



FOCUS on Field Epidemiology

CONTRIBUTORS

Author:

David Bergmire-Sweat, MPH
Amy Nelson, PhD, MPH
FOCUS Workgroup*

Reviewers:

FOCUS Workgroup*

Production Editors:

Tara P. Rybka, MPH
Lorraine Alexander, DrPH
Rachel A. Wilfert, MD, MPH

Editor in chief:

Pia D.M. MacDonald, PhD, MPH

* All members of the FOCUS Workgroup are named on the last page of this issue.

Advanced Data Analysis: Methods to Control for Confounding (Matching and Logistic Regression)

It is unwise to be too sure of one's own wisdom. It is healthy to be reminded that the strongest might weaken and the wisest might err.
—Mahatma Gandhi (1869 - 1948)

Introduction

Amazing. Confusing. Bewildering. Puzzling. Mystifying. Perplexing. Stunning. Confounding.

What do you think of when you read these words? In your endeavors to achieve a higher plane in the practice of epidemiology, you may think of a brilliant sage studying an age-old problem. Or you may think of the words of an ancient text or the history of a lost people. Perhaps you have a lighter heart, and you think of an amazing magician.

In this issue of FOCUS, you will explore the magical capacity of epidemiology to reduce the effects of confounding, granting you the tools to become wiser. Just be sure you do not get cocky about it.

Epidemiology is the foundational science in public health because it uses the power of scientific analysis to provide evidence that an exposure may be associated with developing or preventing a disease. Sometimes, however, the most basic methods of epidemiology are not enough to determine “the causes of happenings,” or whether an exposure is truly associated with a disease. Other exposures or characteristics among the population may be confusing the exposure-disease relationship, a situation known as **confounding**.

For example, it might appear that people in a gastrointestinal outbreak are mostly members of the same dinner club, but many members also went to a city-wide food festival. Food handling practices within the dinner club might be blamed for the outbreak, when in fact, food eaten at the festival was the cause — and a much larger population is at risk of illness. Membership in the dinner club could be a confounder of the relationship between attending the food festival and illness. If we could analyze the data in a way to account for dinner club membership and food festival attendance at the same time, we might be able to determine which of these events was truly associated with the outcome.

In a previous issue of FOCUS, we addressed this type of situation through stratification methods. We could calculate the risk of illness due to the food festival among those who are in the dinner club, and then calculate the risk of illness separately among those who are not in the dinner club. If both groups, or both strata, show that attending the food festival is a significant risk factor for illness, then the food festival can be implicated because illness occurred whether or not people were members of the dinner club.

However, what happens if there are multiple factors that might be confounding the exposure-disease relationship? In the above example, what if we had to stratify by membership in the dinner club *and* by health status (since people with poorer health are more susceptible to illness)? There might be other factors that could be



NORTH CAROLINA
CENTER FOR PUBLIC
HEALTH PREPAREDNESS

The North Carolina Center for Public Health Preparedness is funded by Grant/Cooperative Agreement Number U90/CCU424255 from the Centers for Disease Control and Prevention. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the views of the CDC.

considered potential confounders as well, such as age, occupation, income, exercise, and other lifestyle factors. These other factors may disguise relevant associations or protective effects. Trying to stratify by all these different layers quickly becomes difficult to calculate, and the number of 2x2 tables needed grows unwieldy.

More advanced methods are needed to deal with this type of situation. In this issue of FOCUS, we explore logistic regression and accounting for matched data. Logistic regression is an efficient way to control for many potential confounders at one time. Matching, if done correctly in the study design stage of the investigation, reduces confounding before the analysis even begins.

Confounding Confounders

Most commonly in field epidemiology, we compare two groups by using measures of association, such as the risk ratio (RR) in cohort studies and odds ratio (OR) in case-control studies. If you are not familiar with these measures, please see FOCUS issues covering these topics (*Volume 3 Issue 1* and *Volume 3 Issue 2*) in detail before jumping in here.

In a perfect world, you can collect data, analyze it using a 2x2 table to get a measure of association and confidence interval (CI), or a chi-square and p-value, and the analysis is done. You might know, however, that field epidemiology is rarely this simple.

