

# Adjusting for Confounding

Garrett Fitzmaurice, ScD

From the Department of Biostatistics, Harvard School of Public Health,  
Boston, Massachusetts, USA

## INTRODUCTION

My previous column<sup>1</sup> dealt with the concept of confounding. An important consideration when examining the association between any two variables (e.g., an exposure of interest and a particular disease) is whether the apparent association (or, perhaps, the lack thereof) may be due, at least in part, to the effects of a third variable related to the other two. This phenomenon is referred to as *confounding*, and the third variable that obscures the association of real scientific interest is referred to as a *confounder*. In this column I discuss a tried and trusted method of adjusting for confounding in the analysis: *stratification*.

Figure 1 presents (hypothetical) data from a study that examined the association between vitamin E supplementation and the risk of occurrence of coronary heart disease (CHD). The data are from 1000 participants between the ages of 50 and 55 y who were initially free of CHD and followed for 15 y. Those who developed CHD during the 15-y follow-up were identified; study participants also were classified according to their use of vitamin E supplements (the “exposure” variable). Analysis of the data shown in Figure 1 indicates that the risk of CHD for those who used vitamin E supplements is approximately 0.09 (or 50 of 551) and that the corresponding risk for those who did not is approximately 0.15 (or 65 of 449). This apparent association can be expressed in terms of the odds ratio<sup>2</sup> (OR):

$$OR = (50)(384)/(501)(65) = 0.59$$

This estimated OR suggests that the risk or, more specifically, the odds of developing CHD, is almost halved by vitamin E supplementation. However, a potential problem with this analysis, and the resulting conclusion, is that the two groups may differ in ways other than their use of vitamin E supplements. For example, individuals who use vitamin E supplements may be much less likely to smoke. If that were the case, then there must be some concern that the observed association shown in Figure 1 is due at least in part to the *confounding* effects of smoking because smoking is a well-known major risk factor for CHD.

To address this concern, the data presented in Figure 1 are *stratified* according to whether or not a study participant is a smoker (Figure 2). Focusing on the stratified data shown in Figure 2, the odds of developing CHD among those who use vitamin E supplements, relative to those who do not, can be estimated separately for smokers and non-smokers. For smokers the estimated OR is

$$OR (\text{smokers}) = (11)(200)/(40)(49) = 1.12$$

for non-smokers the estimated OR is

$$OR (\text{non-smokers}) = (39)(184)/(461)(16) = 0.97$$

Correspondence to: Garrett Fitzmaurice, ScD, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston MA 02115, USA. E-mail: fitzmaur@hsph.harvard.edu

		CHD	
		Present	Absent
Vitamin E Supplement	Yes	50	501
	No	65	384

FIG. 1. Hypothetical data on the association between vitamin E supplementation and the risk of CHD.

These results indicate that there is little evidence of an association between vitamin E supplements and risk of CHD after controlling for the effects of smoking history. Why is there an apparent contradiction with the results obtained from the non-stratified or collapsed data shown in Figure 1? The reason for the paradoxical findings is that the vitamin E supplement group contains considerably fewer smokers (9.3% versus 55.5%), thereby decreasing this group's apparent risk of CHD.

## ADJUSTING FOR CONFOUNDING IN THE ANALYSIS

The additional analysis, based on the data shown in Figure 2, illustrates one of the most effective techniques for adjusting for the effects of confounding in the analysis: *stratification*. That is, the effects of confounding can be controlled through stratification on levels of the potential confounder. Specifically, confounding is controlled when the association of interest is evaluated within distinct groups, or strata, constituted by individuals who are relatively homogeneous with respect to the confounding variable.

For example, when we stratify the analysis by smoking history, comparison of the odds of CHD within any strata cannot be confounded by smoking because the strata are, by definition, constituted by individuals with the same smoking history. Thus, each of the stratum-specific estimates of the OR is free of the potential confounding effects of smoking. Although these separate, unconfounded, estimates of association could be reported, it is generally more appealing to provide a single, overall, summary estimate of association that is adjusted for the effects of the confounder.

When the stratum-specific estimates of association are uniform across the levels of the potential confounder, there are several methods for combining these estimates into a summary measure of *adjusted* association. As we will see, these methods are based on the simple notion of taking a weighted average of the stratum-specific estimates. However, in cases where the stratum-specific estimates of association are discernibly different (statistical tests for homogeneity of the stratum-specific estimates are available),

**Smokers:**

		CHD	
		Present	Absent
Vitamin E Supplement	Yes	11	40
	No	49	200

**Non-Smokers:**

		CHD	
		Present	Absent
Vitamin E Supplement	Yes	39	461
	No	16	184

FIG. 2. Hypothetical data on the association between vitamin E supplements and the risk of CHD, stratified by smoking history.

the results should not be combined; instead, the way in which the association differs by levels of the stratifying variable should be reported.

