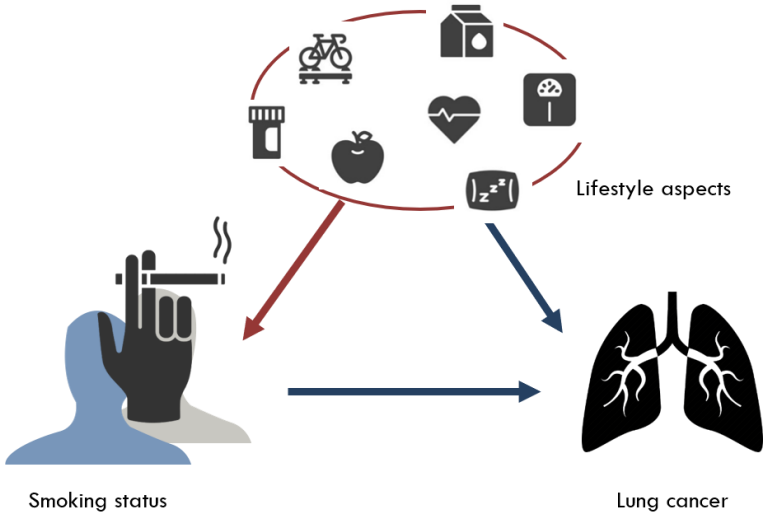
Causality in observational data

E-mail distributed on 21-06-2025

Dear all,

There are two common misconceptions on the ability to draw causal conclusions from observational data, that **(a)** this is not possible, and **(b)** it can be done "easily" with appropriate methods, such as the mediation model of Baron and Kenny (1986), or structural equation modelling (SEM). While the truth is closer to (a), and one should always prefer experimental designs over observational ones, I have personally been more concerned for the misconception of (b). That is, too many researchers oversimplify their causal assumptions and believe that a tool like SEM is sufficiently sophisticated and powerful to infer causality. This is obviously not true. In this mailing, I would like to discuss the problem of inferring causality using an alternative approach, known as **propensity score adjustment**.



1. A classic example

Suppose we would like to estimate the causal impact of smoking (yes or no) on the development of lung cancer (yes or no). Randomly assigning participants to smoking or non-smoking would be unethical, therefore we cannot run an experiment. However, we can make progress with observational smoking data, by first collecting as much information as we can about the demographical, lifestyle, and medical characteristics of a target sample (e.g., 50,000 people sampled randomly in Switzerland), especially characteristics that we suspect are linked to being a smoker and/or developing lung cancer.

One can imagine there are numerous such characteristics, perhaps hundreds, but most likely more than would be practical for a SEM analysis, which requires us to spell out literally all variable relationships. Running a regression also encounters difficulties, as adjusting for too many confounders simultaneously may create multicollinearities, suppression effects, or engage in false extrapolation. Propensity score adjustment instead proposes a different approach.

First, we run a logistic regression predicting smoking status from all measured confounders, and output the predicted probabilities per participant (i.e., the propensity model). These probabilities are known as the **propensity score** (Rosenbaum & Rubin, 1983), in this case the propensity to be a smoker, given confounder characteristics, *P*. Second, we run a logistic regression predicting lung cancer from smoking status, adjusted for the propensity score obtained in the first model (i.e., the outcome model). Now, we can estimate the causal impact of smoking on cancer, controlling for the probability to be a smoker in the first place.

$$P(\text{SMOKING}) = \alpha_0 + \alpha_1 c_1 + \ldots + \alpha_p c_P \quad (1)$$

$$P(\text{CANCER}) = \beta_0 + \beta_1 \text{SMOKING} + \beta_2 P(\text{SMOKING}) \quad (2)$$

While the example is somewhat simplified, the above has numerous advantages over either SEM or the naïve regression approach:

- No need to spell out all variable relationships
- The propensity score model is allowed to be overparametrized[1]
- All confounder information is efficiently compressed into a single score, *P*
- The outcome model with propensity adjustment has more power than SEM or regression for testing the group differences of interest
- No false extrapolation in strata where confounder characteristics between smoking groups do not overlap
- Tailored software or packages are (mostly) not required, it can be run simply as a two-step regression with any existing software.

Propensity score adjustment remains highly uncommon in social sciences, though it has been popular in medical sciences for decades, where it has become a standard approach for confounder adjustment in observational data. In these studies, the target predictor of interest is usually referred to generically as the "treatment" or "exposure".

