

Controlling Confounding by Stratifying Data

In an earlier chapter, we saw that the apparent effect of birth order on the prevalence at birth of Down syndrome (see Fig. 7-3 in Chapter 7) is attributable to confounding. As demonstrated in Figure 7-4, maternal age has an extremely strong relation to the prevalence of Down syndrome. Figure 7-5, which classifies the Down syndrome data simultaneously by birth order and maternal age, shows that there is a maternal-age effect at every level of birth order, but no clear birth order effect at any level of maternal age. The birth order effect in the crude data is confounded by maternal age, which is correlated with birth order.

Figure 7-5 is a graphic demonstration of *stratification*. Stratification is used here to mean the cross-tabulation of data; usually, in this context, stratification refers to cross-tabulation of data on exposure and disease by categories of one or more other variables that are potential confounding variables. Another example of stratification was discussed in Chapter 1, which introduced the concept of confounding. Stratification is an effective and straightforward means to control confounding. In this chapter, we explore stratification in greater detail and present formulas to derive an unconfounded estimate of an effect from stratified data.

AN EXAMPLE OF CONFOUNDING

Consider another example of confounding. The data in Table 10-1 are mortality rates for male and female patients with trigeminal neuralgia, a recurrent paroxysmal pain of the face.

The rate ratio of 1.10 indicates a slightly greater mortality rate for males than for females in these crude data. (The male group may be thought of as the exposed group and the female group as the unexposed group to make this example analogous to other settings in which the exposure variable is a specific agent.) This estimate of the association between being male and death among trigeminal neuralgia

Table 10-1 MORTALITY RATES AMONG PATIENTS WITH TRIGEMINAL NEURALGIA CATEGORIZED BY SEX¹

| | Males | Females |
|--------------------|---------------|---------------|
| Deaths | 90 | 131 |
| Person-years (pyr) | 2465 | 3946 |
| Mortality rate | 36.5/1000 pyr | 33.2/1000 pyr |
| Rate ratio | 1.10 | |
| 90% CI | 0.88-1.38 | |

Data from Rothman and Monson.¹

patients is confounded. Table 10-2 shows the data stratified into two age strata, which are split at age 65. The age stratification reveals several interesting things about the data. First, as might have been predicted, patients in the older age group have much higher death rates than those in the younger age group. The striking increase in risk of death with age is typical of any population of older adults, even adults in the general population. Second, the stratification shows a difference in the age distribution of the person-time of male and female patients; the male person-time is mainly found in the younger than 65 years category, whereas the female person-time is predominantly found in the 65 years or older category. Thus, the female experience is older than the male experience. This age difference lowers the overall death rate for males relative to females, because to some extent comparing the death rate among males with that of females is a comparison of young with old. Third, in the crude data the rate ratio (male/female) was 1.10, but in the two age categories, it was 1.57 and 1.49, respectively. This discrepancy between the crude rate ratio and the rate ratios for each of the two age categories is a result of the strong age effect and the fact that the female patients tend to be older than the male patients. It is a good example of confounding by age, in this case biasing the crude rate ratio downward because the male person-time experience is younger than that of the females.

Stratification into age categories allows us to assess the presence of confounding. It also permits us to refine the estimate of the rate ratio by controlling age confounding. Later we will show how to remove confounding for this trigeminal neuralgia example and examples of other types of data by using stratification.

Table 10-2 MORTALITY RATES AMONG PATIENTS WITH TRIGEMINAL NEURALGIA BY SEX AND AGE CATEGORY¹

| | Age | | | |
|---------------------------------------------|-----------|---------|-----------|---------|
| | <65 Years | | 65+ Years | |
| | Males | Females | Males | Females |
| Deaths | 14 | 10 | 76 | 121 |
| Person-years | 1516 | 1701 | 949 | 2245 |
| Mortality rate (cases/1000 person-years) | 9.2 | 5.9 | 80.1 | 53.9 |
| Rate ratio | 1.57 | | 1.49 | |

Data from Rothman and Monson.¹

Controlling Confounding by Stratifying Data

In an earlier chapter, we saw that the apparent effect of birth order on the prevalence at birth of Down syndrome (see Fig. 7-3 in Chapter 7) is attributable to confounding. As demonstrated in Figure 7-4, maternal age has an extremely strong relation to the prevalence of Down syndrome. Figure 7-5, which classifies the Down syndrome data simultaneously by birth order and maternal age, shows that there is a maternal-age effect at every level of birth order, but no clear birth order effect at any level of maternal age. The birth order effect in the crude data is confounded by maternal age, which is correlated with birth order.

Figure 7-5 is a graphic demonstration of *stratification*. Stratification is used here to mean the cross-tabulation of data; usually, in this context, stratification refers to cross-tabulation of data on exposure and disease by categories of one or more other variables that are potential confounding variables. Another example of stratification was discussed in Chapter 1, which introduced the concept of confounding. Stratification is an effective and straightforward means to control confounding. In this chapter, we explore stratification in greater detail and present formulas to derive an unconfounded estimate of an effect from stratified data.

AN EXAMPLE OF CONFOUNDING

Consider another example of confounding. The data in Table 10-1 are mortality rates for male and female patients with trigeminal neuralgia, a recurrent paroxysmal pain of the face.

The rate ratio of 1.10 indicates a slightly greater mortality rate for males than for females in these crude data. (The male group may be thought of as the exposed group and the female group as the unexposed group to make this example analogous to other settings in which the exposure variable is a specific agent.) This estimate of the association between being male and death among trigeminal neuralgia

Table 10-1 MORTALITY RATES AMONG PATIENTS WITH TRIGEMINAL NEURALGIA CATEGORIZED BY SEX¹

| | Males | Females |
|--------------------|---------------|---------------|
| Deaths | 90 | 131 |
| Person-years (pyr) | 2465 | 3946 |
| Mortality rate | 36.5/1000 pyr | 33.2/1000 pyr |
| Rate ratio | 1.10 | |
| 90% CI | 0.88-1.38 | |

Data from Rothman and Monson.¹

patients is confounded. Table 10-2 shows the data stratified into two age strata, which are split at age 65. The age stratification reveals several interesting things about the data. First, as might have been predicted, patients in the older age group have much higher death rates than those in the younger age group. The striking increase in risk of death with age is typical of any population of older adults, even adults in the general population. Second, the stratification shows a difference in the age distribution of the person-time of male and female patients; the male person-time is mainly found in the younger than 65 years category, whereas the female person-time is predominantly found in the 65 years or older category. Thus, the female experience is older than the male experience. This age difference lowers the overall death rate for males relative to females, because to some extent comparing the death rate among males with that of females is a comparison of young with old. Third, in the crude data the rate ratio (male/female) was 1.10, but in the two age categories, it was 1.57 and 1.49, respectively. This discrepancy between the crude rate ratio and the rate ratios for each of the two age categories is a result of the strong age effect and the fact that the female patients tend to be older than the male patients. It is a good example of confounding by age, in this case biasing the crude rate ratio downward because the male person-time experience is younger than that of the females.

Stratification into age categories allows us to assess the presence of confounding. It also permits us to refine the estimate of the rate ratio by controlling age confounding. Later we will show how to remove confounding for this trigeminal neuralgia example and examples of other types of data by using stratification.

