ıe formulas, but we are actually describing mortality data. This use is ·gitimate because a mortality rate is an incidence rate of death.)

The 90% confidence interval for this pooled estimate of the mortality ıte ratio can be calculated from the fourth variance formula in Table -4.

$$\text{Var}[\ln(IR_{MH})] = \frac{\dfrac{307 \cdot 62{,}119 \cdot 15{,}763}{77{,}882^2} + \dfrac{324 \cdot 6085 \cdot 2780}{8865^2}}{\left(\dfrac{196 \cdot 15{,}763}{77{,}882} + \dfrac{167 \cdot 2780}{8865}\right) \cdot \left(\dfrac{111 \cdot 62{,}119}{77{,}882} + \dfrac{157 \cdot 6085}{8865}\right)}$$

$$= \frac{49.56 + 69.74}{92.04 \cdot 196.30} = \frac{119.30}{18{,}067.4} = 0.00660$$

he corresponding standard error is $0.00660^{1/2} = 0.081$. The 90% confience interval for the pooled rate ratio is calculated as follows:

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

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

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

·ase-Control Studies

or case-control data, we use the following notation for stratum i of a ·ratified 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 ·om the following formula.

$$OR_{MH} = \frac{\sum_i \dfrac{a_i d_i}{T_i}}{\sum_i \dfrac{b_i c_i}{T_i}} \qquad (8\text{–}6)$$

·he data in Table 8–6 are from a case-control study of congenital heart

Table 8–6. Infants with congenital heart disease and Down syndrome and healthy controls, by maternal spermicide use before conception and maternal age at delivery*

	Maternal Age (years), Spermicide Use					
	<35			≥35		
	Yes	**No**	**Total**	**Yes**	**No**	**Total**
Cases	3	9	12	1	3	4
Controls	104	1059	1163	5	86	91
Total	107	1068	1175	6	89	95
Odds ratio		3.39			5.73	

*Data from Rothman.[6]

syndrome among the subset of cases that 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 above strata into a single table, the odds ratio is 3.50. Applying formula 8–6 gives us an estimate of the effect of spermicide use unconfounded by age.

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

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 8–4.

$$G_1 = 2.704 \quad G_2 = 0.905$$
$$H_1 = 0.797 \quad H_2 = 0.158$$
$$P_1 = 0.904 \quad P_2 = 0.916$$
$$Q_1 = 0.096 \quad Q_2 = 0.084$$

$$\text{Var}[\ln(OR_{MH})]$$

$$= \frac{2.704 \cdot 0.904 + 0.905 \cdot 0.916}{2 \cdot [illegible]}$$

$$+ \frac{(2.704 \cdot 0.096 + 0.797 \cdot 0.904) + (0.905 \cdot 0.084 + 0.158 \cdot 0.916)}{2(2.704 + 0.905) \cdot (0.797 + 0.158)}$$

$$+ \frac{0.797 \cdot 0.096 + 0.158 \cdot 0.084}{2(0.797 + 0.158)^2}$$

$$= 0.126 + 0.174 + 0.049 = 0.349$$

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

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

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

Standardization

ndardization is a method of combining category-specific rates into a gle summary value by taking a weighted average of them. It weights category-specific rates using weights that come from a *standard* pop-ation. The weights, in fact, define the standard. Suppose one is stan-rdizing a set of age-specific rates to conform to a specific age stan-rd. One might decide to use the U.S. population in the year 2000 as standard. That choice means that the weights used to average the e-specific rates reflect the distribution of the U.S. population in the ar 2000. Standardization is thus a process of weighting the rates in o or more categories by a specified set of weights.

