Review Article

Minimization of confounding in research in medical sciences

Mollen Akinyi Odero[1], Nancy Odero[1] and N. B. Okelo*[2]

[1]School of Health Sciences, Jaramogi Oginga Odinga University of Science and technology, P. O. Box 210-40601, Bondo-Kenya
[2]School of Mathematics and Actuarial Science, Jaramogi Oginga Odinga University of science and technology, P. O. Box 210-40601, Bondo-Kenya

***Corresponding Author**
N. B. Okelo
School of Mathematics and Actuarial Science,
Jaramogi Oginga Odinga University of science
and technology, P. O. Box 210-40601,
Bondo-Kenya
E-mail: bnyaare@yahoo.com

Keywords:
Confounding,
Medical Science,
statistics

Abstract

Confounding variables are variables that the researcher failed to control, or eliminate, damaging the internal validity of an experiment. A confounding variable, also known as a third variable or a mediator variable, can adversely affect the relation between the independent variable and dependent variable. This may cause the researcher to analyze the results incorrectly. The results may show a false correlation between the dependent and independent variables, leading to an incorrect rejection of the null hypothesis. For example, a research group might design a study to determine if heavy drinkers die at a younger age. They proceed to design a study, and set about gathering data. Their results, and a series of statistical tests, indeed show that people who drink excessively are likely to die younger. Unfortunately, when the researchers do a crosscheck with their peers, the results are ripped apart, because their peers live just as long - maybe there is another factor, not measured, that influences both drinking and living age? In many fields of science, it is difficult to remove entirely all of the variables, especially outside the controlled conditions of a lab. A well-planned experimental design, and constant checks, will filter out the worst confounding variables. For example, randomizing groups, utilizing strict controls, and sound operationalization practice all contribute to eliminating potential third variables. In this work we explore types of confounding and how to control them.

1.Introduction

In statistics, a confounding variable (also confounding factor, a confound, or confounder) is an extraneous variable in a statistical model that correlates (directly or inversely) with both the dependent variable and the independent variable. A perceived relationship between an independent variable and a dependent variable that has been misestimated due to the failure to account for a confounding factor is termed a spurious relationship, and the presence of misestimating for this reason is termed omitted-variable bias. In the case of risk assessments evaluating the magnitude and nature of risk to human health, it is important to control for confounding to isolate the effect of a particular hazard such as a food additive, pesticide, or new drug. For prospective studies, it is difficult to recruit and screen for volunteers with the same background (age, diet, education, geography, etc.), and in historical studies, there can be similar variability. Due to the inability to control for variability of volunteers and human studies confounding is a particular challenge. For these reasons, experiments offer a way to avoid most forms of confounding. As an example, suppose that there is a statistical relationship between ice-cream consumption and number of drowning deaths for a given period. These two variables have a positive correlation with each other. An evaluator might attempt to explain this correlation by inferring a causal relationship between the two variables (either that ice-cream causes drowning, or that drowning causes ice-cream consumption). However, a more likely explanation is that the relationship between ice-cream consumption and drowning is spurious and that a third, confounding, variable (the season) influences both variables: during the summer, warmer temperatures lead to increased ice-cream consumption as well as more people swimming and thus more drowning deaths. While specific definitions may vary, in essence a confounding variable fits the following four criteria, here given in a hypothetical situation with variable of interest "V", confounding variable "C" and outcome of interest "O":

1. C is associated (inversely or directly) with O
2. C is associated with O, independent of V
3. C is associated (inversely or directly) with V
4. C is not in the causal pathway of V to O (C is not a direct consequence of V, not a way by which V produces O)

In a more concrete example, say one is studying the relation between birth order (1st child, 2nd child, etc.) and the presence of Down's Syndrome in the child. In this scenario, maternal age would be a confounding variable:

1. Higher maternal age is directly associated with Down's Syndrome in the child
2. Higher maternal age is directly associated with Down's Syndrome, regardless of birth order (a mother having her 1st vs 3rd child at age 50 confers the same risk)
3. Maternal age is directly associated with birth order (the 2nd child, except in the case of twins, is born when the mother is older than she was for the birth of the 1st child)
4. Maternal age is not a consequence of birth order (having a 2nd child does not change the mother's age)

2. Types of Confounding

In some disciplines, confounding is categorized into different types. In epidemiology, one type is "confounding by indication" which relates to confounding from observational studies. Because prognostic factors may influence treatment decisions (and bias estimates of treatment effects), controlling for known prognostic factors may reduce this problem, but it is always possible that a forgotten or unknown factor was not included or that factors interact complexly. Confounding by indication has been described as the most important limitation of observational studies. Randomized trials are not affected by confounding by indication due to random assignment. Confounding variables may also be categorised according to their source. The choice of measurement instrument (operational confound), situational characteristics (procedural confound), or inter-individual differences (person confound).

