7.1 The nature of confounding variables

A risk factor for the disease under study that is differentially distributed among the groups receiving different interventions in which the disease incidence is being compared is called a confounding factor. Unless the trial is very small, confounding factors are not likely to bias the comparisons between intervention and control groups in randomized trials, as the process of randomization ensures that any such factors, whether known or unknown, will be equally distributed in the different groups (apart from random variation). In studies in which those in the different groups have not been allocated at random, the control of confounding factors is a critical component in the analysis. For example, consider a comparative study of TB incidence in persons who received BCG in a routine vaccination programme and those who were not vaccinated. BCG coverage is often higher in urban areas and, independently of any effect of BCG, those living in urban areas also tend to have a higher incidence of TB because of overcrowding and other environmental factors. In this instance, residential status (rural/urban) could be a confounding factor, and, if it is not taken into account in the analyses, any protective effect of BCG against TB might be underestimated. Consider the hypothetical situation depicted in Table 21.8, which shows the incidence of TB over a 10-year period in BCG-vaccinated and unvaccinated individuals in urban and rural areas.

BCG coverage is appreciably higher in the urban population (80%) than in the rural population (50%). Also, in unvaccinated persons, the incidence of TB is higher in the urban population (20 per thousand over 10 years) than in the rural population (10 per thousand). In consequence, although BCG vaccine efficacy is 50% in both urban and rural areas, the estimate obtained from a comparative study, in which the place of residence is ignored, is only 41%. This difference is due to the confounding effect of the place of residence on the estimate of efficacy (the place of residence being related to both the disease incidence and, independently, to the prevalence of vaccination).
7.2 Adjusting for confounding variables

A powerful way of removing the effect of a confounding variable is to restrict comparative analyses to individuals who share a common level of the confounding variable and then to combine the results across the different levels in such a way so as to avoid bias. Thus, in the example in Table 21.8, if the vaccine efficacy was first estimated separately for rural and urban dwellers, and then the two estimates were to be combined, the estimate of efficacy obtained (50%) would be free of the confounding bias of the place of residence. In general, to control for confounding, the study population is divided into a number of strata. Within each stratum, individuals share a common level of the confounding variable. Estimates of risk, rate or mean differences, or ratios are made within each stratum, and the resulting estimates are then pooled in some way across strata, in order to obtain an overall measure of the effect which is free of any confounding due to the variable on which the stratification was made. Such stratification may be carried out on several confounding variables simultaneously (for example, age and sex).

If it is known, when a study is planned, that it will be necessary to allow for confounding variables in the analysis, it is desirable to give consideration to this at the design stage, both in terms of the information which must be collected and because it will require an increase in the required sample sizes (to achieve the desired statistical power, see Chapter 5). Usually, the necessary increase in sample size to allow for confounding variables is not great (for example, less than 20%), and often the information needed for these sample size calculations is not available before the study starts anyway. Formal methods for calculating sample sizes, allowing for adjustment for confounding variables, are given in Breslow and Day (1987).

7.3 Adjusting risks

7.3.1 Overall test of significance

After stratifying on the basis of the confounding variable(s), the analysis is conducted one stratum at a time, and then the results are pooled. In the $i$th stratum, the data may be depicted, as shown in Table 21.9.

To test the hypothesis that the relative risk is 1 in all strata or equivalently that the risk difference is zero in each stratum, a generalization of the method given in Section 4.2 may be used. The statistical test is known as the Mantel–Haenszel test.
In the \( i \)th stratum:

\[
\text{Expected value of } a_{i} = E(a_{i}) = \frac{m_{1i} n_{1i}}{N_{i}}
\]

\[
\text{Variance of } a_{i} = V(a_{i}) = \frac{n_{1i} n_{2i} m_{1i} m_{2i}}{N_{i}^2 (N_{i} - 1)}
\]

An overall test of the null hypothesis that the relative risk is unity is given by calculating

\[
\chi^2 = (\Sigma a_i - \Sigma E(a_i) - 0.5)^2 / \Sigma V(a_i),
\]

where the summation is over all strata, which may be tested for statistical significance using tables of the chi-squared distribution with one df.