Suppose we have a rate of 10/1000 yr^{-1} for males and a rate of 5/ 00 yr^{-1} for females. We can standardize these sex-specific rates to any ndard that we wish. A reasonable standard might be one that weights les and females equally. We would then obtain a weighted average of two rates that would equal 7.5/1000 yr^{-1}. Suppose the rates re-cted the disease experience of nurses, 95% of whom are female. In t case, we might wish to use as a standard a weight of 5% for males d 95% for females. The standardized rate would then be as follows:

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

all categories had similar rates, the choice of weights would matter le. Suppose that males and females had the same rate, 8.0/1000 yr^{-1}. en the standardized rate, after standardizing for sex, would have to 8.0/1000 yr^{-1} because the standardization would involve taking a ighted average of two values, both of which were 8.0/1000 yr^{-1}. In ch a situation, the choice of weights is not important. When rates do ry over categories, however, the choice of weights, which is to say 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. On the other hand, it may be convenient or even necessary to obtain a single summary value, in which case a standardized rate provides at least some information about how the category-specific information was weighted, by disclosing which standard was used.

Although one 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. If we wish 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 formulas and, thus, 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. Standardization also assigns a weight to each stratum and 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. Thus, in pooling, the weights come from the data themselves, whereas in standardization, the weights can come from outside the data and simply reflect the distribution of the standard, which may correspond to a specific population or be chosen arbitrarily.

Standardization also differs from pooling in that pooling assumes 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 of the 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 obtain 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 (ranging from 1 to K) and PT_i is the number of person-time units in that category, the crude rate is as follows:

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

A is the total number of cases in the population and PT is the total person-time. The crude rate is thus a weighted average of the age-specific rates, where the weights are the same as the denominators for the rates: PT_1, PT_2, . . . PT_K. These are the *natural* weights, or *latent* weights, for the population. If we now 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. Thus, 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.

ıen pooling is a reasonable alternative, simply because standardiza-
n uses a defined set of weights to combine results across strata. This
ıracteristic of standardization provides for better comparability of
ıtified results from one study to another or within a study.
Consider the data on clozapine use and mortality in Table 8–5. We
ained a pooled estimate of the mortality rate difference, using the
ntel-Haenszel approach, of -720×10^{-5} yr^{-1}. Suppose we chose
tead to standardize the rates for age over the two age categories.
ıat age standard might we use? Let us standardize to the age distribu-
ı of current clozapine use in the study, since that is the age distribu-
ı of those who use the drug. There were a total of 68,204 person-years

What is an SMR?

When the standardized rate ratio is calculated using the exposed group as the standard, the result is usually referred to as a *standardized mortality,* or *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, 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. Thus, for the 10–54 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 55–94 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. That is a misnomer, however, as there is nothing indirect about indirect standardization. Indeed, the only feature that distinguishes it from supposedly direct standardization is that for an SMR the standard is always the exposed group. The calculations for any rate standardization, direct or indirect, are basically the same.

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 their age distribution, is the same as the crude rate for current users, which is 532.2/100,000 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 yr^{-1} = −612/100,000 yr^{-1}, slightly smaller in absolute value than the −720/100,000 yr^{-1} 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 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.

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 (that is, 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 comparison with other data. Thus, standardization may be less efficient, but it may provide for better comparability. For a more detailed discussion of standardization, including appropriate confidence interval formulas for standardized results, see Rothman and Greenland.[7]

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 examining the stratum-specific findings. This approach to presenting the data is especially attractive when the effect measure of interest appears to change considerably across the strata. In such a situation, a single summary estimate is less attractive an option than it would be in a situation in which the effect measure is nearly 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, we give here the formulas for calculating *p* values from stratified data. These are straightforward extensions of the formulas presented in Chapter 7 for crude data.

For risk, prevalence, or case-control data, all of which consist of a set of 2 × 2 tables, the χ formula is 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 8–6 gives the

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

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 formula to the data in Table 8–5, 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, for the Appendix, implying an extremely small *p* value.

Measuring Confounding

The control of confounding and assessment of confounding are closely intertwined. It might seem reasonable to assess how much confounding a given variable produces in a body of data before we control for that confounding. The assessment might 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.

