Causal Inference Challenges in the Relationship Between Social Determinants and Cardiovascular Outcomes

Jay S. Kaufman, PhD

Department of Epidemiology, Biostatistics, and Occupational Health, Faculty of Health Sciences, McGill University, Montréal Québec, Canada

ABSTRACT
The effects of social determinants on cardiovascular outcomes are frequently estimated in epidemiologic analyses, but the profound causal and statistical challenges of this research program are not widely discussed. Here, we carefully review definitions and measures for social determinants of cardiovascular health and then examine the various assumptions required for valid causal inference in multivariable analyses of observational data, such as what one would typically encounter in cohorts, population surveys, health care databases, and vital statistics databases. We explain the necessity of the "well-defined exposure" and show how this goal relates to the "consistency assumption" that is necessary for valid causal inference. Well-defined exposure is especially challenging for social determinants of health because they are seldom simple atomistic interventions that are easily conceptualized and measured. We then review threats to valid inference that arise from confounding, selection bias, information bias, and positivity violations. Other causal considerations are reviewed and

Reducing the population burden of cardiovascular disease (CVD) morbidity and mortality requires effective evidenced-based interventions in both social policy and clinical medicine. The latter category of evidence is well represented in medical journals, but evidenced-based approaches to the former causes of health and disease—the social determinants that arise from lifestyle and environment—are less well developed in the biomedical literature. There is little doubt that disease and mortality rates are highly responsive to social conditions, but evaluating the impact of specific public policies on population disease is a complex and daunting task. Clinical interventions can often be evaluated in randomized trials in which all background patient characteristics can be effectively balanced among treatment groups. For the effects of public policies and social conditions, however, valid and precise causal inference can be more challenging. Medicine and public health have made great strides in formalizing causal inference as the set of designs and quantitative techniques that answer the question: what will the expected outcome be if we do thing A in contrast to doing thing B? In medical contexts, these interventions could be pills or surgical procedures, but the same basic logic applies to population-level interventions such as laws, public policies, and social programs. Medical decision making requires causal inference to discern optimal choices, and this need is no less acute in the realm of public health and social policy. This review therefore considers social determinants of health in relation to CVD and reviews the statistical considerations for causal inference in this relationship.

What Are Social Determinants of Health?
The World Health Organization (WHO) defines “social determinants of health” as the set of nonmedical factors that influence disease and wellness. These are the axes of social distinction among people and material conditions in which they are born, grow up, work, eat, live, and grow old: the
explained, such as correct model specification, absence of immortal
time, and avoidance of the “Table 2 Fallacy,” and their application to
social determinants of cardiovascular outcomes are discussed. Fruitful
approaches, including focusing on policy interventions and the “target
trial” frameworks are proposed and provide a pathway for a more
efficacious research program that can more reliably improve popula-
tion health. Valid causal inference in this setting is quite challenging,
but—with clever design and thoughtful analysis—the important role of
social factors in patterning cardiovascular outcomes can be quantified
and reported.

Structures of daily life and the hierarchies of access and privi-
lege that allot resources; status; and access to goods, services,
and amenities. There are many levels at which these de-
terminants can be measured and studied from the granularity
of individuals at specific moments—for example, in the case of
earned income—to the broad sweep of economic systems
and ideologies that shape distributions of income over whole
nations and epochs.

The social determinants of health therefore not only
corporate social and material environments but also the
policies and ideologies that structure and maintain these social
and material environments. The government of Canada lists 9
social determinants of health: income and social status,
employment and working conditions, education and literacy,
childhood experiences, physical environments, social supports
and coping skills, gender (as distinct from biological sex),
culture and race/ethnicity. But each of these is actually a
collection of proximal and distal determinants that exist at
various levels, from the individual to the national or inter-
national policies. For example, an individual’s attained income
is clearly a dynamic interaction among their particular skills,
opportunities, and choices integrated with the broader set of
national and international policies including trade, immigra-
tion, and currency valuations. Incomes also have an inher-
te relativity effect on lifestyle through the economic
processes that determine prices for essential commodities such
as food and housing. Even this perspective, however, presumes
a capitalist economy in which individuals purchase essential
goods and services as commodities, whereas socialist systems
may guarantee some essential aspects of health so that they are
not functions of an individual’s finances. In this way, eco-

conic ideology is a level of social determination of health that
is situated hierarchically above the others.

Given the almost infinitely diffuse nature of social de-
terminants of health, it is therefore impossible to enumerate
them or study them in any exhaustive or comprehensive
fashion. The vast majority of biomedical studies and publica-
tions therefore approach this topic by selecting a few
salient or tangible examples of social determinants of health:
usually factors that are relatively common in the medical
records and survey datasets used by population health
studies. These commonly measured social factors tend to be
proximal to the individuals whose health outcomes are being
studied. Perhaps the most commonly found of these social
factors in the literature are race and ethnicity, sex and gender,
and socioeconomic class or position. Each of these, in turn,
is subject to considerable variation in conceptualization and
measurement.

