A Graphical Catalog of Threats to Validity

Linking Social Science with Epidemiology

Ellicott C. Matthay and M. Maria Glymour

Abstract: Directed acyclic graphs (DAGs), a prominent tool for expressing assumptions in epidemiologic research, are most useful when the hypothetical data generating structure is correctly encoded. Understanding a study's data generating structure and translating that data structure into a DAG can be challenging, but these skills are often glossed over in training. Campbell and Stanley's framework for causal inference has been extraordinarily influential in social science training programs but has received less attention in epidemiology. Their work, along with subsequent revisions and enhancements based on practical experience conducting empirical studies, presents a catalog of 37 threats to validity describing reasons empirical studies may fail to deliver causal effects. We interpret most of these threats to study validity as suggestions for common causal structures. Threats are organized into issues of statistical conclusion validity, internal validity, construct validity, or external validity. To assist epidemiologists in drawing the correct DAG for their application, we map the correspondence between threats to validity and epidemiologic concepts that can be represented with DAGs. Representing these threats as DAGs makes them amenable to formal analysis with d-separation rules and breaks down cross-disciplinary language barriers in communicating methodologic issues.

Keywords: Causal inference; Directed acyclic graph; Epidemiologic methods; Quasi-experimental designs; The Campbell tradition; Threats to validity

(*Epidemiology* 2020;31: 376–384)

Directed acyclic graphs (DAGs) have rapidly gained popularity among epidemiologists. DAGs are a particularly useful tool for causal inference because, when one has correctly specified the hypothetical DAG to encode prior knowledge of the data generating structure, a set of mathematics-based rules can be applied to determine whether and how the causal effect of interest can be identified—for example, which confounders need to be controlled. However, the DAG framework is entirely silent on what the causal model should look like, and epidemiologists often struggle to tie their methodologic training in DAGs to specific, real-world settings.

Nearly 50 years ago, Campbell and Stanley[1] presented a catalog of threats to validity, describing reasons empirical studies may fail to deliver causal effects. Their work, along with subsequent revisions[2,3] (hereafter, "the Campbell tradition"), has been extraordinarily influential in social science training programs,[4,5] but has received less attention in epidemiology. The framework delineates four types of validity—internal, statistical conclusion, construct, and external (Box 1). The Campbell tradition guides researchers to assess alternative explanations for an association besides the causal relation of interest ("threats to validity") when evaluating evidence from a specific study design and analysis, and to incorporate design or analysis features—for example, randomization or masking—to diminish the influence of such threats. The Campbell tradition's most recent set of 37 threats to validity are based on decades of empirical research experience. They are verbal descriptions of biases that can arise in epidemiologic research. Most can be considered possible causal structures and represented in a DAG. Epidemiologic research and,

BOX 1. SHADISH, COOK, AND CAMPBELL'S VALIDITY TYPOLOGY

1. Statistical conclusion validity: appropriate use of statistical methods to assess the relationships among study variables;
2. Internal validity: the extent to which the estimated association in the study sample corresponds to a causal effect from exposure to outcome;
3. Construct validity: the extent to which measured variables capture the concepts the investigator intends to assess with those measures; and
4. External validity: the extent to which study results can be generalized to other units, treatments, observations made on units, and settings of study conduct.

Submitted April 17, 2019; accepted January 15, 2020.
From the Center for Health and Community, University of California, San Francisco, CA.
Supported by Evidence for Action program of the Robert Wood Johnson Foundation.
The authors report no conflicts of interest.
SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: M. Maria Glymour, 550 16th Street, 2nd floor, University of California, San Francisco, San Francisco, CA 94143. E-mail: Maria.glymour@ucsf.edu.

ISSN: 1044-3983/20/3103-0376
DOI: 10.1097/EDE.0000000000001161

in particular, the challenging task of correctly drawing one's DAG, could be enhanced by considering these threats. However, to our knowledge, no comprehensive crosswalk between the Campbell tradition's named biases and DAGs has been published. The DAG for threats deemed plausible in the given study context can be incorporated into the investigator's DAG to provide insight into the problem and to determine effective analytic solutions.

We define the Campbell tradition's named threats to validity. For each threat, we provide the epidemiologic analog, a corresponding DAG, and one or more examples, to illustrate how they might inform the epidemiologist's DAGs. We aim to enhance epidemiologic research by facilitating the use of cross-disciplinary concepts to inform DAG specification and to facilitate cross-disciplinary communication by using DAGs as common language to understand biases in causal research.