Table 10-2 MORTALITY RATES AMONG PATIENTS WITH TRIGEMINAL NEURALGIA BY SEX AND AGE CATEGORY¹

| | Age | | | |
|---------------------------------------------|-----------|---------|-----------|---------|
| | <65 Years | | 65+ Years | |
| | Males | Females | Males | Females |
| Deaths | 14 | 10 | 76 | 121 |
| Person-years | 1516 | 1701 | 949 | 2245 |
| Mortality rate (cases/1000 person-years) | 9.2 | 5.9 | 80.1 | 53.9 |
| Rate ratio | 1.57 | | 1.49 | |

Data from Rothman and Monson.¹

UNCONFOUNDED EFFECT ESTIMATES AND CONFIDENCE INTERVALS FROM STRATIFIED DATA

How does stratification control confounding? Confounding, as explained in Chapter 7, comes from the mixing of the effect of the confounding variable with the effect of the exposure. If a variable that is a risk factor for the disease is associated with the exposure in the study population, confounding will result. Confounding occurs because the comparison of exposed with unexposed people is also a comparison of those with differing distributions of the confounding factor. In the trigeminal neuralgia example, comparing men with women was also a comparison of younger people (ie, men in the study) with older people (ie, women in the study). Stratification creates subgroups in which the confounding factor either does not vary at all or does not vary much. Stratification by nominal scale variables, such as sex or country of birth, theoretically results in strata in which the variables of sex or country of birth do not vary; in actuality, there may still be some residual variability because some people may be misclassified into the wrong strata. Stratification by a continuously measured variable, such as age, will result in age categories within which age can vary, although over a restricted range. With either kind of variable, nominal scale or continuous, a stratified analysis proceeds under the assumption that within the categories of the stratification variable there is no meaningful variability of the potential confounding factor. If the stratification variable is continuous, such as age, the more categories that are used to form strata, the less variability by age there can be within those categories.

In some stratified analyses, the end result is nothing more than the presentation of the data within each of the strata, with estimates of rates, risks, or effect estimates for each stratum. Often, however, the investigator hopes to summarize the relation between exposure and disease over the strata. The methods that do so compare exposed and unexposed subjects within each stratum and then aggregate the information from these comparisons over all the strata. The two basic approaches to aggregate the information over strata are referred to as *pooling* and *standardization*, representing two different methods for combining the data across the strata.

Pooling

Pooling is one method for obtaining unconfounded estimates of effect across a set of strata. When pooling is used, it comes with an important assumption: that the effect being estimated is constant across the strata. With this assumption, each stratum can be viewed as providing a separate estimate, referred to as a *stratum-specific estimate*, of the overall effect. The principle behind pooling is to take an average of these stratum-specific estimates of effect. The average is taken as a weighted average, which is a method of averaging that assigns more weight to some values than to others. In pooling, the weights are assigned so that the strata that provide the most information, which is to say the strata with the most data get the most weight. This weighting is built directly into the formulas for obtaining the pooled estimate. When the data do not conform to the assumption that the effect is constant across all strata, pooling is not applicable. In that situation,

it is still possible to obtain an unconfounded summary estimate of the effect over the strata using *standardization*, which is discussed later.

Cohort Studies with Risk Data or Prevalence Data

Consider risk data; for analytic purposes, prevalence data may be treated the same as risk data. We use the same basic notation as we did for unstratified data, but we add a stratum-identifying subscript, i , which ranges from 1 to the total number of strata. The notation for stratum i in a set of strata of risk data is as follows:

| | Exposed | Unexposed | Total |
|---------------|----------|-----------|----------|
| Cases | a_i | b_i | M_{1i} |
| Noncases | c_i | d_i | M_{0i} |
| Total at risk | N_{1i} | N_{0i} | T_i |

For risk data, we can calculate a pooled estimate of the risk difference or the risk ratio. The pooled risk difference may be estimated from stratified data using Equation 10-1:

$$RD_{MH} = \frac{\sum_i \frac{a_i N_{0i} - b_i N_{1i}}{T_i}}{\sum_i \frac{N_{1i} N_{0i}}{T_i}} \quad [10-1]$$

\sum signifies summation over all values of the stratum indicator i . The subscript "MH" for the pooled risk difference measure refers to Mantel-Haenszel, indicating that the equation is one of a group of equations for pooled estimates that derive from an approach that was originally introduced by Mantel and Haenszel.²

The pooled risk ratio from stratified risk or prevalence data can be calculated by Equation 10-2:

$$RR_{MH} = \frac{\sum_i \frac{a_i N_{0i}}{T_i}}{\sum_i \frac{b_i N_{1i}}{T_i}} \quad [10-2]$$

Example: Stratification of Risk Data

The stratification of risk data is illustrated in the example of the University Group Diabetes Program (see Tables 7-7 and 7-8 in Chapter 7). For convenience, the age-specific data are repeated in Table 10-3.

First, we consider the risk difference. From the crude data (see right part of Table 10-3), the risk difference is 4.5%. Contrary to expectations, the tolbutamide group had a greater risk of death than the placebo group, despite the fact

Table 10-3 RISK OF DEATH FOR GROUPS RECEIVING TOLBUTAMIDE OR PLACEBO IN THE UNIVERSITY GROUP DIABETES PROGRAM IN 1970³

| | Age | | | | | |
|-----------------|-----------|---------|-----------|---------|-------|---------|
| | <55 Years | | 55+ Years | | Total | |
| | Tolb. | Placebo | Tolb. | Placebo | Tolb. | Placebo |
| Deaths | 8 | 5 | 22 | 16 | 30 | 21 |
| Total at risk | 106 | 120 | 98 | 85 | 204 | 205 |
| Risk of death | 0.076 | 0.042 | 0.224 | 0.188 | 0.147 | 0.102 |
| Risk difference | 0.034 | | 0.036 | | 0.045 | |
| Risk ratio | 1.81 | | 1.19 | | 1.44 | |

Data from University Group Diabetes Program.³

that tolbutamide was thought to prevent complications of diabetes that might lead to death. Critics of the study believed this finding to be erroneous and looked for explanations such as confounding that might account for this surprising result. Age was one of the possible confounding factors. By chance, the tolbutamide group tended to be slightly older than the placebo group. This age difference is evident in Table 10-3: 48% (98/204) of the tolbutamide group is at least 55 years old, whereas only 41% (85/205) of the placebo group is at least 55 years old. Older people have a greater risk of death, a relation that is also evident in Table 10-3. Consider the placebo group: The risk of death during the study period was 18.8% for the older age group but only 4.2% for the younger age group. We therefore suspect that the greater risk of death in the tolbutamide group is in part due to confounding by age. We can explore this issue further by obtaining a pooled estimate of the risk difference for tolbutamide compared with placebo after stratifying by the two age strata in Table 10-3.

