eAppendix

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). An additional theme is situations of measurement error that lead to reduced statistical power.

*Threats 10-15*. Low statistical power, violated assumptions of statistical tests, fishing and the error rate problem, inaccurate effect size estimation, extraneous variance, and heterogeneity of units correspond to the same statistical concepts in epidemiology. Low statistical power can lead to type II errors. Violated test assumptions can reduce power and bias a measured association. Inflated type I error due can be avoided by making statistical adjustments to account for multiple comparisons—i.e. to avoid fishing and the error rate problem. Inaccurate effect size estimation (threat 13) is specific to the parameter of interest. It may arise when there is misalignment between the estimator and the target causal quantity—for example, when odds ratios are used to estimate risk ratios1—or simply as a property of the estimator—for example, instrumental variable estimates are biased towards the ordinary least squares estimate in finite samples.2 Extraneous variance in the experimental setting refers to inconsistency in how the treatment or exposure is implemented. Heterogeneity of units refer to high variance in the outcome, and applies to non-differential and independent measurement errors.13 These situations reduce power, whether by extraneous influences in the study setting (e.g. distracting noises) or unmeasured heterogeneity of study participants (e.g. unmeasured genetic differences in risk). 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, these five threats are critical, but are generally not represented with DAGs

*Threat 16*. Unreliability of measures corresponds to nondifferential measurement error.13 If this error is in the exposure or outcome (eFigure 1A), the error introduced by UE and/or UD usually weakens the measured relationship (e.g. between Em and D) and reduces power to detect the true effect of E on D. For example, unmeasured non-adherence to a treatment for tuberculosis (TB) weakens its measured association with disease cure.

If the measurement error is in a confounder (eFigure 1B), residual bias will remain after controlling for the measured confounder—for example, if the association of TB treatment on cure were confounded by socioeconomic status, which was imperfectly measured with the proxy of educational attainment. This is true whether E causally affects D or not.

*Threat 17*. Restriction of the range corresponds to reduced power or attenuated effect estimates due to restrictions on the range of values the exposure or outcome can take, in situations where E has an effect on D. If restriction is on the range of the exposure (eFigure 1C), the relationship between the restricted exposure (Em) and the outcome may be weakened. This threat might occur, for example, when a continuous measure of body mass index is categorized, thereby reducing power to detect an association with heart disease.

If the restriction is on the outcome, bias can result, as in eFigure 1D where conditioning on Dm opens a backdoor path from E to D through UD.23 For example, genetics (UD) and education (E) might both affect intelligence as measured by an IQ test (D), which could be categorized as being above or below some threshold (Dm). Restricting to either strata of Dm would weaken the measured E-D association. Intuitively, if we restrict to a range of high values of D, and both UD and E increase D, then anyone with a low value of E must have a high value of UD; this creates an inverse association between UD and E.

*Threat 18*. Unreliability of treatment implementation refers to attenuated effect estimates due to partial treatment implementation, nonadherence, or noncompliance in studies where the exposure is assigned and the investigator is interested in the effect of a treatment or program delivered under ideal conditions to all participants (as opposed to a pragmatic trial where nonadherence is expected).13 When this pattern is not measured or accounted for, it introduces imprecision and bias. For example, if the assigned treatment (T) is not the exposure all participants actually receive (E), then the observed association will reflect E’s impact on D and be an underpowered and biased measure of T’s effect on D under perfect adherence (eFigure 1E).

eFigure 1: Threats to statistical conclusion validity represented as directed acyclic graphs



Supplemental threats to construct validity

*Threat 24*. Treatment sensitive factorial structure refers to situations where the exposure affects the dimensionality of the outcome but dimensions of the outcome that are measured by the researcher do not change. This threat can correspond to measurement error or differential item functioning over time. This issue might occur, for example, when studying a health literacy intervention’s impact on a multi-component score of attitudes towards physicians, where the untreated participants’ attitudes remain unchanged, but the treated participants’ attitudes change along several dimensions (D1, D2) that are masked when the score is summed or only assesses one component (e.g. Dm in eFigure 2A). This DAG also illustrates why controlling for one dimension in order to isolate the impact of the exposure on the other dimension could actually induce a spurious association between the exposure and the second dimension.

*Threat 25*. Reactive self-report changes can occur, for example, in a study of psychotherapy and depressive symptoms, where participants exaggerate their self-reported symptoms in order to seem more meritorious of treatment, and stop exaggerating once treatment has been assigned, generating a spurious association between the treatment assignment and changes in self-reported symptoms.20 This pattern could be considered a consistency violation, if changes in the motivation to exaggerate (U) are not considered part of the treatment T’s effect, or differential measurement error, if the self-reported symptoms are considered an inaccurate measurement of latent depression (eFigure 2B). This threat is particularly relevant when the treatment is assigned by the investigator and participants are not masked to treatment assignment.

*Threat 32*. Treatment diffusion occurs when the assigned exposure is not the one that was actually received by some participants.13 In this situation, it is important to correctly attribute the measured effect to either the exposure received (E) or the exposure assigned (T) (eFigure 2C). When the diffusion is accurately measured, a clear delineation and correct interpretation of either the intent-to-treat estimate or the per protocol estimate can remedy this threat when appropriate analyses are adopted.

eFigure 2: Supplemental threats to construct validity represented as directed acyclic graphs



Supplemental threats to external validity

*Threat 35*. Context-dependent mediation (eFigure 3A) arises when the target causal quantity is a mediation parameter. This threat is consistent with the widespread recognition in epidemiology that contextual factors may modify the relevance of specific mediators. For example, the controlled direct effect of neighborhood deprivation on risky sexual behavior that is not mediated by economic strain may depend on contextual levels of urban blight.

eFigure 3: Supplemental threats to external validity represented as directed acyclic graphs



Bibliography

1. Zhang J, Yu KF. What’s the relative risk?: a method of correcting the odds

ratio in cohort studies of common outcomes. JAMA. 1998;280:1690–

1691.

2. Bound J, Jaeger DA, Baker RM. Problems with instrumental variables

estimation when the correlation between the instruments and the endogenous

explanatory variable is weak. J Am Stat Assoc. 1995;90:443–450.