REPRESENTING THREATS TO VALIDITY AS DIRECTED ACYCLIC GRAPHS

DAGs are causal models that visually represent background knowledge and assumptions about the relationships between variables.[6] They encode the hypothesized data generating mechanisms and can include features of the study design such as instruments and study implementation such as measurement. DAGs are interpreted with mathematics-based rules that provide a flexible but rigorous method for determining sets of variables that, when measured and adjusted appropriately, are sufficient to control confounding and identify causal effects. Box 2 presents an introduction to key concepts of DAGs and notation used in this article. For a more detailed introduction, we refer the reader to Glymour and Greenland.[7]

For each threat below, the DAG provided is either the archetypal causal structure, or one or more examples of possible causal structures. For threats that may arise with or without a direct causal effect from exposure to outcome (e.g., under the null), we generally exclude the directed edge from exposure to outcome. A directed edge from the exposure to the outcome is included when the threat is only applicable when there is a direct causal effect of the exposure on the outcome. Threats that are redundant, arise less frequently in epidemiology, or are less amenable to DAG representation are presented in the eAppendix; http://links.lww.com/EDE/B634. No human subjects were involved in this research.

BOX 2. INTRODUCTION TO KEY CONCEPTS OF DIRECTED ACYCLIC GRAPHS

DAGs are comprised of variables ("nodes") and directed arrows ("edges") that indicate potential direct causal effects of one node on another. A "path" is a sequence of nodes following edges, not necessarily in the indicated direction, from one node in the graph to another. DAGs are "acyclic," meaning no directed path leads back to the same node. Direct and indirect effects of a node are referred to as its "descendants." A "backdoor path" is a path connecting the outcome and exposure but with an arrow pointing into the exposure. "Colliders" on a path are nodes with at least two directed arrows pointing into them from other nodes on the path. Paths are "blocked" by conditioning on a proposed covariate set if either (1) one or more nodes on the path are in the covariate set or (2) the path contains at least one collider and neither the collider nor any of its descendants are in the covariate set. The "backdoor criterion" can be used to identify the necessary set of variables that must be appropriately measured and controlled for unbiased estimation of the causal effect of interest. It states that a set of variables is sufficient to control confounding if (1) no variables in the set are a consequence of the treatment and (2) conditioning on the set of variables blocks all backdoor paths from the outcome to the treatment.

Example Directed Acyclic Graph:

Notation: DAGs include the following variables: E, the exposure or treatment received; D, the outcome; U, unmeasured confounders, variation, or error, with subscripts referring to the variables they affect; T, treatment assignment (i.e., in the context of randomization, where assigned treatment may not equal actual exposure); S, selection into treatment or into the study; and M, a mediator. Numerical subscripts indicate time of measurement or multiple components of the corresponding variable. The subscript "m" indicates a (possibly incorrect) measurement of the corresponding variable (e.g., we distinguish between depression as a latent construct and depression as measured, perhaps using a scale of depressive symptoms).

THREATS TO INTERNAL VALIDITY

Threats to internal validity are the central concern of most causal analyses, with violations generally corresponding to confounding or failure to meet the backdoor criterion.[7] Confounding (often referred to as "selection into treatment" or merely "selection" among social scientists) constitutes a core threat, but eight others are also delineated and their definitions are presented in Table 1. Several of these threats are only relevant, or most commonly relevant, in extremely weak study designs lacking a contemporaneous control group. In these cases, the Campbell tradition's accounting of reasons these designs are rarely valid helps provide more critical insight into stronger designs. Specifically, history, maturation, regression, testing, and instrumentation are particularly relevant to certain pre–post designs with no comparison. In the DAGs below, this is reflected by backdoor paths involving a time node. The exposure is determined by time (pre vs. post), and time also affects other factors. The result is bias that could be controlled by including a comparable, contemporaneous unexposed group.

TABLE 1. Threats to Internal Validity

Threat No.	Threat Name	Definition
1	Ambiguous temporal precedence	Lack of clarity about which variable occurred first may yield confusion about which variable is the cause and which is the effect.
2	Selection	Systematic differences in respondent characteristics that also affect the outcome can cause bias or be confused with a causal effect.
3	History	Events occurring concurrently with exposure could cause the observed outcomes and be confused with an effect of exposure.
4	Maturation	Naturally occurring changes over time could be confused with an effect of exposure.
5	Regression to the mean	When units are selected for their extreme scores, they will often have less extreme scores on other variables and subsequent assessments, an occurrence that can be confused with an effect of exposure.
6	Testing	Having been tested can affect scores on subsequent exposures to that test, an occurrence that can be confused with an effect of exposure.
7	Instrumentation	The nature of a measure may change over time or conditions in a way that could be confused with an effect of exposure.
8	Attrition	Loss of respondents to treatment or to measurement can produce artifactual effects if that loss is systematically correlated with other study variables.
9	Additive and interactive effects of threats to internal validity	The impact of a threat can be added to that of another threat or may depend on the level of another threat.