You may have multiple exposures that are statistically associated with the disease... So how do you know which one is the real culprit? In other situations, you may have *no* exposures associated with the disease when you collected data to test every exposure imaginable. If either situation happens, you need to delve deeper into the data. There may be a confounder lurking there, either making it appear that exposures are associated with the disease when they really are not, or that there is no association between an exposure and the disease when there really is one.

A confounder is a third variable that distorts the risk ratio or odds ratio of an exposure and an outcome. Confounding is a form of bias that can result in a distortion in the measure of association between an exposure and disease and should be eliminated. (1)

Confounding can occur in an observational epidemiologic study whenever two groups are compared to each other. These groups can be an exposed group and an unexposed group as in a cohort study, or a diseased group and a non-diseased group as in a case-control study. Confounding is a "mixing of effects" when comparing these groups to each other. In any exposure-disease relationship examined, other factors can affect both exposure and disease. These third factors could be confounding the relationship between the exposure and disease you are trying to study.

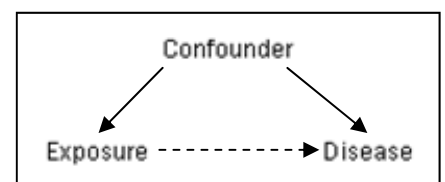
Some common confounders to look out for include age, socioeconomic status, and gender, among others. Here are a few examples:

- Children born later in the birth order (whether born first, second, last, etc) are more likely to have Down's syndrome. Does this mean we can conclude that birth order causes Down's? No, the relationship between birth order and Down's is confounded by mother's age. Older women are more likely to have children with Down's. Older women are also likely to be having children who are late in the birth order (it's a lot more common for a 35-year-old woman to have 6 children than it is for a 21-year-old woman). Mother's age was confounding the association between birth order and Down's syndrome: it looked like there was an association when there really was not. (2)
- Some studies of women's use of hormone replacement therapy (HRT) show no association between HRT use and cardiovascular disease, yet others suggest there is an increased risk. Women of higher socioeconomic status (SES) are more likely to be able to afford HRT, and women of lower SES are generally at higher risk for cardiovascular disease. Differences in SES can be confounding the relationship between HRT and increased risk of cardiovascular disease, that only becomes apparent when controlling for SES among study participants. (3)
- A hypothetical outbreak of gastroenteritis at a restaurant finds that women are at much higher risk of disease than men. However, this association is confounded by eating salad. Women are much more likely to order salad than men, and the salad was contaminated with the agent that caused disease. The relationship between gender and disease was confounded by salad consumption (which was actually the true cause of the outbreak).

If you examine these examples, you will notice that there are two key characteristics of confounders:

- A confounder must be associated with the *disease* being studied, and
- A confounder must be associated with the *exposure* being studied.

In other words, the "triangle" at right must be present for confounding to occur:



Now that you know how to recognize confounding, how do you make it go away? To control for confounding, you must take the confounding variable out of the picture.

There are three ways to do this:

- 1. Restrict the analysis.** Only analyze the exposure to disease relationship among those in one level of the confounding variable. (e.g., Examine the relationship of HRT and cardiovascular disease only among women of high socioeconomic status). The problem with this approach is that you end up throwing out a big chunk of your data (in this example, all women of low SES).
- 2. Stratify.** Analyze the exposure to disease relationship separately in all levels of the confounding variable. (e.g., Examine the relationship of HRT and cardiovascular disease among women of high SES, and separately among women of low SES. SES could even be broken down into multiple levels such as low, average, and high).
- 3. Conduct logistic regression.** Regression is a way of putting all the variables into a mathematical model. Regression is very handy when there are multiple confounders that need to be controlled (maybe SES and gender and age).

Stratification to control for confounding

Stratification can be used to tease out the effects of exposures and confounders on an outcome. For example, in an outbreak of tuberculosis among homeless men, both a homeless shelter and a soup kitchen have been implicated as the place of transmission. The men were likely to spend time in both places. In order to determine which site should be prioritized for intervention, we could examine the association between the homeless shelter and tuberculosis among men who did *not* go to the soup kitchen, and then examine the same relationship separately among men who *did* go to the soup kitchen.

