8

Controlling Confounding by Stratifying Data

Earlier we saw that the apparent effect of birth order on the prevalence at birth of Down syndrome (Fig. 5–2) is attributable to confounding. As demonstrated in Figure 5–3, maternal age has an extremely strong relation to the prevalence of Down syndrome. In Figure 5–4, which classifies the Down syndrome data simultaneously by birth order and maternal age, we can see 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 5–4 is a graphic demonstration of *stratification. Stratification* is basically the cross-tabulation of data; usually, stratification refers to cross-tabulation of data on exposure and disease by categories of one or more other variables that are potential confounding variables. We saw another example of stratification 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 simple formulas to derive an unconfounded estimate of an effect from stratified data.

An Example of Confounding

First, let us consider another example of confounding. The data in Table 8–1 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 effect of being male on the death rate of trigeminal neuralgia patients is confounded, however. Table 8–2 shows the data stratified into two age groups, split at age 65. The age

Table 8–1. Mortality rates among patients with trigeminal neuralgia, by sex*

	Males	Females
Deaths	90	131
Person-years	2465	3946
Mortality rate	36.5/1000 person-years	33.2/1000 person-years
Rate ratio	1.10	
90% CI	0.88–1.38	

*Data from Rothman and Monson.[1]

one might predict, 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 in the under 65 category, whereas the female person-time is predominantly in the 65 or older category. Thus, the female experience is older than the male experience. This age difference would tend to produce a lower overall death rate in males relative to females, because to some extent comparing the death rate among males with that among 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 female patients tend to be older than 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.

Table 8–2. Mortality rates among patients with trigeminal neuralgia, by sex and age category*

	Age (years)			
	<65		*≥65*	
	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	

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. Below, we use this trigeminal neuralgia example and examples of other types of data to obtain unconfounded effect estimates using stratification.

Unconfounded Effect Estimates and Confidence Intervals from Stratified Data

How does stratification control confounding? Confounding, as explained in Chapter 5, 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 comes about 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 above, comparing men with women was also a comparison of younger people (the men in the study) with older people (the 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, but 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, like age, then the more categories that are used to form strata, the less variability by age there will be within those categories.

A stratified analysis can be as straightforward as a presentation of the data within each of the strata. Often, however, the investigator hopes to summarize the relation between exposure and disease in a simple way. The way to do that is to make the essential comparisons within each stratum and then to aggregate the information from these comparisons over all strata. There are two methods to aggregate the information over strata, *pooling* and *standardization*, each with its own formula for combining the data across 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

With this assumption, one can view each stratum as providing a separate estimate (referred to as a *stratum-specific estimate*) of the overall effect. With each stratum providing a separate estimate of effect, the principle behind pooling is simply 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 others. In pooling, weights are assigned so that the strata providing the most information, that is, the strata with the most data, get the most weight. In the formulas presented below, this weighting is built directly into the calculations. When the data do not conform to the assumption necessary for pooling that the effect is constant across all strata, pooling is not applicable. In such a situation, it is still possible to obtain an unconfounded summary estimate of the effect over the strata using *standardization*, which is discussed later in this chapter.

Cohort Studies with Risk Data (or Prevalence Data)

Let us consider risk data. (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 would be 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 the following formula.

$$RD_{\mathrm{MH}} = \frac{\sum_i \frac{a_i N_{0i} - b_i N_{1i}}{T_i}}{\sum_i \frac{N_{1i} N_{0i}}{T_i}} \qquad (8\text{–}1)$$

Σ 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 formula is one of a group of formulas for pooled estimates that derive from an approach originally introduced by Mantel and Haenszel.[2]

The formula for the pooled risk ratio from stratified risk or prevalence

$$RR_{MH} = \frac{\sum_i \frac{a_i N_{0i}}{T_i}}{\sum_i \frac{b_i N_{1i}}{T_i}} \qquad (8\text{–}2)$$

Example: Stratification of Risk Data

To illustrate the stratification of risk data, let us revisit the example of the University Group Diabetes Program (Tables 5–3 and 5–4). For convenience, the age-specific data are repeated here in Table 8–3. First, we consider the risk difference. From the crude data, the risk difference is 4.5%. Contrary to expectations, the tolbutamide group experienced a greater risk of death than the placebo group, despite the fact 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. 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 8–3: 48% (98/204) of the tolbutamide group is at least 55 years of age, whereas only 41% (85/205) of the placebo group is at least 55 years of age. We know that older people have a greater risk of death, a relation that is also evident in Table 8–3. 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. Therefore, we would 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 8–3.

