Epidemiological studies – a cautious view

David Goddard Revised July 2021

Table of contents

How do we know things?

These notes are about the harm to health that may come from too much exposure to a harmful agent. How do we know about those things?

Much of what we know comes from personal experience or from what is told to us by those of status whom we trust. Personal experience is most potent when a harmful effect

happens *quickly* and *always*, e.g. grasping a very hot object, or sniffing strong ammonia gas. Information gained by personal interaction is most believable when the teller is a person of stature and when their line of argument is reasonably consistent with what we already consider to be true.

Sometimes, though, rather than being told, we seek to act on published information. This happens particularly when harmful health effects are slow to appear – things such as cancer, birth defects, or changes to brain function. These published authors invite us to believe that a specified substance or ray has caused this sort of harm, so more control is required than happens now. The rest of what I say will be about such a situation.

However, there is potential for confusion when we attempt to satisfy ourselves whether an exposure causes slow-appearing harmful health effects. This potential for confusion comes for two reasons:

- not all of those exposed suffer the effect – indeed, usually it is a minority; and
- people *without* such exposure also suffer the effect.

As a result, we can never be *certain* that a particular exposure has caused a slow-appearing disease. Nor can we be certain that it hasn't. An assertion about a causal[1] link between exposure and the disease can only honestly be couched in terms of probability. It is a matter of "*how likely*" **not** "*is*" or "*isn't*"

To find out how likely it is that an exposure will cause slow-appearing harmful health effects, we rely on both animal experiments and human studies. The term *epidemiology* embraces health studies on groups of people.

What is epidemiology?

Epidemiology is the study of *which* people in the world get *what* diseases and *why*. A typical unit of epidemiology is a comparative study of two or more groups of people ranging in number from dozens to tens-of-thousands. Mostly, this requires several investigators and substantial funding to enable on-going gathering of data about exposure and health status. The data is analysed for causal trends. In drawing conclusions, the investigators will often say that the findings from their study are likely to apply more generally – to all people in the world in similar circumstances. That generalisation is likely to be accurate *only if*:

- the group studied has features that are mirrored in many other places; and
- there is a firm basis for assuming that the exposure has caused the health effect.

Of course, not all causal factors require epidemiology to bring them out. So, let's for a moment consider a world without epidemiology.

A world without epidemiology?

The study of *individual cases* will satisfy us that grasping a live high-voltage electrical wire causes electrocution, or that ammonia, when heavy in the air, causes eyes to water.

[1] Sometimes my students misread 'causal' [implying a cause] and think I mean 'casual' [unconcerned or happening by chance]. To avoid this risk, I do not use the word 'casual' in these notes.

This is so because the effect happens immediately on exposure, and because it always happens. Also, where a cluster of an extremely unusual cancer occurs in a group that is homogeneously and heavily exposed to a particular agent, then causal association between the cancer and the exposure would most likely be presumed.

However, when persons' exposure to a suspected substance or physical agent is well separated in time from an alleged health effect[2], then any serious effort to associate such human exposure with that effect will always employ epidemiological methods, i.e. it will employ the systematic study of groups, not simply perusal of individual case histories.

Without epidemiology, any assertion that linked smoking and heart disease, human papilloma virus and cervical cancer, or cadmium and kidney disease would be a barely educated guess.

Of course, epidemiology has its limitations. One situation where it loses its potency is with exposures of very low-intensity. We shall come to this later.

Now let us look at some types of epidemiological study. Our purpose is to examine how well each study type helps us to:

- *observe differences* between groups; and
- *explain* why this may be.

Before I speak of the different studies and what they yield, I offer some words of guidance. I do this because most of my students have found some ideas in epidemiology to be difficult to grasp. So there are some things to do so as to make it easier to learn.

Being part of the culture of epidemiology

If we get employed in a new job, we very likely find that things are done there somewhat differently from where we've been. We want to be well thought-of by new workmates. So, as part of fitting-in, we may adopt some of the ways that others speak or even their informal dress code – all part of the superficialities of conveying that "I am one of you".

