Statistics Presentation

Mandy Porter
Public Health Sciences
Topics covered

- What is statistics?
- Variability
- Confidence intervals
- Problem of small numbers
- Standardisation
- Causality
- Questions to ask
What is statistics?

• Science of collecting, summarising, presenting and interpreting data, and of using them to estimate the magnitude of associations and test hypotheses.

• Not only does it provide a way of organising information on a wider and more formal basis, it takes into account intrinsic variation inherent in most biological processes.

Statistics can be difficult and complicated as it is a specialist topic, and it covers quite a range of different statistical methods. However, with all data analysis methods “garbage in, garbage out” applies. If the data or statistical analyses methods are incorrect, then the resulting conclusions could be incorrect and could be misleading.
Statistics is all about trying to find out about unknown factors within population(s). Generally, we cannot obtain all the information from our population so we need to take a sample and generalise our findings to the population in question.

It is clearly important to assess whether your sample is representative. For example, if you want to know how many portions of fruit and vegetables are being eaten in a specific population. Standing outside a fast food outlet and standing outside a fruit and vegetable shop (if you can find one) and asking everybody coming out of the doors about their 5-A-DAY consumption might give quite different results. It is an extreme example but illustrates a point. Survey bias could occur in either case here. People who have just come out of the fast food outlet might be less likely than people in your overall population to eat 5-A-DAY, but people who have just come out of the fruit and vegetable shop might be more likely. Obtaining a representative sample without bias can be very difficult, and often you just don't know how biased your sample is in relation to the underlying measure (5-A-DAY here) in your population, as its value in the population is unknown.
Variability (n=500)

- Height of five year old girls (500 selected for each histogram).

- Averages are 1.0858m and 1.0892m.

- So averages very similar (but different)

Now we are going to pretend that we have data for a population. A population of over 14,000 heights of five year old girls in Hull and ERoY. I’m going to randomly select different samples and compare the mean or average height from these samples in order to look at variability. Variability is the key to statistics. Real data from Child Health System – measured heights and weights of children for school intake year.

Similar shapes (or so called distributions) with height given on horizontal x-axis and number of children given on vertical y-axis.

Average height or mean heights are similar, but different. Would not expect them to be the same. Can you imagine going into the street and measuring heights of (a sample of) individuals? If this was done more than once, it would be unlikely that the mean height would be the same.
Another sample but this time with fewer individuals only 62 children. The averages again are similar, but they are different, with bigger difference between the two samples.

Variability (n=62)

- Height of five year old girls (62 selected for each histogram).
- Averages are 1.0845m and 1.0945m.
- So averages similar (but more different than before).
In fact, difference between two sample means is approximately 0.4cm when 500 children measured but 1cm when 62 children measured. More variability in the mean estimates with fewer individuals.
Confidence intervals

- To help assess variability and the usefulness of an estimate we can use confidence intervals.

- Generally 95% confidence intervals.
Confidence intervals - definition

• There is 95% chance that the true underlying population statistic will fall within the calculated confidence interval.
Confidence intervals - definition

• There is 95% chance that the true underlying population statistic will fall within the calculated confidence interval.

• Therefore 5% that it won’t!
If the confidence interval is sufficiently wide, then presenting a single figure might be very misleading. It could be so wide that it is not providing any useful information above that which might have been estimated by someone with knowledge about the disease/geographical area/condition, etc. However, you would not know that unless the CIs are presented.

