Causal inference under over-simplified longitudinal causal models

Lola Étiévant - Université Claude Bernard Lyon 1.

Vivian Viallon - Université Claude Bernard Lyon 1, International Agency for Research on Cancer.

Epidemiology has successfully identified links between lifestyle exposures and the risk of disease, such as cancer. Biomarker measurements and -omics data provide important opportunities for investigating mechanisms underlying observed exposure-cancer relationships and cancer epidemiology is now increasingly concerned with the identification of such biological mechanisms. Therefore, mediation analysis has become popular. However, the causal validity of such analyses relies on strong assumptions. In particular, the causal model has to be correctly specified. Most often, *e.g.,* when studying lifestyle exposures such as tobacco smoke, alcohol and obesity, but also biomarkers, the true causal model under study involves time-varying risk factors. When repeated measurements are available for these variables, appropriate statistical approaches have been developed for such longitudinal causal models. However, they are not always available in large observational studies and practitioners most often consider simplified models that involve time-invariant variables only. Little is known on the relationship between estimates derived under these misspecified models and causal quantities of interest under the true longitudinal model.

Filling this gap is our main objective; we study total effects and natural direct and indirect effects, and focus on the case of an outcome Y measured at one single time point. Two particular situations are considered regarding the type of available information for the exposures, which include the exposure of interest as well as potential mediators and confounders. First, the situation where available data for the exposures correspond to their "instantaneous" levels at inclusion time in the study. Then, we turn our attention to a more favorable situation where the available information for each exposure corresponds to a summary measure of its levels up to inclusion in the study. In the first situation, our results are mostly negative: in general, causal inference based on single measurements of the exposures leads to quantities that cannot be related to true causal effects under longitudinal models. We obtain more positive results in the situation where summary measures are available: in the absence of confounding, or when only time-invariant confounders exist, inference based on summary measures returns meaningful quantities for total and natural effects. However, as soon as time-varying confounders exist, problems arise. We provide numerical illustrations to quantify some of the identified bias.

These results allow to shed light on the value of some common approaches. For instance, Mendelian Randomization (MR) is often applied to evaluate causal effects of time-varying exposures, while overlooking their time-varying nature. In that case, an implicit assumption is that the exposure measure is some summary variable of lifetime exposure that captures all its effect on the outcome. But then the limitations raised by our results would apply to MR as well. In particular, MR can lead to misleading results when both time-varying confounder and time-varying mediator exist (even if the latter is not considered at all) or in the presence of a time-varying confounder affected by the exposure of interest. Overall, our results are in line with those of previous works, which established the necessity of applying appropriate statistical methods on repeated measurements of exposures when the true causal model is longitudinal.