We obtain a pooled estimate of the risk difference by applying Equation 10-1:

$$RD_{MH} = \frac{\frac{8 \cdot 120 - 5 \cdot 106}{226} + \frac{22 \cdot 85 - 16 \cdot 98}{183}}{\frac{106 \cdot 120}{226} + \frac{98 \cdot 85}{183}} = \frac{1.903 + 1.650}{56.283 + 45.519} = 0.035$$

The result, 3.5%, is smaller than the risk difference in the crude data, 4.5%. Notice that 3.5% is within the narrow range of the two stratum-specific risk differences in Table 10-3, 3.4% for age <55 years and 3.6% for age 55+ years. Mathematically, the pooled estimate is a weighted average of the stratum-specific values, and it will always be within the range of the stratum-specific estimates of the effect. The crude estimate of effect, however, is not within this range. We should regard 3.5% as a more appropriate estimate of the risk difference than the value of 4.5% from the crude data, because it removes age confounding. The crude risk difference differs from the unconfounded estimate of risk difference because the crude estimate reflects a combination of the effect of tolbutamide (which we estimate to be 3.5% from this analysis) and the confounding effect of age. Because the tolbutamide group is older on average than the placebo group, the risk difference in the crude data is greater than the unconfounded risk difference. If the

tolbutamide group had been younger than the placebo group, the confounding would have worked in the opposite direction, resulting in a lower risk difference in the crude data than from the pooled analysis after stratification.

RESIDUAL CONFOUNDING

The two age categories for the data in Table 10-3 may not be sufficient to control all of the age confounding in the data. More strata with narrower boundaries usually can control confounding more effectively than fewer strata with broader boundaries. If age strata (or strata by any continuously measured stratification factor) are broad, there may be confounding within them. A stratified analysis controls only between-stratum confounding, not within-stratum confounding. Within-stratum confounding is often referred to as *residual confounding*. The same term is used to describe confounding from factors that are not controlled at all in a study or from factors that are controlled but are measured inaccurately.

To avoid within-stratum residual confounding, it is desirable to carve the data into more strata and to avoid open-ended strata (eg, age 55+) when possible. On the other hand, stratifying too finely may stretch the data unreasonably, producing small frequencies of events within cells and leading to imprecise results. Finding the best number of strata to use in a given analysis often requires balancing the need to control confounding against the need to avoid random error in the estimation and ends up being a compromise.

The unconfounded estimate of the risk difference, 3.5%, is unconfounded only to the extent that stratification into these two broad age categories removes age confounding. It is likely that some residual confounding remains (see box) and that the risk difference that is fully unconfounded by age is smaller than 3.5%.

We can also calculate a pooled estimate of the risk ratio from the data in Table 10-3 using Equation 10-2:

$$RR_{MH} = \frac{\frac{8 \cdot 120}{5 \cdot 106} + \frac{22 \cdot 85}{16 \cdot 98}}{\frac{226}{226} + \frac{183}{183}} = \frac{4.248 + 10.219}{2.345 + 8.568} = 1.33$$

This result, like that for the risk difference, is closer to the null value than the crude risk ratio of 1.44, indicating that some age confounding has been removed by the stratification. The pooled estimate is within the range of the stratum-specific estimates, as it must be mathematically. Note, however, that for the risk ratio, the stratum-specific estimates for the data in Table 10-3, 1.81 and 1.19, differ considerably from one another. The wide range between them includes the pooled estimate and the estimate of effect from the crude data. When the stratum-specific estimates of effect are almost identical, as they are for the risk differences in the data in Table 10-3, we have a good idea of what the pooled estimate will be just from inspecting the stratum-specific data. When the stratum-specific estimates vary, it is not clear on inspection what the pooled estimate will be.

As stated earlier, the equations used to obtain pooled estimates are premised on the assumption that the effect is constant across strata. The pooled risk ratio of 1.33 for the previous example is premised on the assumption that there is a single value for the risk ratio that applies to both the young and the old stratum. This assumption seems reasonable for the risk difference calculation, for which the two strata gave almost the same estimate of risk difference, but how can we use this assumption to estimate the risk ratio when the two age strata give such different risk ratio estimates? The assumption does not imply that the estimates of effect will be the same or even almost the same in each stratum. It allows for statistical variation over the strata. It is possible to conduct a statistical evaluation, called a *test of heterogeneity* or a *test of homogeneity*, to determine whether the variation in estimates from one stratum to another is compatible with the assumption that the effect is uniform.⁴ In any event, it is helpful to bear in mind that the assumption that the effect is uniform is probably wrong in most situations. It is asking too much to have the effect be absolutely constant over the categories of some stratification factor. It is more realistic to consider the assumption as a fictional convenience, one that facilitates the computation of a pooled estimate. Unless the data demonstrate some clear pattern of variation that undermines the assumption that the effect is uniform over the strata, it is usually reasonable to use a pooled approach, despite the fiction of the assumption. In Table 10-3, the variation of the risk ratio estimates for the two age strata is not striking enough to warrant concern about the assumption that the risk ratio is uniform. If a more formal statistical evaluation of the assumption of uniformity were undertaken for these data (calculating a P value to test the assumption), it would support the view that the assumption of a uniform risk ratio for the data in Table 10-3 is reasonable.

Confidence Intervals for Pooled Estimates

To obtain confidence intervals for the pooled estimates of effect we need variance formulas to combine with the point estimates. Table 10-4 lists variance formulas for the various pooled estimates that we consider in this chapter.

Although the formulas may look complicated, they are easy to apply. Each variance formula corresponds to a particular type of stratified data. First consider the pooled risk difference. For the data in Table 10-3, we calculated an RD_{MH} of 0.035. We can derive the variance for this estimate and a confidence interval by applying the first formula from Table 10-4 to the data in Table 10-3.

$$\begin{aligned} \text{Var}(RD_{MH}) &= \frac{\left(\frac{106 \cdot 120}{226}\right)^2 \left(\frac{8 \cdot 115}{106^2 \cdot 105} + \frac{5 \cdot 98}{120^2 \cdot 119}\right) + \left(\frac{98 \cdot 85}{183}\right)^2 \left(\frac{22 \cdot 69}{98^2 \cdot 97} + \frac{16 \cdot 76}{85^2 \cdot 84}\right)}{\left[\left(\frac{106 \cdot 120}{226}\right) + \left(\frac{98 \cdot 85}{183}\right)\right]^2} \\ &= \frac{3.1681 + 7.4879}{10,363.7} = 0.001028 \end{aligned}$$

This gives a standard error of $(0.001028)^{1/2} = 0.0321$ and a 90% confidence interval of $0.035 \pm 1.645 \cdot 0.0321 = 0.035 \pm 0.053 = -0.018$ to 0.088 . The

Table 10-4 VARIANCE FORMULAS FOR POOLED ANALYSES

$$\text{Risk Difference: } \text{Var}(RD_{MH}) = \frac{\sum_i \left(\frac{N_{1i} N_{0i}}{T_i}\right)^2 \left[\frac{a_i c_i}{N_{1i}^2 (N_{1i} - 1)} + \frac{b_i d_i}{N_{0i}^2 (N_{0i} - 1)} \right]}{\left(\sum_i \frac{N_{1i} N_{0i}}{T_i}\right)^2}$$

$$\text{Risk Ratio: } \text{Var}[\ln(RR_{MH})] = \frac{\sum_i (M_{1i} N_{1i} N_{0i} / T_i^2 - a_i b_i / T_i)}{\left(\sum_i \frac{a_i N_{0i}}{T_i}\right) \left(\sum_i \frac{b_i N_{1i}}{T_i}\right)}$$

$$\text{Incidence Rate Difference: } \text{Var}(ID_{MH}) = \frac{\sum_i (PT_{1i} PT_{0i} / T_i^2) (a_i / PT_{1i}^2 + b_i / PT_{0i}^2)}{\left(\sum_i (PT_{1i} PT_{0i} / T_i)\right)^2}$$