For example, if the standardised mortality rate (SMR) for a rare disease within Hull is 200, then it means that the mortality rate is twice as high as England even after adjusting for differences in the population structure between Hull and England. You might be concerned... However, the SMR might only be based on a very small number of deaths. If you find out that the confidence interval for the SMR ranges from 50 to 400, then this is much more useful than just looking at the SMR figure alone. This suggests that the true SMR for Hull could range between half the mortality rate of England to four times the mortality rate of England. You might have guessed this anyway, even with no background information on the disease or Hull! In which case, the calculation is provided no additional information above common sense. However, that would not have been the case if the single value (of 200) was quoted alone. It would have caused concern, and resources might have been wasted to find out why it was so high (rather than considering the reliability of the statistics). Another similar example of this follows comparing mortality rates between two geographical areas.
Variability – means (95% CI)

- Sample 1 (n=500): 1.0858 (1.0817 to 1.0899)
- Sample 2 (n=500): 1.0892 (1.0852 to 1.0932)

- Sample 1 (n=62): 1.0845 (1.0706 to 1.0984)
- Sample 2 (n=62): 1.0945 (1.0822 to 1.1067)

- Actual (n=14,260): 1.0932

Going back to the heights of the 5 year old girls. We can calculate the 95% confidence intervals for the four samples that we randomly selected. We can also calculate the mean for the ‘population’ (as we actually have data for almost all of 5 year old children in Hull and ERoY – a rarity to have such information on the population).

We are going to look at the width of the confidence interval.
This is the width of the confidence intervals.

To calculate the confidence interval, you can take the means calculated by each of the sample and subtract the values in the table to obtain the lower 95% confidence limit and add the values in the table to the mean to obtain the upper 95% confidence limit.
Confidence interval width

<table>
<thead>
<tr>
<th>WIDTH OF CONFIDENCE INTERVAL</th>
<th>N=500</th>
<th>N=62</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>0.0082</td>
<td>0.0278</td>
</tr>
<tr>
<td>Sample 2</td>
<td>0.0080</td>
<td>0.0245</td>
</tr>
</tbody>
</table>

- Narrow intervals for larger N (third width).

We can see that the width of the CIs are narrower when N is larger.
Going back to the definition, if the true value has 95% chance of falling within 95% CI then narrow intervals much better as gives better indication of true underlying value.

<table>
<thead>
<tr>
<th>WIDTH OF CONFIDENCE INTERVAL</th>
<th>N=500</th>
<th>N=62</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>0.0082</td>
<td>0.0278</td>
</tr>
<tr>
<td>Sample 2</td>
<td>0.0080</td>
<td>0.0245</td>
</tr>
</tbody>
</table>

- Narrow intervals for larger N (third width).
- If true value has 95% chance of falling within 95% CI then narrow intervals much better as gives better indication of true underlying value.

Going back to the definition, if the true value has 95% chance of falling within the 95% CI then a narrow interval is better as we have a more precise estimate of the true underlying value.
However, in this case we are saying that we know the heights of all five year old girls (we have data on almost all of the entire population) so we can compare the sample means and confidence intervals with the true figure. This is the problem with real data. It does not do exactly what you’d like to illustrate the point, and illustrate that there is variability regardless of the size of the sample. We can see that three of the four intervals contain the true underlying value. However, one of the samples with the larger N does not.

This was the actual analyses completed (based on generating computer-generated random numbers to decide which children to include in each of the samples above), and it was not selected to emphasise the point. If I had, I’d have selected a sample with the smaller N which did not contain the actual value. In practice, when undertaking a survey when asking people to complete a questionnaire (with people having the choice to refuse or not having had the opportunity to be selected as they are in a specific group, e.g. homeless), then it is easier to end up with a biased group. Those participating in a health survey, are probably more likely to be healthy or interested in their health so could well have different characteristics in relation to those who refused to participate.

In reality, the larger the number of individuals surveyed, the larger the narrower the confidence interval, and the more confidence you have in the values. Wider confidence intervals give less certainty, and conclusions are less certain.
Just illustrating a few points about small numbers now (in a similar way to the example given before in the notes section). You can see in the table that area A has mortality rate of 83 deaths per 100,000 people compared to 167 for area B. First question is “Are these different?”

<table>
<thead>
<tr>
<th></th>
<th>AREA A</th>
<th>AREA B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death rate (per 100,000)</td>
<td>83</td>
<td>167</td>
</tr>
</tbody>
</table>

• Is there a difference in mortality rate between areas A and B?
Problem of small numbers

<table>
<thead>
<tr>
<th></th>
<th>AREA A</th>
<th>AREA B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death rate (per 100,000)</td>
<td>83</td>
<td>167</td>
</tr>
</tbody>
</table>

- Is there a difference in mortality rate between areas A and B?
- Yes, considerably different.

