JOURNAL SERIES:

Statistics Review Part 10: Causality and Confounding

by Jordan Rush, PharmD, Maria Ajami, PharmD, Kevin Look, PharmD, PhD, and Amanda Margolis, PharmD, MS, BCACP

Objectives

1. Define causality and correlation
2. Describe confounding variables and their criteria
3. Describe methods for minimizing the presence of confounding variables

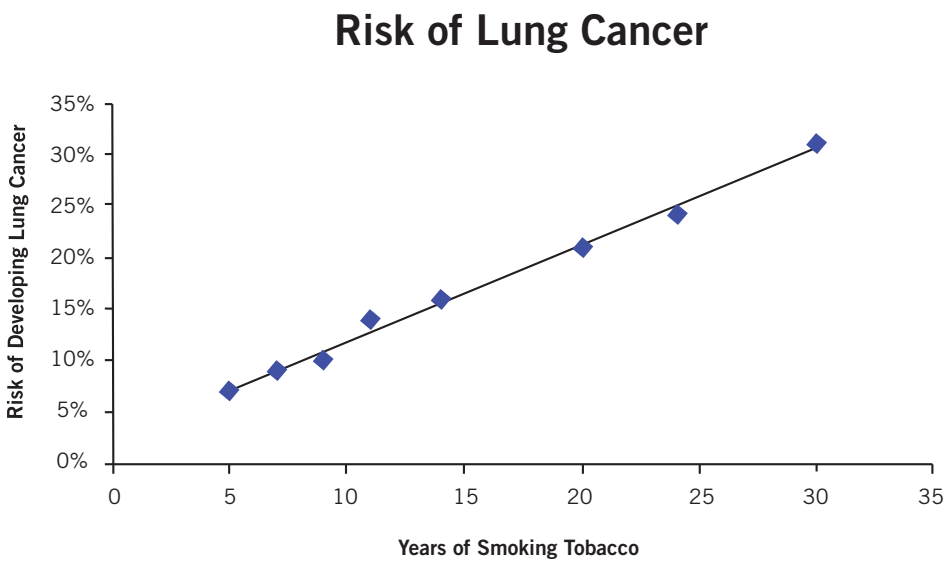
Data interpretation is vital to a pharmacist's ability to understand a study's results and to draw appropriate conclusions from the findings. This review will cover the concept of causality, correlation, and confounding variables and their impact on study design, analysis, and interpretation.

Causality

Causality can be defined as the relationship between a cause and an effect. A causal relationship exists if certain criteria are met. There are a number of principles used to determine causal relationships, including (1) the cause preceded the effect, (2) the cause was related to the effect and (3) there is no plausible alternative explanation for the effect other than the cause.[1] Other criteria can also be used to determine causality but a temporal relationship is the only required characteristic.[2] Figure 1 shows an example of a causal relationship between the number of years a person smokes (independent or causal variable) and the increased risk of developing lung cancer (dependent or outcome variable) [For more on independent and dependent variables, see part 1 of this series].

FIGURE 1. Causal relationship of tobacco smoking leading to an increased risk of developing lung cancer

FIGURE 2. Correlation between the number of years smoking tobacco and the risk of developing lung cancer



Correlation

A correlation shows the strength and direction of the relationship or connection between two variables. Using the previous example in Figure 1, the independent variable, or how many years a person smokes tobacco, could be plotted on the x-axis (horizontal axis) of a graph and the risk of developing lung cancer, the dependent variable, plotted on the y-axis (vertical axis). A trend line could then be derived, which best approximates the linear relationship between the observed data points (Figure 2).

The fit of the trend line can be quantified using the correlation coefficient.[3] The most commonly used correlation coefficient is the "Pearson product-moment correlation coefficient" or more simply the Pearson correlation coefficient.[4] The Pearson correlation coefficient (commonly represented by the letter r) is used for continuous variables and ranges from -1 to +1 depending on the strength and direction of the relationship between the two variables. A positive r value indicates that two variables move in the same direction and a negative r value indicates the two variables move in opposite directions. The strength of the association between the variables is given by the absolute value of

TABLE 1. Criteria for Causality and Criteria for Confounding Variables[5,7]

Criteria for Causality	*Criteria for Confounding Variables*
1. Temporal relationship 2. Strength of relationship 3. No alternative explanation exists	1. Must be associated with the exposure; causally or non-causally 2. Must be an independent risk factor for the outcome of interest 3. Cannot be on the causal pathway between the exposure and outcome