$$\text{Incidence Rate Ratio: } \text{Var}[\ln(IR_{MH})] = \frac{\sum_i M_{1i} PT_{1i} PT_{0i} / T_i^2}{\left(\sum_i \frac{a_i PT_{0i}}{T_i}\right) \left(\sum_i \frac{b_i PT_{1i}}{T_i}\right)}$$

$$\text{Odds Ratio: } \text{Var}[\ln(OR_{MH})] = \frac{\sum_i G_i P_i}{2 \left(\sum_i G_i\right)^2} + \frac{\sum_i (G_i Q_i + H_i P_i)}{2 \left(\sum_i G_i \sum_i H_i\right)} + \frac{\sum_i H_i Q_i}{2 \left(\sum_i H_i\right)^2}$$

where

$$G_i = (a_i d_i / T_i) \quad H_i = (b_i c_i / T_i)$$

$$P_i = (a_i + d_i) / T_i \quad Q_i = (b_i + c_i) / T_i$$

confidence interval is broad enough to indicate a fair amount of statistical uncertainty in the finding that tolbutamide is worse than placebo. It is notable, however, that the data are not very compatible with any compelling benefit for tolbutamide.

A confidence interval can be constructed for the risk ratio estimated from the same stratified data. In that case, an investigator would use the second formula in Table 10-4, setting limits on the log scale, as we did in the previous chapter for crude data. The variance for the logarithm of the RR_{MH} can be calculated as

$$\text{Var}[\ln(RR_{MH})] = \frac{\left(\frac{13 \cdot 106 \cdot 120}{226^2} - \frac{8 \cdot 5}{226}\right) + \left(\frac{38 \cdot 98 \cdot 85}{183^2} - \frac{22 \cdot 16}{183}\right)}{\left(\frac{8 \cdot 120}{226} + \frac{22 \cdot 85}{183}\right) \left(\frac{5 \cdot 106}{226} + \frac{16 \cdot 98}{183}\right)}$$

$$= \frac{3.0605 + 7.5286}{14.466 \cdot 10.913} = \frac{10.5891}{157.88} = 0.0671$$

This result gives a standard error for the logarithm of the RR of $(0.0671)^{1/4} = 0.259$ and a 90% confidence interval of 0.87 to 2.0.

$$RR_L = e^{\ln(1.33) - 1.645 \cdot 0.259} = 0.87$$

$$RR_U = e^{\ln(1.33) + 1.645 \cdot 0.259} = 2.0$$

The interpretation for this result is similar to the interpretation for the confidence interval of the risk difference, which is as expected because the two measures of effect and their respective confidence intervals are alternative ways of expressing the same finding from the same set of data.

As another example, consider again the data in Table 1-2. We can calculate the risk ratio for 20-year risk of death among smokers compared with nonsmokers across the seven age strata using Equation 10-2. This calculation gives an overall Mantel-Haenszel risk ratio of 1.21, with a 90% confidence interval of 1.06 to 1.38. The Mantel-Haenszel risk ratio is different from the crude risk ratio of 0.76, and as discussed in Chapter 1, it points in the opposite direction.

Cohort Studies with Incidence Rate Data

For rate data, we have the following notation for stratum i of a stratified analysis:

| | Exposed | Unexposed | Total |
|---------------------|-----------|-----------|-------|
| Cases | a_i | b_i | M_i |
| Person-time at risk | PT_{1i} | PT_{0i} | T_i |

As we did for risk data, we can calculate a pooled estimate of the rate difference or the rate ratio. The pooled rate difference may be estimated from stratified data using Equations 10-4 and 10-5:

$$ID_{MH} = \frac{\sum_i \frac{a_i PT_{0i} - b_i PT_{1i}}{T_i}}{\sum_i \frac{PT_{1i} PT_{0i}}{T_i}} \quad [10-4]$$

$$IR_{MH} = \frac{\sum_i \frac{a_i PT_{0i}}{T_i}}{\sum_i \frac{b_i PT_{1i}}{T_i}} \quad [10-5]$$

A pooled estimate of the rate ratio may be calculated as follows. Consider the rate data in Table 10-5. These data come from a study of mortality rates among current users and past users of clozapine, a drug used to treat schizophrenia. Clozapine is thought to affect mortality primarily for current users. The experience of past users, who still have many of the indications for using the drug but who have for various reasons

Table 10-5 MORTALITY RATES FOR CURRENT AND PAST CLOZAPINE USERS, OVERALL AND BY AGE CATEGORY⁵

| | Age | | | | | |
|-------------------------------------|-------------|--------|-------------|-------|---------|--------|
| | 10-54 Years | | 55-94 Years | | Total | |
| | Current | Past | Current | Past | Current | Past |
| Deaths | 196 | 111 | 167 | 157 | 363 | 268 |
| Person-years | 62,119 | 15,763 | 6,085 | 2,780 | 68,204 | 18,543 |
| Rate ($\times 10^5$ yr) | 315.5 | 704.2 | 2,744 | 5,647 | 532.2 | 1,445 |
| Rate Difference ($\times 10^5$ yr) | -388.7 | | -2903 | | -912.8 | |
| Rate Ratio | 0.45 | | 0.49 | | 0.37 | |

Data of Walker et al.⁵

discontinued it, was used as the reference for judging the effect of current use. As for the tolbutamide example, the data are stratified into two broad age categories.

The death rates are much greater for older patients than for younger patients, as expected. Among schizophrenia patients, as for the general population, death rates climb strikingly with age. There is also an association between age and current versus past use of clozapine. Among current users, 9% (6085/68,204) of the person-time is in the older age category, whereas among past users, 15% (2780/18,543) of the person-time is in the older age category. This difference is enough to introduce some confounding, although it is not large enough to produce more than a modest amount. Because the person-time for past use has an older age distribution, the age differences will lead to lower death rates among current users. The crude data do indicate a lower death rate among current users, with a rate difference of -912.8 cases per 100,000 person-years. At least some of this difference is attributable to age confounding. We can obtain an estimate of the mortality rate difference that is unconfounded by age (apart from any residual age confounding within these broad age categories) from Equation 10-4:

$$ID_{MH} = \frac{196 \cdot 15,763 - 111 \cdot 62,119}{62,119 \cdot 15,763} + \frac{167 \cdot 2,780 - 157 \cdot 6,085}{6,085 \cdot 2,780} \\ = \frac{-48,864 - 55,396}{12,572.633 + 1908.212} = -720.0 \times 10^{-5} \text{ yr}^{-1}$$

This result is smaller in absolute value than the crude rate difference of -912.8×10^{-5} person-years, as was predictable from the direction of the difference in the age distributions. The amount of the confounding is modest, despite age being a strong risk factor, because the difference in the age distributions between current and past use is also modest. We cannot say that the remaining difference of -720.0×10^{-5} person-years is completely unconfounded by age because our age categorization comprises only two broad age categories, but the pooled estimate removes some of the age confounding. Further control of age confounding might move the estimate further in the same direction, but it is unlikely that age confounding could account for the entire effect of current use on mortality.

