The answer is that between 1990 and 2005 the distribution of the age structure of the Scottish population has changed, and in particular got older. For example, in 1990 12.1% and 1.9% of the population were over 65 and 80 respectively, which had risen to 14.1% and 2.8% by 2005. Thus as the population were older in 2005 compared to 1990, then they are likely to have more mortalities as the risk for older age groups is larger. This results in the crude rates being biased and not comparable. Therefore, as a general (and very important) rule: **Crude rates should not be used to compare populations**.

4.4 Standardisation

Video4.3 - Direct standardisation

Adjusting for population demographics is known as standardisation, which has two general flavours, direct and indirect. The discussion that follows is based on mortality, but the same approaches are used for incidence and prevalence.

4.4.1 Direct standardisation

Direct standardisation is used to compare the rate of disease in 2 separate populations with different demographics. It requires a standard or reference population, and the rates of disease for two separate study populations are transferred to this reference population to make them directly comparable.

Definition The most common standard populations are the **European standard** population and the **World standard** population (Table 22), both of which contain 100,000 people. These are hypothetical populations with an age structure that approximates the average for Europe and the World respectively.

Table 22: The European and World standard populations are shown in 5 year age bands.

Age group	**European standard population**	**World standard population**
Under 5	8000	12000
5-9	7000	10000
10-14	7000	9000
15-19	7000	9000
20-24	7000	8000
25-29	7000	8000
30-34	7000	6000
35-39	7000	6000
40-44	7000	6000
45-49	7000	6000
50-54	7000	5000
55-59	6000	4000
60-64	5000	4000
65-69	4000	3000
70-74	3000	2000
75-79	2000	1000
80-84	1000	500
85+	1000	500
All ages	**100,000**	**100,000**

Estimating mortality rates Let $i = 1, \ldots, G$ denote the G groups for the reference population, which here are defined by $G = 18$ age groups. In this setting $i = 1$ corresponds to under fives, while $i = 18$ corresponds to 85 and over. Further, let:

- y_i be the number of deaths from the population under study in the i^{th} age group.
- n_i be the number of people in the population under study in the i^{th} age group.
- N_i be the number of people in the reference population in the i^{th} age group.

Then the **crude mortality rate** is simply

$$\text{Crude mortality rate } = \frac{\sum_{i=1}^{G} y_i}{\sum_{i=1}^{G} n_i}.$$

The **age specific mortality rate** for the i^{th} group is

$$\text{Age specific mortality rate for group } i \ = \ \frac{y_i}{n_i}.$$

The **expected number of mortalities** from the i^{th} group in the reference population is then:

$$\text{expected number of mortalities for group } i \ = \ \frac{y_i}{n_i} \times N_i.$$

Then, finally the **age standardised mortality rate (ASMR)** is given by:

$$\text{ASMR } = \frac{\sum_{i=1}^{G} \frac{y_i}{n_i} \times N_i}{\sum_{i=1}^{G} N_i}.$$

An approximate variance estimate and hence a 95% confidence interval can be obtained for the age standardised mortality rate based on the assumption that

$$y_i \sim \text{Poisson}(n_i\theta_i),$$

where the probability of death θ_i for the i^{th} age group can be estimated via maximum likelihood methods and is $\hat{\theta}_i = \frac{y_i}{n_i}$. A Poisson distribution is used rather than a binomial distribution as mortality events are rare. Then based on this we have that:

$$\begin{aligned}
\mathbb{V}(ASMR) &= \mathbb{V}\left[\frac{\sum_{i=1}^{G}\hat{\theta}_i N_i}{\sum_{i=1}^{G} N_i}\right] \\
&= \frac{\sum_{i=1}^{G} N_i^2 \mathbb{V}(\hat{\theta}_i)}{\left(\sum_{i=1}^{G} N_i\right)^2} \\
&= \frac{\sum_{i=1}^{G} N_i^2 \hat{\theta}_i/n_i}{\left(\sum_{i=1}^{G} N_i\right)^2},
\end{aligned}$$

where the last line holds because $\mathbb{V}(y_i) = n_i\theta_i$. This formula can then be used to compute a 95% confidence interval for ASMR based on a Gaussian assumption:

$$\text{ASMR} \pm 1.96 \times \sqrt{\mathbb{V}(ASMR)}.$$

Example Recall the lung cancer mortality example. Age specific mortality rates for 2005 are given in Table 23.

