Colliders in Applied Biomedical Research: an educational interactive web application

Miguel Ángel Luque-Fernandez[1,2], *Pedro Femia*[1], *Miguel Ángel Montero-Alonso*[1], *Juan M. Melchor*[1], *Pedro Carmona-Saez*[1], *Juan de Dios Luna del Castillo*[1].

[1]Department of Statistics and Operations Research, University of Granada, Granada, Spain
[2]Department of Non-Communicable Disease Epidemiology. London School of Hygiene and Tropical Medicine. London, U.K.

Classical biomedical research (Epidemiology / Biostatistics) has focused on the control of confounding, but it is only recently that applied researchers have started to focus on the bias produced by colliders. A collider for a certain pair of variables (e.g., an outcome Y and an exposure A) is a third variable (C) that is caused by both. In DAGs terminology, a collider is the variable in the middle of an inverted fork (i.e., the variable C in A → C ← Y). Controlling for, or conditioning an analysis on a collider (i.e., through stratification or regression) can introduce a spurious association between its causes. This potentially explains many paradoxical findings in the medical literature, where established risk factors for a particular outcome appear protective. We used an example from non-communicable disease epidemiology to contextualize and explain the effect of conditioning on a collider. We generated a dataset with 1,000 observations and ran Monte-Carlo simulations to estimate the effect of 24-hour dietary sodium intake on systolic blood pressure, controlling for age, which acts as a confounder, and 24-hour urinary protein excretion, which acts as a collider. We illustrate how adding a collider to a regression model introduces bias. Thus, to prevent paradoxical associations, applied researchers estimating causal effects should be wary of conditioning on colliders. We provide R-code in easy-to-read boxes throughout the manuscript and a GitHub repository (https://github.com/migariane/ColliderApp) for the reader to reproduce our example. We also provide an educational web application allowing real-time interaction to visualize the paradoxical effect of conditioning on a collider http://watzilei.com/shiny/collider/. We investigated a situation where, adding a certain type of variable to a linear regression model, called a “collider”, led to bias with respect to the regression coefficient estimates while still improving the model fit. DAGs are based on subject matter knowledge and are vital for identifying colliders. Determining if a variable is a collider involves critical thinking about the true unobserved data generation process and the relationship between the variables for a given scenario. Then, the decision whether to include or exclude the variable in a regression model using observational data in biomedical research is based on whether the purpose of the study is prediction or explanation/causation. Under the structures we investigated here, adding a collider to a regression model is not advised when one is interested in the estimation of causal effects, as this may open a back-door path. However, if prediction is the purpose of the model, the inclusion of colliders in the models may be advisable if it reduces the model’s prediction error. Most research in Epidemiology and Biostatistics tries to explain how the world works (i.e., it is causal), thus, to prevent paradoxical associations, applied researchers estimating causal effects should be aware of such variables.

Keywords: Colliders.

References.

Pearl, J. Causality: models, reasoning, and inference (2nd ed.). Cambridge: Cambridge University Press. 2009.

Luque-Fernandez MA, Zoega H, Valdimarsdottir U, Williams MA. Deconstructing the smoking-preeclampsia paradox through a counterfactual framework. Eur J Epidemiol. 2016; 31: 613-623.

Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. Int J Epidemiol. 2009; 39:417-420.