Gastroenterology and Hepatology From Bed to Bench

MEDICAL EDUCATION

How to control confounding effects by statistical analysis

Mohamad Amin Pourhoseingholi[1], Ahmad Reza Baghestani[2], Mohsen Vahedi[3]

[1] *Department of Biostatistics, Shahid Beheshti University of Medical Sciences, Tehran, Iran.*

[2] *Department of Mathematic, Islamic Azad University - South Tehran Branch, Iran.*

[3] *Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran*

ABSTRACT

A Confounder is a variable whose presence affects the variables being studied so that the results do not reflect the actual relationship. There are various ways to exclude or control confounding variables including Randomization, Restriction and Matching. But all these methods are applicable at the time of study design. When experimental designs are premature, impractical, or impossible, researchers must rely on statistical methods to adjust for potentially confounding effects. These Statistical models (especially regression models) are flexible to eliminate the effects of confounders.

Keywords: Confounders, Statistical models, Adjustment.

(Please cite as: **Pourhoseingholi MA, Baghestani AR, Vahedi M. How to control confounding effects by statistical analysis. Gastroenterol Hepatol Bed Bench 2012;5(2):79-83.)**

Introduction

Confounding variables or confounders are often defined as the variables correlate (positively or negatively) with both the dependent variable and the independent variable (1). A Confounder is an extraneous variable whose presence affects the variables being studied so that the results do not reflect the actual relationship between the variables under study.

The aim of major epidemiological studies is to search for the causes of diseases, based on associations with various risk factors. There may be also other factors that are associated with the exposure and affect the risk of developing the disease and they will distort the observed association between the disease and exposure under study. A hypothetical example would be a study of relation between coffee drinking and lung cancer. If the person who entered in the study as a coffee drinker was also more likely to be cigarette smoker, and the study only measured coffee drinking but not smoking, the results may seem to show that coffee drinking increases the risk of lung cancer, which may not be true. However, if a confounding factor (in this example, smoking) is recognized, adjustments can be made in the study design or data analysis so that the effects of confounder would be removed from the final results. Simpson's paradox too is another classic example of confounding (2). Simpson's paradox refers to the reversal of the direction of an association when data from several groups are combined to form a single group.

The researchers therefore need to account for these variables - either through experimental design and before the data gathering, or through statistical analysis after the data gathering process. In this case the researchers are said to account for their effects to avoid a false positive (Type I) error

Received: 1 January 2012 *Accepted:* 15 February 2012
Reprint or Correspondence: Mohamad Amin Pourhoseingholi, PhD. Department of Biostatistics, Shahid Beheshti University of Medical Sciences, Tehran, Iran
E-mail: aminphg@gmail.com

(a false conclusion that the dependent variables are in a casual relationship with the independent variable). Thus, confounding is a major threat to the validity of inferences made about cause and effect (internal validity). There are various ways to modify a study design to actively exclude or control confounding variables (3) including Randomization, Restriction and Matching.

In randomization the random assignment of study subjects to exposure categories to breaking any links between exposure and confounders. This reduces potential for confounding by generating groups that are fairly comparable with respect to known and unknown confounding variables.

Restriction eliminates variation in the confounder (for example if an investigator only selects subjects of the same age or same sex then, the study will eliminate confounding by sex or age group). Matching which involves selection of a comparison group with respect to the distribution of one or more potential confounders.

Matching is commonly used in case-control studies (for example, if age and sex are the matching variables, then a 45 year old male case is matched to a male control with same age).

But all these methods mentioned above are applicable at the time of study design and before the process of data gathering. When experimental designs are premature, impractical, or impossible, researchers must rely on statistical methods to adjust for potentially confounding effects (4).

Statistical Analysis to eliminate confounding effects

Unlike selection or information bias, confounding is one type of bias that can be, adjusted after data gathering, using statistical models. To control for confounding in the analyses, investigators should measure the confounders in the study. Researchers usually do this by collecting data on all known, previously identified confounders. There are mostly two options to dealing with confounders in analysis stage; Stratification and Multivariate methods.

