Underestimation of SES in Large Cohorts: A DAG-informed Simulation Study

Ross A. Dunne

2025-08-14

Background. Socioeconomic status (SES) is often modeled linearly or as coarse categories in large cohorts. When true SES→risk relationships are non-linear and samples are selected (healthy-volunteer bias), SES's role can be under-estimated.

Methods. We simulated a latent SES* affecting mediators (BMI, systolic BP, smoking) via non-linear functions and directly affecting a binary outcome. A "biobank-like" sample was generated via selection favoring higher SES and lower risk. We compared typical models (linear/quintile SES; with/without mediator adjustment) to spline models with g-computation of E[Y | do(SES=a)]. We summarized SES attribution via a causal variance share (R^{2} causal) and a two-block Shapley split.

Results. Selection yielded a biobank fraction of **4.9%** (N=9,851/200,000) and reduced prevalence from **5.99%** (population) to **3.57%** (selected). In the selected sample, linear-quintile SES achieved **McFadden R^2 =0.011**; including mediators raised predictive fit (**0.055**) while down-weighting SES as a putative cause. The **causal R^2** was **0.0052** in the population; within the selected sample it was **0.0159** using oracle SES* and **0.0015** using a noisy proxy.

Conclusions. Selection plus functional-form misspecification materially underestimates SES's contribution. Flexible modeling (splines/GAMs) with standardization (or TMLE) restores more of SES's causal role and should be preferred in large cohorts.

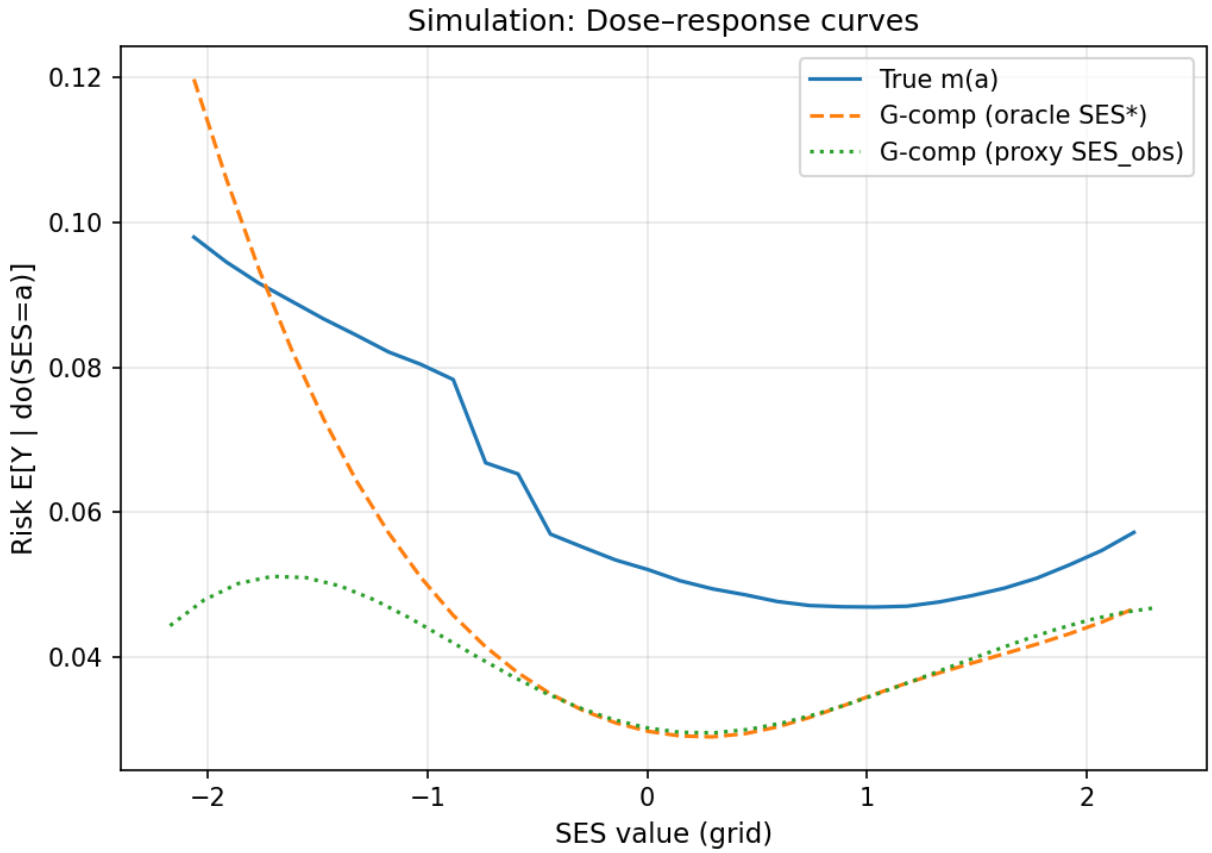

Figure 1: Simulation dose-response curves showing non-linear relationships and selection effects