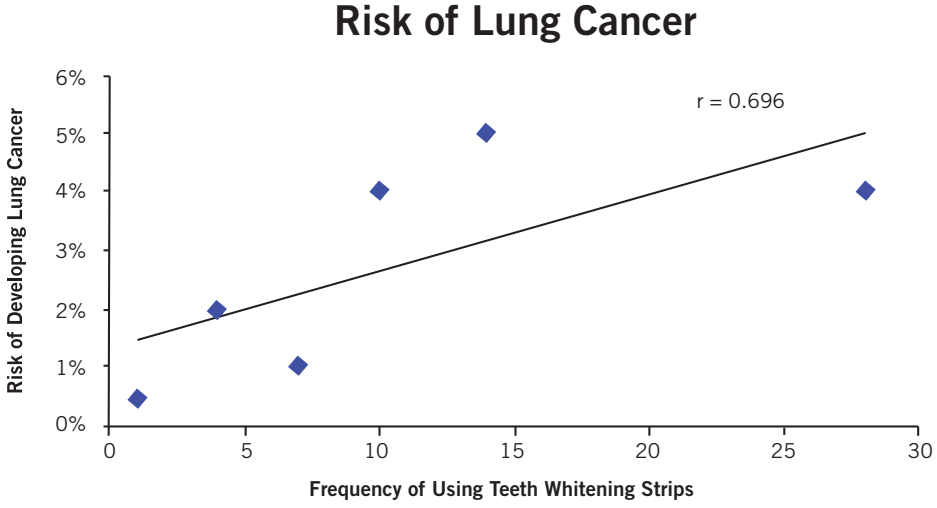
r, regardless of the direction of the effect. If r is +1/- then the two variables have a perfect positive or perfect negative linear relationship. An r value of zero indicates there is no association between the two variables. Often a correlation coefficient value between -0.20 to +0.20 indicates little or no relationship between two variables, an absolute value of 0.20 to 0.50 indicates a moderate relationship, and an absolute value of >0.50 indicates a strong relationship between two variables.[4] In Figure 2, there is a strong relationship between an increased risk of developing lung cancer and each additional year of smoking. If the r value is 0.996 then a strong positive relationship exists between the number of years a person smokes and the risk of developing lung cancer.

Other types of correlation coefficients are used based on the type of data.[4] For example, a Spearman rho correlation coefficient is used for ordinal variables. Although correlations indicate the relationship between two variables, it is very important to realize that correlation does not indicate which variable comes first. That is, a correlation cannot tell us whether smoking causes lung cancer or vice versa. Therefore correlation cannot prove causation.[5]

Confounding

Confounding variables arise when a study investigator relates one studied exposure to an outcome, but finds that the outcome is affected by a third, sometimes unmeasured, factor.[6] This third factor is referred to as a confounding variable. Confounding variables can distort study outcomes because they are correlated with both the exposure and the outcome of interest, and can lead to biased values when estimating the relationship between two variables if the confounding factor is not accounted for. Confounders are not directly on the causal pathway between the exposure and outcome variables, but are related to the studied exposure none-the-less.

FIGURE 3. Correlation between the frequency of using teeth whitening strips and the risk of developing lung cancer



Consider the example presented in Figure 3, which shows a strong correlation (r = 0.696) between the use of teeth whitening strips and an increased risk of developing lung cancer. This relationship does not meet the criteria for causality (Table 1), as there is no direct causal link between these variables because potential alternative explanations exist. This is an example where tobacco smoking acts as a confounding variable (Figure 4). The increased use of tobacco smoking contributed to both the increased use of teeth whitening strips and increased risk of developing lung cancer. Tobacco smoking meets the three criteria in Table 1 for a confounding variable, as it is associated with the exposure (i.e., tobacco smoking can yellow the teeth leading to an increased use of whitening strips), it is an independent risk factor for the outcome of interest (i.e., tobacco smoking causes lung cancer), and it is not on the causal pathway between the exposure and outcome (i.e., whitening strips do not cause tobacco smoking).

Reversing the role of the teeth whitening strips and tobacco smoking shows an example of a non-confounding variable (Figure 5). According to the criteria for confounding variables in Table 1, the potential confounder must be an independent risk factor for the outcome of interest; however, teeth whitening strips are not known to cause lung cancer. Therefore, the use of teeth whitening strips is not a confounding variable in this relationship, even though their use is increased among smokers, who have an increased risk of lung cancer due to smoking.