Definitions are direct quotes or paraphrased from Shadish, Cook, and Campbell.[3]

Threat 1

Ambiguous temporal precedence corresponds to reverse causality in epidemiology (Figure 1A).[8] Ambiguous temporal precedence might arise, for example, when studying the impact of exposure to violence on mental health, because mental disorders may increase an individual's exposure to violence and exposure to violence can cause mental disorders.[9]

Threat 2

Selection is traditional confounding.[10] In its simplest form, this threat can be represented with the DAG in Figure 1B. Selection might occur, for example, when studying the impact of multivitamin consumption on breast cancer, where an association might be explained by predisposition to other healthy behaviors.

Threats 3–4

History can be conceptualized as confounding by concurrent events that are associated with the exposure through their alignment in time (Figure 1C). For example, history may arise when studying the impact of the introduction of a law requiring seatbelt use on subsequent motor vehicle crash injuries, where an association might be explained by a concurrent change in the safety-related design of new motor vehicles (U). Maturation is conceptually and structurally similar: it is confounding by the natural temporal course of the outcome, where the time scale of interest (and confounding pathway) is often age (Figure 1D).[11] For example, maturation might be problematic when studying the impact of an elderly fall prevention program, because an association might be biased by the fact that risk of falls increases with age, irrespective of the program.

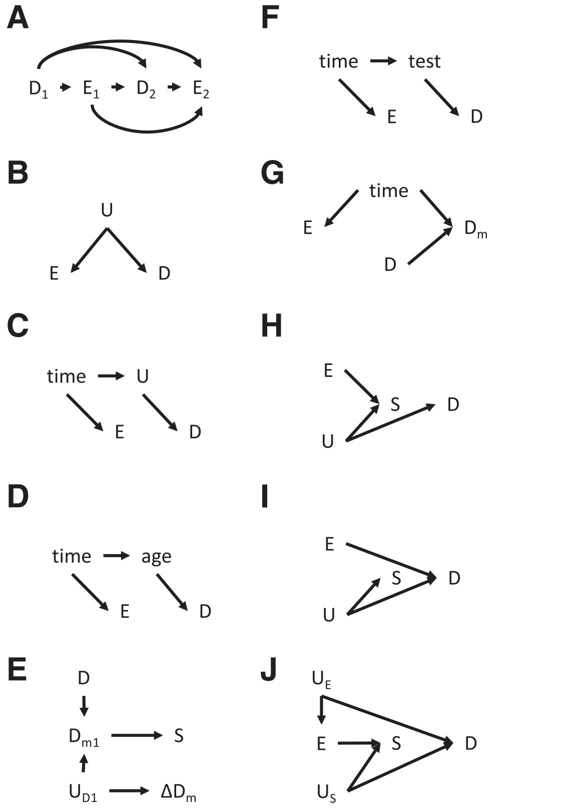

FIGURE 1. Threats to internal validity represented as directed acyclic graphs.

Threat 5

Regression to the mean[12] occurs when participants are selected into treatment based on an extreme measurement of a random variable; less extreme values of the same variable are more likely to be observed in subsequent assessments. This threat is common in situations where participants are treated for their extreme baseline values of the outcome condition. For example, participants might be selected into a study and subsequently treated based on their high baseline blood pressure levels, which would likely decrease with time irrespective of intervention.

In Figure 1E, the measured blood pressure at baseline (D_{m1}) is a combination of true or underlying blood pressure levels (D) and acute variation or random measurement error (U_{D1}). Selection into treatment (S) is determined by whether measured blood pressure at the first time point (D_{m1}) is high. The measured outcome, change in blood pressure from time 1 to time 2 $\left(\Delta D_m = D_{m2} - D_{m1} = \left(D + U_{D2}\right) - \left(D + U_{D1}\right) = U_{D2} - U_{D1}\right)$ is presumed to be a response to treatment. D_{m1} is a collider and S is a descendant of this collider; thus, conditioning on S by studying only subjects with high measured blood pressure at baseline induces a spurious negative association between U_{D1} and D. Among people selected into the study due to high D_{m1}, any whose actual blood pressure was low must had a very high U_{D1}; because they had a high U_{D1}, they are expected to have a low ΔD_m. There would be an association between high overall blood pressure (D) and lower follow-up measures (ΔD_m, U_{D2}), even in the absence of a treatment effect.

Threat 6

Testing can threaten validity when the act of measuring the outcome occurs simultaneous with treatment and affects the measured outcome, such as when weighing a person motivates them to lose weight, irrespective of any weight loss intervention being delivered (Figure 1F).

Threat 7

Instrumentation can be considered a form of confounding by changes in what an instrument is measuring over time.[13] This might occur, for example, in a study of an education policy's impact on Alzheimer disease, where the diagnostic criteria for Alzheimer disease have changed over time, such that a measured change in risk might be incorrectly attributed to the policy change (Figure 1G).

