

Case Study Discussion
Outbreak Investigations Session
Case-Control Studies- To Match or Not to Match?

In this case study, you were asked to decide whether or not you would employ matching to select controls, given a certain scenario. The text below gives some general thoughts on matching including some points to ponder when considering various scenarios.

Why do we match? We employ matching to protect ourselves from reaching the wrong conclusion due to the effect of confounding factors. Our academic training invariably points out mistakes based on biased studies. The motivation to avoid being wrong is powerful, and we are therefore predisposed to employ matching whenever possible. Surprisingly, we are often less aware of the bias introduced by matching itself. As a result, we may not only invest unnecessary resources to obtain matched controls, but may also reach a wrong conclusion¹.

By definition, a confounding factor must be linked to exposure AND to outcome. Common examples of matching include controlling for occupation, neighborhood, or sibship¹⁻³. It is easy to see how these factors could affect a study of lung cancer because they could be associated with an exposure (e.g. smoking) and an outcome (by virtue of exposure to other factors that cause cancer). In foodborne outbreak investigations, the factors that affect outcome are typically those that affect a person's susceptibility to illness, such as age (e.g. infants or elderly), illness (e.g. cancer), or prior medical therapy (e.g. antibiotic treatment) and it is these factors that you need to consider when deciding whether or not to match. The most common pratfall of control selection is to match on a factor that is related only to exposure, but not to outcome, and reduces the likelihood of finding the true source of illness.

Let's take a look at some examples.

When undertaking a case-control study of *Salmonella* patients that are mostly women, there is a temptation to match on gender because women tend to eat different foods than men. The problem with this approach is that the contaminated food may be one of the foods more likely to be eaten by women. Thus for every woman enrolled as a control, it is more likely that she will have eaten the contaminated item, and therefore, more difficult to show an association between that food and illness. In this example, matching on gender biases the study towards the null. Is it better then to select only men for controls? Although this may increase the power for detecting the contaminated food, it will also increase the likelihood of finding a false association. So, while it is tempting to match based on gender, it is also problematic because gender is typically not related to the risk of becoming ill—healthy men and (non-pregnant) women likely have the same risk of getting ill if they eat the same contaminated food.

In outbreaks of listeriosis, most cases are either pregnant or receiving immunosuppressive therapy, and therefore case-control studies are often matched by pregnancy or immunosuppressive therapy. This guarantees that controls had an equal likelihood of becoming ill if they had eaten the contaminated food. Matching controls to listeriosis cases is an interesting example because it is not clear that matching based on pregnancy affects both exposure and outcome. Pregnancy clearly affects outcome but it may not affect food choices (despite our efforts to educate women to avoid high risk foods). The most important result of matching on

pregnancy is that we are confident that we have a group that could have become ill if exposed; otherwise healthy adults are much less likely to get ill. Thus by matching, we are creating a cohort of persons that share the same susceptibility.

In a scenario where most of your cases are infants but most of the controls are adults, age could be a confounding factor because it affects both susceptibility and the types of food eaten. In order to control for age, one could stratify cases and controls into separate age groups and perform an analysis within each group. This analysis would be equally (or more) valid than when matching controls to cases based on age. The disadvantage to this approach is one of sample size – without restricting the age of controls, there is a chance that there would not be enough controls in each age group. Matching would guarantee that there would be at least as many controls as cases in each age group.

It is always necessary to match based on the availability of a contaminated food. For example, if cases are only identified in Connecticut, it makes no sense to obtain controls from Massachusetts because there is no evidence that the contaminated food was available outside of Connecticut. Similarly, if cases only buy food at Hispanic markets (which might be indicated by a high percentage of cases with Hispanic ethnicity, as described in scenario two), it would be reasonable to identify controls who also only buy food at Hispanic markets. Because controls who shop at other markets may not be able to buy the contaminated food, a case-control study may wrongly associate illness with any of the foods available only at the Hispanic market regardless of which one was contaminated. Thus, if the study is designed to determine which of the food at the Hispanic is contaminated, cases and controls both need to have shopped at that market. One could describe this case-control study as, “restricted to a cohort of those shopping at the Hispanic market,” or as a study in which “controls were matched to cases based on where they shop.” The semantics are different but the result is the same.

References:

1. Modern Epidemiology (first edition, by Kenneth Rothman)
2. Field Epidemiology (second edition, edited by Michael Gregg)
3. Essentials of Epidemiology in Public Health (by Ann Aschengrau and George Seage)

Scenarios:

- 80% of your cases are women. Would you match on gender? Why or why not? What are the advantages/disadvantages of your approach?
- 80% of your cases are of Hispanic ethnicity. Would you match on ethnicity? Why or why not? What are the advantages/disadvantages of your approach?
- 80% of your cases are younger than 5 years old. Would you match on age? Why or why not? What are the advantages/disadvantages of your approach?