**MANTEL-HAENSZEL METHOD**

One of the most widely used methods for combining the stratum-specific estimates of association is the Mantel-Haenszel (MH) method.<sup>3</sup> The MH method calculates a pooled, summary measure of association by taking a weighted average of the stratum-specific estimates, with weights that are proportional to the sample size within each stratum. More accurately, the weights are inversely proportional to the variances of the stratum-specific estimates of the OR, thereby giving greater weight to the more precise estimates.

Next we consider the calculation of the pooled estimate of the OR. Suppose there are *K* levels of the stratifying variable (e.g., *K* = 2 for the data presented in Figure 2). The data can then be summarized in terms of *K* two-by-two contingency tables. The four internal cells of each table contain frequency counts of the number of individuals having a particular combination of the two variables. Figure 3 shows such a table, where the rows correspond to an exposure (e.g., vitamin E supplementation) and the columns correspond to disease status (e.g., presence or absence of CHD). The MH formula for the pooled estimate of the OR from a series of *K* two-by-two contingency tables is given by

		Disease	
		Yes	No
Exposure	Yes	<b>a<sub>k</sub></b>	<b>b<sub>k</sub></b>
	No	<b>c<sub>k</sub></b>	<b>d<sub>k</sub></b>

FIG. 3. Stratum-specific two-by-two contingency table.

$$OR_{MH} = \frac{\sum_{k=1}^K (a_k d_k) / n_k}{\sum_{k=1}^K (b_k c_k) / n_k}$$

where *n<sub>k</sub>* is the total number of observations in the *k*th table (i.e., *n<sub>k</sub>* = *a<sub>k</sub>* + *b<sub>k</sub>* + *c<sub>k</sub>* + *d<sub>k</sub>*) and the summation in the formula is taken over the *K* levels of the stratification variable.

For example, the MH pooled estimate of the OR based on the data from the two contingency tables shown in Figure 2 is

$$OR_{MH} = \frac{(11 \times 200) / 300 + (39 \times 184) / 700}{(40 \times 49) / 300 + (461 \times 16) / 700} = 1.03.$$

As expected for the data shown in Figure 2, the pooled estimate of the OR is approximately 1.0 and indicates that there is no evidence of association between vitamin E supplements and risk of CHD after controlling for the effects of smoking history. A confidence interval for the adjusted OR also can be calculated. The limits of the 95% confidence interval for the adjusted OR are (0.61 to 1.65) and include the null value of 1.0 for the OR. The formula for constructing the confidence interval is somewhat complicated; calculation of the confidence interval for the adjusted OR is best handled with the aid of a computer.

**CONCLUSION**

In summary, determining the association between an exposure and a disease is not quite so straightforward as it first appears. When assessing the evidence for an association between the two variables, the effect of potential confounders, including those that were measured and those that were not, must be considered before drawing any definitive conclusions about the association. Confounding that cannot be controlled in the design of a study must be adjusted for in the analysis.

In this column I have described one of the most effective means of controlling for confounding in the analysis, namely stratification on levels of the confounder. I also have described the MH procedure for obtaining a summary measure of adjusted association from the stratum-specific estimates. The MH procedure simply takes a weighted average of the unconfounded, stratum-specific estimates of association. Although stratification is a remarkably robust method of adjusting for confounding, it is less appealing when there are many potential confounding variables, resulting in strata with too few individuals to make meaningful comparisons. An alternative approach for adjusting for confounding in the analysis is to examine the exposure effect in a regression model for the dependence of the disease outcome on the exposure of interest and any potential confounders. For example, an adjusted estimate

of the association between vitamin E supplements and the risk of CHD might be obtained from a logistic regression model<sup>4</sup> for the dependence of the log odds of CHD on use of vitamin E supplements, while controlling from smoking history and other potential confounders (e.g., gender and physical activity) through their inclusion as predictors in the model. Adjusting for confounding via regression models will be the topic of a future column.

## REFERENCES

1. Fitzmaurice G. Confused by confounding? *Nutrition* 2003;19:189
2. Pagano M. Odds, risks and their relatives. *Nutrition* 1995;11:473
3. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959;22:719
4. Pagano M. Logistic regression. *Nutrition* 1996;12:135