Threat 8

Attrition is a form of loss to follow-up and can lead to bias in two ways. The first is through collider stratification bias—that is, a statistical association induced by conditioning on a collider. Previous work has enumerated a range of DAGs involving collider stratification bias.[14,15] As a simple example, this threat might arise in a study of poverty's impact on mortality, because poverty influences participants' ability to stay in the study, although other unmeasured factors also affecting participation (e.g., underlying health status) also influence mortality (Figure 1H). Restricting to those not lost to follow-up then involves conditioning on collider S, inducing a spurious association between poverty and mortality.

Attrition can also lead to bias when retention in the study is affected by a factor that modifies the exposure-outcome association (Figure 1I). In this scenario, restricting to participants who remain in the study does not bias estimates for the subpopulation who remain in the study, but may provide a biased effect estimate for the baseline population.[15] This issue might arise if poverty causes mortality and another factor (e.g., number of children in the household) impacts both participants' retention in the study and the effect of poverty on mortality. This scenario also arises in situations of selection of susceptibles, in which those most responsive to treatment drop out first. Losses are informative and conditioning on S by restricting to subjects who remain in the study may bias the measured effect for participants overall.[15,16]

When evaluating the bias introduced by attrition, it is essential to be specific about the target population of interest. In the second scenario (Figure 1I), loss to follow-up may result in biased effect estimates for the baseline study population, but produce valid estimates for the population remaining in the study. The scenario in Figure 1I does not cause bias under the sharp null that E does not affect D for anyone. In the case of collider stratification bias (Figure 1H), effect estimates will generally be biased for both the baseline population and the subsample who remain in the study and biased even under the sharp null. Considerations of external validity and specifically the task of identifying the population(s) to whom the results apply helps clarify this issue.

Threat 9

Additive and interactive effects of threats to internal validity refer to multiple biases that may sum together, offset one another, or interact in a single study. For example, in Figure 1J, a cohort study might be biased by both attrition (if the study conditions on S) and confounding (by U_E). These biases may interact because the degree of attrition S depends on the strength of the upstream relationship of U_E to E, and the U_E–E relationship also affects the degree of confounding by U_E. This situation is distinct from additive or interactive effect measure modification.

THREATS TO STATISTICAL CONCLUSION VALIDITY

Threats to statistical conclusion validity (Table 2) generally correspond to failures to conduct appropriate statistical inference in epidemiology. This includes ruling out random error, meeting necessary assumptions of the statistical model (e.g., independent and identically distributed observations on units, no interference or spillover), and correctly specifying the statistical model (e.g., the association between age and the outcome is linear). Most discussions of DAGs assume an infinite sample size and therefore disregard the possibility of chance findings or insufficient power. Additionally, because DAGs are nonparametric, many violated assumptions of statistical tests are not represented as DAGs. Thus, most threats to statistical

TABLE 2. Threats to Statistical Conclusion Validity

Threat No.	Threat Name	Definition
10	Low statistical power	An insufficiently powered experiment may incorrectly conclude that there is no relationship between treatment and outcome.
11	Violated assumptions of statistical tests	Violations of statistical test assumptions can lead to either overestimating or underestimating the size and precision of an effect.
12	Fishing and the error rate problem	Repeated tests for significant relationships, if uncorrected for the number of tests, can artifactually inflate statistical significance.
13	Inaccurate effect size estimation	Some statistical estimation approaches systematically overestimate or underestimate the magnitude of a given causal quantity.
14	Extraneous variance in the experimental setting	Some features of an experimental setting may inflate error, making detection of an effect more difficult.
15	Heterogeneity of units	Increased variability on the outcome variable within conditions increases error variance, making detection of a relationship more difficult.
16	Unreliability of measures	Measurement error weakens the relationship between two variables and strengthens or weakens the relationships among three or more variables.
17	Restriction of range	Reduced range on a variable usually weakens the relationship between it and another variable.
18	Unreliability of treatment implementation	If a treatment that is intended to be implemented in a standardized manner is implemented only partially for some respondents, effects may be underestimated compared with full implementation.

Definitions are direct quotes or paraphrased from Shadish, Cook, and Campbell.[3] "Experiment" in these definitions is used generally to refer to exploring the effects of manipulating a variable, not necessarily to a randomized trial.

conclusion validity are not represented as DAGs (threats 10–15). However, some threats to statistical conclusion validity are situations of measurement error or modifications to measured variables that reduce statistical power, and several of these can be informatively represented as DAGs (threats 16–18). Several threats (low statistical power; violated assumptions of statistical tests; fishing and the error rate problem) refer to null hypothesis significance testing, which is increasingly recognized as problematic practice.[17,18] However, these threats are also relevant to estimation, because they imply that estimates may be imprecise, potentially uninformative, or likely to deviate from the population estimate by chance. We present threats to statistical conclusion validity and corresponding epidemiologic concepts represented as DAGs when relevant in the eAppendix; http://links.lww.com/EDE/B634.

