been white but the same in every other respect.

Bias induced by differential survival is a form of collider stratification bias. To see this, consider the causal diagram in Fig. 8, in which race and genetics are independent causes of dementia, and also both independent causes of survival to age 80. Since we can only observe dementia among those who did not die (survival=1), observation of the surviving cohort at the end of follow-up implies restriction on survival. As survival is a collider in the DAG, conditioning on survival via restriction of the population to those who are alive induces an association between race and genotype, even though they were initially independent at baseline (before anyone died).

As with other examples of collider stratification bias, if the genetic risk is observed, then further conditioning on Z in the analysis will remove the bias. However, in the presence of unmeasured common causes of survival and dementia, the estimates of the association between X and Y will generally be biased. An identical process can occur in any study focusing on clinical populations. Everyone in the study was selected into being a patient, and in particular being a patient in the clinical setting from which the sample was drawn. Any unobserved factors that influence the probability of being a patient in that particular clinic have the potential to induce a selection bias process.

The strength of the bias induced by collider stratification depends on a number of parameters. Most obviously, these include the strength of the associations among the variables. Frequently, the unobserved associations are not large enough to induce a serious bias, or at least, the selection bias would be dwarfed by other sources of bias (Greenland 2003). Additionally, the prevalence of survival is an important factor. If overall survivorship is quite high (for example, >90%), the potential effect of collider stratification will likely be modest.

The problem of differential loss to follow up and its effects on causal inference is well known in the survival analysis literature. For example, let us consider a time-to-event analysis in which the Y is diagnosis with Alzheimer's disease. In the vocabulary of survival analysis, subjects who die before developing Alzheimer's are censored at their time of death. It is a basic tenet of survival analysis that valid estimates of the effect of X and Y are obtainable only when censoring and event times are conditionally independent (Kalbfleisch and Prentice 1980; p. 120). If correlation between censoring and failure times are induced (say, by the existence of an unmeasured common cause U), then this assumption is violated. Moreover, the validity of the assumption of conditional independence is unlikely to be empirically demonstrable, since by definition one does not (and in the case of mortality, one cannot possibly) observe the distribution of event times among those who are censored.

Finally, this problem can also be conceptualized as a problem of missing data, where subjects who die prematurely are missing data on the outcome. Estimation of the effect of an exposure and outcome is only valid if all common causes of the outcome and missingness are observed, and can be included in the analysis. In this situation, the data are considered Missing At Random, whereas if an unmeasured common cause of missingness and the outcome exists, the data are considered Not Missing At Random (Little and Rubin 2002).

Improved understanding of survival bias may have application to studies of the hypothesis of "fetal origins of disease." Originally proposed by David Barker and colleagues (Barker 1990), the hypothesis proposes that prenatal and postnatal nourishment programs an infant's metabolic physiology which affects his health risks as an adult. The hypothesis has generated about as much empirical evidence as it has controversy (Paneth and Susser 1995; Tu et al. 2005; Weinberg 2005). Yet, given that the exposures of interest occur early in life and many outcomes of interest occur in older age, selective survivor may have an impact on these results, plausibly *underestimating* the risks associated with low birth weight (a proxy for suboptimal fetal nourishment), depending on the assumptions made about the relations between key variables (Weinberg 2005).

Survivor bias may also be operating in the observations that traditional risk factors for mortality seem to lose their "punch" among older adults. A Danish study of 463 nonagenarians failed to find mortality risks associated with smoking, obesity, education, and alcohol consumption, although mortality was associated with impaired cognition (Nybo et al. 2003). In the White hall Study in the UK, low employment grade, cigarette smoking, elevated blood pressure, and elevated cholesterol were all strongly associated with mortality,

but these associations attenuated noticeably among older men (Marang-van de Mheen et al. 2001). Similar results have been observed in other cohorts (Tate et al. 1998). These results have strong implications on the policies and clinical practices that affect older adults (Howard and Goff 1998; Kaplan et al. 1999). Hernan et al. have hypothesized that this phenomenon may lead to a reversal of the association between cigarette smoking and dementia, based on the age of the cohort (Hernán et al. 2008).