Table 8–3. Risk of death for groups receiving tolbutamide or placebo in the University Group Diabetes Program, overall and by age category (1970)*

	Age					
	<55		≥55		*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	

We obtain a pooled estimate of the risk difference by applying formula 8–1, as follows.

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

The result, 3.5%, is smaller than the risk difference in the crude data, 4.5%. Note that 3.5% is within the narrow range of the two stratum-specific risk differences in Table 8–3, 3.4% for age <55 and 3.6% for age ≥55. Mathematically, the pooled estimate is a weighted average of the stratum-specific values, so 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 the 3.5% as a more appropriate estimate than the estimate from the crude data, as it removes age confounding. The crude risk difference differs from the unconfounded estimate because the crude estimate reflects not only the effect of tolbutamide (which we estimate to be 3.5% from this analysis) but also 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, then 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.

The unconfounded estimate of 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 unconfounded by age is smaller than 3.5%.

We can also calculate a pooled estimate of the risk ratio from the data in Table 8–3, using formula 8–2.

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

This result, like that for the risk difference, is closer to the null value than the crude risk ratio of 1.44, indicating that age confounding has

Residual confounding

The two age categories in Table 8–3 may not be sufficient to control all of the age confounding in the data. In general, more strata, with narrower boundaries, will control confounding more effectively than fewer strata with broader boundaries. If age strata (or strata by any continuous 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 from the beginning.

To avoid within-stratum residual confounding, it is desirable to carve the data into more strata and to avoid open-ended strata (such as 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.

within the range of the stratum-specific estimates, as it must be. Note, however, that for the risk ratio, the stratum-specific estimates for the data in Table 8–3, 1.81 and 1.19, differ considerably from one another. The wide range between them includes not only the pooled estimate but also the estimate of effect from the crude data. When the stratum-specific estimates of effect are nearly identical, as they were for the risk differences in the data in Table 8–3, 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 will not be as clear on inspection what the pooled estimate will be.

As stated above, the formulas to obtain pooled estimates are premised on the assumption that the effect is constant across strata. Thus, the pooled risk ratio of 1.33 for the above 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 strata. This assumption seems reasonable for the risk difference calculation, for which the two strata gave nearly 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 nearly the same, in each stratum. It allows for statistical variation over the strata. It is possible to conduct a statistical evaluation of this assumption, called 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 keep 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 8–3, 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 one undertakes a more formal statistical evaluation of the assumption of uniformity for these data, it would support the view that the assumption of a uniform risk ratio for the data in Table 8–3 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 8–4 lists variance formulas for the various pooled estimates that we consider in this chapter.

Although the formulas 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 8–3, we calculated that RD_{MH} was 0.035. We can derive the variance for this estimate, and thus a confidence interval, by applying the first formula from Table 8–4 to the data in Table 8–3.

$$\text{Var}(RD_{\text{MH}}) = \frac{\left(\frac{106 \cdot 120}{226}\right)^2\left(\frac{8 \cdot 115}{106^2 \cdot 105} + \frac{5 \cdot 98}{120^2 \cdot 119}\right) + \left(\frac{98 \cdot 85}{183}\right)^2\left(\frac{22 \cdot 69}{98^2 \cdot 97} + \frac{16 \cdot 76}{85^2 \cdot 84}\right)}{\left[\left(\frac{106 \cdot 120}{226}\right) + \left(\frac{98 \cdot 85}{183}\right)\right]^2}$$

$$= \frac{3.3761 + 7.5278}{10{,}363.7} = 0.001052$$

This gives a standard error of $(0.001052)^{1/2} = 0.0324$ and a 90% confidence interval of $0.035 \pm 1.645 \cdot 0.0324 = 0.035 \pm 0.053 = -0.018$ to 0.088. The confidence interval is broad enough to indicate a fair amount of statistical uncertainty in the finding that tolbutamide is worse than placebo. Notably, however, the data are not compatible with any com-

