A brief review of epidemiology

Gavin Shaddick

September 24, 2008

1 Vital Statistics and Measurement

1.1 Vital Statistics

The basic measures for assessing the health of a community and its needs for health services are the size and composition of the human population and the counts of vital events occurring within them (births, deaths, morbid disorders, etc). Such measures are known as *vital statistics*.

Indices, that is summary statistics, derived from such raw data in terms of rates, ratios and proportions are of use to many workers in the health field. Understanding these indices and their uses should help towards efficient provision and use of resources and to appropriate preventative measures targeted at susceptible sections of the community. By themselves, the raw data are of little use until they are transformed to standard indices for valid comparisons.

1.2 Population and Demographics

1.2.1 Population Size

There is only one accurate way to estimate the size of the relevant population, and describe its demographic characteristics, and that is to count it at a particular point in time. This is known as a *census* and is usually done once every ten years. The more recent census in the UK took place in 2001.

1.2.2 Inter-censal events

The second major problem of measurement in public health is to record the demographic events that occur within the community: births, deaths, marriages and migration both in and out of the area. If these are measured accurately and continuously then the changes in population from year to year can be determined, and the frequent need for censuses eliminated.

In the United Kingdom, every birth, marriage and death in the country is registered. From these registers it is possible to calculate the rates at which births and deaths are occurring in different areas, social groups, ages and at different times of the year. Immigration and emigration are measured only for the whole country by sample estimates at ports and airports.

Population size estimates between census years are only an approximation. Some estimates, such as the numbers in different occupations, are so unreliable that mortality data for them is only tabulated in census years.

1.3 Incidence, Prevalence and Rates

1.3.1 Rates

A *rate* is defined as the number of events, for example deaths or cases of disease, per unit of population, in a particular time span. To calculate a rate we require:

- a defined period of time
- a defined population, with an accurate estimate of the size of the population during the defined period
- the number of events occurring over the period.

$$\text{rate} = \frac{\text{No. events}}{\text{total person-time at risk}}$$

For a fixed time period Δt, an average size of the population at risk during that period $\bar{N}$ and the number of events A the rate is

$$\text{rate} = \frac{A}{\bar{N} \times \Delta t}$$

1.3.2 Incidence

The *incidence rate* refers to the number of new cases of a particular disease that develop during a specified time interval.

For a fixed time period Δt, a population at risk of size N and the number of new cases of disease A the incidence rate is

$$\text{incidence} = \frac{A}{N \times \Delta t}$$

An incidence rate lies between 0 and ∞, and a yearly incidence rate measures the number of cases per person-year.

Measuring the population at risk may be difficult because of changes in the population over the time period. Since it may not be possible to measure disease-free periods precisely, we often calculate the incidence using the average size of the population. This is reasonably accurate if the population size is stable and the incidence rate is low.

1.3.3 Prevalence

The *prevalence* refers to the number of cases of disease that exist at a specified point in time. It is the proportion of the population who have a disease at a given time. Prevalences may also be calculated over a time period; for example the number of events within a time period.

$$\text{prevalence} = \frac{\text{Number of cases}}{\text{Number at risk}}$$

Diseases with high incidence rates may have low prevalences if they are rapidly fatal.

1.3.4 Crude Mortality Rate

The *crude mortality rate* is usually calculated as deaths per 1000 population per year. Let D be the number of deaths in a given time period of length Δt, and $\bar{N}$ be the average size of the population at risk during that period (often approximated by the number in the population at the mid-point of the time period). Then the crude mortality rate is given by

$$\mathrm{r} = \frac{d}{\bar{N} \times \Delta t} \times 1000$$

1.3.5 Specific Rates

Rates may be required for particular sections of a community and these are referred to as *specific rates*; that is, where the populations are *specified*; that is the denominators. For example, age-specific or age- and sex-specific rates may be used for comparison of different populations. Other common specific rates are area, occupation or social class specific (and combinations of these).