The mortality crossovers are not merely esoteric statistical phenomena. As discussed above, they can compromise etiologic research. In the public realm, these results are occasionally viewed not just as curiosities, but have been presented as promising evidence of a health care system that is working well for black older adults (e.g., Moss L. “Stroke survival rates higher among blacks.” Press Association Limited. July 29 2005). Indeed, an accumulation of similar results could sway health policy decisions.

Implications

Several methodological problems, including those introduced by measurement difficulties, have either technical or design solutions. Some do not, but sensitivity analyses are helpful to assess the direction and magnitude of the likely biases. Finally, some rather common analytic approaches are vulnerable to biases that compromise their usefulness for etiologic research. Specifying causal assumptions and considering the testable (or untestable) implications of those assumptions can help understand and avoid such biases. Failing to draw careful distinctions between statistical associations and causal relationships can be especially problematic in neuropsychological research, because measurement difficulties and selection processes frequently create entirely spurious statistical associations. In some cases, such as baseline adjustment or ceiling effects, these difficulties can be exacerbated by our best efforts to use statistical methods to resolve the problem.

Acknowledgments

Maria Glymour was a Robert Wood Johnson Health and Society Scholar at Columbia University when this was written. The authors gratefully acknowledge funding support from the National Institutes of Health (ES005257) and a Robert Wood Johnson Health and Society Program seed grant.

References

- Angrist JD, Imbens GW, et al. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*. 1996; 91(434):444–455.
- Angrist JD, Krueger AB. Instrumental variables and the search for identification: from supply and demand to natural experiments. *Journal of Economic Perspectives*. 2001; 15(4):69–85.
- Anonymous. Coronary heart disease mortality trends among whites and blacks Appalachia and United States, 1980–1993. *MMWR CDC Surveillance Summaries*. 1998; 47(46):1005–1008.
- Barker DJ. The fetal and infant origins of adult disease. *BMJ*. 1990; 301(6761):1111. [PubMed: 2252919]
- Baron RM, Kenny DA. The moderator mediator variable distinction in social psychological-research —Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*. 1986; 51(6):1173–1182. [PubMed: 3806354]
- Berkman LF. The association between educational attainment and mental status examinations: Of etiologic significance for senile dementias or not. *Journal of Chronic Diseases*. 1986; 39(3):171–175. [PubMed: 3949941]
- Blakely TA. Commentary: Estimating direct and indirect effects—Fallible in theory, but in the real world. *International Journal of Epidemiology*. 2002; 31:166–167. [PubMed: 11914315]
- Carpenter J, Bithell J. Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine*. 2000; 19:1141–1164. [PubMed: 10797513]