Table 23: The European standard population, mortality rates and the expected number of deaths due to lung cancer among Scottish males per 100,000 for 2005 are shown. The data are presented in 5 year age bands.

Age group	**Age specific mortality rate (Scottish males; 2005)**	**European standard population**	**Expected deaths (males; 2005)**
Under 5	0	8000	0.0
5-9	0	7000	0.0
10-14	0	7000	0.0
15-19	0	7000	0.0
20-24	0	7000	0.0
25-29	0	7000	0.0
30-34	0	7000	0.0
35-39	1.6	7000	0.1
40-44	8.2	7000	0.6
45-49	21.1	7000	1.5
50-54	50.3	7000	3.5
55-59	93.7	6000	5.6
60-64	196.1	5000	9.8
65-69	299.6	4000	12.0
70-74	432.3	3000	13.0
75-79	612.0	2000	12.2
80-84	714.1	1000	7.1
85+	655.3	1000	6.6
All ages		**100,000**	**72.0**

The fourth column has been created by multiplying the age specific mortality rates for Scotland by the number of people in the European standard population. The total number of expected deaths is then obtained by summing over the age groups. Thus the final ASMR is 0.00072 or 72 per 100,000.

The ASMR for 1990 is 104.4 per 100,000, giving a reduction in ASMR of 31%, which is much higher than the reduction of 18.2% in crude mortality rate shown earlier.

The lung cancer mortality data were obtained from Public Health Scotland (formerly the Information Services Division (ISD)) of NHS Scotland, and they give confidence intervals as follows:

- **1990**: 104.4 per 100,000 (95%CI: 100.4, 108.5);
- **2005**: 72.0 per 100,000 (95%CI: 98.9, 75.0).

Here, the confidence intervals give some idea of the fluctuation in mortality rate from year to year due to random chance. They are not being used to generalise from sample to population.

Note The absolute value of the standardised rates will change if a different reference population is used. For example, using the world standard population the ASMR in 1990 and 2005 drop to 68.7 and 46.6 per 100,000 respectively. This is because the world population is younger on average than the European one. However, this is still a 32.2% reduction, which is similar to that observed using the European reference population.

4.4.2 Indirect standardisation

Video4.4 - Indirect standardisation

In **Indirect standardisation** we only have 1 study population, and the aim is to compute the number of deaths expected in the study population if the age specific mortality rates from the reference population applied.

Definition Let r_i denote the rate of mortality from the reference population in the i^{th} age group, that is from the reference population

$$r_i = \frac{\text{Number of mortalities from the reference population from the } i^{th} \text{ age group.}}{\text{Population size in the reference population for the } i^{th} \text{ age group}}$$

Then the **expected number of mortalities from the study population** with age sizes $(n_1, \ldots, n_G)$ is given by:

$$E = \sum_{i=1}^{G} n_i r_i.$$

Thus E measures the number of mortalities expected if standardised mortality rates applied, where as $Y = \sum_{i=1}^{G} y_i$ denotes the observed number of mortalities from the study population.

Definition Thus, a measure of mortality risk is called the **Standardised mortality ratio (SMR)**, which is given by

$$\text{SMR} = \frac{Y}{E}.$$

Here:

- An SMR$= 1$ means there were as many deaths as expected, and thus is the null risk.
- An SMR> 1 means there are more deaths than expected. If SMR=1.2, then there were 20% more deaths than expected.
- An SMR< 1 means there are fewer deaths than expected. If SMR=0.9, then there were 10% fewer deaths than expected.

Based on the Poisson model $Y \sim$ Poisson(ER),where R measures disease risk and $\hat{R} = \frac{Y}{E} = SMR$, the variance of the SMR is given by:

$$\begin{aligned}
\mathbb{V}[SMR] &= \mathbb{V}\left[\frac{Y}{E}\right] \\
&= \frac{\mathbb{V}[Y]}{E^2} \\
&= \frac{ER}{E^2} \\
&= \frac{R}{E},
\end{aligned}$$

where R is replaced by $\hat{R} = \frac{Y}{E}$. Thus, based on approximate normality a 95% CI for the SMR is

$$\text{SMR} \pm 1.96 \times \sqrt{\frac{Y}{E^2}}.$$

Example Recall the lung cancer example. The reference population is the Scottish population, and here we focus on the Greater Glasgow and Clyde Health Board (GGCHB) as the study population. The data required to compute the expected number of mortalities is shown in Table 24 below.