Conceptualization and Measurement of the
Social Determinants of Health

Race and ethnicity

Racial and ethnic categories are ubiquitous markers of
social position in multicultural societies because they are
invariably employed in the construction of social hierarchies,
explicitly or implicitly. In the biomedical sphere, race and
ethnicity had traditionally been viewed as relevant to ancestry
and physical characteristics, but the modern conceptualization
has evolved to focus more on these groupings as markers of
self-identification and social affiliation, so that they are best
ascertained by self-report, rather than any physical or bio-

cological criteria. Although some measure of geographic relat-
edness is available from genotyping individuals and comparing
a large number of polymorphic alleles with distributions from
reference populations, this estimation of ancestry components
is distinct from race and ethnicity in many ways, and there is
no strict correspondence between identity and ancestry.
Every country has its own particular history of racial and
ethnic categorizations, and many have established elaborate
administrative management of these categories. This
incommensurability across place and time makes it difficult to
maintain a cohesive and commensurable biomedical literature
on racial and ethnic patterns of health and disease. Given its
recent history of slavery and legal segregation, the United
States has tended to focus most intently on a binary classifi-
cation system of “White” and “Black,” but as the country has
grown increasingly more Hispanic and Asian in the last 50
years, this racial dualism has waned. The United States
government currently recognizes 5 basic racial categories and
additionally applies the dimension of ethnicity as Hispanic or
non-Hispanic to each of these groups. The categories date
from the 1970s but are now in the process of revision, and
new categories are expected in 2024. Canada has a less consistent structure to its collection of racial and ethnic data, sometimes opting for a broad “visual minority” category that curiously excludes Aboriginal populations and sometimes using a large number of check boxes that are a mix of racial and national designations.

Sex and gender

Similar to race and ethnicity, sex and gender serve as important determinants of the social and material circumstances across countries and societies and should therefore almost always be considered in biomedical research, whether as exposure, covariate, effect-measure modifier, or stratification. Gender is the social expression of biologic sex, and both are clearly relevant for health and disease. These categories are traditionally assumed to be assigned at birth into 2 distinct groups by chromosomal and phenotypic traits, but recent years have witnessed a growing challenge to overly simplistic binaries. Nonetheless, the overwhelming practice in biomedical research remains self-report into a binary categorization, whether the variable label is sex or gender. Even so, there is also increasing awareness of individuals that transition from gender assigned at birth and of sexual orientation as a distinct dimension of identity that is relevant to social environment, behaviour, and discrimination, all of which bear on health and well-being. There is also a movement toward the recognition of “intersectionality” between race and gender, which posits some meaningful synergism among these social designations in determining health and disease.

Socioeconomic class or position

Although Marxists define social class dichotomously between workers and owners, modern social epidemiology has been more strongly dominated by a Weberian framework that views social position as a continuous gradation of status and resources, indexed by measures such as attained education or earned income. The number of years of achieved education is often collected on birth certificates and in other vital statistics and offers an ordinal status that does not generally change in adulthood and is not subject to substantial reporting error. One disadvantage is that educational quality may differ widely by time and place, making the number of years spent in school a crude measure with respect to skills and status. There is also a movement toward the recognition of “intersectionality” between race and gender, which posits some meaningful synergism among these social designations in determining health and disease.

Relating Socioeconomic Measures to CVD Outcomes

Across all of epidemiologic science, there are 3 general domains of quantitative inference in which researchers may typically be engaged: surveillance, prediction, and etiologic inference. Surveillance is the set of descriptive activities that are often conducted by health organizations and government agencies to monitor health outcomes of interest in relation to selected dimensions of interest, including socioeconomic measures and geography. Surveillance is about describing the world accurately, quantifying the burden of disease and how this is distributed in the population, and how this burden evolves over time. There may be statistical models employed, but these are largely for the purpose of smoothing, as well as for removing data-collection errors such as imperfections in measurement. There is no consideration of “confounding” in the formal sense of a bias in the estimation of causal effects, as surveillance involves no estimation of causal effects. Even so, descriptive parameters can be biased by imperfections in data gathering and measurement; for example, if some observations are informatively missing. There may also be issues of generalizability from samples, but the quantity of interest is most often a frequency measure of disease, such as a prevalence or incidence, and not a causal effect formed by contrasting levels of an exposure. Surveillance may be made in relation to various scaled axes of space, time, or socioeconomic level, but these are not “exposures” or “treatments” in the sense of contemplating hypothetical interventions on the population under study.

Whereas surveillance depicts the real world at the present time, the goal of prediction is to depict the real world in the future. This involves forecasting how the targeted descriptive quantity, such as a prevalence or incidence, will evolve over time. Prediction is just surveillance extrapolated or modeled into the future. For example, Pearson-Stuttard and colleagues developed a hierarchical Bayesian model to combine age, period, and cohort effects from the last 40 years with expected demographic shifts to project the racial disparities in cardiovascular mortality that will be observed in the United States in 2030. This kind of exercise requires accurate data and appropriate modelling, and so validity necessarily rests on assumptions. But, even so, there is no threat of confounding, as there is no exposure or treatment and therefore no contrast being made. Predictions may have errors if the underlying models are misspecified, but these are prediction errors, not confounding errors.