So it was when, thirty plus years ago, I entered a "Land" of epidemiology – (what is now) the Monash University School of Public Health and Preventive Medicine. I faced a forest of new terminology and felt that, to be 'cool', I needed to speak nonchalantly about significant, standard error or sensitivity; to roll off my tongue ARR, SEM or NNT; to have a formula to calculate chi-square or likelihood ratio. So I fitted in. But, when I came to teach, my students saw terminology more as a barrier than a gateway to understanding. I faced legitimate questions such as, "Why do we have to learn this *sh*t*?', "What's *that* mean?", or assertions such as "Maths freaks me out!". I found that most student mistakes, especially in the use of a formula, were because my student didn't really know (or sometimes didn't care) what he or she was trying to achieve by using it, or because my student didn't understand the notion of ratio, proportion or why dividing by a number is the same as multiplying by its reciprocal.

[2] E.g. Does living near power lines for 10 years increase our likelihood of getting cancer?

My students were smart people with wide interests. Their difficulty with epidemiology was occasioned by the adage, "If you don't use it, you lose it". Many had practised capably in medicine or other health fields for several years, but had no reason to bring their high school maths into their working lives.

So, that's why I say to you what I now do. I suggest you firstly skim-read the next four pages. Then, if you feel confident that you know these things, carry on with the rest of the document. Otherwise come back here and go over the parts that could aid your understanding.

Explaining ideas to others

You know best when you've learnt something if you can explain it to others in a convincing way. And, in explaining, you will feel more empowered if you can vocalise your ideas using terms or analogies that are familiar to your everyday life.

Some of the most profound thoughts can be expressed in simple words. Many of the great books illustrate that. If you tell a person something using your own, plain language, then he or she will know that you have brought the concept into your own personal area – that you are attempting to say: "I have thought about this and here is my perspective: …." It also respects the person with whom you are communicating. It says, "Hey, this is what I believe". Merely reciting a textbook or formal definition is like a proclamation from on high, or a call from far, far away. It lacks the softness or the sort of pleasure-in-giving that a truly personal message can bring. Yet expressing sophisticated or abstract ideas in lay terms can risk embarrassment if what you say comes out all wrong. You can feel less vulnerable, less exposed, if you hide behind a technical term.

The ability to persuade another person depends upon creation of understanding *plus* a sincere attempt to try to see the world from the other person's point of view. You cannot persuade a person who doesn't understand you nor who believes you don't care about what is important to him or her.

When technical terms that occur in pairs sound similar, it can be confusing. For example, *specificity* and *sensitivity* often appear together. It's a bit like seeing a friend with her twin girls. One has to pause a moment to tell the twins apart and address each by her right name.

On using numbers

This short passage may seem to be a particularly strange thing to offer to you who have survived school, a medical course and are now undergoing physician training. However, I have met many medical students, and even doctors, who are not at all relaxed with numbers.

In modern society, numbers tend to speak with a 'loud voice'; politicians, economists and many others take them very seriously indeed. The risk in doing this is that the number itself becomes the focus and we lose sight of the broader pattern to which it belongs.

Realise that numbers are simply adjectives. A word such as *four* or *twenty-one* simply tells us more about people, things or events. The use to which a number is put is what determines the importance of that number.

Some numbers are important enough to remember, and that's easier to do if the number is simplified. The front of a number is usually more important than its rear end. For example, it is nearly always more useful to know whether we have **4**23 or **9**23 dollars left in our bank account than whether we have 42**3** or 42**8**. Very often this gives us latitude to simplify numbers by 'rounding off' their rear ends – thus 21.8 may become 22 or perhaps, depending on the circumstances, a suitable approximation may be 20.