THREATS TO CONSTRUCT VALIDITY

A "construct" is the idea, concept, or theory a researcher intends to capture or measure in a scientific study. Construct validity concerns (Table 3) relate fundamentally to whether study measurements capture the constructs they are intended to capture. This in turn affects the interpretation of results, the attribution of observed effects, and the value of results for guiding future interventions. The tasks of accurate measurement, interpretation, and attribution are essential to make use of results—for example, to replicate relevant features of an intervention. When such threats are recognized, they can be addressed in design or measurement innovations or simply by tempering interpretation of the study's findings.

Several threats described in this section can be conceptualized alternatively as measurement error, confounding, or a consistency violation. Consider the DAG in Figure 2A. Suppose E is completing high school coursework, which affects health outcome D, E_m is having a high school completion credential, and U is passing a general educational development (GED) test. The GED is a US high school credential but does not require the same coursework as a diploma.

If the investigator were interested in the effect of high school credentials on health, then low-construct validity could be conceptualized as a consistency violation. Consistency implies that any variations in conditions leading to the exposure assignment or implementation of the exposure would still result in the same observed outcome.[19] Attempts to replicate the study's findings by intervening on GED tests would be unsuccessful because the resulting changes in credentials would not affect coursework or health.

Alternatively, this issue could be conceptualized as confounding of the E_m–D association, where failure to control for coursework would bias the estimated credentials-health association. Finally, if the investigator were interested in the effect of coursework on health, but what they measured is the credentials-health association, the difference between the true effect and the measured effect could be attributed to measurement error.

Threats 19–21

Inadequate explication of constructs, construct confounding, and confounding constructs with levels of constructs all refer to situations where the named variable (typically the exposure) to which the relationship is attributed does not capture all aspects of the variables that are actually operating to generate the relationship. These threats can be conceptualized as measurement error,[13] confounding, or a violation of consistency (Figure 2A).

Table 3.Threats to Construct Validity

Threat No.	Threat Name	Definition
19	Inadequate explication of constructs	Failure to adequately explicate a construct may lead to incorrect inferences about the causal relationship of interest.
20	Construct confounding	Exposures or treatments usually involve more than one construct, and failure to describe all the constructs may result in incomplete construct inference.
21	Confounding constructs with levels of constructs	Inferences made about the constructs in a study fail to respect the limited range of the construct that was actually studied, i.e., effect estimates are extrapolated beyond the range of the observed data.
22	Mono-operation bias	Any one operationalization (measurement or intervention implementation) of a construct both underrepresents the construct of interest and measures irrelevant constructs, complicating the attribution of observed effects.
23	Mono-method bias	When all operationalizations (measurements or intervention implementations) use the same method (e.g., self-report), that method is part of the construct actually studied.
24	Treatment sensitive factorial structure	The structure of a measure may change as result of treatment. This change may be hidden if the same scoring is always used.
25	Reactive self-report changes	Self-reports can be affected by participant motivation to be in a treatment condition. This motivation may change after assignment is made.
26	Compensatory equalization	When treatment provides desirable goods or services, administrators, staff, or constituents may provide compensatory goods or services to those not receiving treatment. This action must then be included as part of the treatment construct description.
27	Compensatory rivalry	Participants not receiving treatment may be motivated to show they can do as well as those receiving treatment. This action must then be included as part of the treatment construct description.
28	Resentful demoralization	Participants not receiving a desirable treatment may be so resentful or demoralized that they may respond more negatively than otherwise. This response must then be included as part of the treatment construct description.
29	Reactivity to the experimental situation	Participant responses reflect not just treatments and measures but also participants' perceptions of the experimental situation. These perceptions are part of the treatment construct actually tested.
30	Experimenter expectancies	The experimenter can influence participant responses by conveying expectations about desirable responses. These expectations are part of the treatment construct as actually tested.
31	Novelty and disruption effects	Participants may respond unusually well to a novel innovation or unusually poorly to one that disrupts their routine. This response must then be included as part of the treatment construct description.
32	Treatment diffusion	Participants may receive services from a condition to which they were not assigned, making construct descriptions of both conditions more difficult.

Definitions are direct quotes or paraphrased from Shadish, Cook, and Campbell.[3] "Experiment" in these definitions is used generally to refer to exploring the effects of manipulating a variable, not necessarily to a randomized trial.