- An **operational confounding** can occur in both experimental and non-experimental research designs. This type of confounding occurs when a measure designed to assess a particular construct inadvertently measures something else as well.
- A **procedural confounding** can occur in a laboratory experiment or a quasi-experiment. This type of confound occurs when the researcher mistakenly allows another variable to change along with the manipulated independent variable.
- A **person confounding** occurs when two or more groups of units are analyzed together (e.g., workers from different occupations), despite varying according to one or more other (observed or unobserved) characteristics (e.g., gender).

Examples: In risk assessments, factors such as age, gender, and educational levels often have an impact on health status and so should be controlled. Beyond these factors, researchers may not consider or have access to data on other causal factors. An example is on the study of smoking tobacco on human health. Smoking, drinking alcohol, and diet are lifestyle activities that are related. A risk assessment that looks at the effects of smoking but does not control for alcohol consumption or diet may overestimate the risk of smoking [4]. Smoking and confounding are reviewed in occupational risk assessments such as the safety of coal mining [5]. When there is not a large sample population of non-smokers or non-drinkers in a particular occupation, the risk assessment may be biased towards finding a negative effect on health.

The above correlation-based definition, however, is metaphorical at best – a growing number of analysts agree that confounding is a causal concept, and as such, cannot be described in terms of correlations nor associations [1-8].

2.1 Causal Definition

The concept of confounding can be formalized, and managed, when information is available about the data generating model (as in the Figure above). To be more specific, let X be some independent variable, Y some dependent variable, and M a causal model that asserts the cause-effect relationships between variables in the system. To estimate the effect of exposure X on outcome Y, the statistician must suppress the effects of extraneous variables that influence both X and Y. We say that, X and Y are confounded by some other variable Z whenever Z is a cause of both X and Y.

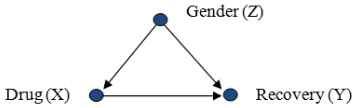
In the causal framework, denote *P(y/do(x))* as the probability of event Y = y under the hypothetical intervention X = x. X and Y are not confounded in causal model M if and only if the following holds:

$$P(y/do(x))=P(y/x)\text{……………………………………………}(1)$$

for all values X = x and Y = y, where *P(y/x)* is the conditional probability upon seeing X = x. Intuitively, this equality states that X and Y are not confounded whenever the observationally witnessed association between them is the same as the association that would be measured in a controlled experiment, with x randomized.

2.2 Minimization of Confounding

Consider the scenario of a physician deciding to administer drug X to a patient with gender Z. The physician knows that gender differences influence a patient's choice of drug as well as their chances of recovery. In this scenario, gender Z is a confound of administering drug X on recovery outcome Y since Z is a cause of both X and Y:



Consequently, we will encounter the inequality:

$$P(y|do(x)) \neq P(y|x) \qquad (2)$$

Since the observational quantity contains information about the correlation between X and Z, and the interventional quantity does not (being an unbiased estimate of the effect of X on Y). Clearly the statistician desires the unbiased estimate, but in cases where only observational data is available, an unbiased estimate can only be obtained by "adjusting" for all confounding factors, namely, conditioning on their various values and averaging the result. In the case of a single confounder Z, this leads to the "adjustment formula":

$$P(y|do(x)) = \sum_z P(y|x,z)P(z) \qquad (3)$$

Which gives an unbiased estimate for the causal effect of X on Y. The same adjustment formula works when there are multiple confounders except, in this case, the choice of a set Z of variables that would guarantee unbiased estimates must be done with caution. The criterion for a proper choice of variables is called the Back-Door [9][10] and requires that the chosen set Z "blocks" (or intercepts) every path from X to Y that ends with an arrow into X. Such sets are called "Back-Door admissible" and may include variables which are not common causes of X and Y, but merely proxies thereof.

2.3 Decreasing the potential for confounding to occur

A reduction in the potential for the occurrence and effect of confounding factors can be obtained by increasing the types and numbers of comparisons performed in an analysis: Increasing the number of confounding factors controlled for increases significance. If measures or manipulations of core constructs are confounded (i.e., operational or procedural confounds exist), subgroup analysis may not reveal problems in the analysis. Additionally, increasing the number of comparisons can create other problems (see multiple comparisons).

Peer review is a process that can assist in reducing instances of confounding, either before study implementation or after analysis has occurred. Peer review relies on collective expertise within a discipline to identify potential weaknesses in study design and analysis, including ways in which results may depend on confounding. Similarly, replication can test for the robustness of findings from one study under alternative study conditions or alternative analyses (e.g., controlling for potential confounds not identified in the initial study).

Confounding effects may be less likely to occur and act similarly at multiple times and locations. In selecting study sites, the environment can be characterized in detail at the study sites to ensure sites are ecologically similar and therefore less likely to have confounding variables. Lastly, the relationship between the environmental variables that possibly confound the analysis and the measured parameters can be studied. The information pertaining to environmental variables can then be used in site-specific models to identify residual variance that may be due to real effects[7].

Depending on the type of study design in place, there are various ways to modify that design to actively exclude or control confounding variables[12].