1.3.6 Other commonly-used rates

infant mortality rate	number of deaths under one year of age after live birth, divided by the number of live births
neonatal mortality rate	number of deaths at 0 to 27 days after live birth, divided by the number of live births
stillbirth rate	number of stillbirths, divided by the total number of births, live and still
perinatal mortality rate	number of stillbirths and deaths at days 0 to 6, divided by the total number of births
birth rate	number of live births per year, divided by total population
fertility rate	number of live births per year, divided by number of women 15-44 (ie of childbearing age)

Table 1: Mortality Data by Age Group for South West of England 2003. *Source: ONS Mortality Statistics Series DH1 no. 36*

Age Group	Population (1000s)	% in age group	Deaths	Age-specific mortality rate
0-4	259.7	5.2	244	
5-14	609.3	12.2	59	
15-24	591.0	11.8	251	
25-44	1319.3	26.4	1219	
45-64	1282.2	25.6	5944	
65-74	470.7	9.4	8668	
75+	467.1	9.3	39480	
All	4999.3	100	55865	

1.4 Standardisation

Table 2: Mortality Data by Age Group for England 2003. *Source: ONS Mortality Statistics Series DH1 no. 36*

Age Group	Population (1000s)	% in age group	Deaths	Age-specific mortality rate
0-4	2848.2	5.7	3682	1.3
5-14	6299.8	12.6	725	0.1
15-24	6304.0	12.6	2634	0.4
25-44	14485.7	29.1	14001	1.0
45-64	11971.3	24.0	62755	5.2
65-74	4158.6	8.3	87714	21.1
75+	3788.3	7.6	331898	87.6
All	49855.9	100	503409	10.1

When comparing populations, we can eliminate the effects of, for example, different age structures by looking at age-specific rates. However, this can be cumbersome, and it is often easier to compare a single summary figure.

1.4.1 Direct Standardisation

For direct standardisation, we use a standard population structure for reference. We then calculate the overall mortality rate that this reference population would have observed if it had the age-specific mortality rates of the population of interest.

Suppose the reference population has population counts $N'_k; k = 1, \ldots, K$ in each age-group k. We calculate the age-specific mortality rates r_k for the population of interest. The directly standardised

rate is given by

$$\text{directly standardised rate} = \frac{\sum_{k=1}^{K} N'_k r_k}{\sum_{k=1}^{K} N'_k}$$

1.4.2 Indirect Standardisation

For indirect standardisation, we take the age-specific rates from the reference population and convert them into the mortality rate we would observe if those reference rates were true for the age-structure of the population of interest. This gives us the expected rate for the population of interest, if age-specific mortality rates were the same as for the reference population.

We calculate the age-specific mortality rates r'_k for the reference population. Suppose the population of interest has population counts $N_k; k = 1, \ldots, K$ in each age-group k. The expected rate of deaths in the population of interest is

$$\text{expected rate} = \frac{\sum_{k=1}^{K} N_k r'_k}{\sum_{k=1}^{K} N_k}$$

and the expected number of deaths in the population of interest is

$$E = \sum_{k=1}^{K} N_k r'_k$$

1.4.3 Standardised Mortality Ratio

We can compare the expected number of deaths, using the indirect standardisation method, with the observed number using the *standardised mortality ratio* (SMR). Let O be the observed number of deaths in the population of interest, and E be the expected number of deaths when indirectly standardised with respect to some reference population.

$$\text{SMR} = \frac{O}{E} \times 100$$

The SMR is a ratio, not a rate or a percentage. An SMR of 100 means that the population of interest has the same number of deaths as we would expect from the reference population. If it is greater than 100, we have more deaths than expected; if it is less than 100 we have less.

