Confounding factor or lurking variable

Confounding variable, also known as confounding factor or lurking variable can be defined as an undesirable variable that has an influence on the relationship between the variables of an experiment. Although they are not the variable of actual interest (i. e. the independent variable), they can influence the outcome of an experiment and they are considered to be undesirable as they could add error to an experiment. A proper designed experiment should aim to decrease or control the influence of such confounding variables in order to avoid type 1 error; an error that raises a ' false positive' conclusion that the independent variables have a casual relationship with the dependent variable. The relationship between the two observed variables is called a spurious relationship, hence a confounding variable is a threat to the validity of inferences made about cause and effect, i. e. the internal validity because the observed effect should be attributed to the independent variable rather than to the confounding variable.

An example can be illustrated by the relationship between ice cream sales and drowning deaths. When these variables are entered into a statistical analysis, they may show a positive and potentially statistically significant correlation. However, it is a mistake to infer a causal relationship (i. e., ice cream causes drowning) because an important confounding variable which causes both ice cream sales and an increase in drowning deaths has not been accounted for: i. e. summertime. Although there is a body of literature of criteria for causality, Pearl claimed that confounding variables cannot be defined in terms of statistical notions alone; some causal assumptions are sometimes necessary. For example, when causal assumptions are being

defined in the form of causal graphs, a simple criterion called backdoor will identify sets of confounding variables.

Types of confounding variables

Confounding variables may also be categorized according to their source: the choice of measurement instrument, situational characteristics, or inter-individual differences.

Solution

There are several ways to combat confounding variables in an experimental design by excluding or controlling it. Here is the following:

Case control studies: by assigning the same confounding variables to both the experimental and control group can control for such confounder, for example, if the cause of multiple infarct dementia is being studied, age and sex could be the confounding variables, therefore these factors should be matched paired between the two participant groups. In addition, randomization is also another solution as having all confounding variables (whether known or unknown) will be equally distributed across all groups.

Cohort studies: this is done by admitting a specific group of participants into the sampling population, for example a specific age range that may affect multiple infarct dementia, therefore only a certain group is chosen for the study design such as male aged 45-50 years old. This would limit the degree of matching between the groups and also cohorts can be comparable in regard to the possible confounding variable.

Stratification: in the example of multiple infarct dementia study, physical activity is hypothesized to be a variable that can prevent this dementia from happening. With age as a possible confounder. The sampling data will then be stratified by age group so that the association between physical activity and dementia can be analyzed per age group. If different age group yields different risk ratios (this can be analyzed by statistical tools called Mantel-Haenszel methods), then age is seen as a confounding variable.

Despite solutions for the controlling and limiting confounding variables, these strategies have limitations too. For example if a participant in the case-control study is a 47year old African-American from Alaska, avid tennis player, vegetarian, working as an engineer and suffer from multiple infarct dementia. Proper matching would require a person of the same characteristics but with the sole difference of being healthy. This is extremely difficult to achieve and there is a risk of over- or undermatching of the study population. Additionally, in a cohort study, too many people may be excluded with this criteria, and in stratification, single strata can get too narrow and contain only a small, non-significant number of samples.

Randomization

One of the most common reasons for the existence of confounding variable is when the experimental design does not randomly assign participants to groups or some types of individual difference such as ability, extroversion, height and weight. For example, studies involving a comparison between men and women are inherently plagued with confounding variables since the social environment for males and females is very different to start with.

However, this does not mean that there is no value in gender comparison studies or other studies that does not employ random assignment but it implies that results interpretation should be done cautiously. In sum, random assignment is a useful and powerful tool in experimental design. Although it does not minimize the overall amount of extraneous variable in an experiment, it aims to equalize the error that may occur as a result of extraneous variable, therefore it can greatly decrease systematic error: error that varies within the independent variable.

Multivariate analyses

Another method for controlling confounding variable is by the use of covariates in multivariate analyses. However, this only gives little information about the strength of the confounding variable compared to stratification methods. Furthermore, confounding variables are not always known or measurable which means residual confounding (term for incompletely controlled confounding) may appear. In an experimental design, covariate adjustment can help to reduce the noise in an outcome variation whilst enabling the manipulation effect to be performed. In sum, successful randomization can minimize confounding variables by bother measured and unmeasured factors, whereas statistical control addresses only confounding variables that have been measured and can introduce more confounding variables and other biases through inappropriate control.