What is the confidence interval for the pooled estimate? To obtain the interval, we use the third variance equation in Table 10-4:

$$\begin{aligned} \text{Var}(ID_{MH}) &= \frac{\left(\frac{62,119 \cdot 15,763}{77,882}\right)^2 \left(\frac{196}{62,119^2} + \frac{111}{15,763^2}\right) + \left(\frac{6085 \cdot 2780}{8865}\right)^2 \left(\frac{167}{6085^2} + \frac{157}{2780^2}\right)}{\left(\frac{62,119 \cdot 15,763}{77,882} + \frac{6085 \cdot 2780}{8865}\right)^2} \\ &= \frac{78.644 + 90.394}{209694871.6} = 8.061 \times 10^{-7} \end{aligned}$$

The square root of the variance gives a standard error of 89.8×10^{-5} person-years, for a 90% confidence interval of $(-720.0 \pm 1.645 \cdot 89.8) \times 10^{-5}$ person-years = -867.7×10^{-5} person-years, -572.3×10^{-5} person-years. The narrow confidence interval is the result of the large numbers of observations in the two strata.

The pooled incidence rate ratio for these same data is calculated from Equation 10-5 as

$$IR_{MH} = \frac{\frac{196 \cdot 15,763}{77,882} + \frac{167 \cdot 2780}{8865}}{\frac{111 \cdot 62,119}{77,882} + \frac{157 \cdot 6085}{8865}} = \frac{39.67 + 52.37}{88.53 + 107.77} = 0.47$$

This value indicates that after control of confounding by age in these two age categories, current users have about one half the mortality rate of past users. (We have been using the notation of incidence rate in these formulas, but we are actually describing mortality data. This use is legitimate because a mortality rate is an incidence rate of death.)

The 90% confidence interval for this pooled estimate of the mortality rate ratio can be calculated from the fourth variance equation in Table 10-4:

$$\begin{aligned} \text{Var}[\ln(IR_{MH})] &= \frac{\frac{307 \cdot 62,119 \cdot 15,763}{77,882^2} + \frac{324 \cdot 6085 \cdot 2780}{8865^2}}{\left(\frac{196 \cdot 15,763}{77,882} + \frac{167 \cdot 2780}{8865}\right) \cdot \left(\frac{111 \cdot 62,119}{77,882} + \frac{157 \cdot 6085}{8865}\right)} \\ &= \frac{49.56 + 69.74}{92.04 \cdot 196.30} = \frac{119.30}{18067.4} = 0.00660 \end{aligned}$$

The corresponding standard error is $(0.00660)^{1/2} = 0.081$. The 90% confidence interval for the pooled rate ratio is calculated as

$$IR_L = e^{\ln(0.47) - 1.645 \cdot 0.081} = 0.41$$

$$IR_U = e^{\ln(0.47) + 1.645 \cdot 0.081} = 0.54$$

This confidence interval is narrow, as is that for the rate difference, because there is a large number of deaths in the study. Thus, the study indicates with substantial precision that current users of clozapine had a much lower death rate than past users.

Case-Control Studies

For case-control data, we use the following notation for stratum i of a stratified analysis:

| | Exposed | Unexposed | Total |
|----------|----------|-----------|----------|
| Cases | a_i | b_i | M_{1i} |
| Controls | c_i | d_i | M_{0i} |
| Total | N_{1i} | N_{0i} | T_i |

The pooled incidence rate ratio is estimated as a pooled odds ratio from Equation 10-6:

$$OR_{MH} = \frac{\sum_i \frac{a_i d_i}{T_i}}{\sum_i \frac{b_i c_i}{T_i}} \quad [10-6]$$

The data in Table 10-6 are from a case-control study of congenital heart disease that examined the relation between spermicide use and Down syndrome among the subset of cases who had both congenital heart disease and Down syndrome. The total congenital heart disease case series comprised more than 300 subjects, but the Down syndrome case series was a small subset of the original series that was of interest with regard to the specific issue of a possible relation with spermicide use.

For the crude data, combining the previous strata into a single table, the odds ratio is 3.50. Applying Equation 10-6 gives us an estimate of the effect of spermicide use unconfounded by age.

$$OR_{MH} = \frac{\frac{3 \cdot 1059}{1175} + \frac{1 \cdot 86}{95}}{\frac{104 \cdot 9}{1175} + \frac{5 \cdot 3}{95}} = \frac{2.704 + 0.905}{0.797 + 0.158} = 3.78$$

Table 10.6 INFANTS WITH CONGENITAL HEART DISEASE AND DOWN SYNDROME, AND HEALTHY CONTROLS, BY MATERNAL SPERMICIDE USE BEFORE CONCEPTION AND MATERNAL AGE AT DELIVERY⁶

| | Maternal Age (years), Spermicide Use | | | | | |
|------------|--------------------------------------|-------|-------|------|----|-------|
| | <35 | | | 35+ | | |
| | Yes | No | Total | Yes | No | Total |
| Cases | 3 | 9 | 12 | 1 | 3 | 4 |
| Controls | 104 | 1,059 | 1,163 | 5 | 86 | 91 |
| Total | 107 | 1,068 | 1,175 | 6 | 89 | 95 |
| Odds Ratio | 3.39 | | | 5.73 | | |

Data from Rothman.⁶

This result is slightly larger than the crude estimate of 3.50, indicating that there was modest confounding by maternal age. We can obtain a confidence interval for the pooled estimate from the last variance formula in Table 10-4:

$$\begin{aligned} G_1 &= 2.704 & G_2 &= 0.905 \\ H_1 &= 0.797 & H_2 &= 0.158 \\ P_1 &= 0.904 & P_2 &= 0.916 \\ Q_1 &= 0.096 & Q_2 &= 0.084 \end{aligned}$$

$$\begin{aligned} \text{Var}[\ln(\text{OR}_{\text{MH}})] &= \frac{2.704 \cdot 0.904 + 0.905 \cdot 0.916}{2(2.704 + 0.905)^2} \\ &+ \frac{(2.704 \cdot 0.096 + 0.797 \cdot 0.904) + (0.905 \cdot 0.084 + 0.158 \cdot 0.916)}{2(2.704 + 0.905) \cdot (0.797 + 0.158)} \\ &+ \frac{0.797 \cdot 0.096 + 0.158 \cdot 0.084}{2(0.797 + 0.158)^2} \\ &= 0.126 + 0.174 + 0.049 = 0.349 \end{aligned}$$

The corresponding standard error is $(0.349)^{1/2} = 0.591$. The 90% confidence interval for the pooled odds ratio is calculated as follows:

$$\text{OR}_L = e^{\ln(3.78) - 1.645 \cdot 0.591} = 1.43$$

$$\text{OR}_U = e^{\ln(3.78) + 1.645 \cdot 0.591} = 10.0$$

STANDARDIZATION

Standardization is a method of combining category-specific rates into a single summary value by taking a weighted average. The weights used in averaging come from a *standard* population or distribution. The weights define the standard. Suppose an investigator is standardizing a set of age-specific rates to conform to a specific age standard. He or she may decide to use the U.S. population of 2010 as the standard. That choice means that the weights used to average the age-specific rates reflect the age distribution of the U.S. population in the year 2010. Standardization is a process of averaging the rates in two or more categories using a specified set of weights.