Clearly, there is a considerable difference.
### Problem of small numbers

<table>
<thead>
<tr>
<th></th>
<th>AREA A</th>
<th>AREA B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death rate (per 100,000)</td>
<td>83</td>
<td>167</td>
</tr>
</tbody>
</table>

- Is there a difference in mortality rate between areas A and B?
- Yes, considerably different.
- Is this difference of concern?

However, the key question is whether this is of concern or not. Are they really different or can such a big difference be due to random variation?
This table gives the actual number of deaths, and it can be seen that there are only one death in Area A and two deaths in Area B. Therefore, there is only a very small number of deaths on which we are basing our estimate rate upon. If there are only one deaths a year (on the basis of the underlying population estimate) then one would expect some years there would be no deaths, some years there would be one death, some years there would be two deaths, etc. That is, there would not be exactly one death every year. This has partly answered the question. Is the difference of concern? No, as there is only a very small number of deaths.
Confidence intervals illustrate this point further. It is extremely useful to examine the confidence intervals as it can be very apparent from them whether there is a problem with small numbers or not. In this case, the 95% CIs are extremely wide. So wide that they are meaningless and not useful. That is, with a little local knowledge (or even no knowledge about the disease or areas I might have guessed the mortality rate fell within this very large range!), I probably could guessed that the mortality would be between 2 and 464 for area A. So comparing the mortality rate between areas A and B will not reveal any important differences. There may be genuine differences present in the mortality rate, but with such a small number of deaths, it would not be possible to know this. Therefore, one would conclude that there is no evidence to suggest that the mortality rates are different between area A and area B.

In order to assess the differences here, you would have to examine the number of deaths over a much longer period of time. Even then it might not be possible to do this, as the number of deaths even over a period of time might be too low to provide useful statistical analyses. Furthermore, you run into problems if examining data over a long period of time as factors influencing the deaths from this particular disease might have changed over time between areas A and B.
After looking at variability and confidence intervals, we’re now moving on to
standardisation. When using standardisation you are generally wanting to compare
different groups, or produce figures that others can use to compare their own figures.

• For example, does the mortality rate differ between area A and area B?

Comparison of two groups

• In many cases, it is useful to compare a particular statistic between two groups, or
you want to produce a figure others can use to compare their own figures.

After looking at variability and confidence intervals, we’re now moving on to
standardisation. When using standardisation you are generally wanting to compare
different groups, or produce figures that others can use to compare their own figures.
The question is frequently of the form, does the mortality rate for area A differ from
that of area B?
Comparison of two groups

• Need to look at reasons for comparison:
  
  – Assessment of the provision of services?

  – Compare area A and area B to assess whether a key factor is influencing any potential differences?

However, whether or not to standardise the data is based on what question you are trying to address.
Provision of services?

• Necessary to report prevalence of ill-health, disease, mortality and risk factors as they stand.

• It is necessary to treat and have the provision to treat the population as it is, regardless of the age and gender structure.

If assessing the provision of resources and services, then you need to treat and have the provision to treat the population as it is, regardless of the age and gender structure. It is necessary to report prevalence and mortality for the population itself and not report standardised rates.
Mortality rate affected by deprivation?

- Want to take into consideration other factors which will influence mortality.
- Age will be influence mortality rate strongly. Gender too (but less so).
- If you find a difference, how do you know whether it is due to deprivation, or a difference in age structures?

Whereas, if you are wanting to ask a question along the lines of “Does deprivation influence the mortality rate?” then you will want to take into consideration other factors which may influence the mortality rate. If you find a difference you will not know whether the difference is really due to deprivation or due to a difference in the age composition of your two groups.
Mortality rate affected by deprivation?