Finally, the third variety of epidemiologic activity, and the most commonly encountered in research grants and published biomedical papers, is etiologic inference, often referred to as causal inference. This is scientific work in which the parameter of interest is the causal relationship between a defined exposure or treatment and an outcome event. In the case of mediation, this work can also target the specific mechanistic pathway through which this causal effect occurs. In causal inference, we are no longer targeting the real world, but instead the focus is a hypothetical world in which the outcome event frequency would change in response to various alternative settings of the exposure or treatment. These contrasting settings are achieved through hypothetical interventions of some kind, although the exact mechanism of these interventions is often left unspecified. It is in conducting
causal inference that one is faced with the possibility of confounding as a bias, as association is not necessarily the same as causation. What this means is simply that observing X and Y to track in the data does not imply that intervening on X will affect Y when it is assigned by external intervention: for example, that shooting a rooster does not prevent the rising of the sun.

### Causal Inference Considerations: The Well-Defined Exposure

Yuusuf and colleagues analyzed over 155,000 individuals from 21 countries in the Prospective Urban Rural Epidemiology (PURE) study, reported in The Lancet, in 2020, that the covariate-adjusted effect of low education (primary school or less vs university) on CVD was a hazard ratio of 1.37 (95% confidence interval [CI], 1.23-1.52). These authors were not asserting that the hazard of incident CVD in the poorly educated is 37% higher than in the university educated, because that would be the crude or descriptive result. Rather, this conditional hazard ratio emerges from a multivariate survival model in which there was simultaneous adjustment for 15 individual and household-level covariates. Adjustment for covariates is intended to reduce the possibility of confounding; for example, if study site predicts education level and also predicts cardiovascular disease, the crude association would not justify the causal interpretation of predicting the outcome of an intervention contrast. By fitting the multivariable model, the authors are clearly targeting something that is only ever directly observed in crossover trials. And value variation in the exposure, but consistency is defined exposure. This is not enough to guarantee that the outcome Y when exposure X is passively observed to be at level x is equal to the potential outcome Y[SET[X = x]] that arises under intervention. This might seem trivial at first, but—in fact—it is an assumption that is fundamental to all causal inference methods, and it is easy to violate in practice. A closely related statement is Rubin’s Stable Unit Treatment Value Assumption (SUTVA), which requires that there be no variation in the exposure, but consistency is defined to encompass the method of assigning X to its specific value x more explicitly.

Consider the example of adiposity as an exposure and cardiovascular disease as the outcome. Suppose we agree on a very specific definition of adiposity, based on a measure like body mass index (BMI) or percent body fat, and hence we have a well-defined exposure. This is not enough to guarantee that Y[SET[X = x]] = Y[SET[X = x]]. For example, suppose that one person is manipulated to BMI = 20 by a regimen of caloric restriction, and another person is manipulated to BMI = 20 by a regimen of vigorous physical activity. The intervention forces the exposure to take exactly the same precisely measured numerical value, but the outcome may be different because of the mechanism by which the exposure was only aspire to vaguely distinguish a hazard ratio of 1 from a hazard ratio of 10, then, of course we can be more relaxed about these treatment details. But note that The Lancet authors reported an estimate to the second decimal place, suggesting that the exposure in this case, as well as all the covariates and outcomes, are assumed to be very precisely defined and measured quantities.
manipulated. That is to say, arriving at $BMI = 20$ via reducing the flow in of calories may have completely different effects on cardiovascular health compared with arriving at $BMI = 20$ via increasing the flow out of calories. A trial protocol would have to specify the mechanism of intervention, and so to meet the consistency assumption, and therefore an observational study must do the same.

For causal inference about socioeconomic exposures, the consistency assumption is perhaps the most difficult to justify in practice. Take the example of years of educational attainment, the most commonly applied measure of social position. It is quite obvious that there is variation in the meaning of this measure over time, as a high school diploma now is much less selective a credential than it was 50 years ago. There are also administrative variations; for example, Québec delivers a secondary school diploma after 11 years and Ontario after 12 years. And there are variations in quality from school to school, even classroom to classroom, and “achieving” a year of education can encompass any of a range of intensities of experience from barely scraping by to being top of the class. Moreover, how would one assign a participant to receive an additional year of schooling? You can lead a horse to water, but you cannot make it drink. Ethics aside, it is easy to imagine a protocol of denying education to someone who wants it, but assigning education to those who elected to drop out is difficult to operationalize in any realistic or meaningful way. For example, truancy laws can force a student to be physically present in the school, but cannot force the student to derive some benefit from that experience. At the same time, that one must be concerned about the question of exposure variation, one must also use reasonable substantive judgements. If one participant is educated in a red schoolhouse, and another is educated in a blue schoolhouse, this is treatment variation that is likely to be completely ignorable. The challenge is to make the protocol for the hypothetical trial specific enough to capture the variation that matters but coarse enough to ignore the variation that does not matter.