Suppose we have a rate of 10/1000 yr^{-1} for males and a rate of 5/1000 yr^{-1} for females. We can standardize these sex-specific rates to any standard that we wish. A reasonable standard may be one that weights males and females equally. We would then get a weighted average of the two rates that would equal 7.5/1000 yr^{-1} . Suppose the rates reflected the disease experience of nurses, 95% of whom are female. In that case, we may wish to use as a standard a weight of 5% for males and 95% for females. The standardized rate would then be

$$0.05 \times 10/1000 \text{ yr}^{-1} + 0.95 \times 5/1000 \text{ yr}^{-1} = 5.25/1000 \text{ yr}^{-1}$$

If all categories had similar rates, the choice of weights would not matter much. Suppose that males and females had the same rate, 8.0/1000 yr^{-1} . The standardized rate, after standardizing for sex, would have to be 8.0/1000 yr^{-1} , because the standardization would involve taking a weighted average of two values, both of which were 8.0/1000 yr^{-1} . In this situation, the choice of weights is not important. When rates do vary over categories, however, the choice of weights, which means the choice of a standard, can greatly affect the overall summary result. If the standard couples large weights with categories that have high rates, the standardized rate will be high, whereas if it assigns large weights to categories with low rates, the standardized rate will be low. Some epidemiologists prefer not to derive a summary measure when the value of the summary is so dependent on the choice of weights. Nevertheless, it may be convenient or even necessary to obtain a single summary value, in which case a standardized rate at least provides some information about how the category-specific information was weighted, by disclosing which standard was used.

Although an investigator can standardize a single set of rates, the main reason to standardize is to facilitate comparisons; therefore, there are usually two or more sets of rates that are standardized. To compare rates for exposed and unexposed people, we would standardize both groups to the same standard. The standardized comparison is akin to pooling. Both standardization and pooling involve comparing a weighted average of the stratum-specific results. With pooling, the weights for each stratum are buried within the Mantel-Haenszel equations, and their values are not immediately obvious. The built-in weights reflect the information content of the stratum-specific data. These Mantel-Haenszel weights are large for strata that have more information and small for strata that have less information. Because the weighting reflects the amount of information in each stratum, the result of pooling is an overall estimate that is optimal from the point of view of statistical efficiency. That efficiency translates to a narrower confidence interval for the effect estimate than what would be obtained using a less efficient approach. Standardization also assigns a weight to each stratum and also involves taking a weighted average of the results across the strata. Unlike pooling, however, in standardization, the weights may have nothing to do with the amount of data in each stratum. In pooling, the weights come from the data themselves, whereas in standardization, the weights can come from outside the data. The standard may correspond to a specific population of interest or may be chosen arbitrarily. The study population itself could be chosen as the standard, which will lead to an efficient analysis that will approximate the efficiency of pooling, but the standard is not required to be based on the study data.

Standardization also differs from pooling in that pooling requires the assumption that the effect is the same in all strata (often called the *assumption of uniformity of effect*). This assumption is the premise from which the formulas for pooling are derived. As explained earlier, even when the assumption of uniformity of effect is wrong, pooling may still be reasonable. We do not necessarily expect that the effect is strictly uniform across strata when we make the assumption of uniformity; rather, it is an assumption of convenience. We may be willing to tolerate substantial variation in the effect across strata as a price for the convenience and efficiency of pooling as long as we are comfortable with the idea that the actual relation of the effect to the stratification variable is not strikingly different

for different strata. When the effect is strikingly different for different strata, however, we can still use standardization to obtain a summary estimate representing the net effect across strata, because standardization has no requirement that the effect be uniform across strata.

CRUDE RATES AND STANDARDIZED RATES

A crude rate may be thought of as a weighted average of category-specific rates, in which the weights correspond to the actual distribution of the population. Consider age for the purpose of discussion. Every population can be divided into age categories. The age-specific rates in a population can be averaged to get an overall rate. If the averaging uses weights that reflect the amount of the population (or person-time) that actually falls into each age category, the weighted average that results is the crude rate. Algebraically, if each age-specific rate is denoted as A_i/PT_i , where A_i is the number of cases in age category i (i ranging from 1 to K) and PT_i is the number of person-time units in that category, the crude rate is as follows:

$$\frac{PT_1 \frac{A_1}{PT_1} + PT_2 \frac{A_2}{PT_2} + \dots + PT_K \frac{A_K}{PT_K}}{PT_1 + PT_2 + \dots + PT_K} = \frac{\sum A_i}{\sum PT_i} = \frac{A}{PT}$$

A is the total number of cases in the population, and PT is the total person-time. The crude rate is a weighted average of the age-specific rates in which the weights are the same as the denominators for the rates: PT_1, PT_2, \dots, PT_K . These are the *natural* weights, or *latent* weights, for the population. If we change the weights from the denominator values of the rates to an outside set of weights drawn from a standard, the resulting standardized rate can be viewed as the value that the crude rate would have been if the population age structure were changed from what it actually is to that of the standard, and the same age-specific rates applied. A standardized rate is a hypothetical crude rate that would apply if the age structure were that of the standard instead of what it happens to be.

Although standardization is preferable to pooling when an effect apparently varies across strata, standardization may be desirable even when pooling is a reasonable alternative, simply because standardization uses a defined set of weights to combine results across strata. This characteristic of standardization provides for better comparability of stratified results from one study to another or in comparing different subgroups within a study. Standardization can guarantee that differences in the distribution of the standardized variable cannot account for any differences in the summary measures of the exposure effect. In contrast, with pooling, the weights are different for every summary measure, because they come from the data that are being summarized, and therefore differences between pooled summary measures may be influenced by differences in the stratification variable. We say that pooled measures are internally unconfounded (ie, comparing the

exposed with unexposed within the measure) but not externally unconfounded (ie, comparing two different summary measures). Standardized estimates that use the same standard weights are both internally and externally unconfounded.

Consider the data on clozapine use and mortality in Table 10-5. We obtained a pooled estimate of the mortality rate difference, using the Mantel-Haenszel approach, of $-720 \times 10^{-5} \text{ yr}^{-1}$. Suppose we chose instead to standardize the rates for age over the two age categories. First we must choose an age standard to use. We might standardize to the age distribution of current clozapine use in the study, because that is a reasonable approximation for the age distribution of those who use the drug. There were a total of 68,204 person-years of current clozapine use, of which 62,119 (91.1%) were in the younger age category. To standardize the death rate for past users to this standard, we take a weighted average of past use as follows:

$$0.911 \times 704.2/100,000 \text{ yr}^{-1} + 0.089 \times 5647/100,000 \text{ yr}^{-1} \\ = 1144/100,000 \text{ yr}^{-1}$$

The standardized rate for current users, standardized to the age distribution of current users, is the same as the crude rate for current users, which is $532.2/100,000 \text{ yr}^{-1}$. The *standardized rate difference* is the difference between the standardized rates for current and past users, which is $(532.2 - 1144)/100,000 \text{ yr}^{-1} = -612/100,000 \text{ yr}^{-1}$, slightly smaller in absolute value than the $-720/100,000 \text{ yr}^{-1}$ that was obtained from the pooled analysis. Analogously, we can obtain the *standardized rate ratio* by dividing the rate among current users by that among past users, giving a result of $532.2/1144 = 0.47$, essentially identical to the result obtained through pooling. The stratum-specific rate ratios did not vary much, so any weighting, whether pooled or standardized, will produce a result close to this value.