Many calculations can be done in one's head. Mental arithmetic is made relatively easy by rounding off numbers. For example, the product of 22.6×37 is, for most purposes, near enough to 20×40. Rounding off numbers, so that operations such as multiplication and division can be done in the head, makes the act of calculating brief. That way, the effort of calculating does not long distract us from the people, things or events that the numbers are telling us about. Certainly, it's helpful to get a rough estimate done in your head before using a calculator.

If a calculator is used for multiplication or division, it may display more digits after a decimal point than are needed. This makes a number appear more complex and more difficult to recognise for what it is. Before using a calculator to multiply or divide numbers, it helps to decide how many digits (if any) we need after the decimal point. This way we may simplify a long number by rounding off any digits that are superfluous.

Appropriate rounding-off also shows that you are aware of the limitations of your measuring instrument. Other professionals will respect your recording of a patient's temperature as 37.5°C, but *not* a systolic blood pressure reading of 151**.6** mmHg nor a sound level measurement of 87**.2** decibels.

Ratio

Quite often, we seek to compare two numbers, e.g. shoppers comparing prices in a store. Formally, comparison can be achieved in two ways. By *subtracting* one number from the other, the *difference* between them tells you *how much more* one is than the other. Alternatively, *dividing* the larger number by the smaller one, tells you *how many times more* one is than the other. If you divide the smaller number by the larger one, you obtain a fraction which tells you *how many times less* is one than the other.

A ***ratio*** is a comparison of two numbers achieved by dividing one by the other.

I'll start with a simple example. Say you have a bowl of fruit with three apples and two oranges. The ***ratio*** of apples to oranges is 3**:**2. For some purposes, it is nice to have the number on the right hand side of the colon, '**:**', equal to 1. This can be achieved by dividing the number on each side of the colon by (in this case) 2, thus, $\frac{3}{2} : \frac{2}{2} = 1.5:1$. Quite often then, the '**:**1' is left off and the ratio is expressed simply as 1.5.

So a ***ratio*** is a comparison of numbers where that comparison is likely to be useful or have some meaning, e.g. big/small, red/black. In epidemiology, a ratio can be the

number of observations with a characteristic of interest compared with the number *without* that characteristic, e.g. 43 **:** 26. Such a ratio is usually easier to remember when the right hand number is scaled down to 1. This is achieved by dividing the left-hand number by the right-hand one, thus, 43 **:** 26 = $\frac{43}{26}$ **:** $\frac{26}{26}$ = 1.65 **:** 1. Usually, in epidemiology, the "**:** 1" is left off and, if that were done here, this ratio would be simply expressed as $\frac{43}{26}$ or 1.65.

Some ratios are more sophisticated, e.g. relative risk and odds. The common unit of sound level measurement, the decibel, is a ratio; zero decibels is *not* zero sound.

Comparing numbers by *difference* maintains the units of measurement, e.g. dollars and cents for price. Comparing numbers by *ratio* cancels out the units of measurement.

Proportion

When something is part of a whole and we want to say how large a part it is, then we use words like half, three-quarters or a percentage. The general term used for such words is *proportion*.

Proportion is the number of items or observations with a characteristic of interest divided by the total number of items or observations. An example of proportion is prevalence or probability; so is sensitivity and specificity. In epidemiology, a proportion typically appears as a decimal fraction between 0 and 1, or as a percentage.

To come back to the bowl of fruit, the ***proportion*** of apples in the bowl is $\frac{3 \text{ apples}}{5 \text{ pieces of fruit}}$ or simply $\frac{3}{5}$. The proportion of oranges in the bowl is $\frac{2 \text{ oranges}}{5 \text{ pieces of fruit}}$ or simply, $\frac{2}{5}$. A *proportion* is a part divided by the whole.

In conversing about proportion, it often seems easier to talk in percentages than decimal fractions with a value less than one. However, the simplest form of many equations uses the decimal fraction rather than the percentage. Please make sure you know if your equation calls for a decimal fraction rather than a percentage, and be able to slip quickly from one to the other in your own mind, e.g. 5% = 0.05, 20% = 0.20, 75% = 0.75.