Table 8–4. Variance formulas for pooled analyses

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

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

Incidence rate difference: $$\mathrm{Var}(ID_{\mathrm{MH}}) = \frac{\sum_i (PT_{1i}PT_{0i}/T_i)^2 \, (a_i/PT_{1i}^2 + b_i/PT_{0i}^2)}{\left(\sum_i (PT_{1i}PT_{0i}/T_i)\right)^2}$$

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

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

where

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

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

One might also construct a confidence interval for the risk ratio estimated from the same stratified data. In that case, one would use the second formula in Table 8–4, setting limits on the log scale, as we did in the previous chapter for crude data.

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

This result gives a standard error for the logarithm of the RR of $(0.0671)^{1/2} = 0.259$ and a 90% confidence interval of 0.87–2.0.

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

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

The interpretation for this result is similar to that for the confidence interval of the risk difference, which is as one would expect since 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 formula 8–2. This calculation gives an overall Mantel-Haenszel risk ratio of 1.21, with a 90% confidence interval of 1.06–1.38. The Mantel-Haenszel risk ratio not only is different from the crude risk ratio of 0.76 but, as noted 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 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 the following formula.

$$ID_{\mathrm{MH}} = \frac{\sum_i \frac{a_i PT_{0i} - b_i PT_{1i}}{T_i}}{\sum_i \frac{PT_{1i}PT_{0i}}{T_i}} \qquad (8\text{–}4)$$

A pooled estimate of the rate ratio may be estimated as follows:

$$IR_{\mathrm{MH}} = \frac{\sum_i \frac{a_i PT_{0i}}{T_i}}{\sum_i \frac{b_i PT_{1i}}{T_i}} \qquad (8\text{–}5)$$

Table 8–5. Mortality rates for current and past clozapine users, overall and by age category*

	Age (years)					
	10–54		*55–94*		*Total*	
	Current	**Past**	**Current**	**Past**	**Current**	**Past**
Deaths	196	111	167	157	363	268
Person-years	62,119	15,763	6085	2780	68,204	18,543
Rate ($\times 10^5$ years)	315.5	704.2	2744	5647	532.2	1445
Rate difference ($\times 10^5$ years)	−388.7		−2903		−912.8	
Rate ratio	0.45		0.49		0.37	

*Data from Walker et al.[5]

As an illustration, consider the rate data in Table 8–5. These data come from a study of mortality rates among current users and past users of clozapine, a drug used to treat schizophrenia. As clozapine is thought to affect mortality primarily for current users, the experience of past users was used as the reference by which to judge 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 one would expect: among schizophrenia patients, just 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 estimate the mortality rate difference that is unconfounded by age (apart from any residual age confounding within these broad age categories) from formula 8–4.

$$ID_{MH} = \frac{\dfrac{196 \cdot 15{,}763 - 111 \cdot 62{,}119}{77{,}882} + \dfrac{167 \cdot 2780 - 157 \cdot 6085}{8865}}{\dfrac{62{,}119 \cdot 15{,}763}{77{,}882} + \dfrac{6085 \cdot 2780}{8865}}$$

$$= \frac{-48.864 - 55.396}{12{,}572.633 + 1908.212} = -720.0 \times 10^{-5}\ \text{yr}^{-1}$$

This result is smaller 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 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 formula in Table 8–4.

$$\text{Var}(ID_{MH}) = \frac{\left(\dfrac{62{,}119 \cdot 15{,}763}{77{,}882}\right)^2\left(\dfrac{196}{62{,}119^2} + \dfrac{111}{15{,}763^2}\right) + \left(\dfrac{6085 \cdot 780}{8865}\right)^2\left(\dfrac{167}{6085^2} + \dfrac{157}{2780^2}\right)}{\left(\dfrac{62{,}119 \cdot 15{,}763}{77{,}882} + \dfrac{6085 \cdot 2780}{8865}\right)^2}$$

$$= \frac{78.644 + 90.394}{209{,}694{,}871.6} = 8.061 \times 10^{-7}$$

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 $= -867.7 \times 10^{-5}$ 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 formula 8–5 as follows.

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

This value indicates that after control of confounding by age in these two age categories, current users have half the mortality rate of past