A productive approach to avoiding violations of the consistency assumption in research on the social determinants of CVD is to focus on concrete policies that modify the exposure. Continuing with the example of attainment of education, some obvious candidates are compulsory schooling laws and grants for financial assistance of needy students. Such policy changes can be mobilized as “instrumental variables” that effectively randomize participants into additional educational attainment, although this requires additional assumptions of homogeneity or else making only local inference (i.e., changing the target population). Considering other indicators of social position, income poses many of the same consistency assumption dilemmas. The ways one achieves income (labor, inheritance, etc.) may have different implications for health. As with education, an advantageous strategy would similarly be to find policies that affect income such as changes to tax law or establishment of aid programs. Different ways of achieving education or income may still have heterogeneous effects on health, but there are several advantages to selecting policies as the exposures. The first is that the effect can be described more specifically in the paper as the effect of the policy, not of observed income or education more broadly. Second, the protocol for enacting a public policy will be concrete and detailed, unlike the pathways taken by diverse individuals to their attained status. Finally, an exogenous policy change (i.e., uncorrelated with characteristics of the participants) will also address the assumption of exchangeability, described in the following section.

### Causal Inference Considerations: The Exchangeability Assumption

Another crucial causal inference assumption is “exchangeability,” which is more intuitive. It is simply the assumption that there is no confounding in expectation: the same scenario that we would encounter in a randomized trial. To see why a randomized trial is the gold standard for causal inference, one must simply recall that the causal effect is the contrast between 2 potential outcomes: $Y|SET[X = x_1]$ vs $Y|SET[X = x_2]$, in which $Y$ is the outcome. “$vs$” is any mathematical contrast, such as a difference or a ratio, and $x_1$ and $x_2$ are 2 selected values of the exposure, such as active medication and placebo. This is not necessarily the same as the observed contrast, $Y[X = x_1]$ vs $Y[X = x_2]$, which are just the mean $Y$ values among those found through passive observation to have exposure values of $x_1$ and $x_2$. How does this difference arise? The problem is the presence of any third variable, a confounder, which is correlated with the exposure and is a cause of the outcome $Y$ (or a proxy for a cause of $Y$). For example, suppose that we look in an administrative dataset and find that people taking angiotensin converting enzyme (ACE) inhibitors are more likely to die of CVD. Is this evidence that ACE inhibitors cause cardiovascular mortality? Of course not, because the exposure and the outcome share a common cause: elevated blood pressure. Therefore, the observed correlation between exposure and outcome can arise simply because of this shared influence, represented by arrows leading into these nodes in Figure 1.

We know that this is not a causal effect because if we randomly assign ACE inhibitors in a trial, we break the normal mechanism of ACE inhibitors being assigned by indications such as high blood pressure and replace this with a mechanism in which they are assigned only by a coin flip, as shown in Figure 2.

![Figure 1](https://example.com/figure1.png)

The absence of any pathway in this diagram between exposure and outcome, except for the direct causal pathway, means that association equals causation in this setting. The goal in an observational study is therefore to achieve a structural scenario like this in which the exposure has no common sources of variation with the outcome. This can occur by design, such as in natural experiments or by statistical modelling if all common causes can be measured and adjusted. The assumption of “exchangeability” that is necessary for causal inference is simply that this independence of mechanisms for setting the exposure and the outcome has been achieved. The term derives from the fact that in a randomized trial, the baseline characteristics of the exposed and the unexposed groups, with respect to all arrows entering into the exposure node, will be equal in expectation. Therefore, the 2 groups are “exchangeable” with one another in terms of these characteristics that could be confounders if they were instead imbalanced.

In studies of the causal effects of social determinants of health, such as education and income, meeting the exchangeability assumption is a serious challenge. These
exposures have complex and historical causes that are difficult to enumerate and measure and are frequently missing from health databases. This makes the prospect for achieving exchangeability via statistical control in regression modelling very limited. Instead, some more realistic progress can be made with the same advice given here in relation to the consistency assumption, which is to seek out natural experiments that provide exogenous shocks to the exposure, distinct from the usual mechanisms by which people achieve their educations and incomes. Policy changes provide such opportunities, as they have an impact on social determinants while being plausibly unrelated to individual health outcomes through any common causes. Investigators can also refine their questions to focus on sources of variation that do not derive from complex historical processes: for example, by looking not at income but at changes in income over time.44 This pivot to looking at the effect of change instead of the effect of status also allows for the comparator to be the same person at a different time, instead of a different person at the same time, which balances all time-fixed characteristics such as genetics and demographics.45 As described here in relation to the consistency assumption, this approach may compromise generalizability and precision and may require a change in the target population, but when threats to validity arise primarily from lack of exchangeability, this can be a very powerful strategy. It has the additional advantages of being closer to the kind of protocol that could be implemented in a trial and closer to the kind of changes that might be of interest to policymakers.