Let’s take an example and work through the numbers. An outbreak has taken place at a reception, and both the cookies and the punch have been implicated in the initial analyses. We suspect that one of these food items is confounding the other, but cannot tease out the effects without stratifying because many people consume both cookies and punch.

After conducting a case-control study, the overall data look like this:

Cookie Exposure

	Cases	Controls	Total
Cookies	37	21	58
No Cookies	13	29	42
Total			100

OR = (37x29)/(21x13) = 3.93; 95% CI, 1.69 – 9.15
 P= 0.001*

Punch Exposure

	Cases	Controls	Total
Punch	40	20	60
No Punch	10	30	40
Total			100

OR = (40x30)/(20x10) = 6.00; 95% CI, 2.83 – 12.71
 P= 0.0004*

Both cookies and punch have a high odds ratio for illness and confidence interval that do not include 1. We can stratify to try and tease out these effects.

To stratify by punch exposure, we want to know:

- Among those who did not drink punch, what is the effect measure (OR) for the association between cookies and illness?
- Among those who did drink punch, what is the effect measure for the association between cookies and illness?

If cookies are the culprit, then there should be an association between cookies and illness, regardless of whether anyone drank punch.

Stratification of the cookie association by punch exposure:

Did have punch

	Cases	Controls	Total
Cookies	35	17	52
No Cookies	5	3	8
Total			60

OR = (35x3)/(17x5) = 1.3; 95% CI, 0.17 – 7.22
 P= 1.0*

Did not have punch

	Cases	Controls	Total
Cookies	2	4	6
No Cookies	8	26	34
Total			40

OR = (2x26)/(4x8) = 1.63; 95% CI, 0.12 – 13.86
 P= 0.63*

To stratify by cookie exposure, we want to know:

*Note: Exact p-values and confidence intervals (CI) were calculated in tables with cell sizes less than 5 using SAS 9.13. Chi square statistical test applied to detect differences between exposed and unexposed groups and whether they were ill/not ill according to the case definition.

- Among those who did not eat cookies, what is the effect measure for the association between punch and illness?
- Among those who did eat cookies, what is the effect measure for the association between punch and illness?

If punch is the culprit, then there should be an association between punch and illness, regardless of whether anyone ate cookies.

Stratification of the punch association by cookie exposure:
 $OR = (35 \times 4) / (17 \times 2) = 4.12$; 95% CI, 0.52 - 48.47

Did have cookies

	Cases	Controls	Total
Punch	35	17	52
No Punch	2	4	6
Total			58

$P = 0.18^*$

$OR = (5 \times 26) / (3 \times 8) = 5.42$; 95% CI, < 0.80 - 40.95

Did not have cookies

	Cases	Controls	Total
Punch	5	3	8
No Punch	8	26	34
Total			42

$P = 0.08^*$

Stratifying allows us to examine two risk factors independently of each other. In this example, we can see that cookies were not really a risk factor independent of punch (stratified OR's ≈ 1), while punch remained a potential risk factor independent of cookies (the odds ratios are large and p-values close to significant). The next step would be to figure out what happened in the purchase, preparation, and storage of that punch that could have made people sick (ie. an environmental health investigation).

Remember those statistics guys, Mantel and Haenszel from the last issue of FOCUS? Well, they devised a clever way to control for confounding using stratified analyses. In the stratified analysis above, we are left with two stratum-specific estimates of the association between punch and illness: 4.1 and 5.4. But it would be much more convenient to have just one estimate; after all, both odds ratios we calculated are estimates of the same "true" odds ratio measuring the association between punch and illness. The Mantel-Haenszel odds ratio takes an association (like our punch and illness), stratifies it by a potential confounder

(like our cookies), and then recombines them by averaging them into one estimate that is "controlled" for the stratifying variable. This recombined odds ratio is the pooled or **common odds ratio**. This is a very handy technique to use when you know you have a confounder but want to present only one point estimate. It also works well when stratifying by variables that have more than two potential values, such as the amount of a food item consumed (two helpings or one bite, or not at all), or the number of times a person went swimming. But you probably would not stratify by each level of a continuous variable since the numbers would be too small. Some exposures are associated with disease, but only once a critical amount of exposure is reached. In those instances, people who were exposed at a low level could be confounding the ability to identify the association because they did not get sick even though they were exposed.