Controlling for Confounding Variables

Confounding variables are problematic because they may introduce bias into the conclusions of the study if they are not accounted for in the study design or data analysis.[7] Identifying, measuring, and controlling for confounding variables can help reduce bias in a study.

Randomization during study design is the preferred way to minimize the effect of confounders. Through randomization, the chance of bias due to confounding factors is minimized and is the only way

FIGURE 4. **Tobacco smoking is a confounding variable in this example of the potential association of using teeth whitening strips and the increased risk of developing lung cancer**

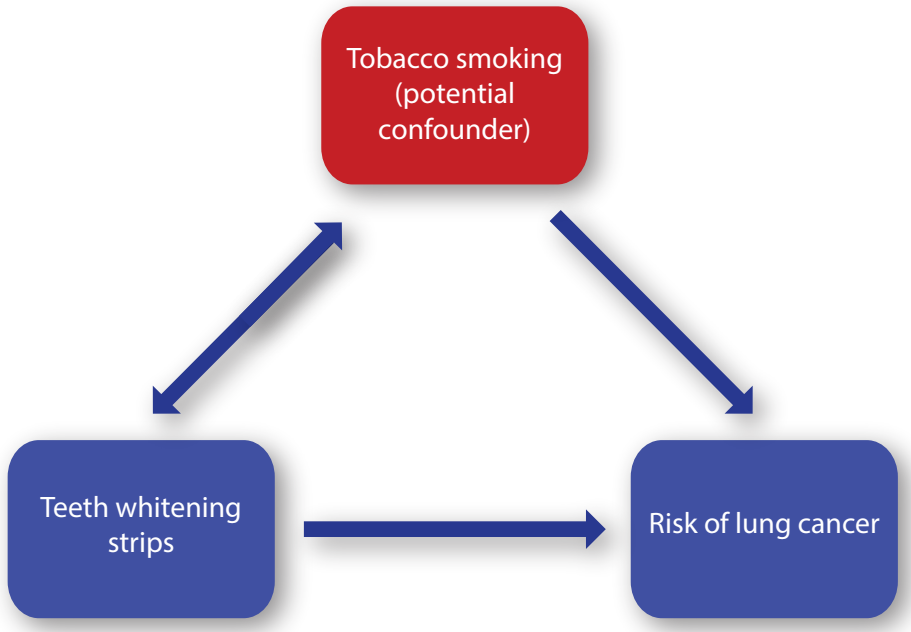
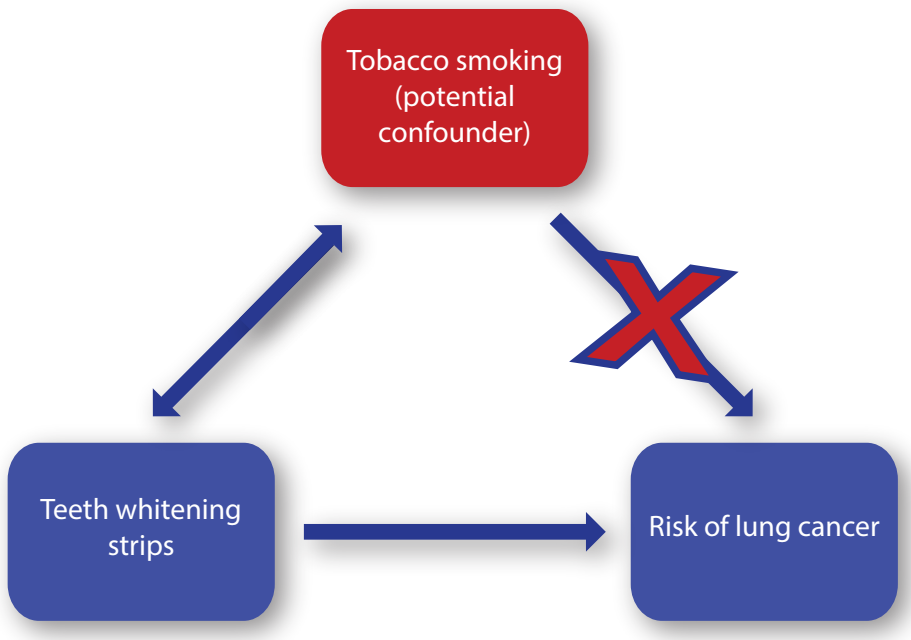


FIGURE 5. **Teeth whitening strips are not a confounder in this example of the potential association of tobacco smoking and the risk of developing lung cancer**



to completely remove unknown and unmeasured confounding variables from a study. Randomization, however, can be cost-prohibitive to a study. Observational studies are more likely to fall victim to the effect of confounding variables due to the lack of randomization and lack of objective inclusion criteria within the study.[6] A study may have more than one confounding variable identified, and these variables may be a mix of measurable and non-measurable covariates.