In one recent example of this kind of study, Hamad and colleagues46 took advantage of variation in American state-level compulsory schooling laws as a natural experiment that generated differences in the minimum required years of required schooling by state and year. Using the policy change as an "instrumental variable," such as the coin flip in a randomized trial, they found that increased educational attainment was a cause of reduced smoking, improved lipids, and reduced incidence of heart disease. In a similar vein, Carter et al.47 used random inheritance of alleles that increase cardiovascular risk and showed that these effects were modified by educational status. The authors’ premise is that genetic effects are not confounded under the assumption of random assortment of genetic variants during meiosis, but even if this assumption is satisfied, the results only pertain to that fraction of the variation that is genetic not to social and behavioural causes of cardiovascular disease that might be correlated with educational level through complex sociohistorical pathways.
Causal Inference Considerations: The Positivity Assumption

The positivity assumption is a practical necessity in causal analysis and is also relatively intuitive. This assumption is merely that there must be data to support the hypothetical contrast on which the causal estimate is constructed. For example, suppose that we are estimating the effect of a medication on incidence of hyperglycemia, but the medication assignment is not randomized, and so must be adjusted for all variables that could plausibly be common causes of the therapy and the outcome. Pregnancy is clearly a cause of incident hyperglycemia, and so we specify it as a potential confounder that requires adjustment. The problem, however, is that this medication carries a counter-indication that it may not be administered during pregnancy. The result of this prohibition is that there are no observed cases of a pregnant woman being assigned the treatment. Statistical control for a covariate implies that one estimates the exposure contrast within each stratum of the covariates and then combines these stratum-specific estimates into a summary measure that is conditional on the covariate. But when we try to do this with an empty cell, the procedure will fail. Thus, the adjustment fails because the positivity assumption was violated. When conducting a propensity score analysis, and covariates are all combined into a single score, the lack of overlap over some range of scores also represents a local positivity violation, and in that context is referred to as a lack of “common support.”

The previous example was a deterministic violation of positivity, as treatment was prohibited in one stratum, but empty cells can also arise stochastically. These random violations arise when at some level of the exposure, outcome or confounders, nobody happens to be observed, even though, in principle someone could be. This is a common problem when data are sparse or the number of covariates and interactions is large. These random violations can often be resolved by enlarging categories, although at some risk of residual confounding. One can also interpolate, if there are populated cells on either side of the empty cell. A solution to all positivity violations is to change the target population. For example, in the medication example here, we could simply redefine the effect as that in women who are not pregnant. This is the approach most commonly applied in the propensity score setting, when areas of nonoverlap are discarded, maintaining common support for causal contrasts but over a more narrowly defined population.

As for the previously discussed biases, positivity assumptions can pose particularly acute challenges for causal studies of the effects of socioeconomic exposures on cardiovascular disease. The reason is that structured social relations will often lead social indicators to be highly correlated, making mutual adjustment impossible because of the resulting sparsity. When this is caused by linked social processes, Oakes has coined the term “structural confounding.” The classic example is described by Messer et al., who sought to study the independent effects of socioeconomic status and race on preterm delivery in North Carolina. Cognizant of theories of “intersectionality,” which posit that different social groups experience socioeconomic variation in distinct ways, the authors conditioned simultaneously on both race and poverty at the neighbourhood level. Upon close inspection, however, they found that in their dataset there were almost no poor White neighbourhoods and almost no rich Black neighbourhoods, making this mutual adjustment conceptually impossible. This was not a chance event or a defect in data collection, but rather an enduring structural feature of racial residential segregation in the American South. The authors had to look at the tabular data to appreciate this violation, however, as a regression analysis would simply fit lines through empty space and provide (meaningless) parameter estimates. This is also why Rosenbaum and Rubin first recommended checking overlap in propensity score models, as regression can be so deceptive in smoothing over these regions of empty space and providing estimates based on statistical assumptions even when there are few or no data to support them. In research on social determinants of health, however, “intersectionality” is likely to be a persistent consideration, suggesting the benefit of making some compromise between rigidly parametric regression models and nonparametric cross classifications. For example, when some cross-classified cells become sparse, “borrowing strength” from adjacent cells or the overall mean can help to balance between concerns of precision and validity.