Both pooling and standardization can be used to control confounding. Because they are different approaches and can give different results, it is fair to ask why we would want to use one rather than the other. Both involve taking weighted averages of the stratum-specific results. The difference is where the weights come from. In pooling, the data determine the weights, which are derived mathematically to give statistically optimal results. This method gives precise results (ie, relatively narrow confidence intervals), but the weights are statistical constructs that come out of the data and cannot easily be specified. Standardization, unlike pooling, may involve weights that are inefficient if large weights are assigned to strata with little data and vice versa. On the other hand, the weights are explicit. Ideally, the weights used in standardization should be presented along with the results. Making the weights used in standardization explicit facilitates comparisons with other data. Standardization may be less efficient, but it may provide for better comparability. A more detailed discussion of standardization, including appropriate confidence interval equations for standardized results, can be found in Rothman, Greenland and Lash⁷ (see pages 265-269 of that text).

In a stratified analysis, another option that is always open is to stratify the data and to present the results without aggregating the stratum-specific information over the strata. Stratification is highly useful even if it does not progress beyond the examination of the stratum-specific findings. This approach to presenting the data is especially attractive when the effect measure of interest appears to change

WHAT IS AN SMR?

When the standardized rate ratio is calculated using the exposed group as the standard, the resulting standardized rate ratio is usually referred to as a *standardized mortality ratio* or *standardized morbidity ratio* (SMR). The standardized rate ratio for clozapine that is calculated using the age distribution of current users as the age standard is an example of an SMR. An SMR can be expressed as the ratio of the total number of deaths in the exposed group, which was 363 in the clozapine example, divided by the number expected in the exposed group if the rates among the unexposed prevailed within each of the age categories. For the 10- to 54-year-old age group, if the rate among past users of 704.2/100,000 yr⁻¹ had prevailed among the 62,119 person-years experienced by current users, there would have been 437.4 deaths expected in that age category. Similar calculations give 343.6 deaths expected in the 55- to 94-year-old age category. The figure for total expected deaths is 437.4 + 343.6 = 781.0. The SMR is the ratio of observed to expected deaths, which is 363/781.0 = 0.47. This result is algebraically identical to standardization based on taking a weighted average of the age-specific rates and taking the age distribution of current users as the standard.

The SMR is sometimes claimed to result from a method of standardization called *indirect standardization*, as opposed to *direct standardization*. Direct standardization is what we have been describing as standardization. Indirect standardization is a misnomer. The method is actually the same as direct standardization, but it has one additional feature, which is that the standard is always the exposed group. It is sometimes described differently, but mathematically the calculations are the same as direct or ordinary standardization, with the proviso that the standard is the exposed group.

considerably across the strata. In this situation, a single summary estimate is less attractive an option than in a situation in which the effect measure is almost constant across strata.

CALCULATION OF P VALUES FOR STRATIFIED DATA

Earlier, we gave the reasons why estimation is preferable to statistical significance testing. Nevertheless, for completeness, the formulas for calculating *P* values from stratified data are given here. These formulas are straightforward extensions of the formulas presented in Chapter 9 for crude data.

For risk, prevalence, or case-control data, all of which consist of a set of 2 × 2 tables, chi can be calculated as follows:

$$\chi = \frac{\sum_i a_i - \sum_i \frac{N_{1i} M_{1i}}{T_i}}{\sqrt{\sum_i \frac{N_{1i} N_{0i} M_{1i} M_{0i}}{T_i^2 (T_i - 1)}}$$

Applying this formula to the case-control data in Table 10-6 gives the following chi statistic:

$$\chi = \frac{(3+1) - \left(\frac{12 \cdot 107}{1175} + \frac{4 \cdot 6}{95} \right)}{\sqrt{\frac{107 \cdot 1068 \cdot 12 \cdot 1163}{1175^2 \cdot 1174} + \frac{6 \cdot 89 \cdot 4 \cdot 91}{95^2 \cdot 94}}} = 2.41$$

This result translates to a *P* value of 0.016 (see Appendix).

For rate data, the corresponding formula is as follows:

$$\chi = \frac{\sum_i a_i - \sum_i \frac{PT_{1i} M_i}{T_i}}{\sqrt{\sum_i M_i \frac{PT_{1i} PT_{0i}}{T_i^2}}}$$

Applying this equation to the data in Table 10-5, we obtain the following:

$$\chi = \frac{(196+167) - \left(\frac{62,119 \cdot 307}{77,882} + \frac{6085 \cdot 324}{8865} \right)}{\sqrt{\frac{307 \cdot 62,119 \cdot 15,763}{77,882^2} + \frac{324 \cdot 6085 \cdot 2780}{8865^2}}} = -9.55$$

This result is too large in absolute value to be found in the Appendix table, implying an extremely small *P* value.

MEASURING CONFOUNDING

The control of confounding and the assessment of confounding are closely intertwined. It may seem reasonable to assess how much confounding a given variable produces in a body of data before we control for that confounding. The assessment may indicate, for example, that there is not enough confounding to present a problem, and we may therefore ignore that variable in the analysis. It is possible to predict the amount of confounding from the general characteristics of confounding variables, that is, the associations of a confounder with both exposure and disease. To measure confounding directly, however, requires that we control it: The procedure is to remove the confounding from the data and then see how much has been removed.

An example of the measurement of confounding can be found in Tables 1-1 and 1-2 (see Chapter 1). In Table 1-1, we have a risk of death over a 20-year period of 0.24 among smokers and 0.31 among nonsmokers. The crude risk ratio is 0.24/0.31 = 0.76, indicating a risk among smokers that is 24% lower than that among nonsmokers. As was indicated in Chapter 1 and earlier in this chapter, this apparent protective effect of smoking on the risk of death is confounded by age, which can be seen from the data in Table 1-2. The age confounding can be

removed by applying Equation 10-2, which gives a result of 1.21. This value indicates a risk of death among smokers that is 21% greater than that of nonsmokers. The discrepancy between the crude risk ratio of 0.76 and the unconfounded risk ratio of 1.21 is a direct measure of the age confounding. Were these two values equal, there would be no indication of age confounding in the data. To the extent that they differ, it indicates the presence of age confounding. The age confounding is strong enough in this instance to have reversed the apparent effect of smoking, making it appear that smoking is related to a reduced risk of death in the crude data. This biased result occurs because the smokers tend to be younger than the nonsmokers, and the crude comparison between smokers and nonsmokers is to some extent a comparison of younger women with older women, mixing the smoking effect with an age effect that negates it. By stratifying, the age confounding can be removed, revealing the adverse effect of smoking. The direct measure of this confounding effect is the comparison of the pooled estimate of the risk ratio with the crude estimate of the risk ratio.

A common mistake is to use statistical significance tests to evaluate the presence or absence of confounding. This mistaken approach to the evaluation of confounding applies a significance test to the association between a confounder and the exposure or the disease. The amount of confounding, however, is a result of the strength of the two associations between the confounder and both exposure and disease. Confounding does not depend on the statistical significance of these associations, only the magnitude of the associations. Furthermore, a significance test evaluates only one of the two component associations that give rise to confounding. A common situation in which this mistaken approach to evaluating confounding is applied is in the analysis of randomized trials, when baseline characteristics are compared for the randomized groups. Baseline comparisons are useful, but they often are conducted with the sole aim of checking for statistically significant differences in any of the baseline variables as a means of detecting confounding. A better way to evaluate confounding in a trial or any study is to control for the potential confounder and determine the extent to which the unconfounded result differs from the crude, potentially confounded, result.