The calculations are illustrated in Table 21.10, with data on disease incidence rates in vaccinated and unvaccinated individuals in three areas—urban, semi-urban, and rural.

\[
\Sigma a_{i} = 530 ; \Sigma E(a_{i}) = 738 ; \Sigma V(a_{i}) = 284.21
\]

Thus:

\[
\chi^2 = \frac{(530 - 738 - 0.5)^2}{284.21} = 151.49
\]

Thus, there is very strong evidence against the null hypothesis, as \( p < 0.000001 \) (from tables of the chi-squared distribution).

### 7.3.2 Pooled estimate of risk difference

If it is considered that the risk *difference* (rather than the risk ratio) is likely to be constant across different strata, a pooled estimate of the common risk difference may be required. This is obtained by taking a weighted average of the
risk differences in each stratum, weighting each by the inverse of its variance (as this may be shown to give the 'best' estimate of the common risk difference).

In the $i$th stratum, the risk difference is:

\[
d_{\mathrm{i}}=p_{\mathrm{li}}-p_{2\mathrm{i}}=\left(\left(a_{\mathrm{i}} / n_{\mathrm{1i}}\right)\right)-\left(\left(c_{\mathrm{i}} / n_{\mathrm{2i}}\right)\right)
\]

and the variance of the risk difference is:

\[
V\left(d_{\mathrm{i}}\right)=\left\{p_{\mathrm{i}}\left(1-p_{\mathrm{i}}\right)\left(\frac{1}{n_{\mathrm{li}}}+\frac{1}{n_{\mathrm{2i}}}\right)\right\}
\]

(as given also in Section 4.2), where:

\[
p_{\mathrm{i}}=\frac{n_{\mathrm{1i}} p_{\mathrm{1i}}+n_{\mathrm{2i}} p_{\mathrm{2i}}}{n_{\mathrm{1i}}+n_{\mathrm{2i}}} = \frac{a_{\mathrm{i}}+c_{\mathrm{i}}}{n_{\mathrm{1i}}+n_{\mathrm{2i}}}
\]

Now, let $w_{\mathrm{i}} = 1/V\left(d_{\mathrm{i}}\right)$.

The pooled estimate of the common risk difference is given by $d = \Sigma w_{\mathrm{i}} d_{\mathrm{i}} / \Sigma w_{\mathrm{i}}$. For the data in the example given in Table 21.10, the computations for the common risk difference are shown in Table 21.11.

Pooled estimate of the common risk difference $d = \Sigma w_{\mathrm{i}} d_{\mathrm{i}} / \Sigma w_{\mathrm{i}} = -0.0083$.

### 7.3.3 Pooled estimate of risk ratio

A pooled estimate of the common risk ratio $R$ across strata may be obtained, using the following formulae.

In the $i$th stratum, the risk ratio is given by:

\[
R_{\mathrm{i}}=\frac{a_{\mathrm{i}}}{n_{\mathrm{1i}}} / \frac{c_{\mathrm{i}}}{n_{\mathrm{2i}}} = \frac{a_{\mathrm{i}}}{n_{\mathrm{1i}}} \frac{n_{\mathrm{2i}}}{c_{\mathrm{i}} n_{\mathrm{1i}}}
\]

A pooled estimate across all strata is given by:

\[
R = \frac{\Sigma a_{\mathrm{i}} / n_{\mathrm{1i}}}{\Sigma c_{\mathrm{i}} / n_{\mathrm{2i}}}
\]
Thus, for the example in Table 21.10:

\[
R = \left\{ \left[ (160)(4000)/20\,000 \right] + (170)(16\,000)/40\,000 \right\} / \left\{ \left[ (80)(16\,000)/20\,000 \right] + [(240)(24\,000)/40\,000] + [(400)(40\,000)/80\,000] \right\} = 0.49.
\]

### 7.3.4 Confidence intervals

The easiest way of obtaining CIs on the estimates of the common risk difference or the common risk ratio is to use the ‘test-based’ method (Miettinen, 1976).