- Want to take into consideration other factors which will influence mortality.
- Age will be influence mortality rate strongly. Gender too (but less so).
- If you find a difference, how do you know whether it is due to deprivation, or a difference in age structures?
- One solution is to standardise.

One way to do this is to standardise.
What is standardisation?

- Creating an ‘adjusted’ measure (e.g. mortality rate) which is adjusted to take into consideration of differences in the age and gender structure.
- The age/gender structure of a ‘standard’ population is used.
- Direct and indirect standardisation.

You can calculate a measure which is not standardised, e.g. crude mortality rate. However, the standardised mortality rate is essentially the crude mortality rate adjusted to take into account the differences in the age and gender structure. You can standardised to factors other than age and gender, but this is unusual. You need to decide on a ‘standard’ population and have data on this population. What information is required depends on whether direct or indirect standardisation is used.
Not going into details, but for direct standardisation you calculate the mortality rates in each age-gender group of your sample and apply the rates observed to a standard population. This results in a standardised mortality rate or a mortality rate adjusted for the age-gender structure.

More details are available on website together with examples of direct and indirect standardisation.
For indirect standardisation, you do the reverse and apply mortality rates from the ‘standard’ population and apply them to the study group to calculate the ‘expected number of deaths’ and this figure is then compared to the actual or observed number of deaths. This results in a standardised mortality or morbidity ratio. If the SMR is 100 then the mortality rates are the same in the study and standard populations. If the SMR is greater than 100 then the mortality rate in the study population is higher, i.e. there are more actual deaths than would have been expected based on the age-gender composition of your study group.
Direct or indirect?

- Often not really important.
- Can combine multiple SMRs (from indirect standardisation) into single value. For example, can use data from SMRs for all wards to calculate single SMR for Hull.
- Often method depends on available data and relative accuracy.
- Indirect better with small numbers.

With regard to available data, for instance, if you don’t have the mortality rates for your study group then you cannot use direct standardisation. Similarly, if you don’t have the age-gender specific mortality rates for a standard population then you cannot use indirect standardisation.
Does not really matter greatly what standard population is. However, if you are examining trends over time, then you would need to use the same standard population over time.

Nationally produced standardised mortality rates use the European Standard Population for directly standardised mortality rates, and England for indirectly standardised mortality ratios. For example, those on the NHS Health and Social Care Information Centre Indicator Portal (available on http://www.indicators.ic.nhs.uk/webview/ or http://www.indicators.ic.nhs.uk/webview/)
However, need to bear some things in mind when select population. For example, is good data available and will this information be available in the future if you are making comparisons in the future? What have others used? National data are generally use the European Standard Population (ESP) or England’s population.

However, the calculation of the standardised rate will differ depending on the standard population used. As the ESP has a much lower proportion of the elderly in the England’s population, you might end up with a directly standardised mortality rate of 50 deaths per 100,000 when using the European Standard population but 70 deaths per 100,000 when using England as the standard population.
Now a little on relationships and causality.

Just because you observe a relationship or association between two factors does not necessarily mean there is causality.

This is a common problem. Many people assume an association denotes causality, it does not!
Just “unlucky” with the research. When a relationship does not exist, if performing a statistical test, one in twenty times you will find a relationship when testing at the 5% significance level.
The people in your study might not be representative of your target population. They may be more interested in their health so more interested in participating in a particular study. Those overweight might be more interested in participating in a survey that examines the effect of a diet programme.

This is a difficult problem to assess. How biased is your sample? Did you ask a representative sample of people? Or are they different from your overall population? How different are those who participating in the survey compared to those that refused? How can you know when you have no or very little information on those who refused to participate!
A true relationship may not exist, but you may have found one indirectly through another factor. A third variable controls both or mediates between the two events. For example, a relationship between the prevalence of mental health and the mortality from lung cancer. There may not be a true relationship between the two, but prevalence of mental health is associated with a higher smoking prevalence which in turn is associated with lung cancer. Smoking is a confounder. The association that you found between the prevalence of mental health and the mortality from lung cancer is artificial as it is mediated by another factor.
Is relationship causal?