2. Assumptions and extensions

The propensity score approach seems simple enough, but one should not gloss over several important assumptions. First and foremost, there should be **no unmeasured confounders** (Robins et al., 1992; VanderWeele & Vansteelandt, 2009), an assumption shared by all methods for causal inference, and one that is far too often neglected in SEM and simple causal models like Baron and Kenny. Second, while the propensity model is allowed to be overparametrized, no confounder should perfectly separate the treatment groups. In other words, propensity scores should not be systematically 0 or 1. This requires an inspection of the distribution of the propensity scores and their overlap between treatments. Third, the propensity model must be *correct*. Even if all relevant confounder variables are

[1] When the group of confounders is large they are often said to be "high-dimensional"

included, some effects may not be linear, but curvilinear, or involved in interactions. The propensity model must include these effects for the resulting propensity scores to be valid.

Fourth, while the approach does not require the detail of SEM, one should still reflect on the causal nature of the different confounders. Some may in fact be mediators, and others common outcomes of treatment and outcome. In the latter case, adjusting for such information can lead to Berkson's paradox, a.k.a., collider bias, and may produce spurious causal associations. Fifth, it is not entirely appropriate the fit the outcome model without further adjustments to the standard errors. That is, the uncertainty of the propensity model needs to be taken into account, which is typically done by the use of so-called "sandwich" or robust variance estimators, or via bootstrapping (Vansteelandt, Suetens & Goetghebeur, 2009).

Bearing the assumptions in mind, the idea behind propensity adjustment has been extended in several ways, sometimes going under different names. When *P* is used to weight the observations, rather than as a covariate, the method is typically referred to as **Inverse Probability Weighting (IPW)**[2] (Robins, Hernan & Brumback, 2000). Other studies divide *P* into bins and stratify the analysis categorically along these groups (Lunceford & Davidian, 2004). Another extension is to adjust the outcome model not just for *P* but also for the most important confounders from the propensity model, making the estimator "doubly robust", in the sense that even if the propensity model is misspecified, the presence of the confounder in the outcome models still allows for valid causal inference (Leon, Tsiatis and Davidian, 2003). When the treatment variable is continuous rather than categorical, the propensity model is sometimes called **G-estimation** (Robins, Mark & Newey, 1992). These methods have also been applied to quasi-experimental designs, where the target predictor of interest cannot be observed or manipulated directly but is operationalized by a third, indirect variable (experimental or observational), called an **instrumental variable**. Finally, propensity score adjustment is relevant in the context of **time-varying confounding** models, where a treatment, its confounders/mediators, and its outcome are measured repeatedly over time.

3. Reflecting on confounding and causality

Even if you do not plan to use propensity score adjustment, it is always useful to reflect on the causality in your data, and the possibility of confounding in particular. Too often the concern for confounding is restricted to subject-level confounding in observational designs, when in fact this may occur at multiple levels of measurement and also in experimental designs. For example, one may randomize participants to reading either abstract or concrete words and exclude confounding due to age, gender, reading speed, etc., but this design neglects that abstract words tend to be longer on average than concrete words (stimulus confounding). Similarly, an experiment that was conducted during winter may produce different results than one during summer. Randomized condition assignment of participants cannot exclude seasonal confounding.

Finally, it should also be clear that "simple" methods for inferring causality should not be blindly trusted. For mediation, for example, we are rarely in a situation where a simple X-M-Y path makes for an appropriate analysis, even in experimental data. There may not even be a plausible temporal contingency between the three variables (they may all be trait-like constructs), or the study did not measure them sequentially. SEM appears to be much more sophisticated (and in many ways it is) but here too many researchers take faith in their diagram as representing the causal truth, when

[2] Sometimes called Inverse Probability of Treatment Weighting (IPTW)

in fact it imposes numerous constraints on relationships that may not be realistic. Before using such methods, it is strongly advised to acquaint yourself with their assumptions.

References

Baron, R. M., & Kenny, D. A. (1986). Moderator-mediator variables distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51(6)*, 1173–82.

Leon, S., Tsiatis, A. A., & Davidian, M. (2003). Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics*, *59(4)*, 1046–1055.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, *23(19)*, 2937–2960.

Robins, J. M., Blevins, D., Ritter, G., & Wulfsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology*, *3(4)*, 319–336.

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11(5)*, 550–560.

Robins, J. M., Mark, S. D., & Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 479–495.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70(1)*, 41–55.

VanderWeele, T., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. Statistics and its Interface, 2, 457–468.

Vansteelandt, S., Mertens, K., Suetens, C., & Goetghebeur, E. (2009). Marginal structural models for partial exposure regimes. *Biostatistics*, *10(1)*, 46–59.

--

Ben Meuleman, Ph.D.
Statistician
Swiss Center for Affective Sciences
University of Geneva | Campus Biotech
Chemin des Mines 9 | CH-1202 Genève
ben.meuleman@unige.ch | +41 (0)22 379 09 79