The approximate 95% CI on the risk difference is given by:

\[
d \pm \frac{1.96}{\sqrt{x}}
\]

Thus, in the example in Tables 21.10 and 21.11, the confidence limits are:

\[
-0.0083(1+1.96/\sqrt{151.49})=-0.0096 \text{ to } -0.0070
\]

The 95% CI on the logarithm of the relative risk is given by:

\[
\log_e R \pm \frac{1.96}{\sqrt{\chi^2}}
\]

In the example, the confidence limits are:

\[
\log_e(0.49)(1 \pm 1.96/\sqrt{151.49})=-0.8269 \text{ to } -0.5998
\]

And thus the confidence limits on the relative risk are 0.44 to 0.55.
7.4 Adjusting rates

The computations for adjusting rates are very similar to those for adjusting risks and involve only some changes to the formulae given in Section 7.3.

Suppose the results observed in the \(i\)th stratum are as shown in Table 21.12.

7.4.1 Overall test of significance

In the \(i\)th stratum, \(e_{1i}\) is the number of individuals who developed disease in group 1.

\[
\text{Expected value of } e_{1i} = E(e_{1i}) = e_i y_{1i} / y_i
\]

\[
\text{Variance of } e_{1i} = V(e_{1i}) = e_i y_{1i} y_{2i} / y_i^2
\]

An overall test of significance (that the common rate ratio is unity or the common rate difference is zero) is given by:

\[
\chi^2 = \left(\left|\sum e_{1i} - \sum E(e_{1i})\right| - 0.5\right)^2 / \sum V(e_{1i})
\]

where the summation is over all strata.

The value calculated should be looked up in tables of the chi-squared distribution with one df.

7.4.2 Pooled estimate of rate difference

In the \(i\)th stratum, the rate difference is:

\[
d_i = r_{1i} - r_{2i} = (e_{1i} / y_{1i}) - (e_{2i} / y_{2i})
\]

Its estimated variance is:

\[
V(d_i) = (r_{1i} / y_{1i} + r_{2i} / y_{2i})
\]

Let \(w_i = 1 / V(d_i)\) then the estimate of the common rate difference across all strata is
given by:

\[ d = \sum w_i d_i / \sum w_i \]

### 7.4.3 Pooled estimate of rate ratio

In the \( i \)th stratum, the rate ratio is:

\[ R_i = r_{1i} / r_{2i} = (e_{1i} / y_{1i}) / (e_{2i} / y_{2i}) = e_{1i} y_{2i} / (e_{2i} y_{1i}) \]

A pooled estimate of the common rate ratio is given by:

\[ R = \Sigma (e_{1i} y_{2i} / y_i) / \Sigma (e_{2i} y_{1i} / y_i) \]

### 7.4.4 Confidence intervals

The 95% CI on the common rate difference is given by:

\[ d \left(1 \pm 1.96 / \sqrt{\chi^2}\right) \]

The 95% CI on the logarithm of the common rate ratio is given by:

\[ \log_e R \left(1 \pm 1.96 / \sqrt{\chi}\right) \]

**Example:** In Table 21.13, the numerical computations are illustrated, as before, with data on the disease incidence in vaccinated and unvaccinated individuals in three areas: urban, semi-urban, and rural.

**Table 21.13** Disease incidence rates in urban, semi-urban, and rural areas, according to vaccination status

<table>
<thead>
<tr>
<th>Area</th>
<th>Vaccinated</th>
<th>Unvaccinated</th>
<th>Both groups</th>
<th>( E(n_i) )</th>
<th>( V(n_i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases ( n_i )</td>
<td>Person-years ( y_i )</td>
<td>Cases ( n_i )</td>
<td>Person-years ( y_i )</td>
<td>Cases ( n_i )</td>
</tr>
<tr>
<td>Urban (1)</td>
<td>80 8000</td>
<td>40 2000</td>
<td>120 10000</td>
<td>96 19.2</td>
<td></td>
</tr>
<tr>
<td>Semi-urban (2)</td>
<td>85 12000</td>
<td>120 8000</td>
<td>205 20000</td>
<td>123 49.2</td>
<td></td>
</tr>
<tr>
<td>Rural (3)</td>
<td>80 20000</td>
<td>200 20000</td>
<td>280 40000</td>
<td>140 70.0</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>245 40000</td>
<td>360 30000</td>
<td>605 70000</td>
<td>359 138.4</td>
<td></td>
</tr>
</tbody>
</table>

Example: In Table 21.13, the numerical computations are illustrated, as before, with data on the disease incidence in vaccinated and unvaccinated individuals in three areas: urban, semi-urban, and rural.