Table 24: Population estimates for the GGCHB, mortality rates per 100,000 among Scottish males and the expected number of deaths due to lung cancer per 100,000 in the GGCHB are given for 2005. The data are presented in 5 year age bands.

Age group	**GGCHB population (2005)**	**Age specific mortality rate (Scottish males; 2005)**	**Expected deaths in GGCHB (males; 2005)**
Under 5	23,434	0	0.0
5-9	24,116	0	0.0
10-14	26,197	0	0.0
15-19	29,816	0	0.0
20-24	34,643	0	0.0
25-29	31,517	0	0.0
30-34	29,779	0	0.0
35-39	31,430	1.6	0.5
40-44	32,262	8.2	2.6
45-49	29,916	21.1	6.3
50-54	25,707	50.3	12.9
55-59	24,109	93.7	22.6
60-64	18,893	196.1	37.0
65-69	17,020	299.6	51.0
70-74	14,459	432.3	62.5
75-79	10,672	612.0	65.3
80-84	6,580	714.1	47.0
85+	3,726	655.3	24.4
All ages	**41,4276**		**332.2**

For each row of the table the expected number of deaths in each age group is computed via the calculation:

$$\text{Expected number of deaths in group } i \;=\; \frac{G\text{GHB population} \times \text{rate of death}}{100,000}.$$

So for the 75-79 age group it is

$$\text{Expected number of deaths in 75-79 group} \;=\; \frac{10,672 \times 612.0}{100,000} \;=\; 65.3.$$

These age specific expected counts are then summed over ages to get 332.2 expected mortalities. The observed number of deaths in GGCHB in 2005 was 449, giving an SMR of

$$\text{SMR} = \frac{449}{332.2} \;=\; 1.351,$$

which corresponds to a 35.1% increase compared with the Scottish average in 2005. A 95% confidence interval is computed as follows:

$$\text{SMR} \quad \pm \quad 1.96 \times \sqrt{\frac{Y}{E^2}}$$

$$1.351 \quad \pm \quad 1.96 \times \sqrt{\frac{449}{332.2^2}}$$

$$(1.226 \quad , \quad 1.476).$$

As a comparison to see how high the risk is in GGCHB, the SMR was computed for all 14 health boards in Scotland. The results are summarised in Table 25; note that the SMR has been multiplied by 100, as it is sometimes presented on that scale.

Table 25: The SMR and accompanying 95% CI are shown for every Health board in Scotland.

Health board	**SMR**	**95% CI**
Argyll and Clyde	102.5	(88.8, 118.3)
Ayr and Arran	98.1	(84.3, 114.0)
Borders	67.4	(47.0, 93.7)
Dumfries and Galloway	97.1	(78.0, 120.9)
Fife	85.1	(71.8, 100.8)
Forth Valley	98.0	(81.8, 117.3)
Grampian	75.1	(64.6, 87.2)
Greater Glasgow	135.1	(123.2, 148.2)
Highland	82.5	(66.7, 102.0)
Lanarkshire	105.5	(92.8, 120.0)
Lothian	97.8	(87.4, 109.4)
Orkney	111.6	(55.7, 199.7)
Shetland	52.6	(17.0, 122.5)
Tayside	102.9	(89.5, 118.2)
Western Isles	79.0	(39.4, 141.4)

Also note that here the 95% CI for GGCHB is slightly different as the normal approximation described above was not used. GGCHB is by far the worst. Why?

4.5 Deprivation

Video4.5 - Deprivation

Socio-economic deprivation measures poverty, and populations who are poorer typically exhibit worse health on average than more affluent populations. The reason for this is because, on average, people from poorer populations are more likely to smoke, drink, eat unhealthy food and lack regular exercise. Note, we are not saying that all poor people do this, it is just that the prevalence of these factors will be higher among poorer populations.