Stratification and Effect Measure Modifiers

We have seen a stratification that results in no association (cookie exposure stratified by punch intake), and a stratification that results in a consistent, strong association (punch exposure stratified by cookie intake). But there is another possible outcome. We could find that one stratum showed no association with an odds ratio near 1.0, while the other stratum showed that there was an association. This occurrence is known as **effect measure modification**. In this case, the third variable is not confounding the association between exposure and outcome and is not a bias we want to eliminate. Instead, we want to identify and present estimates separately for each level. So if gender is an effect measure modifier, you should give 2 odds or risk ratios: 1 for men and 1 for women. We identify effect measure modification by stratification, the same technique we use to identify confounding, but are looking for the measure of effect to be very different (heterogeneous) between the 2 or more strata. For example:

- Among the elderly, gender is an effect modifier of the association between nutritional intake and osteoporosis. Nutritional intake (particularly calcium) is associated with osteoporosis outcome among women. Among men, this association is not so strong because men's bone mineral content is different than women's and is not affected as much by nutritional intake.
- In developing countries, sanitation is an effect modifier of the association between breastfeeding infants and infant mortality. In conditions that are very unsanitary, breastfeeding has a strong effect in reducing infant mortality. In conditions that are much cleaner, however, infant mortality is not a very different between breastfed and bottle-fed infants. (Sanitation could be thought of as a proxy for socio-economic status in this example).

Matching

In an earlier FOCUS issue we mentioned the technique of matching characteristics of cases and controls to reduce confounding. If you need a refresher on matching, see the FOCUS issue on case control studies (*Volume 3 Issue 2*). In cohort studies, unexposed persons are matched to exposed persons on the desired characteristics. In case-control studies, controls are matched to cases on the desired characteristics. Generally, you must account for matching when you conduct the analysis of matched data; otherwise you end up introducing bias. Think about this for a moment: if you choose the controls to be the same age, gender, and occupation as cases by matching, you are purposefully introducing selection bias into your study! That’s okay if the matching reduces confounding, as long as we account for matching in the analysis and the matched variables are not exposures of interest.

Let’s say we are conducting a study in a high school where a number of students have reported a strange smell and sudden illness. We test the association between smelling an unusual odor in the school and a set of symptoms (nausea, vomiting, fainting), and we match cases and controls on gender, grade, and hallway. Why match on these factors? There is a precedent for an “outbreak” of illness with unusual odors in buildings, often without any defined cause even after much environmental and epidemiological investigation. Some hypothesize that this type of illness is psychogenic, meaning that once a few people smell the odor and feel ill, others in the building react to seeing this with some panic, causing a contagious illness that is spread by hysteria rather than a true cause. The factors we match on are potential confounders. Gender is a confounder because women tend to be more reactive in this type of situation; grade level is a way of controlling for age, as older and younger students may react differently; and matching on hallway controls for the actual odor observed, as students in a hallway with a chemistry classroom may perceive an unusual odor that is different from students in a hallway near the cafeteria.

Because we have matched case-control pairs, we set up the 2x2 table differently: we examine *pairs* in the table, so we have cases along one side and controls along the other, and the table cells contain pairs. The generic set-up is seen in Table 1.

Table 1: Analysis of matched pairs for a case control study.

		Controls		
		Exposed	Not Exposed	Total
Cases	Exposed	e	f	e + f
	Not Exposed	g	h	g + h
	Total	e + g	f + h	

Cell “e” contains the number of matched case-control pairs where both the case and the control were exposed. This is a concordant cell (and so is cell “h”) because the case and the control have the same exposure status. Cell “f” contains the number of matched case-control pairs where the cases were exposed but the control of the pair was not exposed. This is a discordant cell (and so is cell “g”) because the case and the control have a different exposure status. Because we want to contrast the exposure between cases and controls, only the discordant cells (f and g) give us useful data.

A chi-square for matched data – from yet another statistics guy named McNemar – can be easily calculated from these data using a statistical computing program. The calculation examines discordant pairs, and results in a McNemar chi-square value and p-value. If the p-value is <0.05 (you know this drill), you can conclude that there is a statistically significant difference in exposure between cases and controls.

We can also use the table of discordant pairs to calculate a measure of association. Table 2 uses example data from the sudden illness outbreak mentioned above.