1.4.4 Confidence Intervals for SMRs

Suppose that deaths are independent of one another, and occur randomly in time so we can use the Poisson distribution with unknown parameter λ:

$$O \sim \text{Poisson}(\lambda)$$

with

$$E[O] = \lambda \qquad \text{var}(O) = \lambda$$

and so we can approximate $\text{var}(O) \approx O$. Since the expected number of deaths is calculated from a large sample we can treat it as a constant (since its variance is small enough to be negligible) so

$$\begin{aligned} E[\text{SMR}] &= E\left[\frac{100 \times O}{E}\right] \\ &= \frac{100\lambda}{E} \\ \text{var}(\text{SMR}) &= \text{var}\left(\frac{100 \times O}{E}\right) \\ &= \frac{100^2\lambda}{E^2} \end{aligned}$$

Provided the number of deaths is large enough, say $O > 10$, we can use the Normal approximation and obtain a $100(1-\alpha)\%$ confidence interval for the SMR:

$$\text{SMR} \pm z_{\alpha/2} \times 100 \times \frac{\sqrt{O}}{E}$$

2 Observational Studies

In an experiment or clinical trial we make some intervention and observe the result. In an observational study we observe the existing situation and try to understand what is happening. Observational studies are often used when an experiment would be impractical or unethical.

When examining the association between two variables, bias may be introduced by *confounding*. *Confounders* are factors that produce confounding. A confounder is a variable which is strongly associated with the response and effect of interest, and is not on the causal pathway between them. In medicine, an important task is to determine the *cause* of a disease so that we may devise treatment or prevention. The interpretation of observational studies is more difficult than a randomised trial as bias due to confounding may influence the measure of interest. If careful consideration is taken beforehand to identify and measure important confounders that may differ between exposure groups, these can be incorporated in the analysis and the groups can be adjusted to take account of differences in these baseline characteristics. Of course, if further factors are present and not included these will bias the results.

2.1 Cross-sectional Studies

An investigation will often start with a cross-sectional study, where a large number of people are asked the same questions, for example to relate current self reported health status to self reported lifestyle factors. While this type of study can take place relatively quickly and cheaply it relies on peoples willingness to fill in questionnaires, their honesty and their memory and other factors. Furthermore, when the condition of interest is rare a very large sample is needed.

2.2 Cohort studies

Cohort studies as observational rather than intervention studies observe the progress of individuals over time (they are said to be *longitudinal* rather than cross-sectional). A study may follow two groups, one exposed to some risk factor and the other not to see if, for example, exposure influences the occurrence of certain diseases. Cohort studies are useful for investigating and determining aetiological factors.

Cohort studies as observational studies suffer from the same problems as cohort experiments (see Section 5.3). For rare conditions very large samples will be required to observe any cases. One insurmountable problem with cohort studies is that if the cohort has to be set up there is a long delay before analysis can take place. For this reason some cohorts have been set up as a resource to be used by future researchers.

2.3 Case controlled studies

Another solution to the problem of the small number of people with the condition of interest is the case-control study. In this we take a group of people with the disease, the cases, and a second group without the disease, the controls. We then find the exposure of each subject to the possible causative factor and see whether this differs between the two groups. The advantage of this method of investigation is that it is relatively quick as the event of interest already having happened, and cheap as the sample size can be small. However, there are difficulties in the selection of cases, the selection of controls, and obtaining the data. Because of these and other problems, case-control studies often produce contradictory and conflicting results.

2.3.1 Problems with case controlled studies

Selection of cases Case selection usually receives little consideration beyond a definition of the type of disease and a statement about the confirmation of the diagnosis. However, we must careful of confusing the cause of a disease with the process used to detect it; this is called *ascertainment bias*.

Selection of Controls The selection of controls is one of the most difficult areas of study design. We want a group of people who do not have the disease in question, but who are otherwise comparable to our cases. We must first decide the population from which they are to be drawn. There are many sources of controls, two obvious ones being the general population and patients with other diseases. The latter is usually preferred because of its accessibility. However, these two populations are not the same. While it is easier to use hospital patients as controls, there may bias introduced because the factor of interest may be associated with other diseases.