STRATIFICATION BY TWO OR MORE VARIABLES

For convenience of presentation, the examples in this chapter have used few strata with only one stratification variable. Nevertheless, stratified analysis can be conducted with two or more stratification variables. Suppose that an investigator wished to control confounding by sex and age simultaneously, with five age categories. The combination of age and sex categories will produce 10 strata. All of the methods discussed in this chapter can be applied without any modification to a stratified analysis with two or more stratification variables. The only real difficulty with such analyses is that with several variables to control, the number of strata increases quickly and can stretch the data too far. Controlling five different variables with three categories each in a stratified analysis would require $3 \times 3 \times 3 \times 3 \times 3 = 243$ strata. With so many strata, many of them would contain few observations and would end up contributing little or no information to the data summary. When the numbers within strata become very small, and in particular when zeroes

become frequent in the tables, some tables may not contribute any information to the summary measures, and some of the study information is effectively lost. As a result, the analysis as a whole becomes less precise. Consequently, stratified analysis is not a practical method to control for many confounding factors at once. Fortunately, it is rare to have substantial confounding by many variables at once.

STRATIFICATION AFTER MATCHING

When matching is used in study design to control confounding, the matching should be taken into account in the data analysis. As described in Chapter 7, the implications of matching are very different in case-control studies and in cohort studies. In cohort studies, matching on potential confounding factors prevents confounding by creating a balance of risk factors for the outcome in the compared cohort. As a result, the matching can be ignored in the analysis without introducing any bias; matching has done its job in the selection of subjects, which suffices to prevent confounding by the matched factors. Even so, it is worthwhile to take the matching into account in the analysis by stratifying by the matched sets of subjects. Doing so does not remove any confounding, which has already been prevented, but it can lead to narrower confidence intervals than would be obtained if the matching had been ignored.

For case-control studies, unlike cohort studies, matching by one or more factors that are related to exposure will result in selection bias, which must be removed in the data analysis (see Chapter 7). Stratification by the matched sets (each set consisting of a case and its matched controls is an individual stratum) can accomplish this goal. If some matched sets have the same values for all the matching factors, they can be lumped together into one stratum, which may narrow the resulting confidence interval. Usually, epidemiologists employ a specialized regression model rather than stratification to remove the bias introduced by matching in case-control studies. This model is the *conditional logistic* model, which is a version of the logistic regression model that conditions on the sets that comprise a case and all its matched controls (see Chapter 12 for a discussion of logistic models). Conditional logistic models provide the same or almost the same result as stratified analysis, but they have the advantage of allowing the investigator to include in the regression model other confounders that were not matched, something not easily accomplished in a stratified analysis.

IMPORTANCE OF STRATIFICATION

The equations in this chapter may look imposing, but they can be applied readily with a hand calculator, a spreadsheet, or a pencil and paper. Consequently, the methods described to control confounding are widely accessible without heavy reliance on technology. These are not the only methods available to control confounding. In Chapter 12, we discuss multivariate modeling to control confounding. Multivariate modeling requires computer hardware and software but offers the possibility of convenient methods to control confounding not merely for a single variable but simultaneously for a set of variables. The allure of these multivariate

methods is almost irresistible. Nevertheless, stratified analysis is preferable and should always be the method of choice to control confounding. This is not to say that multivariate modeling should be ignored; it does have its uses. Stratification, however, is a preferred approach, at least as the initial approach to data analysis. Stratification has several advantages over multivariate analysis:

1. With stratified analysis, the investigator can visualize the distribution of subjects by exposure, disease, and the potential confounder. Strange features in the distributions of the major variables, such as data that have been miscoded during programming, can become immediately apparent. Regression models do not divulge this kind of information as readily.
2. Not only the investigator, but the consumer of the research as well can visualize the distributions. Indeed, from detailed tables of stratified data, a reader can check the calculations or conduct his or her own pooled or standardized analysis.
3. Fewer assumptions are needed for a stratified analysis, reducing the possibility of obtaining a biased result.

It should be standard practice to examine the data by categories of the primary potential confounding factors, that is, to conduct a stratified analysis. A multivariate analysis rarely changes the interpretation produced by a competent stratified analysis. The stratified analysis can keep the researcher and the reader better informed about the nature of the data. Even when it is reasonable to conduct a multivariate analysis, it should be undertaken only after the researcher has conducted a stratified analysis and has a good appreciation for the confounding in the data or lack of it by the main study variables.

QUESTIONS

1. In Table 10-3, the crude value of the risk ratio is 1.44, which is between the values for the risk ratio in the two age strata. Could the crude risk ratio have been outside the range of the stratum-specific values, or must it always fall within the range of the stratum-specific values? Why or why not?
2. The pooled estimate for the risk ratio from Table 10-3 was 1.33, also within the range of the stratum-specific values. Does the pooled estimate always fall within the range of the stratum-specific estimates of the risk ratio? Why or why not?
3. If you were comparing the effect of exposure at several levels and needed to control confounding, would you prefer to compare a pooled estimate of the effect at each level or a standardized estimate of the effect at each level? Why?
4. Prove that an SMR is directly standardized to the distribution of the exposed group; that is, prove that an SMR is the ratio of two standardized rates that are both standardized to the distribution of the exposed group.

5. Suppose that an investigator conducting a randomized trial of an old and a new treatment examines baseline characteristics of the subjects (eg, age, sex, stage of disease) that may be confounding factors and finds that the two groups are different with respect to several characteristics. Why is it unimportant whether these differences are "statistically significant"?

6. Suppose one of the differences in question 5 is statistically significant. A significance test is a test of the null hypothesis, which is a hypothesis that chance alone can account for the observed difference. What is the explanation for baseline differences in a randomized trial? What implication does that explanation have for dealing with these differences?

7. The larger a randomized trial, the smaller the expected confounding. Why? Explain why the size of a study does not affect confounding in nonexperimental studies.

8. Imagine a stratum of a case-control study in which all subjects were unexposed. What is the mathematic contribution of that stratum to the estimate of the pooled odds ratio (see Equation 10-6)? What is the mathematic contribution of that stratum to the variance of the pooled odds ratio (see bottom equation in Table 10-4)?

REFERENCES

1. Rothman KJ, Monson RR. Survival in trigeminal neuralgia. *J Chron Dis.* 1973;26:303-309.
2. Mantel N, Haenszel WH. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst.* 1959;22:719-748.
3. University Group Diabetes Program. A study of the effects of hypoglycemic agents on vascular complications in patients with adult onset diabetes. *Diabetes.* 1970;19(Suppl. 2):747-830.
4. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott-Raven; 2008: Chapter 15, Introduction to Stratified Analysis, pp 279-280.
5. Walker AM, Lanza LL, Arellano F, Rothman KJ: Mortality in current and former users of clozapine. *Epidemiology* 1997;8:671-677.
6. Rothman KJ: Spermicide use and Down syndrome. *Am J Public Health.* 1982;72:399-401.
7. Rothman KJ, Greenland SL, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott-Raven; 2008: Chapter 15, Introduction to Stratified Analysis, pp 265-269.