Example GGCHB has been split into 271 intermediate geographies (IG) and prenatal smoking data are available for the year 2011. Figure 21 shows the proportion of pregnant women who smoke plotted against the percentage of people defined to be income deprived for each IG in 2011.

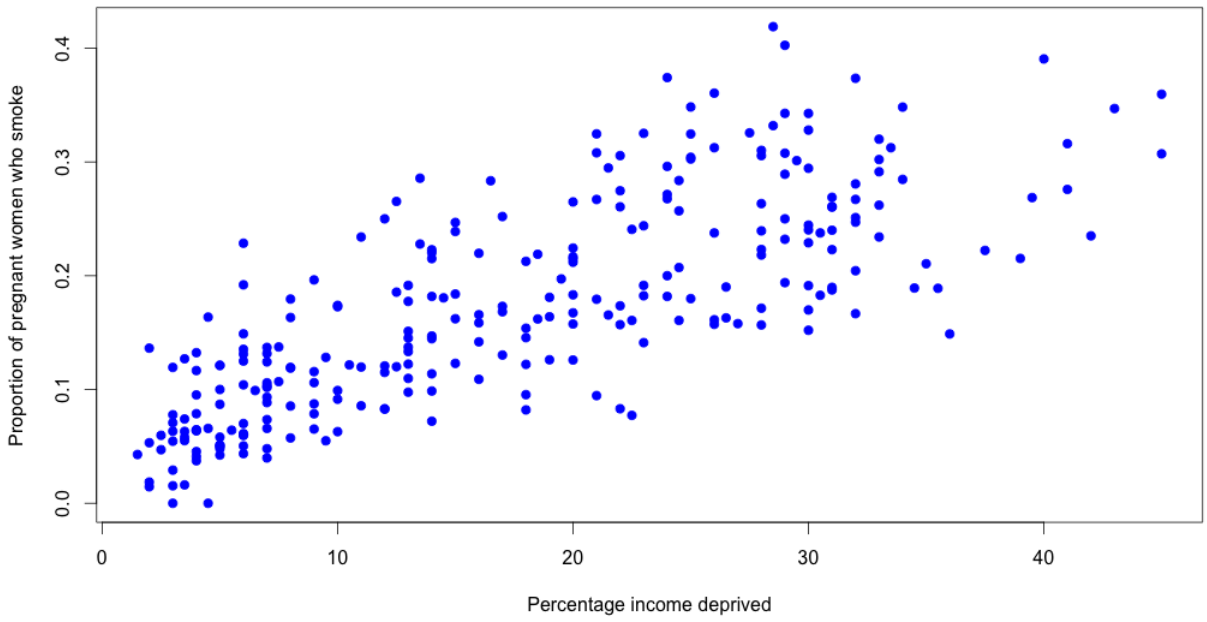

Figure 21: The proportion of pregnant women who smoke is plotted against the percentage of people defined to be income deprived for every IG in the GGCHB in 2011.

The relationship between them is strong, with a correlation coefficient of 0.77.

This begs the question how do you measure socio-economic deprivation? The use of income deprivation is just one measure of area level deprivation, what about educational attainment, house prices, etc.? Deprivation is a multidimensional quantity that is difficult to measure.

Two common measures are the Carstairs measure and the Scottish Index of Multiple Deprivation (SIMD). You can find more information on both those indices in technical reports posted on the moodle page. Those reports are not examinable.

Definition One common index used to reflect material deprivation is the **Carstairs index**. For any postcode sector (e.g. G12 8, KA25 7), this is a score based on 4 variables routinely collected at the census:

- Overcrowding: persons in private households living at a density of more than one person per room as a proportion of all persons in private households.
- Male unemployment: proportion of economically active males who are seeking work.
- Social class 4 or 5: proportion of all persons in private households with head of household in social class 4 or 5, the two lowest social classes.
- No car: proportion of all persons in private households with no car.

Postcode sectors are commonly split into quintiles or deciles for analysis. However, this index is limited in several ways:

- It only considers male and not female unemployment, what if in a family the wife works and the husband stays at home to look after the children.

- Car ownership means something different in the countryside than in the middle of a city where public transport is good and parking is bad.

