Dealing with multiple confounders in observational studies: beyond logistic regression

Annibale Cois - Research Associate, School of Public Health and Family Medicine, University of Cape Town.

Estimation of existence and strength of causal relationships between exposures and outcomes is arguably the most common objective of studies in the health sciences, and randomized controlled trials (RCTs) are considered the gold standard approach at this end. The main reason of such a central role of RCTs compared to other study designs is their ability to effectively reduce the likelihood that the observed relationship exposure-outcome is biased by the presence of extraneous factors.

Two necessary — albeit not sufficient — conditions for an extraneous factor ("confounder") to produce such a bias are (Figure 1a):

1. the confounder is a risk factor for the outcome;
2. the confounder is associated with the exposure, i.e. its distribution is different among individuals with different exposure status.

![Figure 1: Schematic illustration of confounding control. Arrows represent causal effects, double arrows associations of any nature. E = exposure, C = confounder, O = outcome.](image)

If successful, the random allocation of subjects to the exposure which characterise RCTs ensures a balanced distribution of known and unknown confounding factors between exposed and non-exposed subjects. This is equivalent of removing the association between the exposure and all potential confounders (Figure 1b), and therefore, the possibility of confounding itself. In this case, the effect of the exposure on the outcome can be directly estimated by simply comparing outcomes between exposed and unexposed subjects (1).

Despite this unique ability to control for both known and unknown confounders, RCTs do not lack limitations, and often observational studies are the only practicable method of gaining insights on causal relationships. However, in absence of random assignment, individuals with different exposure status might differ systematically in characteristics other than the exposure of interest. If some of these characteristics are risk factors for the outcome, the observed relationship exposure-outcome may be biased: what we observe is not only the effect of the exposure on the outcome, but the combined effect of the exposure and the confounder(s) or even that due to the confounder alone (2, 3).

Accounting for multiple confounding factors in observational studies

In epidemiology, accounting for systematic differences between exposure groups in presence of multiple confounders has historically largely relied on regression methods. Among these, logistic regression is certainly the most common approach to deal with binary outcomes (e.g. "diseased" vs. "non diseased").
Regression uses mathematical modelling to estimate the effect of confounders on the outcome, and to "remove" this effect statistically. This is equivalent of removing (or, more realistically, reducing) the association between confounder and outcome, thus eliminating the second necessary condition for confounding (Figure 1c).

However, logistic regression performs poorly when the number of confounders is large compared to the number of individuals who experienced the outcome of interest (4), and, unlike randomization, cannot deal with unmeasured confounders. In these cases the estimates of the effect of the exposure on the outcome can be severely biased.

To overcome these limitations, two "families" of analytical methods are growing in popularity among epidemiologists as useful integrations of multiple regression: propensity score matching and Monte Carlo sensitivity analysis.

Propensity score matching is based on the concept of propensity score, defined as the probability of a subject being in an exposure group given the value of all measured confounders. It aims at balancing the distribution of confounding factors between exposure groups, thus mimicking — from the point of view of measured confounders — the results of a successful random assignment to the exposure (Figure 1b) (3).

Monte Carlo sensitivity analysis, on the other hand, is used to assess the sensitivity of the results of a study with regard to unknown, or unmeasured, confounding factors. It is based on the estimation of the effect of a large number of potential confounders on the observed relationship exposure-outcome, and takes advantage of the unprecedented availability of low-cost computing power.

In the following sections a concrete application of both of these techniques to deal with a typical situation in which regression methods alone are of limited usefulness is presented. Both examples are drawn from a recently published article by the author (5).

The effect of problem drinking on TB disease

In 2003, the South African Demographic and Health Survey (6) collected a broad range of demographic and health-related information in a representative sample of the South African population. Data from the 8115 adult respondents were analysed in order to estimate the strength of the association between lifetime problem drinking and prevalence of past diagnosis of tuberculosis (TB), taking into account a large number of potential confounders. Details on data and method, and a discussion on the epidemiological meaning and limitations of the results are reported in the original study (5). In this article we focus on the conceptual aspects of the analysis, neglecting unessential details.

The analytical problem can be summarised as the estimation of the "true" prevalence odds ratio (POR) of past TB diagnosis (TB+) between problem drinkers (PD+) and moderate drinkers/abstainers (PD-).

The challenges for this analysis are due to two features:

1. The large number of measured potential confounders relative to the total number of subjects reporting the outcome of interest.
2. The absence in the dataset of information about a known potentially powerful confounder, namely HIV status.

(Too) many measured confounders

Past diagnosis of TB was reported by 205 subjects. Considering the pattern of missing data, this value corresponds to less than 5.5 positive outcomes for each of the 31 confounders, well below the threshold which makes multivariate logistic modelling advisable (7-8 events per confounder) (4). A propensity score matching procedure was employed to deal with this problem.

First, the probability of being a problem drinker given the values of all potential confounders (the propensity score, PS, of being a problem drinker) was estimated for each subject in the sample. This was done building a probit regression model with all known confounders as predictors.