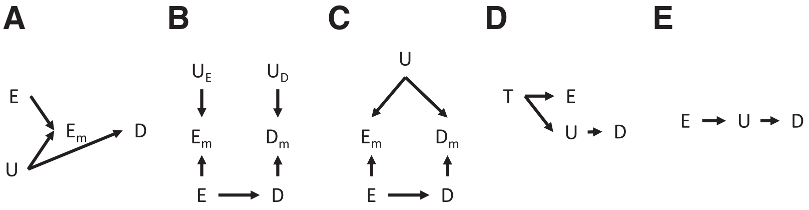

FIGURE 2. Threats to construct validity represented as directed acyclic graphs.

Continuing with the same example of high school credentials and health, construct confounding might occur if coursework co-varied too closely with credentials to be controlled separately. Failure to consider both coursework and passing a GED test as part of credentials would be inadequate explication of constructs. If conclusions drawn about the association then refer to levels beyond the range actually observed (e.g., extrapolation from a study of high school credentials to doctoral degrees), then the threat is confounding constructs with levels of constructs or alternatively, restriction of the range (eFigure 1C; http://links.lww.com/EDE/B634 and eFigure 1D; http://links.lww.com/EDE/B634).

Threats 22–23

Mono-operation bias and mono-method bias most commonly refer to nondifferential or differential measurement error[13] (Figure 2, B and C, respectively), but they can also be conceptualized as confounding or consistency violations (Figure 2A). Concerns about measurement error with respect to construct validity relate to the fact that, unless it can be measured and accounted for, sources of measurement error must be considered as part of the variable to which the association is attributed. For example, exclusive reliance on self-reported exposure to community violence might inadvertently incorporate respondent outlook as part of the exposure (mono-operation bias). If the outcome (e.g., perceived wellbeing) is also self-reported, respondent outlook could also induce a spurious correlation (mono-method bias).

Threat 24

Treatment sensitive factorial structure: see eAppendix; http://links.lww.com/EDE/B634.

Threat 25

Reactive self-report changes: see eAppendix; http://links.lww.com/EDE/B634.

Threats 26–28

Compensatory equalization, compensatory rivalry, and resentful demoralization arise when participants or others respond to treatment assignment in unexpected ways. This is important in unmasked studies because the response may influence the outcome and any outcome differences between treated and untreated may partially reflect the compensatory responses (Figure 2D).[20] For causal questions about the effects of E on D using T as an instrumental variable (e.g., in an randomized controlled trial), this is a threat to the exclusion restriction, and T is no longer a valid instrument for the effect of E on D. This structure does not bias the intent-to-treat estimates of the effects of T; it leads to a serious misinterpretation, however, because a nonzero intent-to-treat estimate does not imply any effect of E on D. These threats might arise, for example, in a study of a weight loss program, where those assigned to the control condition pursue other weight loss services to compensate, or become extra motivated or demotivated to lose weight.

Threats 29–31

Reactivity to the experimental situation, experimenter expectancies, and novelty and disruption effects involve failure to consider a response to exposure as a component of the exposure (Figure 2E). Continuing the weight loss program example, researchers may assume any effects relate to a program feature such as dietary recommendations or physical activity regimen, whereas exposed participants' outcomes may alternatively be affected by knowing they are participating in a weight loss program itself, by investigator's expectations that they will lose weight, or by the novel experience of participating, being part of a weight loss program that interrupts their daily routines. Similar to reactive self-reports changes, this threat is particularly relevant when participants are not masked to exposure.

Failure to include the experimental situation or experimenter expectancies as part of the exposure construct can be considered a consistency violation. This results in a misinterpretation of results as indicating that the program is effective. It is especially problematic because expectancies will likely not be stable in future implementations of the intervention. Alternatively, a measured exposure that incorporates the experimental situation or experimenter expectancies could be considered measurement error in the exposure of interest (i.e., dietary restrictions or physical activity recommendations).

Threat 32

Treatment diffusion: see eAppendix; http://links.lww.com/EDE/B634.

THREATS TO EXTERNAL VALIDITY

External validity concerns relate to the populations and places to which study results can be generalized, and the fact that the causal relationship of interest may interact with participant characteristics, settings, the types of outcomes measured, or treatment variations. Most often, threats to external validity (Table 4) are addressed in the interpretation of results, in which the investigator must clearly delineate the target population to whom the results refer (e.g., with respect to sociodemographics or geography) and judge the extent to which the findings are relevant to individuals, treatments, outcomes, and settings beyond the ones studied. However, external validity concerns can also be addressed with design or analytic features such as oversampling of underrepresented groups or modeling causal interactions.