**Overall test of significance:**
\[
\chi^2 = \left( \left| \Sigma e_{1i} - \Sigma E(e_{1i}) \right| - 0.5 \right)^2 / \Sigma V(e_{li})
\]

\[
= \left( |245-359| - 0.5 \right)^2 / 138.40 = 93.08 (p < 0.000001)
\]

The estimation of the common rate difference is shown in Table 21.14.

\[
d = \frac{\sum w_i d_i}{\sum w_i} = -0.0066
\]

The 95% CI on the common rate difference is:

\[
d \left( 1 \pm \frac{1.96}{\sqrt{\chi^2}} \right) = -0.0066 \left( 1 \pm \frac{1.96}{\sqrt{93.08}} \right)
\]

\[
= -0.0079 \text{ to } -0.0053
\]

The estimate of the common rate ratio is:

\[
R = \frac{\Sigma (e_{1i} y_{2i})}{\Sigma (e_{2i} y_{1i})} \quad \text{or} \quad R = \frac{\left( \frac{80}{8000} \right) + \left( \frac{85}{12000} \right) + \left( \frac{80}{20000} \right)}{\left( \frac{40}{8000} \right) + \left( \frac{120}{12000} \right) + \left( \frac{200}{20000} \right)} = \frac{90}{204} = 0.44.
\]

<table>
<thead>
<tr>
<th>Area (l)</th>
<th>Vaccinated</th>
<th>Unvaccinated</th>
<th>Both groups</th>
<th>E(e1i)</th>
<th>V(e1i)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases (e1i)</td>
<td>Person-years (y1i)</td>
<td>Cases (e2i)</td>
<td>Person-years (y2i)</td>
<td>Cases (e1)</td>
</tr>
<tr>
<td>Urban (1)</td>
<td>80</td>
<td>8000</td>
<td>40</td>
<td>2000</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>96</td>
<td>19.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semi-urban (2)</td>
<td>85</td>
<td>12000</td>
<td>120</td>
<td>8000</td>
<td>205</td>
</tr>
<tr>
<td></td>
<td>123</td>
<td>49.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rural (3)</td>
<td>80</td>
<td>20000</td>
<td>200</td>
<td>20000</td>
<td>280</td>
</tr>
<tr>
<td></td>
<td>140</td>
<td>70.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>245</td>
<td>40000</td>
<td>360</td>
<td>30000</td>
<td>605</td>
</tr>
<tr>
<td></td>
<td>359</td>
<td>138.4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

https://med.libretexts.org/Bookshelves/Nursing/Book%3A_Field_Trials_of_Health_Interventions_-_A_Toolbox_(Smith_Morrow…

Updated: Sat, 03 Sep 2022 16:01:45 GMT
Powered by

8
The 95% confidence limits on the logarithm of the common rate ratio is:

\[
\log_e R(1+1.96/\sqrt{n}) = \log_e (0.44)(1+1.96/\sqrt{93.08}) = -0.988 \text{ to } -0.654
\]

Taking the anti-logarithm, the 95% confidence limits on the common rate ratio are 0.37 to 0.52.

### 7.5 Adjusting means

If the outcome variable is a quantitative measure, other than a risk or rate, adjustment for the effects of a confounding variable involves performing a stratified t-test.

A numerical example is given in Table 21.15 where the comparison is between subjects using mosquito nets (intervention group) and those not using them (control group), and the outcome measure \(x\) is the number of episodes of malaria over a period of 1 year. In this example, age is considered as the confounding variable, and the stratification has been made by dividing the study subjects into three age groups. The size of each subgroup has been made small to simplify the computations for illustrative purposes.

The data may be represented algebraically for those in the \(i\)th stratum, as shown in Table 21.16.