Causal Inference Considerations: Selection Bias

Traditionally, selection biases refer to situations when the dataset used for the analysis changes in ways that distort the causal estimates. This is less intuitive to many researchers because it requires clarity on the target population for inference as distinct from the study population and the subset of the study population that is actually available in a given analysis. For example, the target population might be all adult Canadians, the study population might be the participants in a given cohort study, and the analytic sample may exclude some people with missing values or who are excluded or lost to follow-up. This framework gives rise to notions of external validity (correct causal inference for all Canadians) as distinct from internal validity (correct causal inference for people in the analysis, even if they are very different from all Canadians). One common mechanism by which selection bias occurs is when, through the design of the study or the analysis of data, there is conditioning on a common consequence of the exposure and the outcome, which is referred to as a “collider.” Any conditioning on a “collider”—for example, by conducting the analysis only among people diagnosed with CVD or by adjusting a regression model for having CVD—can generate a bias, and this happens frequently in practice. Figure 3 considers a study on the relationship between treating a patient with ACE inhibition (ACEI) therapy and heart disease. Income is a confounder because it affects both treatment and outcome and therefore must be adjusted so that the estimated causal effect is not biased. Next, consider an investigator with a focus on social determinants of health and who therefore conducts a study on the relation between income and receipt of ACEI therapy. Heart disease is affected by both the exposure (income) and the outcome (ACEI therapy), and so it is a “collider” and must not be adjusted for or stratified on. Researchers would be inadvertently conditioning on heart disease: for example, if they conducted their study in a clinical setting among diagnosed patients. Such a study would be biased in the sense that the observed
association between income and ACEI therapy would not have a causal interpretation, even in the absence of confounding.

A striking example of how severe this bias can become in practice is found in a paper by Canto et al.\(^60\) that appeared in *JAMA*. These authors examined the relationship among 5 major coronary heart disease risk factors (hypertension, smoking, dyslipidemia, diabetes, and family history) with in-hospital mortality among one-half million patients with first myocardial infarction but without previous CVD. The severe selection comes from the fact that patients with previous heart disease were excluded, so if hypertension or smoking or other risk factors had already caused disease, those observations were eliminated. In addition, patients had to be hospitalized to experience in-hospital mortality, so those who did not survive to admission were also dropped. These exclusions eliminated more than three-quarters of the cohort, leading to such severe selection bias that the traditional risk factors were observed to be protective against mortality. The authors did not appear to be aware of selection bias as a threat to validity and interpreted their study as showing a truly protective effect of these traditional risk factors on in-hospital mortality. A similarly catastrophic misinterpretation of selection bias endures in the "obesity paradox" literature, in which many researchers interpret obesity as having protective effects because of an analogous conceptual error.\(^61\)

Selection bias is clearly a threat to validity in studies of social exposures and cardiovascular outcomes, although there seems no reason to suspect that it would be worse than in the wider biomedical literature as a whole. Indeed, because clinical recruitment represents a particularly dangerous form of conditioning, as one must be diagnosed and treated in order to be enrolled, population-based studies may be less severely affected. One still has to worry, however, about differential recruitment into cohorts\(^62\) as well as the potential bias associated with dropping prevalent cases of disease at cohort initiation if these are not small numbers.\(^63\) There is also the risk that these selection processes might operate differentially across social categories, creating spurious heterogeneity in the estimated exposure effects.\(^64\)

**Figure 3.** (A) Angiotensin converting enzyme (ACE)-inhibition therapy is the exposure and heart disease is the outcome. Income is a confounder and must be adjusted for to eliminate confounding. The effect of ACE-inhibition therapy adjusted for income has a causal interpretation. (B) Income is the exposure, and ACE-inhibition therapy is the outcome, and heart disease is a “collider” (as arrows from exposure and outcome “collide” at that node). The effect of income on ACE-inhibition therapy adjusted for heart disease does not have a causal interpretation. The 2 panels have the exact same structure but differ in which variable is the exposure and which is the covariate.

### Additional Causal Inference and Statistical Considerations

#### Measurement error

The 3 main structural biases in modern epidemiologic theory are often taken to be confounding bias, selection bias, and information bias. The last of these involves measurement error and misclassification, which can occur to exposures, outcomes, or covariates: each with different potential consequences.\(^65\) Moreover, things can get worse if misclassification of exposure is differential by outcome status or vice versa. Obviously, validity rests on assumptions of accurate measurement, and there is a wide assortment of techniques for diagnosing and correcting biases associated with such errors, although, in practice, there is often also a lack of conceptual clarity in the measure. For example, one might use BMI as a measure of obesity, but does it suffer from measurement error? To answer this question, one must have a "gold standard" assessment of obesity, and this is elusive. Many exposures and diseases commonly modelled in epidemiologic research similarly lack clear gold standard anchors with which to assess and correct for information biases. This is nowhere more true than in the study of social determinants of health, in which very crude proxies, such as education and income, are meant to stand in for broader social processes. As noted in the section on the consistency assumption, years of attained education is not even an especially good measure of education, and how far this is from an accurate measure of overall social position is impossible to judge with no accepted gold standard. This is why serious sociologic knowledge is required for the careful study of social determinants of health, but, even so, it is difficult to achieve consensus about what underlying constructs are being proxied by simple social indicators, and how one can begin to address information bias in relation to the distance between indicator and theoretical gold standard.\(^66\)