Case-control studies assign confounders to both groups, cases and controls, equally. For example if somebody wanted to study the cause of myocardial infarct and thinks that the age is a probable confounding variable, each 67 years old infarct patient will be matched with a healthy 67 year old "control" person. In case-control studies, matched variables most often are the age and sex. Drawback: Case-control studies are feasible only when it is easy to find controls, *i.e.*, persons whose status vis-à-vis all known potential confounding factors is the same as that of the case's patient: Suppose a case-control study attempts to find the cause of a given disease in a person who is 1) 45 years old, 2) African-American, 3) from Alaska, 4) an avid football player, 5) vegetarian, and 6) working in education. A theoretically perfect control would be a person who, in addition to not having the disease being investigated, matches all these characteristics and has no diseases that the patient does not also have-but finding such a control would be an enormous task.

Cohort studies

A degree of matching is also possible and it is often done by only admitting certain age groups or a certain sex into the study population, creating a cohort of people who share similar characteristics and thus all cohorts are comparable in regard to the possible confounding variable. For example, if age and sex are thought to be confounders, only 40 to 50 years old males would be involved in a cohort study that would assess the myocardial infarct risk in cohorts that either are physically active or inactive. Drawback: In cohort studies, the overexclusion of input data may lead researchers to define too narrowly the set of similarly situated persons for whom they claim the study to be useful, such that other persons to whom the causal relationship does in fact apply may lose the opportunity to benefit from the study's recommendations. Similarly, "over-stratification" of input data within a study may reduce the sample size in a given stratum to the point where generalizations drawn by observing the members of that stratum alone are not statistically significant.

Double blinding

Conceals from the trial population and the observers the experiment group membership of the participants. By preventing the participants from knowing if they are receiving treatment or not, the placebo effect should be the same for the control and treatment groups. By preventing the observers from knowing of their membership, there should be no bias from researchers treating the groups differently or from interpreting the outcomes differently.

Randomized controlled trial

A method where the study population is divided randomly in order to mitigate the chances of self-selection by participants or bias by the study designers. Before the experiment begins, the testers will assign the members of the participant pool to their groups (control, intervention, parallel), using a randomization process such as the use of a random number generator. For example, in a study on the effects of exercise, the conclusions would be less valid if participants were given a choice if they wanted to belong to the control group which would not exercise or the intervention group which would be willing to take part in an exercise program. The study would then capture other variables besides exercise, such as pre-experiment health levels and motivation to adopt healthy activities. From the observer's side, the experimenter may choose candidates who are more likely to show the results the study wants to see or may interpret subjective results (more energetic, positive attitude) in a way favorable to their desires.

Stratification

As in the example above, physical activity is thought to be a behaviour that protects from myocardial infarct; and age is assumed to be a possible confounder. The data sampled is then stratified by age group – this means, the association between activity and infarct would be analyzed per each age group. If the different age groups (or age strata) yield much different risk ratios, age must be viewed as a confounding variable. There exist statistical tools, among them Mantel–Haenszel methods, that account for stratification of data sets.

Controlling for confounding by measuring the known confounders and including them as covariates is multivariate analyses such as regression analysis. Multivariate analyses reveal much less information about the *strength* or *polarity* of the confounding variable than do stratification methods. For example, if multivariate analysis controls for antidepressant, and it does not stratify antidepressants for TCA and SSRI, then it will ignore that these two classes of antidepressant have *opposite* effects on myocardial infarction, and one is much *stronger* than the other.

All these methods have their drawbacks:

1. The best available defense against the possibility of spurious results due to confounding is often to dispense with efforts at stratification and instead conduct a randomized study of a sufficiently large sample taken as a whole, such that all potential confounding variables (known and unknown) will be distributed by chance across all study groups and hence will be uncorrelated with the binary variable for inclusion/exclusion in any group.
2. Ethical considerations: In double blind and randomized controlled trials, participants are not aware that they are recipients of sham treatments and may be denied effective treatments[8]. There is resistance to randomized controlled trials in surgery because patients would agree to invasive surgery which carries risks under the understanding that they were receiving treatment.

References

[1] Johnston S. C. Identifying Confounding by Indication through Blinded Prospective Review. *Am J Epidemiol 2001;* 154 (3): 276-284.

[2] Pelham B. *Conducting Research in Psychology*. Belmont: Wadsworth, 2006.

[3] Steg L. and Buunk A.P. *Applied Social Psychology: Understanding and managing social problems*. Cambridge, UK: Cambridge University Press, 2008.

[4] Tjonnel and, Anne; Morten Gronbæk, Connie Stripp and Kim Overvad, *American Society for Nutrition American Journal of Clinical Nutrition* 1999; 69 (1): 49-54.

[5] Axelson O. Confounding from smoking in occupational epidemiology, *British Journal of Industrial Medicine* 1989; 46: 505-07.

[6] Pearl J, Paradox S. Confounding, and Collapsibility In Causality: Models, Reasoning and Inference (2nd ed.). New York, NY, USA: Cambridge University Press, 2009.

[7] Vander Weele K., Shpitser T.J. On the definition of a confounder. *Annals of Statistics* 2013; 41: 196-220.

[8] Greenland S., and Robins S., Confounding and Collapsibility in Causal Inference. *Statistical Science* 1999; 14(1): 29-46.