An estimate of the common difference in response between the intervention and control groups is obtained by calculating a weighted average of the differences within each stratum:

\[
d = \sum w_i d_i / \sum w_i
\]

where \(w_i = \left[ n_{1i} n_{2i} / (n_{1i} + n_{2i}) \right] \)

\[
\text{Difference}
\]

\[
d_{-i} = \overline{x}_{-1i} - \overline{x}_{-2i}
\]

Thus, in the example:

\[
\begin{align}
d &\approx (-0.75)(32/12) + (-0.45)(132/23) + (-0.20)(50/15) \quad \land \quad ((32/12) + (132/23) + (50/15)) \quad \land \quad -5.25/11.75 \quad \land \quad -0.45. \quad \text{\end{align}}
\]
An overall test of significance is obtained by calculating a test statistic as:

\[
\Sigma w_{i} d_{i} / \left[ s \sqrt{\Sigma w_{i}} \right]
\]

where \( s = \sqrt{\frac{\sum (n_{1i} - 1)s^2_{1i} + \sum (n_{2i} - 1)s^2_{2i})}{\sum (n_{1i} + n_{2i} - 2)} } \), and the value of the test statistic can be compared with tables of the t-distribution with \( \sum(n_{1i} + n_{2i} - 2) \) df.

In the example:

\[
s = \sqrt{\frac{23.0265}{44}} = 0.7234
\]

The test statistic (44 df) is:

\[
-5.25 \times 0.7234 \times \sqrt{11.74} = -2.12
\]

The absolute value is larger than 2.02, which is the tabulated 5% value for t with 44 df. Thus, there is statistically significant evidence regarding the efficacy of intervention; the reduction in the average number of episodes of malaria is estimated as 0.45 per child per year. The 95% confidence limits on the difference are given by:

\[
\left( \frac{\Sigma w_{i} d_{i}}{\Sigma w_{i}} \right) \pm \frac{t s}{\Sigma w_{i}}
\]

where \( t \) is taken from tables of the t-distribution for 95% confidence limits with \( \sum(n_{1i} + n_{2i} - 2) \) df.

Thus, the 95% confidence limits are:

### Table 21.15: Attacks of malaria in children of different ages in those using (intervention) and not using (control) mosquito-nets

<table>
<thead>
<tr>
<th>(Stratum) Age group</th>
<th>Attacks of malaria/ child (a)</th>
<th>No. of children (n)</th>
<th>Mean (x)</th>
<th>Standard deviation (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) &lt;2y I*</td>
<td>1,0,2,3,1,2,1,0</td>
<td>8</td>
<td>1.25</td>
<td>1.0351</td>
</tr>
<tr>
<td>C*</td>
<td>2,3,1,2</td>
<td>4</td>
<td>2.00</td>
<td>0.8165</td>
</tr>
<tr>
<td>(2) 2–5y I</td>
<td>0,1,2,1,1,0,2,2,1,0</td>
<td>12</td>
<td>1.00</td>
<td>0.7385</td>
</tr>
<tr>
<td>C</td>
<td>2,2,1,1,1,1,1</td>
<td>11</td>
<td>1.45</td>
<td>0.5222</td>
</tr>
<tr>
<td>(3) 4–5y I</td>
<td>1,0,1,1</td>
<td>5</td>
<td>0.80</td>
<td>0.4472</td>
</tr>
<tr>
<td>C</td>
<td>1,1,2,0,1,1,0,2,1,1</td>
<td>10</td>
<td>1.00</td>
<td>0.6667</td>
</tr>
</tbody>
</table>

* I, Intervention group; C, control group.

### Table 21.16: Algebraic representation of data in Table 21.15 for those in the ith stratum

<table>
<thead>
<tr>
<th>Intervention group</th>
<th>No. in group</th>
<th>Mean (x)</th>
<th>Standard deviation (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (I)</td>
<td>n_{1i}</td>
<td>x_{1i}</td>
<td>s_{1i}</td>
</tr>
<tr>
<td>2(C)</td>
<td>n_{2i}</td>
<td>x_{2i}</td>
<td>s_{2i}</td>
</tr>
</tbody>
</table>
\[-0.45 \pm 2.02 / (0.723) / \sqrt{(11.74)} = 0.88 \text{ episodes/year/child.} \]

If it is thought that the intervention is likely to affect the response measured in a relative, rather than an absolute, fashion (i.e. a constant percentage reduction in the number of malaria attacks, rather than a constant absolute reduction in the number of malaria attacks), then it would be appropriate to transform the data initially by taking logarithms of the number of attacks (or, say, loge(number of attacks + 0.1) to avoid zero numbers) and to perform the calculation on the transformed values.