Mismeasurement and mis-specification

Although it is important to spot confounding variables in a study there is often a risk of having a statistically controlled but imperfectly measured

factor that may confound an association of the variables. This is termed residual confounding which describes the mismeasurement and an example is given to illustrate this. In a study example, it was found that people with higher rate of vacation is correlated with lower risk of mortality. Several explanations can account for this as vacation mitigates stress, diminishes anger and encourage more exercise. On the other hand, healthier people might be more likely to travel so vacation may not be a genuine causal factor but only a marker of initial health status that naturally predicts longevity. Consequently, vacation may remain to be a significant predictor even after adjusting for baseline health status as the covariate. It is therefore easy to construct a series of potential confounders but many would lack plausibility. For example, people with more friends may have more vacations and friend was indeed the predictor variable instead, low-stress working environment and wide range of food (I. e. completeness of diet) may all attribute to prolong life too. However, because plausibility is a highly subjective factor for considering whether enough potential founders are included. To identify confounders Priori knowledge of the likely causal pathways are required. The major drawback of this is that observational studies imply that the strength of any causal inference will depend on the biologic plausibility of the putative factor, and the implausibility of uncontrolled potential confounders. In addition, observations contain some judgmental component which varies from experimenters. For example, vacation may prolong longevity because sick people tend to travel less, to deal with this. Measurements of participants' initial health may be used as an adjustment but this however cannot be assessed without error. Moreover, health can be measured in so many different ways and not all can be

included and controlled for. This raises more and more questions such as: can the use of initial stress test be used to capture aspects of health confounded by vacation? Is body mass index relevant? Consequently, even if the optimal measure of confounder is used it is measured with error and adjustment for it may not eliminate the effects of vacations.

From the statistical analysis perspective, poorly measured confounding variables causes more problems as its effect may not be linear, by assuming linearity on the outcome as specified by the model by entering confounding variables as a covariate in standard regression models may not fully adjust for the confounder effects are not linear on that scale.

Mediators and confounders

There is a common conflict that different causal explanations can be possible when adjustment is used to reduce or eliminate the predictive power of the independent variable. For example, a confounding variable may sometimes be a marker of some causal factors but it is not directly involved in the causal chain from one variable to another and there is a problem of over-adjustment. Considering an example on the hypothesis that high blood pressure (BP) reaction to stress causes Hypertension. To test this hypothesis, a longitudinal study should be conducted where BP reactivity and resting BP levels of a large group of participants should be measured. Result findings should report that excessive reactivity to be the risk factor for later hypertension but the problem is reactivity may just be a marker for elevated BP resting level and it is not important per se. consistent with this problem, those participants with higher resting BP may correlate with high BP

reactivity scores. To control for the current confounding variable, the initial resting BP levels should be adjusted by regression analysis which llustrates whether BP reactivity is attributed to any predictive information beside just the initial resting BP level. This may show that reactivity is no longer a very predictive factor and most of the variation in the follow-up BP levels may be accounted for by the initial resting levels. However, this does not mean that reactivity is not causally related to future BP status, i. e. if increased reactivity preceded initial increase in resting BP level, it could also be responsible in part for the initial increase in resting BP level. This is a situation whereby a single variable may have both confounding and mediating roles simultaneously. The example of vacation and mortality is used to illustrate this: assuming that people who go on more vacations are less likely to die over a 5-year longitudinal study, including a factor: initial health status in the regression model could eliminate this association. Alternatively, if people in poor health take fewer vacations then this elimination may reflect the removal of a confounding variable by health status. However, if the participants' tendencies to go on vacation are constant over the 5 year period then health status will reflect the cumulative health impact of a lifetime's vacation habits. This shows that health status will contribute partly as a mediator of vacationing effects. This confusion between a mediator and a confounder will be less apparent if the risk factor is not stable over time. For example if the participant has only just started having vacations, then these will not be reflected in the initial health status and may have higher opportunity to predict subsequent health with initial health status as a covariate in the analysis. However, if these changes become out of control, it can create a quasi-experimental design. For

example, if people take vacations due to change in their company policy rather than the reason of making friends or have spare time, and other group have less vacation for the same reason. Then in this case, it is possible to assess the effect of vacation independently of initial health status.

In sum, indiscriminate adjustment of covariates may result in erroneous conclusions and many socialdemographic variables can be mediated by other factors such as low income, unfulfilling jobs, no friends etc. moreover, there may also be other intermediate variables like self-determinations and release of stress hormones that may affect the results. Considering the wide range of variables listed, any inaccurate measures of them may lead to a reduction or elimination of predictive power. Moreover, by controlling a mediator may produce further confounding variables, which will then increase or decrease the associations of the independent and dependent measures. Furthermore, it may even create a new spurious association when in fact no effect is present.

In sum, despite the number of limitations discussed in this critical review, they have an important role in behavioural research as randomized trials are sometimes found to be impractical and unethical. In spite of the hazardous statistical control of confounding variables will gain insight into special cautions in drawing conclusions and writing research in the future.