And, just as counted numbers can be compared as ratios, proportions can too, e.g. relative risk (risk ratio).

Dividing – numerators and denominators

A common indicator of dividing is to put the number *to be* divided in the numerator, and the number by which the numerator is divided into the denominator. Thus 6 divided by 3 may be written $\frac{6}{3}$, or 1 divided by 2 may be written $\frac{1}{2}$. With algebra, using symbols in place of numbers, *x* divided by *y* may be written $\frac{x}{y}$. If the numerator and denominator are the same, e.g. $\frac{6}{6}$ or $\frac{x}{x}$, then the division of numerator by denominator equals 1.

So, $\frac{x}{x} - \frac{y}{x}$ can be expressed as $1 - \frac{y}{x}$ or it can be given a common denominator, $\frac{x-y}{x}$.
Or, turning it around, $\frac{x-y}{x} = 1 - \frac{y}{x}$.

Reciprocals

In ordinary conversation, the adjective *reciprocal* refers to the giving of something in return, a mutual exchange. In maths, though, when you start with a number, its ***reciprocal*** is *what you must multiply this number by in order to get 1*. For example, the reciprocal of 2 is ½ because 2 × ½ = 1. Similarly, the reciprocal of 4 is $\frac{1}{4}$, and the reciprocal of 10 is $\frac{1}{10}$ or 0.1, and the reciprocal of $\frac{4}{3}$ is $\frac{3}{4}$ because $\frac{4}{3} \times \frac{3}{4} = 1$. Similarly, by replacing numbers with symbols, we have the reciprocal of x is $\frac{1}{x}$ because $x \times \frac{1}{x} = 1$, and the reciprocal of $\frac{x}{y}$ is $\frac{y}{x}$.

When you multiply by a fraction, e.g. 5 × ½ = 2½, it's fairly easy to see what happens and why. However, it is less intuitive when you divide by a fraction, e.g. 5 ÷ ½. This division asks "How many halves go into 5?" It's actually something that mothers do all the time. You have five oranges but more than five mouths to feed. So, what do you do? Cut them in half! How many half oranges do you get out of five oranges? Ten, i.e. 5 × 2 halves. This means that dividing by $\frac{1}{2}$ is the same as multiplying by its reciprocal, i.e. $\frac{2}{1}$ or, simply, 2.

So 5 ÷ ¼ = 5 × 4 = 20, and $4 \div \frac{1}{10} = 4 \times 10 = 40$. Too, $6 \div \frac{3}{4} = 6 \times \frac{4}{3} = 8$. Similarly, with symbols: $x \div \frac{1}{y} = x \times y = xy$. Or $\frac{a}{b} \div \frac{c}{d} = \frac{a}{b} \times \frac{d}{c}$. And $\frac{a}{b} \times \frac{d}{c}$ may be condensed to $\frac{ad}{bc}$.

Equations

Equations help us to calculate a number that we want to know from related numbers that we already know. Equations are usually written in symbols to make them generally available for whatever appropriate numbers we'd like to insert.

The 'look' of an equation may, at first seem formidable. For example, an equation at the top of the next page can be used when a patient is given a diagnostic test. Before the test is done, we make our own estimate of the probability that our patient has the disease of interest. In symbols, let's call that p_1. Its numerical value will be between 0 and 1. We choose a test that is known to distinguish people with this disease from those that don't have it. How well the test does this is measured by what is called the positive likelihood ratio which we'll give the symbol L. The values of L can be anything from 1 to double figures; the higher the value of L, the better its ability to distinguish diseased from not.

If the result of the test is positive, we may calculate the probability p_2, that our patient has the disease based on the numerical values of p_1 and L.