- Is the observed relationship causal?
  - TEMPORALITY (exposure preceded disease)

Basically just because you have a relationship you cannot assume that it is causal. Even with the above criteria and consideration of the quality of the data on the previous slide, you’ve accumulated a lot of evidence but you might be proved wrong later.

Whilst many of these seem obvious, most people don’t consider them when they observe an association and it’s easy to automatically assume causality.

I’ve termed the following in terms of an exposure such as smoking causing a disease such as lung cancer, but it does not have to be an exposure or disease.

The exposure must precede the disease or at least happen simultaneously. This is crucial. For instance, ex-smokers often have a higher mortality rate than current smokers following surgery. Does this mean that quitting smoking is bad? No, it just means that their exposure occurred before they quit. They smoked a high number of cigarettes prior to the surgery, and were told that they could not have the surgery without quitting.

Whereas the current smokes did not smoke as many cigarettes. Therefore, they were more healthy than the ex-smokers so the mortality rate for the current smokers was lower than the ex-smokers.
Strength of the relationships is important. I was doing statistics and really only considered getting into the medical field when I did an MSc project on data from Pakistan. The odds of obtaining oral submucous fibrosis, a very nasty disease which limits opening of the mouth, is 88 times greater for those who chew tobacco and areca nut products compared to those who did not. Not conclusive evidence of causality but makes you think!

Is relationship causal?

- Is the observed relationship causal?
  - TEMPORALITY (exposure preceded disease)
  - STRENGTH (effect large so can rule out other factors)
It helps if there is a proper scientific reason why the exposure might cause the disease.
It helps if there is evidence that intervention has some effect. Stopping the exposure should stop the disease completely or partially.
Is relationship causal?

- Is the observed relationship causal?
  - TEMPORALITY (exposure preceded disease)
  - STRENGTH (effect large so can rule out other factors)
  - PLAUSIBILITY (credible scientific mechanism)
  - EXPERIMENTAL EVIDENCE (physical intervention shows results consistent with association)
  - BIOLOGICAL GRADIENT (more exposure results in more disease)

The causality argument is helped if more exposure results in more disease, for example, heavy smokers are more likely to get lung cancer.

It helps the argument if you have consistency in that others have found a similar relationship.
Is relationship causal?

• Is the observed relationship causal?
  – TEMPORALITY (exposure preceded disease)
  – STRENGTH (effect large so can rule out other factors)
  – PLAUSIBILITY (credible scientific mechanism)
  – EXPERIMENTAL EVIDENCE (physical intervention shows results consistent with association)
  – BIOLOGICAL GRADIENT (more exposure results in more disease)
  – CONSISTENCY (other research has found same results)

Also if others have found similar research projects have resulted in similar conclusions.
Is relationship causal?

• Is the observed relationship causal?
  – TEMPORALITY (exposure preceded disease)
  – STRENGTH (effect large so can rule out other factors)
  – PLAUSIBILITY (credible scientific mechanism)
  – EXPERIMENTAL EVIDENCE (physical intervention shows results consistent with association)
  – BIOLOGICAL GRADIENT (more exposure results in more disease)
  – CONSISTENCY (other research has found same results)
  – SPECIFICITY (exposure associated with specific disease as opposed to wide range of diseases)

Also if the exposure is associated with a specific disease as opposed to a wide range of disease. Confounding or biases may be issues if it is related to a wide range of diseases.
Is relationship causal?

• Is the observed relationship causal?
  – TEMPORALITY (exposure preceded disease)
  – STRENGTH (effect large so can rule out other factors)
  – PLAUSIBILITY (credible scientific mechanism)
  – EXPERIMENTAL EVIDENCE (physical intervention shows results consistent with association)
  – BIOLOGICAL GRADIENT (more exposure results in more disease)
  – CONSISTENCY (other research has found same results)
  – SPECIFICITY (exposure associated with specific disease as opposed to wide range of diseases)
  – COHERENCE (consistent with natural history of disease)

Also helps the causality argument if there is coherence in that results are consistent with the natural history of the disease, for example, lung cancer rates are higher in countries where more people smoke.
Is relationship causal?