Definition The **Scottish Index of Multiple Deprivation (SIMD)** was developed in response to the known limitations of the Carstairs score. It is an area-based measure applied to 6505 small areas in Scotland, known as data zones. The data zone geography covers the whole of Scotland and nests within local authority boundaries. Data zones are groups of Census output areas which have populations of between 500 and 1,000 household residents, and some effort has been made to respect physical boundaries such as rivers when defining these zones. In addition, they have compact shape and, as far as possible, contain households with similar social characteristics. The index is based on data from 7 domains:

- income
- employment
- crime
- education
- health
- housing
- access to services

Data from these 7 domains are weighted to produce a deprivation score, and areas are ranked from most affluent (high) to most deprived (low). An interesting map showing the SIMD can be found online

Example Figure 22 shows the SMR for coronary heart disease by age group and split by SIMD decile, where decile 1 is the most deprived while decile 10 is the least deprived category.

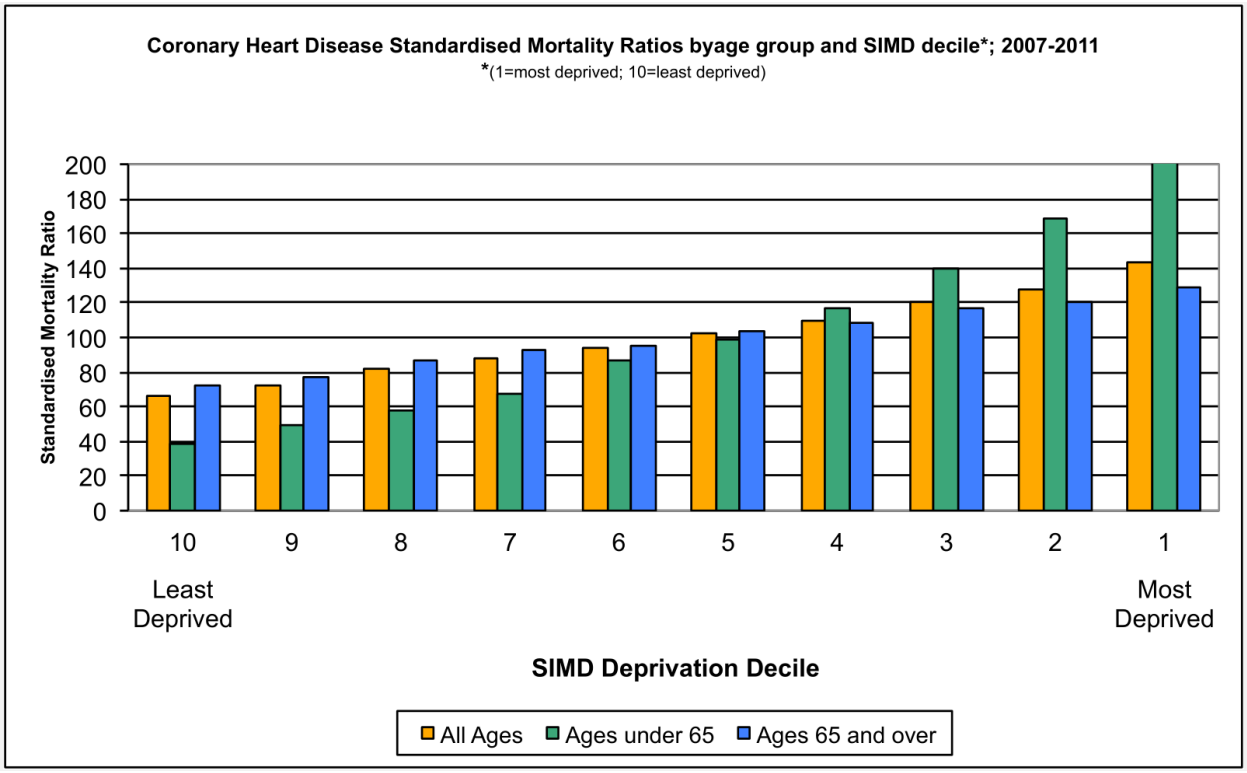

Figure 22: A barchart displaying the SMR for coronary heart disease by age group and SIMD decile.