The equation used to calculate p_2 directly from p_1 and L is:

$$p_2 = \frac{L}{\frac{1}{p_1} + L - 1}$$

The equation might look formidable, but just pause and let it speak to you. 'Take a history' from it. Look at the way the terms are laid out on the right-hand side of the equation to form a pattern, just as, say, the pattern of fever, rhinorrhoea, cough, conjunctival suffusion and Koplik spots signal the pre-eruptive stage of measles.

By examining the individual terms on the right-hand side of this equation, you will see that if p_1 is small, say less than 0.1, then $\frac{1}{p_1}$ (>10) will dominate the denominator, and hence diminish the value of p_2 , unless *L* is large. If *L* is *very, very* large, its presence in *both* numerator and denominator will cause the right-hand side of the equation to approach 1.

Studies that reveal differences

Ecological studies

When I first heard the term 'ecological' given to a type of epidemiological study, I imagined forests and wetlands and furry, burrowing animals. In fact, an ecological study simply means the comparison of two (or more) naturally-occurring communities. You may thus compare two countries, two suburbs, or a group of foundry workers with a group of workers in the extrusion area. The investigator asks the question: "How does this community differ from that community in respect of exposures and health outcomes?" The study compares previously-gathered descriptive statistics of each group; it does *not* go down to individual level.

As a result, an ecological study can, relatively cheaply, identify differences between communities and maybe suggest possible causes. The study reveals approximately how many people within each group had the exposure of interest and how many got the disease, but its weakness is that we *don't* know how many *exposed people got the disease*. So, although it usefully identifies inequalities between two groups; it cannot effectively penetrate why those inequalities exist. Nevertheless, it can usefully point to situations where a more penetrating form of study may be appropriate.

Cross-sectional studies (surveys)

Surveys are common; the most obvious example is political polling and, on a large scale, the fifth-yearly Australian census.

Like an ecological study, a survey that compares two communities or groups may identify current differences between groups in regard to exposures and health outcomes, and so it may provide suggestions of association between them that can be followed up. However, because the information about both the health outcomes and the exposures is collected all at the same time, clear links or associations cannot usually be established.

So, *ecological studies* and *surveys* can describe differences between groups or communities but not usually explain them. Now we come to studies that are designed to probe the reasons for differences. All of these study types require collection of detailed information from individuals. This involves time and substantial cost.

Studies that explore the *reason* for differences

The reason for a difference in a health outcome following exposure to a substance or form of radiation can be either a harmful or a therapeutic effect. Harmful effects may be studied either by starting with people that have already experienced harm and asking them about past factors that may be thought to have caused the harm. The results are compared with the results from a group who are healthy and who are also asked the same questions about their exposure in the past. This is called a *case-control* study. Alternatively, a group with what is thought to be a harmful exposure may be followed up to see how many individuals suffer harm to health when compared with a group without that exposure. This is called a *cohort* study. It is an exposure-control study, i.e. in its simplest form, one group has the exposure, the other group doesn't.

Yet, this research work is important to the extent that it leads to new action in preventive toxicology, or that it confirms the value of action taken already.

Case-control studies

A case-control study begins with cases of a nominated disorder and control subjects who do *not* have the disorder. It then looks backward to identify possible precursors or risk factors. It's a bit like a 'reverse quiz' that says "Here is the answer; now, what is the question?" Case-control studies are of medium cost and are an effective way to investigate *rare disease*.

Bias in case-control studies

Case-control studies have strong potential for bias. If that occurs, the findings may be idiosyncratic, i.e. not fairly generalisable to all people in the world in similar circumstances. *Without* the ability to generalise, the findings of a study are a mere curiosity, *not* a widely usable set of facts. Unless the studied group reasonably fairly represents a wider population, then one cannot generalise reliably from it.

Bias is of two main types:

- bias in selection of participants to act as controls; and
- bias of recall.