#### Model misspecification

Causal effects are inherently counterfactual, as they invoke contrasts among outcomes that would be observed under
different exposure settings in the same person. Because no counterfactuals can be directly observed in the data, they must always be modelled statistically. These models must meet causal assumptions of consistency and exchangeability, as described earlier, but, in addition, statistical models invoke other assumptions of form, scale, and error distribution. For example, a linear regression model operates under the assumption that all conditional means of the outcome fall on a straight line. If this is not true in the real world, this assumption can be relaxed with addition of polynomial terms or other flexible devices. This requires careful attention on the part of the modeller to all the assumptions and their plausibility in each application. Modelled contrasts also fall on a specific scale, which has implications for the interpretation of joint effects. For example, 2 terms in a logistic model interact multiplicatively, so that the odds ratio for the jointly exposed is the product of the 2 odds ratios. Again, this can be relaxed, but analysts must check and confirm such assumptions and make modifications when they are violated. For the study of social determinants of health on cardiovascular disease, the same concerns apply: neither more nor less important than for other research programs. There is a strangely pervasive habit across all of biomedicine of reporting disparities on the ratio scale instead of the absolute scale, which has implications for heterogeneity assumptions and its interpretation of social inequalities. If one views absolute inequalities as more salient in relation to some theory of social justice, such models could be considered to be misspecified.

Immortal time or ambiguities about time-0

Causal inference assumes exchangeability so that an analogy can be made to the covariate balance achieved in a randomized trial. But another crucial feature of a randomized trial is an unambiguous “time-0” at which point the treatment is initiated. This distinction is necessary because pretreatment covariates must be balanced but not post-treatment covariates. Moreover, outcome events that occur before treatment is initiated cannot logically be attributable to the treatment. This whole trial framework is the bedrock of formal causal inference, but it is very tricky to implement for social determinants of health, which tend not to have a clear time-0 for initiation. Income, for example, exerts effects even before birth and can then be volatile over the life course and can be affected by health status, leading to problems of “reverse causation,” in which the causal connection between income and cardiovascular health can run in both directions. The various causal assumptions are related to one another, so one could also view the need for a time-0 as necessary to meet the consistency assumption described here. For example, the counterfactual exposure status for a rich person being poor might have to extend before birth to the social conditions of parents, given intergenerational transmission of privilege. Again, this concern speaks to the advantage of considering policy changes as social treatments, as they are discrete in time, and their timing is independent of participant covariates. For studies of individual social attributes, the time-0 concern can also be minimized by requiring a “target trial” framework, in which a protocol specifies exactly what is done to the treatment groups and when.

Interference. An often neglected component of Rubin’s SUTVA assumption is that one subject’s exposure must not affect another subject’s outcome. Such an effect is termed “interference” and is a typical feature of contagious processes. For example, a vaccine given to a husband lowers the risk of infection for his wife, even though she did not receive the vaccine herself. Interference is not a typical concern in clinical interventions such as medications or surgeries, as cardiovascular treatments do not typically cause or prevent disease in other people, although an argument can be made for exposures such as second-hand smoke and for familial influences on health behaviours. Some researchers have noted that obesity seems to be transmitted as a form of contagion, through conformity of energy intake and expenditure. Social position is also clearly contagious through the sharing of material resources and the generation of privileged connections and opportunities. One approach to minimizing this problem in studies of social determinants of CVD involves aggregating social position at the family level, especially using family income instead individual income. Restricting to 1 observation per household can also be helpful in avoiding clustering or correlated outcomes because of shared environments and endogenous behavioural patterns.

Table 2 fallacy. As noted by Westreich and Greenland, it is common for authors to provide modelled parameter estimates for covariates and then to interpret these causally. This is erroneous because the study is designed with respect to an exposure or treatment, and the model is usually specified correctly only in relation to that selected variable (as in Fig. 3). For example, suppose that point-time treatment A and covariate B are both included in a regression model and that this model meets the causal assumptions necessary to interpret estimated effects of A causally. For example, to be a valid confounder of the effect of variable A, it is assumed that B

![Figure 4](image-url)
occurs temporally before A. In that case, the estimate for B conditional on A cannot have a causal interpretation because it would be adjusted in the model for a variable that is a consequence, rather an antecedent.

**Time-dependent confounding**

The discussion here is premised on time-point exposures, but when treatment is sustained over time, it can affect covariates that become predictors or modifiers of subsequent doses of treatment.9 This sets up the scenario of time-dependent confounding, which requires special causal inference models such as marginal structural models or the parametric g-formula.80 Standard tabular or regression models cannot provide valid inference in this situation. For example, suppose we are studying the effect of income over the life course on cardiovascular mortality. What do we do with a covariate such as incident CVD, as shown in Figure 4? We want the causal effect of income, and from Figure 4 we see that CVD is a cause of the outcome and of income at time t + 1 and therefore must be adjusted for. But it is a consequence of income at time t and therefore must not be adjusted for.