Example - The Glasgow Effect The link between socio-economic circumstances and health is well known, and there is an increasing evidence base supporting the hypothesis of a Scottish Effect, and more specifically a Glasgow Effect, the terminology used to identify higher levels of mortality and poor health found in Scotland and Glasgow beyond that explained by socio-economic circumstances.

For example, a paper by the MRC Social and Public Health Sciences Unit here in Glasgow (*International differences in self-reported health measures in 33 major metropolitan areas in Europe. European Journal of Public Health 2012;22:40-7*), produced Figure 23. The researchers have used a method called *logistic regression* and plotted the residuals of this regression model for 33 major metropolitan areas in Europe. The response of the model is self reported bad/very bad general health for men and the model adjusts for age, education and social class. The zero line represents the average across all 33 metropolitan areas and it is clear that there is significant variation between those areas. Glasgow appears to suffer from the worst heath, even after adjusting for socio-economic deprivation. You can find more information about the Glasgow effect in a publication from the Scottish government: The Scottish Health Survey: The Glasgow Effect (the document is also posted on moodle as a pdf) and from a short video clip.

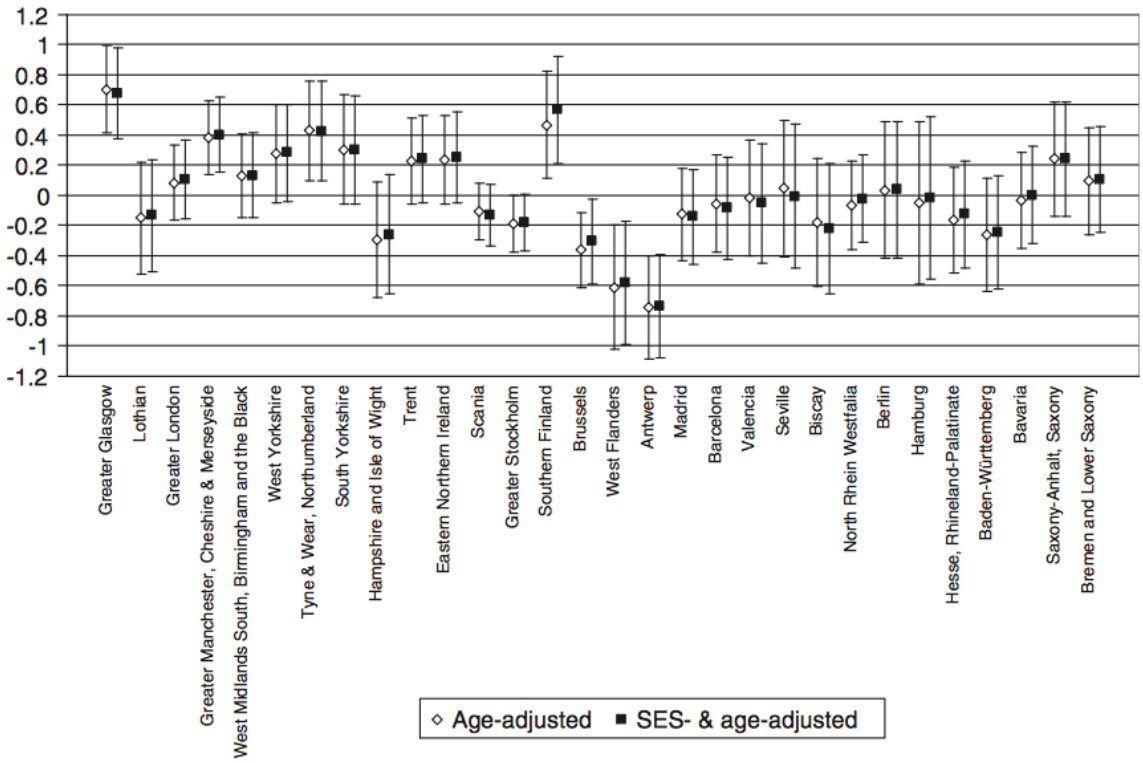

Figure 23: Logistic regression residuals and 95% confidence intervals for self rated bad/very bad general health for men are shown for several European metropolises.

4.6 Measuring the association between a risk factor and disease

Video4.6 - Measuring the association between a risk factor and disease

Suppose a large population of N individuals have been categorised as positive or negative for a potential risk factor and as positive or negative for some disease, as outlined in Table 26 below.