Choosing the sample Having defined the population we must choose the sample. There are many factors which affect exposure to risk factors, such as age and sex. The most straightforward way is to take a large random sample of the control population, ascertain all the relevant characteristics, and then adjust for differences during the analysis, using multivariate methods. The alternative is to try to match a control to each case, so that for each case there is a control of the same age, sex, etc. Having done this, then we can compare our cases and controls knowing that the effects of these intervening variables are automatically adjusted for. If we wish to exclude a case we must exclude its control, too, or the groups will no longer be comparable. We can have more than one control per case, but some of the analysis becomes complicated. Matching on some variables does not ensure comparability on all. The more we match for, the fewer intervening variables there are to worry about. On the other hand, it becomes more and more difficult to find matches. Matching for more than age and sex can be very difficult. Having decided on the matching variables we then find in the control population all the possible matches. If there are more matches than we need, we should choose at random. If no suitable control can be found, we can do two things. We can widen the matching criteria, say age to within ten years rather than five, or we can exclude the case.

Interpreting the results There are difficulties in interpreting the results of case-control studies. One is that the data collection in the case-control design is usually retrospective, that is, we

are starting with the present disease state, e.g. lung cancer, and relating it to the past, e.g. history of smoking. We may have to rely on the unreliable memories of our subjects to recall past events. Interviewers will very often know whether the interviewee is a case or control and this may well affect the way questions are asked.

2.4 Prospective vs retrospective

Cohort studies are often referred to as *prospective* studies and case controlled studies as *retrospective* studies, because of the way the models operate either forwards or backwards through time. Confusingly, the terms prospective and retrospective are also applied to when the data were collected. Data collected prospectively is collected at the time it was current, for example recording baseline data on a cohort when the cohort was set up. Data collected retrospectively is collected later, once it is clear what data items are required.

2.5 Ecological Analysis

Ecology is the study of living things in relation to their environment. In epidemiology, an *ecological study* is one where the disease is studies in relation to the characteristics of the communities in which people live; so areas or communities are the unit of interest. For example, we might compare lung cancer mortality rates for a set of areas, with measures of social deprivation for those areas, and average levels of smoking in the areas. Sometimes variables are intrinsic characteristics of the area, such as social deprivation, and sometimes they are individual characteristics aggregated over the areas, such as average smoking level.

Relationships found in ecological studies are area-level relationships; we cannot conclude that there is a relationship at the individual person level. To do so is called the *ecological fallacy*. Ecological studies can be useful to generate hypotheses, but are never sufficient on their own without corroborative evidence from elsewhere.

2.6 Meta-analysis

When several studies have been conducted on one topic it is possible to pool the results and draw firmer conclusions. This process is called *meta-analysis*. Each separate study gives a single estimate of the treatment effect. We assume the these are all estimates of the same population parameter and, if the assumptions are satisfied, we may combine them to give a common estimate.

We first need to clearly define the question of interest and ensure that only relevant studies are included. We need to include *all* relevant studies, not just those that have been published. We also need to ensure that the studies are measuring the same effect, on comparable populations.

2.7 Evidence from Observational Studies

There are many problems in using observational designs. We have no better way to tackle these questions and so we must make the best of them and look for consistent relationships which stand up to the most severe examination. We can also look for confirmation of our findings indirectly, from animal models, plausible biological mechanisms or from laboratory experiments. A large number of different studies, encompassing different types of study, methods of analysis and study populations, which all give similar results lend weight to the conclusions, as do temporality of disease and risk factor, and very strong relationships. However, we must accept that perfect proof is impossible in these issues, and often we must act on the balance of evidence from a wide range

of sources. For example, the weight of evidence, biological plausibility and consistency over many studies all argue in favour a causal link between smoking and lung cancer.

3 Relative Risks and Odds Ratios

Let D and $\bar{D}$ denote 'disease' and 'not disease', and E and $\bar{E}$ denote 'exposed' and 'unexposed' respectively. Then the data for a comparative cohort study, or a case-control study can be written as in Table 3. In a population study, the total sample size n_{++} is fixed; in a cohort study, the margins n_{+0} and n_{+1} are fixed; and in a case-control study, the margins n_{0+} and n_{1+} are fixed.

