Underestimation of SES effects 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
linear and samples are selected (healthy-volunteer bias), SES's role can be under-
estimated.

Methods. We simulated a latent SES* affecting mediators of cardiovascular disease(BMI, syst
linear functions and directly affecting a binary outcome. A "biobank-like" sample was generated
computation of E[Y | do(SES=a)]. We summarized SES attribution via a causal variance share (R^2
block Shapley split.

Results. Selection yielded a biobank fraction of **4.9%** (N=9,851/200,000) and reduced pre
quintile SES achieved **McFadden R^2 =0.011**; including mediators raised predictive fit (**0.0
weighting SES as a putative cause. The **causal R^2** was **0.0052** in the population; within

Conclusions. Selection plus functional-form misspecification materially underestimates SES

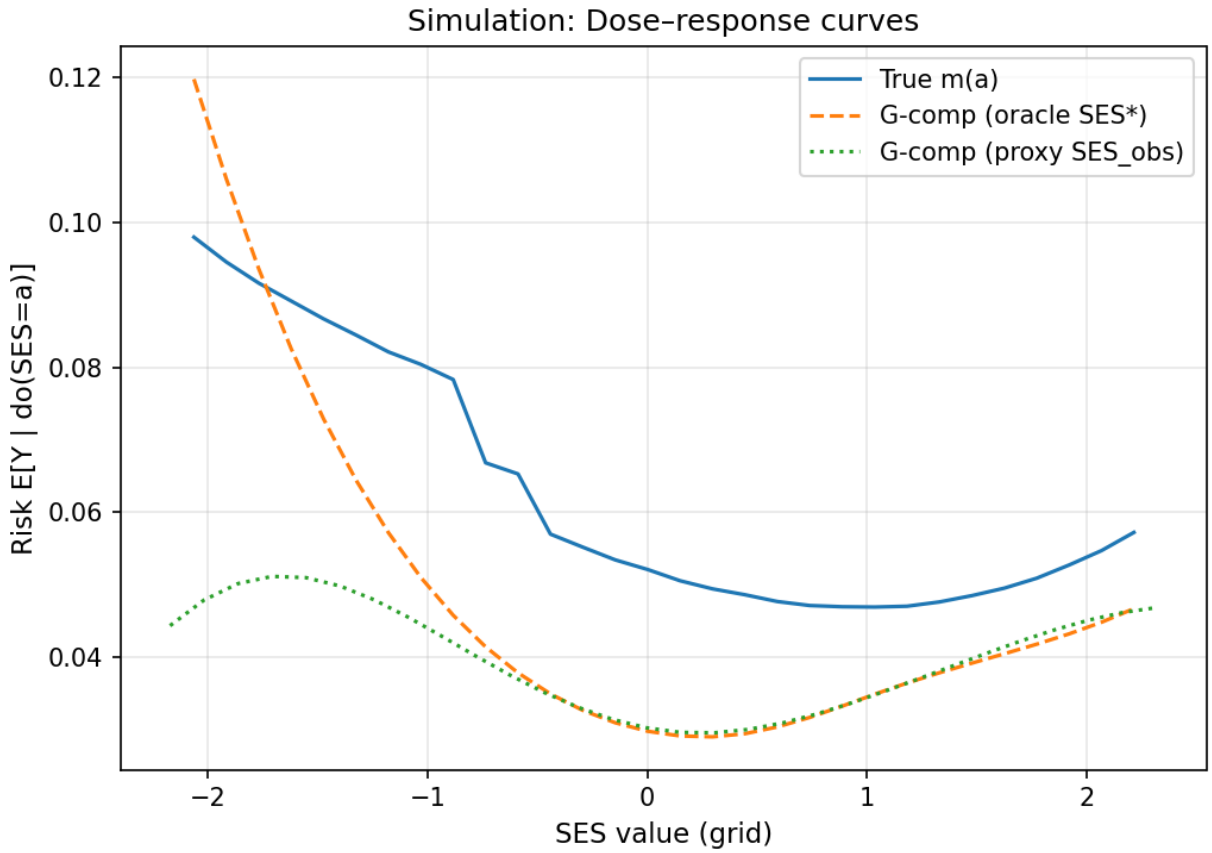

Figure 1: Simulation dose-response curves showing non-linear relationships and selection effects