When planning an observational study, investigators should think of potential confounding variables and measure them whenever possible. Participants could then be matched either individually or in groups on the factor suspected to be a confounder to minimize bias.

Confounding variables are often identified during data analysis. During data analysis, those variables that are identified and measured can be controlled or adjusted for (e.g., using regression techniques, which will be discussed in the next article of this series). Additionally, variables can be stratified during the data analysis.[1,8] Age is one potential confounding variable that can easily be measured and controlled for in a study using stratification.[6] For example, subjects may be grouped by age range; e.g., 18-30 years old, 31-40 years old, and >41 years old. The drawback to using stratification for data analysis is the potential for complicated interpretation of the data. As data are stratified into smaller groups, the 'big picture' is more difficult to interpret in a way that is easily understandable.[6]

Summary

Many experiments attempt to determine causality between two variables. However variables may only be correlated and the criteria for causality should be considered. Additionally, confounding variables can skew a study's results by biasing the relationship between an independent variable and a particular outcome. Confounders in studies can be controlled through prior identification and minimization of effect through study design and analytical techniques. Pharmacists need to be critical readers of the literature and be vigilant for potential bias due to confounding in published studies. ●

Practice Questions

1. True or False: Assuming exercise has been correlated with a decreased risk of developing cardiovascular disease, the expected r-value would have a positive value due to the beneficial effects of exercising.
   True
   False

2. Criteria for confounding variables include all of the following except.
   a. Must be an independent risk factor for the outcome of interest
   b. Cannot be on the causal pathway between an exposure and outcome
   c. A strong relationship exists between the two variables
   d. The variables must be associated with the exposure

3. True or False: Randomization is one way to account for confouding variables during study design.
   True
   False

Answers:

1. **False** Since exercise is correlated with a decreased risk of developing cardiovascular disease, the r-value is expected to be negative.

2. **c** The strength of the relationship is a criteria for causality, not confounding (Table 1).

3. **True** Randomization is the preferred method to minimize the effects of confounders.

Jordan Rush is a Senior Pharmacy Administrative Resident at the University of Wisocnsin Hospital and Clinics in Madison, WI. Maria Ajami is a Senior Pharmacy Administration Resident at the University of Wisconsin Hospital and Clinics in Madison, WI. Kevin Look is an Assistant Professor in the Social & Administrative Sciences Division at the University of Wisconsin School of Pharmacy, Madison, WI. Amanda Margolis is a Lecturer at the UW-Madison School of Pharmacy and a Clinical Pharmacist at the William S. Middleton Memorial Veterans Hospital, Madison, WI..

References and suggestions for further review

1. Shapiro S. Causation, bias and confounding: a hitchhiker's guide to the epidemiological galaxy. Part 2: Principles of causality in epidemiological research: confounding, effect modification and strength of association. J Fam Plann Reprod Health Care. 2008; 34(3): 185-190.
2. Shapiro S. Causation, bias and confounding: a hitchhiker's guide to the epidemiological galaxy. Part 1: Principles of causality in epidemiological research: time order, specification of the study base and specificity. J Fam Plann Reprod Health Care. 2008; 34(2): 83-87.
3. Taylor R. Interpretation of the correlation coefficient: a basic review. JDMS. 1990;1:35-39.
4. Urdan T. Correlation. In: Statistics in Plain English. 3rd ed. New York, NY: Taylor & Francis Group;2010:79-92.
5. Shadish WR, Cook TD, Campbell DT. Experiments and generalized causal inference. In: Experimental and quasi-experimental designs for generalized causal inference. 1st ed. Boston, MA: Houghton Mifflin; 2001:1-12.
6. Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet 2002;359:248-52.
7. Gordis L. More on causal inferences: bias, confounding, and interaction. In: Epidemiology. 4th ed. Philadelphia, PA: Saunders Elsevier; 2009:247-261.
8. Shapiro S. Causation, bias and confounding: a hitchhiker's guide to the epidemiological galaxy. Part 3: Principles of causality in epidemiological research: statistical stability, dose- and duration-response effects, internal and external consistency, analogy and biological plausibility. J Fam Plann Reprod Health Care. 2008; 34(4): 261-264.