Table 3: Notation for cohort or case-control study.

		Exposure to Risk Factor		
		Unexposed ($\bar{E}$)	Exposed (E)	
Disease	Absent ($\bar{D}$)	n_{00}	n_{01}	n_{0+}
	Present (D)	n_{10}	n_{11}	n_{1+}
		n_{+0}	n_{+1}	n_{++}

3.1 Risk and Relative Risk

The risk of an event is the probability that an event will occur within a given period of time. The risk of an individual having a disease ($P(D)$) is estimated by the frequency with which that condition has occurred in a similar population in the past, that is its incidence: n_{1+}/n_{++}.

To compare risks for people with and without a particular risk factor, we look at the ratio. Suppose the risk for the exposed group is $\pi_1 = P(D|E)$ and the risk for the unexposed group is $\pi_0 = P(D|\bar{E})$. Then the *relative risk* is

$$\text{RR} = \frac{\pi_1}{\pi_0}$$

For a cohort study, we have

$$\text{RR} = \frac{n_{11}/n_{+1}}{n_{10}/n_{+0}} = \frac{n_{11}n_{+0}}{n_{+1}n_{10}}$$

We cannot estimate the relative risk directly from a case-control study, as the n_{++} individuals are not selected at random from the population.

For diagnostic and screening tests (Section 4) the relative risk can be used to compare the risk of having the disease for test positive and test negative:

$$\text{RR} = \frac{P(D|T)}{P(D|\bar{T})} = \frac{\text{PPV}}{1-\text{NPV}}$$

3.1.1 Population Attributable Risk

If a relative risk is large but very few people are exposed to the risk factor then the effect on the population will not be large, despite serious consequences to the individual. The effect of a risk factor on community health can be measured by *attributable risk*, which is related to the relative risk and the percentage of the population affected.

The attributable risk is the proportion of the population risk that can be associated with the risk factor:

$$\text{AR} = \frac{\text{population risk} - \text{unexposed risk}}{\text{population risk}} = \frac{P(D) - P(D|\bar{E})}{P(D)} = \frac{\theta(\text{RR} - 1)}{1 + \theta(\text{RR} - 1)}$$

where θ is the proportion of the population who are exposed to the risk factor, $P(E)$.

3.2 Odds and Odds Ratio

An alternative to the risk is the *odds* of an event. If the probability of an event is p then the odds is

$$\frac{p}{(1-p)}$$

This can be useful since it is not constrained to lie between 0 and 1. We often use the *log odds*:

$$\log\left(\frac{p}{(1-p)}\right)$$

Another way to compare people with and without a particular risk factor is to use the *odds ratio*. The odds ratio for disease given exposure is

$$\text{OR} = \frac{\text{odds of disease given exposed}}{\text{odds of disease given unexposed}} = \frac{P(D|E)\left(1 - P(D|\bar{E})\right)}{P(D|\bar{E})\left(1 - P(D|E)\right)}$$

For a 2×2 table as in Table 3 the odds ratio is

$$\text{OR} = \frac{n_{00}n_{11}}{n_{10}n_{01}}$$

For confidence intervals, we use the log odds ratio and transform back, since

$$\text{var}(\log(\text{OR})) = \frac{1}{n_{00}} + \frac{1}{n_{01}} + \frac{1}{n_{10}} + \frac{1}{n_{11}}$$

and for a large enough sample we can assume that the log odds ratio is normally distributed. If the confidence interval for the odds ratio does not contain 1, this is equivalent to rejecting the null hypothesis $H_0 : \text{OR} = 1$.

The odds ratio is a useful measure since it is independent of the prevalence of the condition.