Then, each problem drinker was "matched" with one or more moderate drinkers/abstainers, based on similar values on the propensity scores. The result of this operation was the creation of two groups of exposed (PD+) and unexposed (PD-) subjects with similar distribution of propensity score, as shown in Figure 2.
The interest of this procedure (which underlies the usefulness of the use of PS in observational studies) is that in the absence of unmeasured confounders, balancing the distribution of PS across exposure groups is equivalent to balancing the distribution of all confounding factors (7), without having to take into account each confounder separately.

This result is evident from Figure 3: before matching, the distribution of confounders was very different between PD+ and PD−, while the residual imbalance between matched groups was almost negligible (standardised bias ≤ 3.5% across all variables).

As a consequence of this balancing, data could be analysed as if they were the result of a RCT, and the effect of problem drinking on TB disease simply estimated by comparing the prevalence of TB+ among problem drinkers and among moderate drinkers/abstainers. This comparison was carried out using conditional logistic regression to take into account the matching procedure, without the need of introducing further covariates in the model.
The result of the analyses indicated that the odds of past diagnosis of TB among lifetime problem drinkers were almost double the odds of TB among moderate drinkers/abstainers (POR: 1.97, 95% CI: 1.40 to 2.77). This value was largely in line with current results from the literature, recently reviewed by Rehm et al., and within the 95% CI of their pooled estimate (8). By way of comparison, the POR estimated using logistic regression with direct adjustment for confounding was 1.69 (95% CI: 1.07 to 2.67), significantly lower than the PS estimate and with a lower confidence limit close to the null.

**What about unmeasured confounders?**

Propensity score matching, under certain conditions, is more robust to omitted or incorrectly measured confounders compared to direct adjustment using multivariate regression (9). However, the possibility of bias, even severe, cannot be ruled out, especially when the unmeasured confounder is strongly associated with the outcome and the exposure.

In the TB-alcohol study, a source of special concern was the unmeasured factor "HIV status", known for being strongly causally related with TB disease, and possibly associated with problem drinking via socio-economic pathways.

To address this issue, a sensitivity analysis in the form of a Monte Carlo simulation was performed. Adapting a procedure suggested by Ichino et al. (10), a large set of hypothetical confounders was randomly generated, and the percentage reduction on the observed POR in respect to the value calculated in the absence of the confounder was estimated in each case. The results of the repeated simulations were then jointly depicted in a contour plot, with the axes representing the degree of association of the simulated confounders with the exposure and the outcome. The result is depicted in Figure 4.

![Figure 4: Sensitivity of the estimated prevalence odds ratio of TB to unmeasured confounders. Estimated percent reduction of the prevalence odds ratio of TB in problem drinkers vs. moderate drinkers/abstainers when hypothetical confounders are introduced in the models. Axes represent the adjusted odds ratio of the association — irrespective of the direction — of the hypothetical confounder with problem drinking (OR_{Problem}) and TB (OR_{TB}). Colours represent couples of values for OR_{Problem} and OR_{TB} producing the same reduction in the prevalence odds ratio. Source: (5)](image-url)

The visual inspection of the figure allows us to visually identify lower bounds for the strength of the association that potential unmeasured confounders should have with the outcome and the exposure to offset the observed POR or to reduce its value below any specific threshold. Within the limitations of a "brute force" method relying on the random generation of a relatively large—but obviously not exhaustive—number of possible confounders, the figure suggests that only hypothetical unmeasured factors with an extremely strong association with both TB and problem drinking (ORs > 6) could reduce the POR by about 50% or more, thus accounting for all the observed effect of problem drinking on TB. Even confounders characterised by ORs of 4 could...
account only for about 35% of the observed effect. By way of comparison, the strength of the association between each of the 31 measured confounders (represented by + in the figure) and both TB and problem drinking placed them in the safe area of the graph (potential bias < 25%).

In conclusion, while traditional regression methods have been proven as a powerful research tool, no technique is suitable for all circumstances. Two common situations in which regression techniques alone are likely to produce biased results is when the outcome is rare and the number of measured confounders is large; and/or when important confounders are neglected.

Analysing briefly an example in which both circumstances are present simultaneously, this article has shown how propensity score matching associated with Monte Carlo sensitivity analysis can be considered an interesting complement to traditional multivariate modelling.

The POR estimate produced by PS analysis was more in line with the results from the literature, and more precise (narrower confidence interval) than the POR estimated by logistic regression. Moreover, the PS model allowed for the direct assessment of the balance of the observed covariates between exposed and unexposed subjects, thus giving further reassurance about the limited amount of bias due to the measured confounders. The sensitivity analysis usefully integrated these findings and offered reasonable support to their robustness against confounding due to unmeasured factors.

As a final remark, it is worth noticing that while both PS and Monte Carlo analyses are useful complements to regression methods in some cases, they cannot be thought as substitutes. This fact is self-evident regarding Monte Carlo procedures, but it is true also for PS: the estimation of the PS itself usually relies on regression, and the final analyses of the matched groups are often done through some form of regression, as in our case.

Annibale Coi - Research Associate, School of Public Health and Family Medicine, University of Cape Town. Areas of interest: causal modelling, epidemiology of non-communicable diseases. annibale.cois@uct.ac.za

References: