JOURNAL OF CLINICAL RESEARCH BEST PRACTICES

Vol. 9, No. 12, December 2013 "Happy Trials to You"

Making Sense of Biostatistics: Case-Control Studies

By Lynn M. Ackerson

In a simple, two-arm randomized clinical study, people are randomly assigned to the intervention and control arms, with the expectation that the characteristics of the people in the two groups (e.g., age, gender, race) are the same for everything except their intervention status. The statistics used to analyze data from randomized clinical trials rely on this assumption. Unfortunately, it is not always possible to randomize people to an intervention or exposure, such as smoking. In these instances, we must rely on observational studies that do not allow for control over which people get which intervention or exposure.

One such study design is a case-control study. Case-control studies are particularly useful when you want to study the relationship between a characteristic (e.g., an exposure, such as cigarette smoking or X-rays during pregnancy) and an outcome when: (1) the outcome is rare, (2) there is a long lag time between the exposure and the development of the outcome, or (3) you are studying multiple exposures for a single disease. For example, if we are interested in knowing if *in utero* exposure to Drug A increases the risk of breast cancer in women, due to the long latency period, it isn't feasible to design a randomized trial of pregnant women and wait for their daughters to grow up. However, we could choose a group of women with breast cancer (cases) and a similar group of women without breast cancer (controls), and find out whether or not their mothers took Drug A during their pregnancies.

The first step in a case-control study is to choose a group of people who have the outcome of interest (the case group) and a comparable group of people who do not have the outcome of interest (the control group). Unlike randomized clinical trials, case-control studies are thus retrospective, because the outcomes are always known when study subjects are chosen. You want the control group to be as similar to the case group as possible in every respect except the condition of interest. In the randomized trial of pregnant women described above, the randomization process greatly increases the probability that the two study groups will be similar on almost all variables, but we can't do this in an observational study. Instead, we choose variables we believe might be strongly related to the exposure and the outcome, and make sure the people in the case and control groups are the same on these few variables. We call this "matching," and, while it would be great to match on every conceivable variable, the process becomes increasingly more difficult to implement as the number of matching variables increases.

Two common methods of matching are one-to-one (or n-to-one) matching and frequency matching. For one-to-one matching, each case is matched to a control based on factors strongly associated with both the exposure and the outcome. Common matching variables are age, race, gender and smoking status. Be careful in choosing the matching variables, since, once you've matched on them, you cannot test to see if they are related to the outcome, since, in the matching process, you have forced each case-control pair to have the same value on each matching variable, and so the prevalence of that variable will always be the same in the two groups.

While it is optimal to have a single matched control for each case, choosing two or more controls for each case can increase the power of and/or reduce the cost of a study when: (1) the disease is rare and possible controls are plentiful or (2) it is much more expensive to collect the exposure information from the case group than the control group. However,

beyond four controls per case, the power to detect a significant relationship between exposure and outcome levels off.

With frequency matching, case and control individuals are not matched on an individual basis as is done with one-to-one matching. Instead, the *distribution* of matching variables is the same in the two groups. For example, if 35% of the case group are female, then 35% of control group are also female.

Once the cases and controls are chosen, we determine whether each person had the characteristic of interest at some time in the past, preferably before the cases developed the outcome of interest. The usual statistic for this type of study is the odds ratio (OR). The OR compares the odds of having the outcome among people in the case (exposed) group vs. the odds of having the outcome among people in the control (unexposed) group. (Conversely, given the outcomes, it can compare the odds of having the exposure among the cases vs. the odds of having the exposure among the controls.) If the odds ratio equals 1, there is no relationship between the exposure and the outcome, since the odds of exposure are the same for both outcome groups. If the odds ratio is greater than 1, then those in the case group are more likely to have the exposure than those in the control group. If it is less than 1, then those in the case group are less likely to have the exposure than those in the control group, i.e., the exposure may be protective. To learn more about odds and ORs, refer to an earlier column by Dechert.[1]

If the matching variables are strongly related to both the exposure and the outcome, comparing cases to their individually matched controls greatly increases the power to detect a relationship between the exposure and outcome, compared to just analyzing the data as though the people in the case and control groups were chosen completely independently of each other. Within each matched case-control pair, the two people are exactly the same on a number of important variables (the matching variables), so any difference between the two people would have to be due to some other variable, presumably the exposure. Without the matching, the difference between any single case and its single control could just as likely be due to one of the matching variables as to the exposure of interest.

Case-control studies are subject to various biases, so they must be carefully designed. The most common biases occur at the point of selection of the cases and controls, and also when the exposure measurements are made. Since we know the outcomes, we do not also want to know exposure status when we decide who should be in the case and control groups, since it could affect our choice of whom to include in the study. Determination of exposure status should occur only after the two study groups have been chosen. We also must ensure that the controls we choose are appropriate. For example, there can be a problem when the cases are chosen from hospitalized patients, while the controls are chosen from non-hospitalized people in the community. The differences between the two groups could be due to many other factors besides their exposure status, such as prevalence of comorbidities or activity level.

Case-control studies might also have recall bias. This occurs if the people in one group are more likely to remember whether they had the exposure than people in the other group. For example, a mother whose child was born with a birth defect might have spent a great deal of time trying to remember every possible exposure during her pregnancy that might have caused the birth defect, including every over-the-counter medication she might have taken. In contrast, the mother of the healthy baby probably would not have spent as much time doing this, so might be less likely to remember whether or not she had a specific exposure of interest.

Reference

1. Dechert RE. Making Sense of Biostatistics: What are the Odds?, J Clin Res Best Practices, Vol 6 (6), June, 2010.

Author

Lynn M. Ackerson, PhD, is a Biostatistician in the Division of Research, Kaiser Permanente Medical Care Program. Contact her at 1.510.891.3556 or Lynn.M.Ackerson@kp.org.