- Catalano R, Bruckner T, et al. Ambient temperature predicts sex ratios and male longevity. *Proceedings of the National Academy of Sciences of the United States of America*. 2008; 105(6): 2244–2247. [PubMed: 18250336]
- CDC. Trends in ischemic heart disease deaths rates for blacks and whites—United States, 1981–1995. *MMWR CDC Surveillance Summaries*. 1998; 47(44):945–949.
- Chay KY, Powell JL. Semiparametric censored regression models. *Journal of Economic Perspectives*. 2001; 15(4):29–42.
- Clogg CC, Petkova E, et al. Statistical methods for comparing regression coefficients between models. *American Journal of Sociology*. 1995; 100(5):1261.
- Cole SR, Hernán MA. Fallibility in estimating direct effects. *International Journal of Epidemiology*. 2002; 31(1):163–165. [PubMed: 11914314]
- Cole SR, Hernan MA, et al. Marginal structural models for estimating the effect of highly active antiretroviral therapy initiation on CD4 cell count. *American Journal of Epidemiology*. 2005; 162(5):471–478. [PubMed: 16076835]
- Collie A, Maruff P, et al. The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test–retest intervals. *Journal of the International Neuropsychological Society*. 2003; 9(3):419–428. [PubMed: 12666766]
- Corti MC, Guralnik JM, et al. Evidence for a Black–White crossover in all-cause and coronary heart disease mortality in an older population: The North Carolina EPSE. *American Journal of Public Health*. 1999; 89(3):308–314. [PubMed: 10076478]
- Currie, J. *Welfare and the well-being of children*. Harwood Academic; Chur Switzerland: 1995.
- Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*. 2007; 16(4):309–330. [PubMed: 17715159]
- Elo IT, Preston SH. Estimating African–American mortality from inaccurate data. *Demography*. 1994; 31(3):427–458. [PubMed: 7828765]
- Glymour, MM. Natural experiments and instrumental variables analyses in social epidemiology. In: Oakes, JM.; Kaufman, JS., editors. *Methods in social epidemiology*. Jossey-Bass; San Francisco: 2006a.
- Glymour, MM. Using causal diagrams to understand common problems in social epidemiology. In: Oakes, JM.; Kaufman, JS., editors. *Methods in social epidemiology*. Jossey-Bass; San Francisco: 2006b.
- Glymour, MM.; Greenland, S. Causal diagrams. In: Rothman, KJ.; Greenland, S.; Lash, TL., editors. *Modern epidemiology*. Lippincott Williams & Wilkins; Philadelphia: 2008. p. 183–210.
- Glymour MM, Kawachi I, et al. Does childhood schooling affect old age memory or mental status? Using state schooling laws as natural experiments. *Journal of Epidemiology and Community Health*. 2008; 62:532–537. [PubMed: 18477752]
- Glymour MM, Weuve J, et al. When is baseline adjustment useful in analyses of change? An example with education and cognitive change. *American Journal of Epidemiology*. 2005; 162(3):267–278. [PubMed: 15987729]
- Greene, WH. *Econometric analysis*. Prentice-Hall; Upper Saddle River: 2000.
- Greenland S. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*. 2000; 29(4):722–729. [PubMed: 10922351]
- Greenland S. Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology*. 2003; 14(3):300–306. [PubMed: 12859030]
- Greenland S, Pearl J, et al. Causal diagrams for epidemiologic research. *Epidemiology*. 1999; 10(1): 37–48. [PubMed: 9888278]
- Hernán MA, Alonso A, et al. Cigarette smoking and dementia: potential selection bias in the elderly. *Epidemiology*. 2008; 19(3):448. [PubMed: 18414087]
- Hernán MA, Brumback B, et al. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*. 2000; 11(5):561–570. [PubMed: 10955409]
- Hernan MA, Robins JM, et al. Statistical issues arising in the Women’s Health Initiative—Discussion. *Biometrics*. 2005; 61(4):922–930.