I'll take these in turn. Many case-control studies are conducted by telephone questionnaire. Partly because of the frequent and intrusive telephone calls made for commercial purposes by call-centres, people getting 'cold-called' to be controls in a case-control study are now more likely than not to say "No". This means that those who agree to be controls in the study are not a random group – it's possible that they have special characteristics. This is called a *selection bias*, i.e. a bias that occurs when there is a difference between the characteristics of the people participating in a study and the characteristics of those who would have been eligible but who did not participate.

In case-control studies, the study participants are questioned about their *past* experiences. People with serious disease may be more reflective ("Why did this happen to me?") than the control subjects who are well. The past is then differently (selectively or more vividly) recounted by people in one group than by those in the other. This is termed *recall bias*; it typically comes out as under-reporting of exposure when questions about the past are asked of members of the *control* group. The bias will be greater where there is a self-interest (e.g. compensation) associated with recalling particular experiences.

Cohort studies

The form of cohort study that is simplest and easiest to understand is where a group *with* exposure to a particular substance (or family of substances) is identified along with a comparison group that does *not* have that exposure. The members of each group are followed forward in time, perhaps for ten or twenty (or more) years. At the end of the period of study, the proportion of those that suffered a disease of interest in the exposed group is compared with the proportion of those that suffered that disease in the *non*-exposed group.

Of course, things are almost never that simple. Cohort studies are expensive and time-consuming, so there are typically many comparison groups. For example, there may be several levels of exposure-intensity compared. Several disease outcomes may be studied. Exposed and unexposed groups may be further subdivided according to age, sex, smoking habits or other concurrent exposures. Sometimes there is no control group; the occurrence of disease in the exposed group is instead thoughtfully compared with population statistics. Sometimes, the assignment of participants to 'exposed' and 'non-exposed' groups is done on the basis of historical records.

Potentially, one strong feature of the cohort design is that exposure can be measured. Unfortunately, unless continuous monitoring is performed, truly representative readings are forever difficult to obtain.

A cohort study is also prey to bias. Selection bias occurs when study participants are lost to follow-up as the study proceeds, so that their health outcome is unknown. The personal characteristics and lifestyles of those who remain in a study will often differ from those who drop out. As well as this, a bias of measurement occurs if:

- exposure measurements are not representative, or
- participants are followed for a period that is too short for the disease of interest to become manifest, or
- the disease of interest is ill-defined (e.g. multiple chemical sensitivity), or difficult to diagnose, or if its occurrence is not reliably recorded.

Intervention studies

All substances have potential for harm but some, at low exposure, can have a therapeutic effect. Substances *without* therapeutic effect can be studied only by observing what happens to the health of people who are ordinarily exposed, e.g. in a particular occupation. Thus case-control and cohort studies are referred to as *observational*.

Where exposure to a substance has potential for benefit, people who stand to benefit may actually be selected for deliberate exposure, i.e. to receive a measured dose of a substance on a daily basis for a period of time. This is known as an *intervention* study.

The classical type of intervention study is a clinical trial of a new drug. The members of one group get the new drug and members of another matched group get conventional therapy. Because (ideally) the groups differ essentially by just a single factor – i.e. whether a study participant gets the new drug or not – an intervention study is a powerful probe of cause and effect.

With some drugs, the range of exposure at which they are therapeutic can overlap with the range at which they are toxic. Therefore, an intervention study may look for benefit or harm (or both). Let's for discussion today consider a new anticancer drug being investigated for neurological side-effects (neuropathy) if therapy is sustained for long periods, i.e. alteration one's sense of feeling for hot and cold.

The analysis of the data obtained in an intervention study will produce a number (or multiple numbers) that summarise the result. Let's say that of 200 people randomly assigned to take the drug on trial, 20 experienced neuropathy – changes to their sense of feeling. In a control group of 250 – who received different therapy – there were 15 people with this neuropathy. In other words, for those taking the drug on trial, there was a proportion of $\frac{20}{200} = 0.1$ of people experienced neuropathy, compared with $\frac{15}{250} = 0.06$ in the control group.