Many threats to external validity relate to effect measure modification.[21] Effect measure modification is scale-dependent, so if the exposure and another variable both affect the

Table 4.Threats to External Validity

Threat No.	Threat Name	Definition
33	Interaction of the causal relationship with units	An effect found with certain kinds of units might not hold if other kinds of units had been studied.
34	Interactions of the causal relationship with settings	An effect found in one kind of setting may not hold if the study were conducted in another setting.
35	Context-dependent mediation	An explanatory mediator of a causal relationship in one context may not mediate in another.
36	Interaction of causal relationship with outcomes	An effect found on one kind of outcome observation may not hold if other outcome observations were used.
37	Interaction of the causal relationship over treatment variations	An effect found with one treatment variation might not hold with other variations of that treatment, or when that treatment is combined with other treatments, or when only part of that treatment is used.

Definitions are direct quotes or paraphrased from Shadish, Cook, and Campbell.[3]

outcome, it will occur on the additive scale (difference measures), the multiplicative scale (ratio measures), or both. Effect measure modification can therefore be represented on DAGs by including the modifying variable with an arrow pointing into the outcome.

Threats 33–34

Interaction of the causal relationship with units and interactions of the causal relationship with settings are both forms of effect measure modification that can arise when individual characteristics or contextual factors (respectively) affect the outcome (Figure 3A) or affect a mediator M of the E–D association (Figure 3B). Such effect measure modification threatens external validity when the distributions of these factors (U) differ between the study population and the population to which inference is being made. For example, in a study of the impact of neighborhood deprivation on risky sexual behavior, the measured effect may depend on the cultural background of the study participants or on other features of the contextual environment such as urban blight. Failure to measure and account for these modifiers when generalizing the study results to another population constitutes a threat to external validity.

Threat 35

Context-dependent mediation: See eAppendix; http://links.lww.com/EDE/B634.

Threat 36

Interaction of the causal relationship with outcomes refers to the fact that a cause-effect relationship may exist for one outcome (e.g., 5-year all-cause mortality) but not another seemingly related outcome (e.g., self-rated health). Whether we expect the established causal relation to extend to a new outcome depends on the causal structure linking the two outcomes. In some cases, multiple constructs may arise from a single latent variable or have a shared mechanism of action, thus the exposure would be expected to affect both outcomes (Figure 3C). In other cases, the outcomes are apparently unrelated, and we would not necessarily expect the same association with the exposure or confounding variables (Figure 3D).

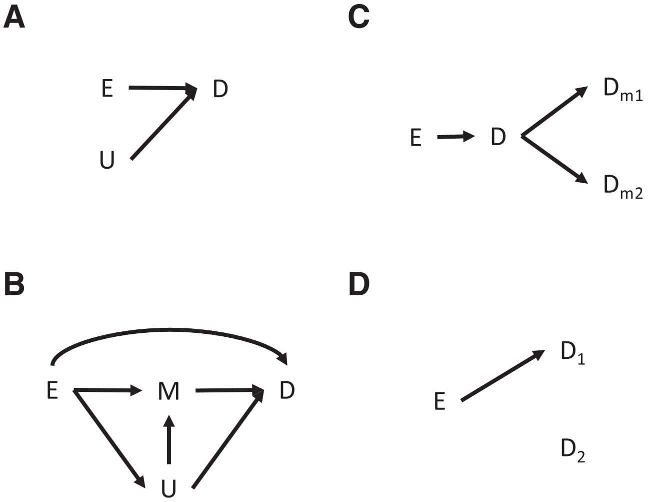

FIGURE 3. Threats to external validity represented as directed acyclic graphs.

Threat 37

Interaction of the causal relationship over treatment variations means that variations in the exposure do not result in the same observed outcome. If differences in the impacts of two distinct but related exposures are clearly defined, differences in their impacts are logical and perhaps expected. If, however, the exposure variations are intended to represent the same underlying variable, such a threat may constitute measurement error or a violation of consistency (Figure 2A).[22]

DISCUSSION AND CONCLUSIONS

To our knowledge, this is the first comprehensive map of the correspondence between Shadish, Cook, and Campbell's threats to validity with DAGs. Although the Campbell tradition and DAGs arise from distinct practices, there is no direct conflict between them. Both approaches will be helpful to applied researchers—the Campbell tradition's to recognize "what can go wrong" in real studies and DAGs to represent these difficulties in a formal language that can immediately inform whether effects of interest are identifiable and lead to new insights on how to deal with threats. To the extent that the Campbell tradition reflects those challenges that most commonly threaten empirical research,[5] the DAGs presented here can be considered a base library for DAG development. This library may be particularly useful, given that causal inference training in DAGs does not typically emphasize how to represent common problems in applied studies as DAGs.

