BY I. ELAINE ALLEN
AND JULIA E. SEAMAN

Know There Are Unknowns

How suspected confounding variables influence models

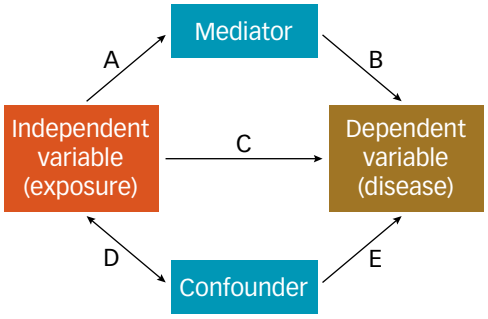
DO THE ITEMS people carry in their pockets or purses make them more likely to develop cancer?

In an epidemiological study on the development of lung cancer in the population, is carrying matches an important variable?[1] No, but it is an indication of a variable that, if not measured, may confound the results. The confounder is whether the individual smokes and does not carry matches, although there probably is a strong relationship between the two.

Here's another question: What caused London's cholera epidemic in 1854? John Snow, the first person to recognize that cholera was caused by waterborne and not airborne bacteria, mapped the locations of the epidemic's victims. From there, he determined the common factor in people contracting cholera was their use of a well on Broad Street. Snow removed the well's pump handle so no one could access the water, and the epidemic died out.

The pump handle and well were not the cause of the epidemic, but it was a confounder in Snow's analysis of cholera. By stratifying the population by whether they were drinking from the well, Snow showed the disease pathway as being waterborne, not airborne.[2]

Pathways / FIGURE 1



One more question: What causes waves of increases in repairs and returns for manufactured products? International companies monitoring product quality noticed repairs and returns increased corresponding to certain batches. The batches corresponded to calendar days that were holidays in the countries where the manufacturing plants were located. Regularly scheduled, experienced workers familiar with manufacturing the products often would take vacation days on the holidays, and less-experienced replacements would create batches of products, which were more prone to errors.

Identifying and incorporating known confounders is relatively simple. Accounting for unknown, but suspected, confounders is more difficult.

What is it?

By definition, a confounder is a variable or factor related to the outcome, but it is not part of the direct causal relationship. It also produces a differential effect based on the observed independent variables.

Confounders are somewhat similar to bias. Unlike confounders, however, bias always involves some type of measurement error that can affect the results by overestimating variability or producing results in particular direction. A confounder, if not identified, will result in a misinterpretation of results.

What isn't a confounder? A factor that is a variable directly related to the outcome but that does not produce a differential result based on the other independent or dependent variable. This is shown as path C in Figure 1.

For example, if a variable lies on the disease pathway (path C) but also produces an effect through a factor (paths A and B), that intermediate factor is not a confounder but, rather, a mediator.[3] In contrast, a confounding variable may be affected by and differentially affect the measured independent variable (path D) while also influencing the dependent variable (path E). A confounder and independent variable are not completely independent, while a mediator is dependent on its independent variable.

Serum cholesterol levels show a direct link to the risk of developing heart disease and directly relate to the disease pathway. Within this model, however, there is the possible confounding variable of diet. Diet can have a downstream effect on serum cholesterol levels and a separate effect on the disease outcome. When including diet, serum cholesterol levels become a mediator variable because there are multiple biological steps between food consumption and the change of serum levels, and the levels do not have any effect on diet.

Extraneous variables may be accounted for but not controlled. If you are measuring your amount of sleep and how food and beverages affect it, but you are awakened during the night by dogs barking, the dogs are an extraneous variable because they do not affect or mediate through the independent variables, and they are not controllable as part of your experiment. They are undesirable variables because they add error to your experiment.

One way statisticians view confounders is through conditional probability. Using Bayes theorem, you can evaluate whether variable B is confounding and affects the outcome of variable A when they occur together. You can also evaluate whether they are independent. Here, the question being answered is whether the probability of A only, P(A), changes when B occurs, P(A|B), and how that is quantified. If the two events or variables are independent, P (A|B) = P(A). Because the confounding variable, B, affects A, P (A|B) will be different than P(A), and the magnitude depends on the strength of the relationship.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Known or unknown

Dealing with confounding can be relatively easy if, as in the case of smoking and lung cancer, you know the likely confounder. An easy solution is to stratify your results and analyze the data set with confounders (smokers) and without (nonsmokers) separately, or you could use a statistical technique to adjust for confounding by the amount of cigarettes an individual smokes.

Dealing with unknown confounders is trickier. There is an apparent association between a risk factor, or an intervention, and an outcome being influenced by an unknown confounder. While this is particularly true of observational studies, it surfaces in all types of experimental designs.

One obvious way to try and eliminate the effect of an unknown confounder is randomization. This ensures that known and unknown confounders are randomly distributed between treatment groups and may minimize the effect.

But if the unknown confounder has a strong influence on the outcome variable, such as in smoking and lung cancer, the outcome may be masked. The true relationship may exist but is found to be nonsignificant because the confounder is differentially affecting independent and dependent variables.

Advantages of randomization are that it is relatively quick and easy, and it doesn't affect the analysis of the results in that no special statistical analyses must be performed. Also, it may not affect the sample size of the experiment. A disadvantage is that it is simple, so it may not capture all the unknown confounding and may bias overall results.

DoE to eliminate confounding

Randomization is a powerful way to eliminate some types of confounding, but incorporating good design of experiment (DoE) methods also can lower the possibility of a confounding variable. In a factory setting, the same materials may come from various suppliers, but there will be variability inherent in the supplies.

For example, there are several machine operators or assemblers working on the same product. They also introduce variability into the defect rate of products. You can test two or more operators under the same conditions to create the product.