Table 2: Example data for sudden illness in a high school. Controls matched to cases on gender, grade, and hallway within the school.

		Controls		
		Smell	No Smell	Total
Cases	Smell	6	12	18
	No Smell	4	5	9
	Total	10	17	

We can calculate the odds ratio directly from this data.

$$OR = \frac{\text{\# pairs with exposed cases and unexposed cases}}{\text{\# pairs with unexposed cases and exposed controls}} = f / g = 12 / 4 = 3.0$$

Interpretation: The odds of having sudden onset of nausea, vomiting, or fainting if students smelled an unusual odor in the school is 3.0 times the odds of having sudden onset of these symptoms if students did not smell an unusual odor in the school, controlling for gender, grade, and location in the school.

An important note about matching: once you have matched on a variable, you cannot use that variable as a risk factor in your analysis! Cases and controls will have the exact same level of matched variables (because you chose them purposefully to be that way), so they are useless as risk factors. The moral of the story: Do not match on any variable you suspect might be a risk factor! If you match, be sure you have a good reason for doing so.

An Introduction to Logistic Regression

Topping off your field epidemiology sophistication regarding data analysis is a primer on logistic regression. As with many statistical terms, the words sound complicated but the process is fairly simple. Logistic regression is merely a mathematical process that results in an odds ratio, made possible to everyday epidemiologists by the advent of the computer. The special thing about logistic regression is that you can control for confounders. We are not talking about one confounder, as in stratification methods; we are talking about controlling for as many confounders as you want (provided you have a large enough sample size, but we'll discuss that later). Logistic regression is a mathematical model that can give an odds ratio that is controlled for multiple confounders. This odds ratio is known as the "adjusted" odds ratio, because its value has been adjusted for the confounders. Some epidemiologists use logistic regression all the time and cannot live without it, while others find that 2x2 tables are really all they need.

As in a 2x2 table, the outcome variable (sick or not sick) and the exposure variable (exposed or not exposed) must both be dichotomous. Other variables (the confounders you want to adjust for) can be dichotomous, categorical with several levels, or continuous.

We have to address the fact that regression involves advanced mathematics, but don't worry, this discussion will not get complicated. The bottom line is that logistic regression uses an equation called a *logit function* to calculate the odds ratio. Let's re-visit the example from earlier in which both the cookies and the punch have been implicated in initial analyses as the source of an outbreak at a reception. We suspect that one of these food items is confounding the other, but cannot tease out the effects without stratifying because many people consume both cookies and punch. The variables are:

- SICK (where the value is 1 if ill, 0 if not ill)
- PUNCH (1 if drank punch, 0 if did not drink punch)
- COOKIES (1 if ate cookies, 0 if did not eat cookies)

In general, the equation would look like this:

$$\text{Logit (OUTCOME)} = \text{EXPOSURE} + \text{CONFOUNDER1} + \text{CONFOUNDER2} + \text{CONFOUNDER3} + \dots \text{ (etc)}$$

In our case, the outcome is the variable SICK. The exposure is the variable PUNCH, and the confounder is the variable COOKIES. The equation looks like this:

$$\text{Logit (SICK)} = \text{PUNCH} + \text{COOKIES}$$

The computer uses the math behind logistic regression (not covered here, but easy enough for you to look up if you are interested) to give the results as odds ratios.

Each variable on the right side of the equation (these are dependent variables) will have its own odds ratio. In our example, the odds ratio given for PUNCH would be the

Logistic Equation and Odds Ratios:

$$\text{logit (SICK)} = \text{PUNCH} + \text{COOKIES}$$

PUNCH OR:

Odds of disease in those who had punch compared to those who did not have punch (controlled for cookies)

COOKIES OR:

Odds of disease in those who had cookies compared to those who did not have cookies (controlled for punch)

odds of becoming ill if punch was consumed compared to the odds of becoming ill if punch was not consumed, controlling for COOKIES. If we had other confounders in the equation (as dependent variables), the odds ratio for PUNCH would also be controlling for those variables.

Moreover, there is an odds ratio for COOKIES. This odds ratio is the odds of becoming ill if cookies were consumed compared to the odds of becoming ill if cookies were not consumed, controlling for PUNCH.