3.2.1 Odds ratio in Case-control Studies

This is particularly useful in a case-control study. Above, we have given the formula for the odds ratio for disease given exposure. From a case-control study we can calculate the odds ratio for exposure given disease:

$$\begin{aligned}
\text{OR}_{E|D} &= \frac{n_{00}n_{11}}{n_{10}n_{01}} \\
&= \frac{P(E|D)P(\bar{E}|\bar{D})}{P(E|\bar{D})P(\bar{E}|D} \\
&= \frac{P(D|E)P(E)P(\bar{D}|\bar{E})P(\bar{E})P(\bar{D})P(D)}{P(D)P(\bar{D})P(\bar{D}|E)P(E)P(D|\bar{E})P(\bar{E})} \\
&= \frac{P(D|E)P(\bar{D}|\bar{E})}{P(\bar{D}|E)P(D|\bar{E})} \\
&= \text{OR}_{D|E}
\end{aligned}$$

which is the same as the odds ratio for disease given exposure. So although we cannot calculate a relative risk for a case-control study, we can estimate the odds ratio, provided that the cases and controls are random samples from the same population.

3.2.2 Relationship to relative risk

The odds ratio gives a reasonable estimate of the relative risk when the proportion of subjects with the disease is small.

Then the risk of disease for an exposed person is

$$\begin{aligned} \pi_1 &= \frac{P(D \cap E)}{P(E)} \\ &\approx \frac{P(E|D)P(D)}{P(E|\bar{D})} \end{aligned}$$

Similarly,

$$\pi_0 \approx \frac{P(\bar{E}|D)P(D)}{P(\bar{E}|\bar{D})}$$

So

$$\begin{aligned} \text{RR} &\approx \frac{P(E|D)P(D)P(\bar{E}|\bar{D})P(\bar{D})}{P(E|\bar{D})P(\bar{D})P(\bar{E}|D)P(D)} \\ &= \text{OR} \end{aligned}$$

So in practice the odds ratio from a case-control study can often be interpreted as if it were the relative risk.

3.2.3 Paired Samples

If the data in the table is paired, we can find the conditional odds ratio. Like the McNemar test, it uses the off-diagonals only:

$$\text{OR}_{\text{paired}} = \frac{\text{case exposed, control unexposed}}{\text{case unexposed, control exposed}}$$

and can be used in a matched case-control study as an estimate of the relative risk.

For confidence intervals, we note that if we define p to be the proportion of mismatched pairs with case exposed and control unexposed, then $\text{OR}_{\text{paired}} = p/(1-p)$. We can first find a confidence interval for p, and then transform the upper and lower limits to get a confidence interval for $\text{OR}_{\text{paired}}$.

3.3 Meta-analysis

Odds ratios from a number of different studies can be combined in a meta-analysis using the *Mantel-Haenszel* method. This is suitable for *fixed effects*; that is, we assume that the effect of the risk factor is constant across all studies.

Suppose we have S studies, indexed by $s = 1, \ldots, S$, each with a table of the form Table 3, where we extend the notation in the obvious way to n_{ijs} to represent the number with disease status i, exposure status j in study s. Each study has an estimated odds ratio of

$$\text{OR}_s = \frac{n_{00s}n_{11s}}{n_{10s}n_{01s}}$$

Table 4: Results from several case-control studies on smoking and lung cancer

Lung cancer		Controls	
smoker	non-smoker	smoker	non-smoker
83	3	72	14
90	3	227	43
129	7	81	19
412	32	299	131
1350	7	1296	61
60	3	106	27
459	18	534	81
499	19	462	56
451	39	1729	636
260	5	259	28

The Mantel-Haenszel odds ratio estimator is given by

$$\mathrm{OR}_{\mathrm{MH}} = \frac{\sum_{s=1}^{S} \frac{n_{00s} n_{11s}}{n_{++s}}}{\sum_{s=1}^{S} \frac{n_{10s} n_{01s}}{n_{++s}}}$$

Similarly, the Mantel-Haenszel relative risk estimator

$$\mathrm{RR}_{\mathrm{MH}} = \frac{\sum_{s=1}^{S} \frac{n_{11s} n_{+0s}}{n_{++s}}}{\sum_{s=1}^{S} \frac{n_{+1s} n_{10s}}{n_{++s}}}$$

The variances of both these estimators are complex and best left to a computer to calculate.