- Is the observed relationship causal?
  - TEMPORALITY (exposure preceded disease)
  - STRENGTH (effect large so can rule out other factors)
  - PLAUSIBILITY (credible scientific mechanism)
  - EXPERIMENTAL EVIDENCE (physical intervention shows results consistent with association)
  - BIOLOGICAL GRADIENT (more exposure results in more disease)
  - CONSISTENCY (other research has found same results)
  - SPECIFICITY (exposure associated with specific disease as opposed to wide range of diseases)
  - COHERENCE (consistent with natural history of disease)
  - ANALOGY (similar results exist that we can draw a relationship to)

Also helps if there is other evidence, scientific mechanisms that can be examined as further evidence.

So just because you find a relationship does not necessarily mean that it is causal!
Questions to ask

• What is the aim of this piece of work and audience?

Just to conclude. When interpreting statistics or requesting information or statistics from others there are a number of key questions to address. The main one is what is the precise aim of the piece of work. Answering this question very clearly and precisely is essential and no real data can be obtained or produced without this question being addressed. As seen earlier when comparing to areas, do you want to know whether a key statistic differs between two areas perhaps due to a specific difference in the areas or whether you are interesting in determining ‘need’ to allocate resources. It may surprise you to know that the majority of people don’t have a clear idea of their aim?
Questions to ask

• What is the aim of this piece of work and audience?
• What data has been used, and is it reliable, accurate, complete, …?

What is the data source? The questions you ask about the quality of the data will depend on what the data is, where the dataset has come from, etc., and will differ depending on the project, but some of the questions below might give an idea of some areas for questions.

For survey data, how likely are the responders going to be similar to the non-responders? More interest in their health and therefore less likely to smoke, eat poorly, not exercise, etc? More likely to be diabetics in survey which is assessing prevalence of diabetes as they are interested in topic? What was the proportion of non-responders? How many participated in the survey relative to the number asked to participate? Is this figure quoted? If the percentage is low or very low, it is more likely that there is a biases within the sample obtained.

Is data from GP system which may only be completed for those who regularly attend their GP? Is information more likely to be completed by large practices where there are specific staff to enter data? Did all practices, pharmacies or dentists participate or provide data? Are those who did not provide information, more likely to not have that information readily available? Is it compulsory to provide the data nationally? Do those using a specific IT system have the data is more readily available? Is there a financial incentive or penalty for providing the information?
Questions to ask

• What is the aim of this piece of work and audience?
• What data has been used, and is it reliable, accurate, complete, …?

What is the coding accuracy likely to be? Has their been a change to the coding system or IT system? Has the data source and methods remained the same or very similar over time? Can trends over time be examined? Has hospital coding of disease (reason for admission) likely to have improved over time? Are people more aware of this disease (asthma, diabetes, obesity, eating disorders, etc) now than they were previously? Is this disease more likely to be included as a secondary cause now than it was before? Are those working in A&E more likely to code alcohol as a secondary cause now than they were ten years ago? Is ethnicity coding accurately or consistency on the records? How complete is the coding? Do 10% or 99% of the admissions have ethnicity coded?

Is body mass index self-reported from height and weight or measured? Did people underestimate their weight when asked? Did people even know their weight? What was the drop out rate for a survey that followed people up over a period of time? Did too many drop out of the survey? Did only the keenest stayed in the survey who are likely to be different to those who failed to participate? Was data missing for many people as it was a sensitive question and people did not want answer that question? Did people really want to give accurate information on the number of sexual partners for a health survey looking at cervical cancers, or give accurate information on drug and alcohol consumption on a general health and lifestyle survey?
Questions to ask

- What is the aim of this piece of work and audience?
- What data has been used, and is it reliable, accurate, complete, …?