Most DAGs can be boiled down to confounding, collider stratification bias, or measurement error, but the more detailed stories cataloged by Shadish, Cook, and Campbell help researchers to comprehend and recognize specific challenges in study design, statistical analysis, measurement, generalization, and interpretation. DAGs can lend clarity to problems that can easily go undetected or cause persistent confusion when left un-graphed.[23] They can also help avoid intuitively appealing but erroneous methodological decisions. For example, explication of regression to the mean (threat 5) highlights why studies involving treatment of participants for their extreme baseline outcomes may be problematic.[23] An intuitive solution is then to include a contemporaneous untreated group and to adjust for differences in measured baseline outcomes. The corresponding DAG in Figure 1E clearly shows why controlling for measured baseline outcomes may mistakenly induce a spurious association. This example highlights how pairing assessment of threats to validity with DAGs can enhance researchers' ability to identify and rule out alternative explanations for an association and to conduct valid studies.

We note several caveats of the present work: The 37 threats are not a collectively exhaustive list of all the ways studies can go wrong. Threats to validity are categorized into four buckets, but these categories are not ironclad; in fact, they have evolved with successive editions.[1–3] They simply provide a useful conceptual organization. Therefore, ruling out all named threats does not necessarily imply that the association can be interpreted causally. Additionally, we present simple DAGs corresponding to the various threats. In real applications, more complex causal structures are likely appropriate.

A primary goal of the epidemiologist's work is to draw causal inferences about the relationships between exposures and outcomes. Tools from other disciplines can enhance the work of epidemiologists, not only by informing causal diagrams. Efforts to map causal inference concepts across disciplines are growing[24,25] and offer researchers the opportunity to collaborate more effectively and to understand and leverage a broader range of tools and concepts useful when addressing causal research questions.

REFERENCES

1. Campbell DT, Stanley JC. *Experimental and Quasi-Experimental Designs for Research*. 11th ed. Chicago, IL: R. McNally College Publishing Company; 1973.
2. Cook TD, Campbell DT. *Quasi-Experimentation: Design and Analysis for Field Settings*. Chicago, IL: Rand McNally; 1979.
3. Shadish WR, Cook TD, Campbell DT. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton-Mifflin; 2002.
4. Cook TD. Twenty-six assumptions that have to be met if single random assignment experiments are to warrant "gold standard" status: a commentary on Deaton and Cartwright. *Soc Sci Med*. 2018;210:37–40.
5. Shadish WR. Campbell and rubin: a primer and comparison of their approaches to causal inference in field settings. *Psychol Methods*. 2010;15:3–17.
6. Pearl J. *Causality*. New York, NY: Cambridge University Press; 2009.
7. Glymour MM, Greenland S. Causal diagrams. In: Rothman KJ, Lash TL, Greenland S, eds. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008:183–209.
8. Hill AB. The environment and disease: association or causation? *Proc R Soc Med*. 1965;58:295–300.
9. Swanson JW, McGinty EE, Fazel S, Mays VM. Mental illness and reduction of gun violence and suicide: bringing epidemiologic research to policy. *Ann Epidemiol*. 2015;25:366–376.
10. Hernán MA, Robins JM. *Causal Inference*. Boca Raton, FL: CRC; 2010.
11. Kuh D, Ben-Shlomo Y, Lynch J, Hallqvist J, Power C. Life course epidemiology. *J Epidemiol Community Health*. 2003;57:778–783.
12. Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. *Int J Epidemiol*. 2005;34:215–220.
13. Rothman KJ, Greenland S, Lash TL. Chapter 9: validity in epidemiologic studies. In: *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008:128–147.
14. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15:615–625.
15. Howe CJ, Cole SR, Lau B, Napravnik S, Eron JJ Jr. Selection bias due to loss to follow up in cohort studies. *Epidemiology*. 2016;27:91–97.
16. Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. *Stat Methods Med Res*. 2012;21:243–256.
17. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology*. 1990;1:43–46.
18. Greenland S, Poole C. Rejoinder: living with statistics in observational research. *Epidemiology*. 2013;24:73.
19. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*. 2009;20:3–5.
20. Day SJ, Altman DG. Blinding in clinical trials and other studies. *BMJ*. 2000;321:504.
21. Greenland S, Lash TL, Rothman KJ. Chapter 5: concepts of interaction. In: *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008:71–83.
22. VanderWeele TJ, Hernán MA. Causal inference under multiple versions of treatment. *J Causal Inference*. 2013;1:1–20.
23. Glymour MM. Using causal diagrams to understand common problems in social epidemiology. In: Oakes MJ, Kaufman JS, eds. *Methods in Social Epidemiology*. San Francisco, CA: Jossey-Bass & Pfeiffer Imprint, Wiley; 2006:393–428.
24. Gunasekara FI, Carter K, Blakely T. Glossary for econometrics and epidemiology. *J Epidemiol Community Health*. 2008;62:858–861.
25. Steiner PM, Kim Y, Hall CE, Su D. Graphical models for quasi-experimental designs. *Sociol Methods Res*. 2017;46:155–188.