1. Stratification

Objective of stratification is to fix the level of the confounders and produce groups within which the confounder does not vary. Then evaluate the exposure-outcome association within each stratum of the confounder. So within each stratum, the confounder cannot confound because it does not vary across the exposure-outcome.

After stratification, Mantel-Haenszel (M-H) estimator can be employed to provide an adjusted result according to strata. If there is difference between Crude result and adjusted result (produced from strata) confounding is likely. But in the case that Crude result dose not differ from the adjusted result, then confounding is unlikely.

2. Multivariate Models

Stratified analysis works best in the way that there are not a lot of strata and if only 1 or 2 confounders have to be controlled. If the number of potential confounders or the level of their grouping is large, multivariate analysis offers the only solution.

Multivariate models can handle large numbers of covariates (and also confounders) simultaneously. For example in a study that aimed to measure the relation between body mass index and Dyspepsia, one could control for other covariates like as age, sex, smoking, alcohol, ethnicity, etc in the same model.

2.1. Logistic Regression

Logistic regression is a mathematical process that produces results that can be interpreted as an odds ratio, and it is easy to use by any statistical package. The special thing about logistic regression is that it can control for numerous confounders (if there is a large enough sample size). Thus logistic regression is a mathematical model that can give an odds ratio which is

controlled for multiple confounders. This odds ratio is known as the adjusted odds ratio, because its value has been adjusted for the other covariates (including confounders).

2.2. Linear Regression

The linear regression analysis is another statistical model that can be used to examine the association between multiple covariates and a numeric outcome. This model can be employed as a multiple linear regression to see through confounding and isolate the relationship of interest (5). For example, in a research seeking for relationship between LDL cholesterol level and age, the multiple linear regression lets you answer the question, "How does LDL level vary with age, after accounting for blood sugar and lipid (as the confounding factors)? In multiple linear regression (as mentioned for logistic regression), investigators can include many covariates at one time. The process of accounting for covariates is also called adjustment (similar to logistic regression model) and comparing the results of simple and multiple linear regressions can clarify that how much the confounders in the model distort the relationship between exposure and outcome.

2.3. Analysis of Covariance

The Analysis of Covariance (ANCOVA) is a type of Analysis of Variance (ANOVA) that is used to control for potential confounding variables**.** ANCOVA is a statistical linear model with a continuous outcome variable (quantitative, scaled) and two or more predictor variables where at least one is continuous (quantitative, scaled) and at least one is categorical (nominal, non-scaled). ANCOVA is a combination of ANOVA and linear regression. ANCOVA tests whether certain factors have an effect on the outcome variable after removing the variance for which quantitative covariates (confounders) account. The inclusion of this analysis can increase the statistical power.

Practical example

Suppose that, in a cross-sectional study, we are seeking for the relation between infection with *Helicobacter*.*Pylori* (HP) and Dyspepsia Symptoms. The study conducted on 550 persons with positive H.P and 440 persons without HP. The results are appeared in 2*2 crude table (table 1) that indicated that the relation between infection with H.P and Dyspepsia is a reverese association (OR=0.60, 95% CI: 0.42-0.94). Now suppose that weight can be a potential confounder in this study. So we break the crude table down in two stratum according to the weight of subjects (normal weight or over weight) and then calculate OR's for each stratum again. If stratum-specific OR is similar to crude OR, there is no potential impact from confounding factors. In this example there are different OR for each stratum (for normal weight group OR= 0.80, 95% CI: 0.38-1.69 and for overweight group OR= 1.60, 95% CI: 0.79-3.27).

This shows that there is a potential confounding affects which is presented by weight in this study. This example is a type of Simpson's paradox, therefore the crude OR is not justified for this study. We calculated the Mantel-Haenszel (M-H) estimator as an alternative statistical analysis to remove the confounding effects (OR= 1.16, 95% CI: 0.71-1.90). Also logistic regression model (in which, weight is presented in multiple model) would be conducted to control the confounder, its result is similar as M-H estimator (OR= 1.15, 95% CI: 0.71-1.89).