Each variable on the right side of the equation is controlling for all the other variables on the right side of the equation. This makes logistic regression very useful for epidemiologists. If you are not sure if one of several variables is a confounder, you can examine them all at the same time. However, there are two important warnings:

- Do not put too many variables in the equation. If you have a sample size of 30 or less (i.e., 30 observations or less), you should only put one dependent variable in the equation. In this case, you might as well use a 2x2 table. A 2x2 table will give you the exact same odds ratio as logistic regression with only one dependent variable. A loose rule of thumb is that you can add one variable to the equation for every 25 observations. So if you are analyzing 50 study subjects, you could include 2 dependent variables in the regression model. In the punch and cookies example, you have 200 cases and controls, which means that you can have up to 8 dependent variables in the regression model to explain the outcome/observations.
- You cannot control for confounders you did not measure. If a child's attendance at a particular daycare was a confounder of the SICK-PUNCH relationship, but you do not have data on children's daycare attendance, you will not be able to control for it. Similarly, there could be a confounder you just did not think to collect information on so you cannot control for it. This is always possible, and is part of the reason we can never be 100% certain of our odds ratios or risk ratios calculations.

Logistic regression can also easily account for matching in the analysis of data, using a special method called *condi-*

tional logistic regression. The computer calculates odds ratios in much the same way as McNemar's test, but the results are "conditioned" on the matching variables. This can also be done in Epi Info. Interpretation of matched odds ratios (MORs) using conditional logistic regression is the same as matched odds ratios calculated from tables.

Many field epidemiologists will never need to use logistic regression when analyzing data from outbreaks or other investigations. However, logistic regression is extremely helpful in managing confounding or potentially confounding variables, and it is particularly useful when analyzing large datasets and studies designed to establish risk factors for chronic conditions, cancer cluster investigations, assessments of environmental exposures, and other situations where numerous confounding factors could obscure the relationships between risk factors associated with disease outcomes. Many software packages can simplify data analysis using logistic regression, including SAS, SPSS, STATA, and Epi Info.*

Examples: Logistic regression

In 1997, attendees of a wedding reception at a private residence complained of a diarrheal illness diagnosed as cyclosporiasis. In univariate analysis (where each exposure is analyzed separately in a 2x2 table), eating raspberries was the exposure most strongly associated with risk for illness; moreover, it was the only exposure significantly associated with risk for illness in a multivariable logistic regression analysis. In the end, investigators determined that the raspberries had not been washed. (4)

*The following software packages are frequently used by epidemiologists for data analysis, including logistic regression. For more information on specific software, visit their websites on the internet. This is not a comprehensive list, and UNC does not specifically endorse any particular software package.

SAS – Cary, North Carolina
<http://www.sas.com/index.html>

SPSS – Chicago, Illinois
<http://www.spss.com/>

STATA – College Station, Texas
<http://www.stata.com>

Epi Info – Atlanta, Georgia
<http://www.cdc.gov/EpiInfo/>

Episheet – Boston, Massachusetts
<http://members.aol.com/krothman/modepi.htm>
 (Episheet cannot do logistic regression but is useful simpler analyses, e.g., 2x2 tables and stratified analysis.)

To assess the relationship between obesity and concern about food security, the Washington State Department of Health analyzed statewide data from the 1995-1999 Behavioral Risk Factor Surveillance System. A variable indicating concern about food security was analyzed using a logistic regression model, with income and education included as potential confounders. Persons who reported being concerned about food security were more likely to be obese than those who did not report such concerns (adjusted OR = 1.29, 95% CI: 1.04–1.83). (5)

Examples: Matching and conditional logistic regression

In 2002, a foodborne *Salmonella* Newport outbreak was detected among 47 people from 5 different states. Controls obtained through random-digit dialing were matched by age group with cases from New York, Michigan, and Pennsylvania. Logistic regression was conducted to calculate odds ratios controlling for confounders (these were not specified, but could have been gender and state). The investigation revealed that cases were more likely to have eaten ground beef when compared to controls (MOR = 2.3, 95% CI: 0.9–5.7), and were more likely to have eaten raw or undercooked ground beef when compared to controls (MOR = 50.9, 95% CI: 5.3–489.0). No specific contamination event was identified, and a public health alert was issued to remind consumers about safe food-handling practices. (6)