Was it a telephone interview, face-to-face interview or self-completed questionnaire? How might this have influenced the findings? Were the interviewers trained for consistency in asking the questions? Were leading questions asked? Could interview style have influenced responses? What about people whose first language is not English? Were they given the opportunity to participate in the survey? What about those not able to complete the questionnaire? Were carers asked? How might literacy skills have influenced participation?

If those asked to participate were randomly selected from telephone records, what about those without telephones? This is a frequent survey method, but does it include telephone calls to mobiles? Who in the household was selected when the telephone was answered? If people were asked to complete a survey in the street or a shopping centre, who would be around at that time? Females with children or retired people might be more likely to be around during the day compared to working men. What day of the week or times of the day were people surveyed? Could this influence the response rate?
Say mortality rate is 1,000 deaths per 10,000 people or equivalently 10%. Is this sensible? Clearly, depends on what is being examined, but it is an important question to ask. If it is an ordinary student population then this would be incorrect. However, if it is patients in a home for the terminally ill then the mortality rate could be correct.

Questions to ask

- What is the aim of this piece of work and audience?
- What data has been used, and is it reliable, accurate, complete, …?
- Do the results look sensible?
Questions to ask

- What is the aim of this piece of work and audience?
- What data has been used, and is it reliable, accurate, complete, …?
- Do the results look sensible?
- What factors might influence this result and have these been taken into account?

Are the age groups very much different? Are there more males than females in the population? Has standardisation been undertaken? If mortality rate differs between area A with chemical factory in centre and area B in much better area, why is this? Is it chemical factory or is it just that area A is much more deprived? Are other risk factors likely to differ, e.g. prevalence of smoking? When examining associations between two factors, it is very important to take into consideration key differences between the groups. An adjusted analysis is often undertaken that adjusts for a number of factors simultaneously. However, even in these sophisticated analyses, it might not be possible to adjust for all the key factors (as data might not be collected on them or there may be unknown factors).

For example, a large project I worked on at the University of Edinburgh involved looking at readmission and mortality rates following coronary angioplasty, comparing rates between two hospitals in Edinburgh and Glasgow. One might have expected both of these to be higher in Glasgow because risk factors such as smoking tend to be higher there. However, both rates were significantly higher in Edinburgh. However, after adjusting the analysis for a range of risk factors including the severity of the disease, no significant difference in readmission or mortality rates was found. The most severe cases and seriously ill patients were referred to the hospital in Edinburgh from all the hospitals across Scotland, so in practice, Edinburgh’s patients had a much higher rate of readmission and mortality simply because they were the most seriously ill patients. Once the severity of the patients’ condition was taken into consideration, no difference was found in the rates.
This is a problem with hospital standardised mortality/morbidity ratios (HSMRs). These have been calculated for all hospitals across the country, with the purpose of comparing hospital quality of care. The analyses adjusted to take into consideration the severity of the patients. However, this would depend on the quality of the coding. If the coding was relatively poor with regard to one factor relating to severity, for example, if the number of comorbidities for each patient was not recorded properly, then the analysis could not be adjusted for that particular factor, and the resulting rates may be misleading.

If data is from an existing database and not collected for the specific purpose of addressing the aim of the research project, then it is quite possible that one key influential factor might be missing from the analysis. If this is the case, then any conclusions could be misleading. For example, you would need to include smoking status when looking at factors influencing lung cancer.

Questions to ask

• What is the aim of this piece of work and audience?
• What data has been used, and is it reliable, accurate, complete, …?
• Do the results look sensible?
• What factors might influence this result and have these been taken into account?
We've already provided some information earlier on confidence intervals and the problems with small numbers, so hopefully you will have appreciated the usefulness and importance of presenting and interpreting confidence intervals. If they are not presented, you would need to ask if you felt that small numbers might be a problem with this specific piece of work.

**A few questions to ask**

- What is the aim of this piece of work and audience? Who is in sample?
- What data has been used, and is it reliable, accurate, complete, …?
- Do the results look sensible?
- What factors might influence this result and have these been taken into account?
- Problem with small numbers? CIs given?
Thank-you