The results of this example clearly indicated that if the impacts of confounders did not account in the analysis, the results can deceive the researchers with unjustified results.

Table 1. The crude contingency table of association between H.Pylori and Dyspepsia

	Dyspepsia (positive)	**Dyspepsia (negative)**
H.Pylori (positive)	50	500
H.Pylori (negative)	60	380

Table 2. The contingency table of association between H. Pylori and Dyspepsia for person who are in normal weight group

	Dyspepsia (positive)	Dyspepsia (negative)
H.Pylori (positive)	10	50
H.Pylori (negative)	50	200

Table 3. The contingency table of association between H. Pylori and Dyspepsia for person who are in over weight group

	Dyspepsia (positive)	Dyspepsia (negative)
H.Pylori (positive)	40	450
H.Pylori (negative)	10	180

Conclusion

Confounders are common causes of both treatment/exposure and of response/outcome. Confounding is better taken care of by randomization at the design stage of the research (6).

A successful randomization minimizes confounding by unmeasured as well as measured factors, whereas statistical control that addresses confounding by measurement and can introduce confounding through inappropriate control (7-9).

Confounding can persist, even after adjustment. In many studies, confounders are not adjusted because they were not measured during the process of data gathering. In some situation, confounder variables are measured with error or their categories are improperly defined (for example age categories were not well implied its confounding nature) (10). Also there is a possibility that the variables that are controlled as the confounders were actually not confounders.

Before applying a statistical correction method, one has to decide which factors are confounders. This sometimes is a complex issue (11-13). Common strategies to decide whether a variable is a confounder that should be adjusted or not, rely mostly on statistical criteria. The research strategy should be based on the knowledge of the field and on conceptual framework and causal model. So expertise' criteria should be involved for evaluating the confounders. Statistical models (especially regression models) are a flexible way of investigating the separate or joint effects of several risk factors for disease or ill health (14). But the researchers should notice that wrong assumptions about the form of the relationship between confounder and disease can lead to wrong conclusions about exposure effects too.

References

1. Elwood JM, ed. Causal Relationships in Medicine. Oxford: Oxford University Press; 1988. P.332.

2. Agresti A, ed. An introduction to categorical data analysis. New Jersey : Wiley ; 2007. P.51.

3. Mayrent SL, ed. Epidemiology in Medicine. New York: Lippincott Williams & Wilkins; 1987.

4. Christenfeld NJ, Sloan RP, Carroll D, Greenland S. Risk factors, confounding, and the illusion of statistical control. Psychosom Med 2004; 66:868-75.

5. Maldonado G, Greenland S. Simulation study of cofounder-selection strategies. Compares a number of data based strategies for selecting variables to include in regression models when the aim is to control confounding. Am J Epidemiol 1993; 138:923–36.

6. Wunsch G. Confounding and control. Demographic Research 2007; 16:97-120.

7. Greenland S. Quantifying biases in causal models: classical confounding vs. collider-stratification bias. Epidemiology 2003; 14:300-6.

8. Cole SR, Hernan MA. Fallibility is estimating direct effects. Int J Epidemiol 2002; 31:163-65.

9. Greenland S, Brumback BA. An overview of relations among causal modelling methods. Int J Epidemiol 2002; 31:1030-37.

10. Blair A, Stewart P, Lubin JH, Forastiere F. Methodological issues regarding confounding and exposure misclassification in epidemiological studies of occupational exposures. Am J Ind Med 2007; 50:199-207.

11. McNamee R. Confounding and confounders. Contrasts competing definitions of a confounder, including those based on data and those based on

notions of comparability. Occup Environ Med2003; 60:227-34.

12. Greenland S, Morgenstern H. Confounding in health research. Annu Rev Public Health 2001; 22:189-212.

13. Greenland S, Pearl J, Robins JM. The problem of identifying confounders of an exposure-disease relationship is addressed through causal diagrams. Causal diagrams for epidemiological research. Epidemiology1999; 10:37-47.

14. McNamee R. Regression modelling and other methods to control confounding. Occup Environ Med 2005; 62:500-506.