A case-control study was conducted to determine risk factors for developing typhoid fever (*Salmonella* serotype Typhi) in an outbreak in Tajikistan affecting over 10,000 people in 1996-1997. Cases were culture positive for the organism and controls were matched to cases by age and neighborhood. Using 2x2 tables, illness was associated with drinking unboiled water in the 30 days before onset (MOR = 6.5, 95% CI: 3.0–24.0), obtaining drinking water from a tap outside the home (MOR = 9.1, 95% CI: 1.6–82.0), and eating food from a street vendor (MOR = 2.9, 95% CI: 1.4–7.2). When all variables were included in conditional logistic regression analysis to tease out the affects of each factor controlling for the others, only drinking unboiled water (MOR = 9.6, 95% CI: 2.7–34.0) and obtaining water from an outside tap (MOR = 16.7, 95% CI: 2.0–138.0) were significantly associated with illness. Routinely boiling water in one's home for drinking was protective (MOR = 0.2, 95% CI: 0.05–0.5). (7)

Conclusion

In this issue of FOCUS, we have discussed the complicated topic of controlling for confounding. The goal of this issue was to gain an understanding of using matched study design and logistic regression with multiple variables to control for confounding. With a little practice, these methods will come as easily to you as creating a 2x2 table.

CONTACT US:

The North Carolina Center for Public Health Preparedness
The University of North Carolina at Chapel Hill
Campus Box 8165
Chapel Hill, NC 27599-8165

Phone: 919-843-5561
Fax: 919-843-5563
Email: nccphp@unc.edu

FOCUS Workgroup:

- Lorraine Alexander, DrPH
- Meredith Anderson, MPH
- David Bergmire-Sweat, MPH
- Kim Brunette, MPH
- Anjum Hajat, MPH
- Pia D.M. MacDonald, PhD, MPH
- Gloria C. Mejia, DDS, MPH
- Amy Nelson, PhD, MPH
- Tara P. Rybka, MPH
- Rachel A. Wilfert, MD, MPH

REFERENCES:

1. Gregg MB. *Field Epidemiology*. 2nd ed. New York, NY: Oxford University Press; 2002.
2. Hecht CA, Hook EB. Rates of Down syndrome at livebirth by one-year maternal age intervals in studies with apparent close to complete ascertainment in populations of European origin: a proposed revised rate schedule for use in genetic and prenatal screening. *Am J Med Genet*. 1996;62:376-385.
3. Humphrey LL, Nelson HD, Chan BKS, Nygren P, Allan J, and Teutsch S. Relationship between hormone replacement therapy, socioeconomic status, and coronary heart disease. *JAMA*. 2003;289:45..
4. Centers for Disease Control and Prevention. Update: Outbreaks of Cyclosporiasis – United States, 1997. *MMWR Morb Mort Wkly Rep*. 1997;46:461-462. Available at: www.cdc.gov/mmwr/PDF/wk/mm4621.pdf. Accessed December 12, 2006.
5. Centers for Disease Control and Prevention. Self-reported concern about food security associated with obesity – Washington, 1995–1999. *MMWR Morb Mort Wkly Rep*. 2003;52:840-842. Available at: www.cdc.gov/mmwr/preview/mmwrhtml/mm5235a3.htm. Accessed December 12, 2006.

If you would like to receive electronic copies of FOCUS on Field Epidemiology, please fill out the form below:

- NAME: _____
- DEGREE (S): _____
- AFFILIATION: _____
- E-MAIL ADDRESS: _____
- May we e-mail any of your colleagues? If so, please include their e-mail addresses here:

Please fax to: (919) 919-843-5563

or mail to: North Carolina Center for Public Health Preparedness
The University of North Carolina at Chapel Hill
Campus Box 8165
Chapel Hill, NC 27599-8165

Or go online: <http://www.sph.unc.edu/nccphp/focus/>

UPCOMING TOPICS!

- Collecting Specimens in Outbreak Investigations
- **Laboratory Diagnosis: An Overview**
- **Laboratory Diagnosis: Molecular Techniques**

We are on the web!

<http://www.sph.unc.edu/nccphp>