Table 26: 2x2 table summarising disease and risk factor.

	Disease		
	Yes	**No**	**Total**
Risk factor present	A	B	$\mathbf{A+B}$
Risk factor absent	C	D	$\mathbf{C+D}$
Total	$\mathbf{A+C}$	$\mathbf{B+D}$	$\mathbf{N}$

There are several ways of quantifying the impact of the risk factor on the disease. How to interpret those different risks is a common source of confusion and it is an important role of the statistician (that is you!) to confidently explain those risks to non-statisticians and to use the appropriate measure in the right situation. Risks can get confusing enough that Sir David Spiegelhalter devised an online tool explaing risk measure you

might encounter in a Journal paper to absolute risk. The following key measures of the impact of the risk factor on disease are commonly used:

Definition The **relative risk (RR)** of disease is given by

$$RR = \frac{P(\text{disease in group with risk factor})}{P(\text{disease in group without risk factor})} = \frac{\frac{A}{A+B}}{\frac{C}{C+D}}.$$

Thus a value of 1 is the null relative risk and represents no difference between disease risk for people with and without the risk factor. In contrast, a relative risk of 2 means people with the risk factor are twice as likely to get the disease than those without. In terms of inference, the distribution of RR is skewed, and thus $\log(RR)$ is approximately normal with variance.

$$\mathbb{V}[\log(RR)] \;=\; \frac{1}{A} - \frac{1}{A+B} + \frac{1}{C} - \frac{1}{C+D},$$

which yields a 95% confidence interval on the log scale of

$$\left(\log(RR) - 1.96\sqrt{\mathbb{V}[\log(RR)]},\;\; \log(RR) + 1.96\sqrt{\mathbb{V}[\log(RR)]}\right)$$

based on the assumption of normality. Then this interval can be exponentiated to get it back to the relative risk scale, yielding an interval of the form

$$\left(\exp(\log(RR) - 1.96\sqrt{\mathbb{V}[\log(RR)]}),\;\; \exp(\log(RR) + 1.96\sqrt{\mathbb{V}[\log(RR)]})\right).$$

Defintion The **Odds ratio** for disease is given by

$$OR = \frac{\text{odds of disease in group with risk factor}}{\text{odds of disease in group with no risk factor}} = \frac{\frac{A}{B}}{\frac{C}{D}} = \frac{AD}{BC}.$$

As with the relative risk, confidence intervals can be computed based on a normal approximation on the log scale. This time the variance is

$$\mathbb{V}[\log(OR)] \;=\; \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D},$$

which yields a 95% confidence interval on the log scale of

$$\left(\log(OR) - 1.96\sqrt{\mathbb{V}[\log(OR)]},\;\; \log(OR) + 1.96\sqrt{\mathbb{V}[\log(OR)]}\right).$$

Then exponentiating gives the interval on the odds ratio scale as

$$\left(\exp(\log(OR) - 1.96\sqrt{\mathbb{V}[\log(OR)]}),\;\; \exp(\log(OR) + 1.96\sqrt{\mathbb{V}[\log(OR)]})\right).$$

Definition The **attributable risk (AR)** is the difference between the absolute risks of the population with and without the risk factor. It given by

$$AR = P(\text{disease with risk factor}) - P(\text{disease without risk factor}) = \frac{A}{A+B} - \frac{C}{C+D}.$$

Here $P_1 = \dfrac{A}{A+B}$ is the probability of disease with the risk factor and $P_2 = \dfrac{C}{C+D}$ is the probability of disease without the risk factor. While the relative risk RR indicates the relative change in risk associated with some risk factor, AR indicates the absolute change in risk. It is usually important from an individuals point of view to consider both RR and AR.