- Hines LM, Stampfer MJ, et al. Genetic variation in alcohol dehydrogenase and the beneficial effect of moderate alcohol consumption on myocardial infarction. *New England Journal of Medicine*. 2001; 344(8):549–555. [PubMed: 11207350]
- Holland PW. Statistics and causal inference. *Journal of the American Statistical Association*. 1986; 81(396):945–960.
- Horton AM. Neuropsychological practice effects * age: A brief note. *Perceptual & Motor Skills*. 1992; 75(1):257–258. [PubMed: 1528677]
- Howard G, Goff DC. A call for caution in the interpretation of the observed smaller relative importance of risk factors in the elderly. *Annals of Epidemiology*. 1998; 8(7):411–414. [PubMed: 9738686]
- Judd CM, Kenny DA. Process analysis—estimating mediation in treatment evaluations. *Evaluation Review*. 1981; 5(5):602–619.
- Kalbfleisch, JD.; Prentice, RL. *The statistical analysis of failure time data*. Wiley; New York: 1980.
- Kaplan GA, Haan MN, et al. Understanding changing risk factor associations with increasing age in adults. *Annual Review of Public Health*. 1999; 20:89–108.
- Katan MB. Commentary: Mendelian randomization, 18 years on. *International Journal of Epidemiology*. 2004; 33(1):10–11. [PubMed: 15075137]
- Kaufman JS, Cooper RS. Commentary: Considerations for use of racial/ethnic classification in etiologic research. *American Journal of Epidemiology*. 2001; 154(4):291–298. [PubMed: 11495850]
- Kaufman J, Maclehorse R, et al. A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. *Epidemiologic Perspectives & Innovations*. 2004; 1(1):4. [PubMed: 15507130]
- Kennedy, P. *A guide to econometrics*. MIT; Cambridge: 1998.
- Lawlor DA, Harbord RM, et al. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*. 2008; 27(8):1133–1163. [PubMed: 17886233]
- Little J, Khoury MJ. Mendelian randomisation: a new spin or real progress. *The Lancet*. 2003; 362(9388):930–931.
- Little, RJ.; Rubin, DB. *Statistical analysis with missing data*. Wiley; Hoboken: 2002.
- MacKinnon DP, Lockwood CM, et al. A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*. 2002; 7(1):83–104. [PubMed: 11928892]
- Manly JJ, Jacobs DM, et al. Effect of literacy on neuropsychological test performance in nondemented, education-matched elders. *Journal of the International Neuropsychological Society*. 1999; 5(3): 191–202. [PubMed: 10217919]
- Manly, JJ.; Jacobs, DM., et al. Reading level attenuates differences in neuropsychological test performance between African American and White elders. Vol. vol. 8. Cambridge University Press; Cambridge: 2002. p. 341-348.
- Marang-van de Mheen PJ, Shipley MJ, et al. Decline of the relative risk of death associated with low employment grade at older age: the impact of age related differences in smoking, blood pressure and plasma cholesterol. *Journal of Epidemiology and Community Health*. 2001; 55(1):24–28. [PubMed: 11112947]
- Mark SD, Robins JM. A method for the analysis of randomized trials with compliance information: an application to the Multiple Risk Factor Intervention Trial. *Controlled Clinical Trials*. 1993; 14(2): 79–97. [PubMed: 8500308]
- Markides KS, Machalek R. Selective survival, aging and society. *Archives of Gerontology and Geriatrics*. 1984; 3(3):207–222. [PubMed: 6395820]
- McArdle JJ, Woodcock RW. Expanding test–retest designs to include developmental time-lag components. *Psychological Methods*. 1997; 2(4):403–435.
- McCaffrey RJ, Westervelt HJ. Issues associated with repeated neuropsychological assessments. *Neuropsychology Review*. 1995; 5(3):203–221. [PubMed: 8653109]
- Mohtashemi M, Levins R. Qualitative analysis of the all-cause Black–White mortality crossover. *Bulletin of Mathematical Biology*. 2002; 64(1):147–173. [PubMed: 11868334]

