Imperial College London

MSc Machine Learning & Data Science

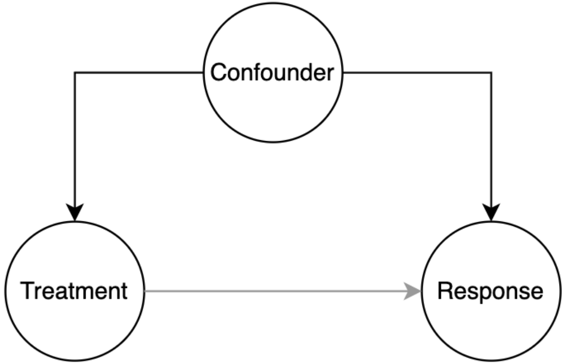
Ethics of Data Science

Ethics of Data Science – Part II

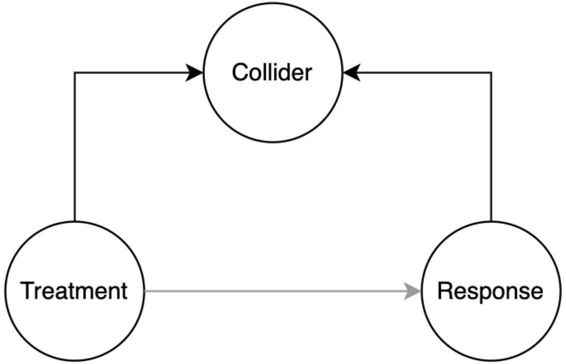
Colliders

Why not condition on everything observed?

e.g., prognostic factor



e.g., post-treatment effect

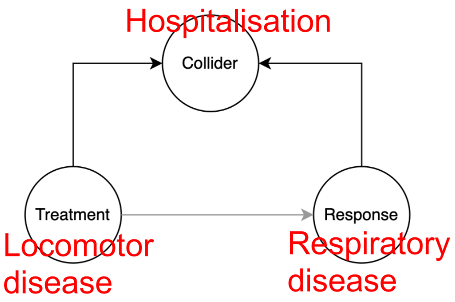


Not conditioning on a confounder introduces bias.

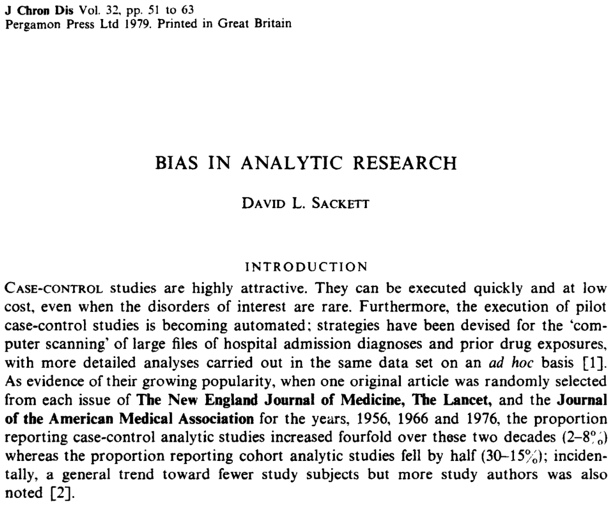
But conditioning on a collider also introduces bias.

Sackett’s admission bias

- Among 257 hospitalized patients, locomotor disease is found to be associated to respiratory disease
- Biologically plausible: inactivity leads to respiratory illness
- In the general population (incl. both hospitalized and non-hospitalized patients), no association exists.
- If you’re in the hospital, you’re there for a reason!



Treatment here used for any covariate hypothesized to have a causal effect on another variable.



J Chron Dis Vol. 32, pp. 51 to 63
Pergamon Press Ltd 1979. Printed in Great Britain

BIAS IN ANALYTIC RESEARCH

DAVID L. SACKETT

INTRODUCTION

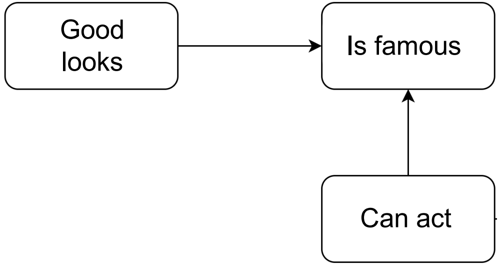
CASE-CONTROL studies are highly attractive. They can be executed quickly and at low cost, even when the disorders of interest are rare. Furthermore, the execution of pilot case-control studies is becoming automated; strategies have been devised for the ‘computer scanning’ of large files of hospital admission diagnoses and prior drug exposures, with more detailed analyses carried out in the same data set on an *ad hoc* basis [1]. As evidence of their growing popularity, when one original article was randomly selected from each issue of **The New England Journal of Medicine, The Lancet,** and the **Journal of the American Medical Association** for the years, 1956, 1966 and 1976, the proportion reporting case-control analytic studies increased fourfold over these two decades (2–8%) whereas the proportion reporting cohort analytic studies fell by half (30–15%); incidentally, a general trend toward fewer study subjects but more study authors was also noted [2].

The good-looking bad actor paradox

- Collider bias is a very important type of selection bias.
- It is present everywhere, not just in biostatistics.

- Among famous actors, “good looks” appear to be inversely correlated with ability to act.
- That is because a famous actor that cannot act must be famous for some other reason. Therefore, among bad actors, good looks are over-represented.

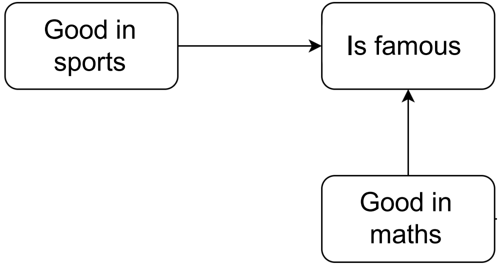
- *Note: other cognitive biases might be present here.*



The good in sports / bad in maths bias

- Collider bias is a very important type of selection bias.
- It is present everywhere, not just in biostatistics.

- If a prestigious school accepts students either on exceptional sporting ability or exceptional academic ability, then among that school population sporting ability and academic ability will be negatively correlated (if you’re bad in maths, then you must be good at sports).



Good in sports

Is famous

Good in maths

Does this paradox apply to a modern DS pipeline?

- Consider a situation where you are trying to understand the relationship between properties of an advertising campaign and sales.
- You look back at historical data to infer which ad works best.
- You only have data on the probability of a visitor that logged in to then buy, as a function of which ad (if any) they had seen.
- You ignore that, fit a random forest and then explore with SHAP the effect of the different properties of the advertising campaign on sales.

- The complexity of the model or the explainable AI technique used to assess the effect of a “treatment” variable does not make it more or less susceptible to collider or unobserved confounder bias.