Definition The **Population Attributable Risk (PAR)** attempts to quantify risk from a 'Public Health' point of view. It is defined as *the proportion of all cases of the disease in the population which are attributable to exposure to the risk factor.*

If the exposure were removed from the population, a total of NP_2 individuals would be expected to develop disease over a period of time. With the exposure present, the actual number developing disease is $(A+C)$. Thus, the difference, $(A+C) - NP_2$, is the number of cases in the population that can be attributed to the exposure. Expressed as a proportion of all cases in the population, this gives

$$PAR = \frac{(A+C) - NP_2}{(A+C)}.$$

Example Suppose that for a less rare disease we have

$$P_1 = 0.02, \quad P_2 = 0.01, \quad \text{then} \quad RR = 2, \quad \text{and} \quad AR = 0.01.$$

The relative risk is 2 meaning you are twice as likely to have the disease given you have the risk factor compared to not. The attributable risk is 1%, so 1% more people with the risk factor will contract the disease. Now suppose you have a much rarer disease with

$$P_1 = 0.000002, \quad P_2 = 0.000001, \quad \text{then} \quad RR = 2, \quad \text{and} \quad AR = 0.000001.$$

You are still twice as likely to contract the disease if you have the risk factor, but only 0.0001% more people with the risk factor will contract the disease compared to those without it.

Example One of the first large-scale case-control studies of smoking and lung cancer was carried out by Doll and Hill in England and Wales from 1948-50 and published in the British Medical Journal in 1950 (preliminary report) and 1951 (final report). They interviewed 1357 males with lung cancer in selected hospitals in England (all under 75 years of age, not too ill to be interviewed) and 1357 controls of the same age from the non-cancer wards of the same hospitals. They obtained, among many other things, a detailed smoking history from each subject. Table 27 looks at the most recent amount smoked before the current period of illness.

Table 27: The number of cigarettes smoked before lung cancer diagnosis is shown for cases and controls. Controls have been matched by age and the number of cigarettes smoked are reported in 6 categories ranging from 0 to 50+.

Cigarettes smoked	0	1-4	5-14	15-24	25-49	50+	**Total**
Cases	7	49	516	445	299	41	**1357**
Controls	61	91	615	408	162	20	**1357**

Then the odds ratio for the 15-24 group compared with the non-smoking group is:

$$OR \;=\; \frac{445/408}{7/61} \;=\; 9.50,$$

so a very large increase in the odds of lung cancer if you smoke 15-25 cigarettes a day. A 95% confidence interval for this odds ratio is created by first calculating the variance of the log odds ratio as:

$$\mathbb{V}[\log(OR)] \;=\; \frac{1}{445}+\frac{1}{408}+\frac{1}{7}+\frac{1}{61} \;=\; 0.164.$$

Then the 95% CI is given by:

$$\Big(\exp(\log(9.50)-1.96\sqrt{0.164}),\exp(\log(9.50)+1.96\sqrt{0.164})\Big) \;=\; (4.29,21.01).$$

The full set of results for all the smoking levels (relative to non-smokers) are shown in Table 28.

Table 28: Odds ratios and 95% CIs for being diagnosed with lung cancer for different severities of smoking are shown. Smoking severity ranges from smoking 0 to 50+ cigarettes a day.

Cigarettes smoked per day	Odds ratio	95% CI
0	1.0	
1-4	4.7	(2.0, 11.1)
5-14	7.3	(3.3, 18.1)
15-24	9.5	(4.3, 21.0)
25-49	16.1	(7.2, 36.0)
50+	17.9	(6.9, 46.2)

None of the C.I.s contain 1.0 and there is clear evidence of a 'dose-response' relationship. The evidence of association between smoking and lung cancer in this study is strong. In 1950 there were 4 other case control studies published in the USA, all showing similar results.

4.7 Confounding

Video4.7 - Confounding I

Confounding is a concept apparently so complex, not even Sheldon Cooper from the big bang theory is up to speed with it. We illustrate the idea of confounding with the following example.

Example Suppose you conduct a case-control study into the effect of alcohol consumption on lung cancer, and recruit a set of cases with lung cancer and controls without lung cancer and ask them about their alcohol consumption. Then the estimated odds ratio, say splitting alcohol into drinker / non-drinker would give a very significant result, suggesting that alcohol consumption is a risk factor for lung cancer. However, it is known that no such causal effect exists, so why would the data tell you otherwise? The answer is due to confounding. On average, people who drink are more likely to smoke than people who do not drink, and people who smoke are more likely to get lung cancer. Thus, if you ignore the important confounding variable smoking, then you observe a spurious relationship between alcohol consumption and lung cancer. Thus the golden rule in epidemiology is: **Association does not imply causation.**