When you examine the products from two or more operators, you may see significant differences in quality of a product, but there may be a confounding interaction between the supplier and operator that has not been captured in the data.

When stratified by the supplier for each test in Table 1, it is easy to see that while operator B has a significantly larger defect rate (16% overall for B vs. 12% overall for A), this rate is being confounded by the supplier because it is having an impact on the operator (B uses more supplies from supplier 3) and the defect rate (supplier 3 has more defective materials).

Therefore, to accurately test operator efficiencies, the confounder must be included in the experimental design as a factor. Another way to handle this for modeling the defect rate is to include an interaction effect of operator X supplier in the model.

Advantages are that this design can be specified in advance and can be incorporated into a randomization scheme for the study. A disadvantage is that it may require a larger sample size to have

Operators' defect rates / TABLE 1

Test	Operator	Overall defect rate	Supplier	Defect rate
1	A	12%	1	3%
2	A		1	2%
3	A		3	7%
1	B	16%	2	3%
2	B		3	7%
3	B		3	6%

In a factory setting, the **same materials may come from various suppliers,** but there **will be variability inherent** in the supplies.

sufficient data to analyze each possible confounder—in this instance, each of three suppliers—and to incorporate interaction terms. Also, the more complex the model, the more difficult it can be to interpret.

Matching to control

If randomization or a true controlled DoE is not possible, matching or a case control study can eliminate known (and perhaps unknown) confounders.

Matching addresses issues of confounding in the design stage of the study and may allow for fewer independent variables or interaction terms in the analysis stage. It has been shown to be more efficient than including many covariates to control for confounders when the matching criteria strongly affect the outcome and the independent variables.

Individual matching: In this type of matching, controls are matched to cases on one or more attributes, such as age, gender, duration of disease and smoking status. Each case or control pair has identical values on the matching factors. This requires more complex analysis than unmatched data because each matched set can be viewed as an individual data point for analyses. Here, a paired analysis is appropriate.

Frequency matching: Here, matching is on "cell" instead of on an individual case. A cell is a summarized group of observations in a tabulation of data.

For example, frequency matching may be done on age and sex distributions. If 25% of cases are 20 to 30-year-old females, then controls are selected so 25% are also 20 to 30 years old and female. This does not require using a paired analysis because you take a random sample of controls in that group or cell (20 to 30-year-old females).

The downside is that the investigator may not know the distribution of cells in advance, and controls can only be assigned after the study is complete.

Advantages of this technique are that it controls confounders even in small matched samples, although frequency matching may require larger sample sizes. The biggest disadvantage is that it is time consuming and, while the sample size can be small, finding appropriate matches can be difficult.

Knowing it's there

While it can be difficult to predict and correct for confounding variables, knowing they exist and that they can influence the models is as important as actually finding them.

Although most models are careful to avoid bias, confounding variables can lead to misleading or even false conclusions. Bias involves systematic measurement error, while confounding involves a missing piece in the middle of the puzzle.

The best way to eliminate the effect of confounders is to know and incorporate as much as possible about the mechanism and relationship between one variable and the others, both independent and dependent.

This means understanding, for example, how a biological disease manifests and interacts with a body, or how the supply chain, operators and final product are intertwined. QP

REFERENCES

1. "Bias and Accounting," College of Emergency Medicine, www.collemergencymed.ac.uk/cem/research/technical_guide/biasconfound.htm.
2. "Epidemiology—Trying to Establish Cause," Kimball's Biology Pages, http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/E/Epidemiology.html (case sensitive).
3. K.W. Murray, "Understanding Confounding in Research," *Pediatrics in Review*, Vol. 31, No. 3, 2010, pp. 124-126.

BIBLIOGRAPHY

Benedetti, Ettore, et al., "Criteria for the Design and Biological Characterization of Radiolabeled Peptide-Based Pharmaceuticals," *Drug Development,* Vol. 18, No. 5, 2004, pp. 279-295.

Fewell, Zoe, et al., "The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies: A Simulation Study," *American Journal of Epidemiology,* Vol. 166, No. 6, 2007, pp. 646-655.

Joby, John, "Improving Quality Through Patient-Provider Communication," *Journal of Health Care Marketing*, Vol. 11, No. 4, 1991, pp. 51-60.

Johnston, S.C., "Identifying Confounding by Indication through Blinded Prospective Review," *American Journal of Epidemiology,* Vol. 154, No. 3, 2001, pp. 276-284.

Schlesselman, J.J., "Assessing Effects of Confounding Variables," *American Journal of Epidemiology*, Vol. 108, No. 1, 1978, pp. 3-8.

Mento, A.J., and R.P. Steel, "Conducting Quality Circles Research: Toward a Comprehensive Perspective," *Public Productivity Review*, Spring 1985.

Szatmari, Peter, et al., "Conducting Genetic Epidemiology Studies of Autism Spectrum Disorders: Issues in Matching," *Journal of Autism and Developmental Disorders*, Vol. 34, No. 1, 2004, pp. 49-57.

Weinberg, C.R., "Toward a Clearer Definition of Confounding," *American Journal of Epidemiology*, Vol. 137, No. 1, 1993, pp. 1-8.

I. ELAINE ALLEN is research director of the Arthur M. Blank Center for Entrepreneurship, director of the Babson Survey Research Group, and professor of statistics and entrepreneurship at Babson College in Wellesley, MA. She earned a doctorate in statistics from Cornell University in Ithaca, NY. Allen is a member of ASQ.

JULIA E. SEAMAN is a doctoral student in pharmacogenomics at the University of California, San Francisco, and a statistical consultant for the Babson Survey Research Group at Babson College. She earned a bachelor's degree in chemistry and mathematics from Pomona College in Claremont, CA.