As an example of the measurement of confounding, let us return to the data in Tables 1–1 and 1–2. In Table 1–1, we have risks 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

ted both in Chapter 1 and earlier in this chapter, this apparent protec-
/e effect of smoking on the risk of death is confounded by age, which
n be seen from the data in Table 1–2. The age confounding can be
moved by applying formula 8–2, which gives a result of 1.21. This
ılue indicates a risk of death among smokers that is 21% greater than
at of nonsmokers. The discrepancy between the crude risk ratio of 0.76
ıd the unconfounded risk ratio of 1.21 is a direct measure of age con-
unding. Were these two values equal, there would be no indication of
nfounding in the data. To the extent that they differ, it indicates the
esence of age confounding. The age confounding is strong enough, in
is instance, to have reversed the apparent effect of smoking, making it
ıpear that smoking is related to a reduced risk of death in the crude
ıta. This biased result occurs because smokers tend to be younger than
ınsmokers, so the crude comparison between smokers and non-
ıokers is to some extent a comparison of younger women with older
ɔmen, mixing the smoking effect with an age effect that negates it. By
ratifying, the age confounding can be removed, revealing the adverse
ect of smoking. The direct measure of this confounding effect is a
mparison of the pooled estimate of the risk ratio with the crude esti-
ate of the risk ratio.

A common mistake is to use statistical significance tests to evaluate
e presence or absence of confounding. This mistaken approach to the
aluation of confounding applies a significance test to the association
tween a confounder and the exposure or the disease. The amount of
nfounding, however, is a result of the strength of the associations be-
een the confounder and both exposure and disease. Confounding
ıes not depend on the statistical significance of these associations. Fur-
ermore, a significance test evaluates only one of the two component
ations that give rise to confounding. Perhaps the most common situa-
ın in which this mistaken approach to evaluating confounding is ap-
ied is the analysis of randomized trials, when "baseline" characteris-
s are compared for the randomized groups. Baseline comparisons are
eful, but often they are conducted with the sole aim of checking for
ıtistically significant differences in any of the baseline variables, as a
eans of detecting confounding. A better way to evaluate confounding,
a trial as in any study, would be to control for the potential con-
ınder and determine whether the unconfounded result differs from
e crude, potentially confounded result.

ratification by Two or More Variables

r convenience of presentation, the examples in this chapter have used
v strata with only one stratification variable. Nevertheless, stratified
alysis can be conducted with two or more stratification variables. Sup-

ously, 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. Thus, to control 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 end up contributing little 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. Thus, 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.

The Importance of Stratification

The formulas in this chapter may look imposing, but they can be applied readily with a hand calculator or a spreadsheet or even a pencil and paper. Consequently, the methods described here to control confounding are widely accessible without heavy reliance on technology. These are not the only methods available to control confounding. In Chapter 10, we discuss multivariable modeling to control confounding. Multivariable 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 multivariable methods is nearly irresistible. Nevertheless, stratified analysis is preferable and should always be the method of choice to control confounding. This is not to say that multivariable modeling should be ignored: it does have its uses. Nevertheless, stratification is the preferred approach, at least as the initial approach to data analysis. Following are the main advantages of stratification over multivariable 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 become immediately apparent. These distributions are obscure when conducting multivariable modeling.
2. Not only the investigator but also the consumer of the research

tables of stratified data, a reader will be able to 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 rimary potential confounding factors, that is, to conduct a stratified nalysis. It is rare that a multivariable analysis will change the interretation produced by a stratified analysis. The stratified analysis will eep both the researcher and the reader better informed about the nature of the data. Even when it is reasonable to conduct a multivariable nalysis, it should be undertaken only after the researcher has conucted a stratified analysis and, thus, has a good appreciation for the onfounding in the data, or lack of it, by the main study variables.

Questions

1. In Table 8–3, 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 8–3 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 (such as age, sex, or stage of disease) that might 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 impli-

7. The larger a randomized trial, the less the possibility for 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 mathematical contribution of that stratum to the estimate of the pooled odds ratio (formula 8–6)? What is the mathematical contribution of that stratum to the variance of the pooled odds ratio (bottom formula in Table 8–4)?

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: *Modern Epidemiology,* Second Edition. Lippincott-Raven, Philadelphia, 1998, pp 275–279.
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 S: *Modern Epidemiology,* Second Edition. Lippincott-Raven, Philadelphia, 1998, pp 260–265.