- Nybo H, Petersen HC, et al. Predictors of mortality in 2,249 nonagenarians—The Danish 1905-cohort survey. *Journal of the American Geriatrics Society*. 2003; 51(10):1365–1373. [PubMed: 14511155]
- Oakes JM. The (mis)estimation of neighborhood effects: Causal inference for a practicable social epidemiology. *Social Science & Medicine*. 2004; 58(10):1929–1952. [PubMed: 15020009]
- Paneth N, Susser M. Early origin of coronary heart-disease (the Barker hypothesis). *British Medical Journal*. 1995; 310(6977):411–412. [PubMed: 7873937]
- Pearl, J. *Causality*. Cambridge University Press; Cambridge: 2000.
- Peto R. The horse-racing effect [letter]. *Lancet*. 1981; 2(8244):467–468. [PubMed: 6115215]
- Powell JL. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*. 1984; 25(3):303–325.
- Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992; 3(2):143–155. [PubMed: 1576220]
- Robins JM, Greenland S. Identification of causal effects using instrumental variables—Comment. *Journal of the American Statistical Association*. 1996; 91(434):456–458.
- Rosenfeld CS, Roberts RM. Maternal diet and other factors affecting offspring sex ratio: A review. *Biology of Reproduction*. 2004; 71(4):1063–1070. [PubMed: 15229140]
- Rousseeuw, PJ.; Leroy, AM. *Robust regression and outlier detection*. Wiley; New York: 1987.
- Rubin DB. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*. 1997; 127(8):757–763. [PubMed: 9382394]
- Rust K, Rao J. Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*. 1996; 5(3):283–310. [PubMed: 8931197]
- Salthouse TA, Schroeder DH, et al. Estimating retest effects in longitudinal assessments of cognitive functioning in adults between 18 and 60 years of age. *Developmental Psychology*. 2004; 40(5): 813–822. [PubMed: 15355168]
- Schaie KW. The impact of methodological changes in gerontology. *International Journal of Aging & Human Development*. 1992; 35(1):19–29. [PubMed: 1506115]
- Schaie, KW.; Hofer, SM. Longitudinal studies in aging research. In: Birren, JE.; Schaie, KW., editors. *Handbook of the psychology of aging*. Academic; San Diego: 2001. p. 53-77.
- Smith GD. Genetic epidemiology: An ‘enlightened narrative’. *International Journal of Epidemiology*. 2004; 33(5):923–924.
- Smith GD, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*. 2004; 33(1):30–42. [PubMed: 15075143]
- Spirtes, P.; Glymour, C., et al. *Causation, prediction, and search*. MIT; Cambridge: 2001.
- Tate RB, Manfreda J, et al. The effect of age on risk factors for ischemic heart disease: The Manitoba Follow Up Study, 1948–1993. *Annals of Epidemiology*. 1998; 8(7):415–421. [PubMed: 9738687]
- Thornton R. The Navajo-US population mortality crossover since the mid-20th century. *Population Research and Policy Review*. 2004; 23(3):291–308.
- Thorvaldsson V, Hofer SA, et al. Effects of repeated testing in a longitudinal age-homogeneous study of cognitive aging. *Journals of Gerontology Series B-Psychological Sciences and Social Sciences*. 2006; 61(6):P348–P354.
- Tombaugh TN. Trail making test A and B: Normative data stratified by age and education. *Archives of Clinical Neuropsychology*. 2004; 19(2):203–214. [PubMed: 15010086]
- Tu YK, West R, et al. Why evidence for the fetal origins of adult disease might be a statistical artifact: The “reversal paradox” for the relation between birth weight and blood pressure in later life. *American Journal of Epidemiology*. 2005; 161(1):27–32. [PubMed: 15615910]
- Weinberg CR. Invited commentary: Barker meets Simpson. *American Journal of Epidemiology*. 2005; 161(1):33–35. [PubMed: 15615911]
- Wesnes K, Pincock C. Practice effects on cognitive tasks: a major problem? *Lancet Neurology*. 2002; 1(8):473. [PubMed: 12849328]
- Williams JE, Massing M, et al. Racial disparities in CHD mortality from 1968–1992 in the state economic areas surrounding the ARIC study communities. *Annals of Epidemiology*. 1999; 9(8): 472–480. [PubMed: 10549880]

- Winship C, Morgan SL. The estimation of causal effects from observational data. *Annual Review of Sociology*. 1999; 25:659–706.
- Wolfe CDA, Smeeton NC, et al. Survival differences after stroke in a multiethnic population: follow-up study with the south London stroke register. *British Medical Journal*. 2005; 331(7514):431–433. [PubMed: 16055452]
- Yanez ND, Kronmal RA, et al. The effects of measurement error in response variables and tests of association of explanatory variables in change models. *Statistics in Medicine*. 1998; 17(22):2597–2606. [PubMed: 9839350]
- Yanez ND, Kronmal RA, et al. A regression model for longitudinal change in the presence of measurement error. *Annals of Epidemiology*. 2002; 12(1):34–38. [PubMed: 11750238]

Fig. 1.
Causal diagram illustrating causal structures that generate statistical associations

Fig. 2. Causal structures for valid natural experiments compared to invalid natural experiments. **a** Z is a valid instrument for the effect of X on Y ; **b** Z is not a valid instrument, because there is a direct pathway from Z to Y , not mediated by X ; **c** Z is not a valid instrument, because there is an unmeasured common cause of Z and Y

A. Direct and indirect effects. **B. Reverse causation from Z to X.** **C. Mismeasured mediator** **D. Unmeasured confounder of the mediator** **Fig. 3.**

a–d Contrasting causal structures under which the conventional approach to estimating direct and indirect effects would succeed or fail

Fig. 4.
Bias in the estimated regression slope when cognitive function is measured with a ceiling

Fig. 5.
Approaches to estimating practice effects

Fig. 6.
Differential measurement error

Fig. 7.
The prevalence characteristic changes with selective survival

Fig. 8.
Survivor bias