This structure leads to a situation in which traditional models cannot provide valid inference, and through the 1980s, this problem remained unsolved in epidemiologic research, until the development of the so-called g-methods (in which “g” stands for “general”) proposed by Robins.39 Given that income and other social determinants of health are generally dynamic over the life course, these methods are frequently needed, unless one elects to focus on the causal effect of a point-time shock such as implementation of a policy. Some measures, such as attained years of education, are stable over time and therefore also immune to these concerns. But for valid inference regarding treatments that can change over time, these more advanced models can become essential.

There are 3 main implementations of g-methods, each of which solves this problem of time-dependent confounding in a slightly different way. The first and most common approach is to use marginal structural models estimated with inverse probability of treatment weighting.42 The weights are functions of the predicted probability of exposure and therefore are analogous to propensity scores.85 When the study population is reweighted in this way, a pseudopopulation is generated that has the characteristic that covariates and exposure are uncorrelated. This implies that there is no longer a problem of time-dependent confounding in the pseudopopulation, and so the crude association between exposure and outcome in this pseudopopulation will have a causal interpretation. The second approach is to use the parametric g-formula, which is simply a structural outcome model for the sequential data. This structural, parametric model is then used to predict the outcome under each hypothetical intervention while standardizing sequentially for the confounders.84 This approach raises concerns about model specification, and so “doubly robust” methods have emerged that essentially combine these first 2 approaches: 1 that models the exposure, and the other that models the outcome. A combined estimator is constructed so that the user has 2 ways to obtain valid inference: either by getting the exposure model right or getting the outcome model right.85 The third approach, g-estimation, is much less common in the literature. Like marginal structural models, g-estimation is based on fitting a model for the exposure, but it necessitates a search for parameters that satisfy an independence condition and therefore is often more cumbersome to apply in practice.

**Additional assumptions for mediation**

The discussion here focuses on estimation of the total average causal effect, but there are many settings in which investigators seek instead a path-specific effect such as the controlled direct effect or the natural direct and indirect effects.32 These are causal effect parameters that refer to hypothetical interventions on both a treatment and some post-treatment variable or variables that distinguish among several causal paths between the exposure and the outcome. These mediation models are now standard tools in the epidemiologic toolkit, and many papers devoted to social determinants and cardiovascular outcomes now use this framework. Although the details of these models and their extensions is outside the scope of the current review, suffice it to say that more assumptions, in addition to those listed here, are generally required. For example, the single exchangeability assumption discussed earlier is replaced by 2 noconfounding assumptions in the case of controlled effects, and 3 noconfounding assumptions in the case of natural controlled effects.38 Some of these mediation models also invoke novel assumptions beyond those discussed here, such as the “cross-worlds assumption” and “no recanting witness” assumptions of the natural-effects models.39 Some of these assumptions remain controversial, with doubts over whether they can ever realistically be satisfied, but such doubts have not hindered the rapid proliferation of these techniques.

**Discussion and Conclusions**

The study of social determinants of cardiovascular disease is an important research program because of the dramatic impact of these factors on incidence and mortality. Nonetheless, the assumptions necessary to treat these exposures as causal treatments pose a significant challenge in design and analysis. Many social determinants, such as race and political ideology, are almost impossible to express as manipulable treatments.50 Others, such as education and income, are responsive to interventions, and therefore—in principle—can be studied causally as treatments.51 Nonetheless, the challenges to doing so fruitfully are formidable and require careful attention in design and analysis. In relation to several of the assumptions, and especially consistency and exchangeability, it was highlighted earlier that discrete policy changes offer substantial advantages over using the observed social status of individuals measured at 1 point in time. In addition to policy innovations, there are exogenous shocks that occur as natural experiments, such as job loss or lottery wins.92 There are also advantages to be found in the use of self-controlled designs, in which 1 person is compared at 2 times, instead of 2 people compared at 1 time.93 Target trial formulations were also recommended, and when data are longitudinal and exposure changes over time, one must also be attentive to the threat of time-dependent confounding.94 Finally, it is worth noting that although causal inference is important, a great volume of the useful social epidemiology of cardiovascular disease is
descriptive, showing trends and burdens, but not seeking to specify quantitative treatment effects.\(^9\)

**Ethics Statement**

This material is the authors’ own original work, which has not been previously published elsewhere. The discussion is appropriately placed in the context of prior and existing research. All sources used are properly disclosed and correctly cited. The author takes public responsibility for the content of this review.

**Patient Consent**

The author confirms that patient consent is not applicable to this article, as it is a review article that is conceptual and therefore does not rely on any original data analysis. No institutional review board approval was sought for the writing of this review.

**Funding Sources**

No funding was provided for this article.

**Disclosures**

The author has no conflicts of interest to disclose.

**References**


62. Huang JY. Representativeness is not representative: addressing major inferential threats in the UK Biobank and other big data repositories. Epidemiology 2021;32:189-93.


64. Reeves A, Elliott MR, Lewis TT, Karvonen-Gutiérrez CA, Herman WH, Harlow SD. Study selection bias and racial or ethnic disparities in estimated age at onset of cardiometabolic disease among midlife women in the US. JAMA New Open 2022;5:e2240665.