If the two proportions had been the same or almost the same, then we would have simply concluded that our study found whether people had one drug or the other made *no* difference to the incidence of neuropathy. However, these two proportions are clearly different.

So a comparison can be made. This comparison may simply be written as a ratio, thus:

$$0.1 : 0.06$$

This is interesting but those 'ugly' numbers are hard to carry around in one's head. What makes it easier to grasp is if we ask, *how many times* more likely is neuropathy among those taking the drug on trial than in those taking other therapy. This is called the *relative risk* or risk ratio[3].

To obtain it, we convert the number on the right hand side of the colon to 1. This is done by dividing the number each side of the colon by (in this case) 0.06 thus:

$$\text{i.e. } \frac{0.1}{0.06} : \frac{0.06}{0.06} = 1.67 : 1$$

[3] In statistics, the term *risk* differs from its use in occupational health. In statistics, it refers to the proportion of events of interest that occurred during a period of study. Such information, drawn from what has already happened, may be used to predict what could happen were the situation to persist, i.e. the information gathered during the period of study can be used to work out the probability of future occurrences.

Typically, the "**:**1" is left off and the relative risk is (in this case) expressed simply as 1.67.

This is the estimate, obtained from comparison of the two groups, of how much more likely it is that neuropathy will occur among those taking the drug on trial than among those taking the other therapy. It is an *estimate* of the effect of the association between exposure to the drug on trial and neuropathy.

This estimate may be generalised to other people who also take this new drug. However, for them, the relative risk is unlikely to be *exactly* 1.67. This brings me to the topic of confidence intervals.

Confidence intervals[4] - *so **VERY** important. Please read carefully.*

Commonly, a team of investigators will compare the health outcomes of two groups of people and generalise those findings to all people in the world in similar circumstances. By reasoning in this way, the team is regarding the members of each group to be a sample of a much larger population. In pursuing this line of argument, a careful investigative team will try hard to ensure that the groups that were studied and generalised from do indeed resemble what might be expected elsewhere in the world.

However, of two or more samples drawn from a much larger population, it is rare to find that they each have identical characteristics. Instead, they generally differ from each other, often just a little, occasionally quite markedly. This, in turn means that no sample is a perfect miniature of the population from which it is drawn. It is likely to be just a little different but occasionally a lot different.

So, the sample finding of a relative risk of 1.67 is most unlikely to be the *exact* value of the relative risk that applies worldwide. Therefore, it becomes necessary to calculate a *range of values* within which we can reasonably claim that worldwide reality[5] lies – such reality being impracticable to ascertain by direct measurement.

What we have then is a range of values, drawn from a sample, within which we believe the world value of the relative risk most likely lies. This range is called a confidence interval. When that interval is calculated by a method that gives the correct answer 95% of the time, it is called a 95% confidence interval. The numerical values at either end of the interval are known as the 95% confidence limits. The width of a confidence interval is determined by the size of the sample and the level of confidence required (e.g. 99%

---

[4] Here, the term *interval* means 'a space between limits'. Nowadays, in ordinary speech, the term is more often applied to a period of time, e.g. a theatre interval. However, *interval* originally meant 'between the walls' (inter-vallum), a space between two concentric ramparts of a mediaeval castle.

[5] Normally, when we speak of '*reality*' we mean what is actual, definite, existing, even tangible, plain to see or touch. We exclaim "Get real!" when someone shows too little regard for what is likely or obvious. Normally, reality is the *opposite* of what is abstract, hidden, or difficult to grasp.

However, in epidemiological research, an investigator is *trying to find reality*. When the research is started (and even sometimes when it is finished) reality is **not** plain to see or touch. Reality is **not** what is known but *what is being searched for*. It is hidden or veiled which is therefore a confusing twist to its normal meaning. It is important to be aware of this perversity, this contradiction, when we use the imperfect vehicle of epidemiology to elucidate reality.

instead of 95%). The smaller the sample or the higher the level of confidence that you demand, the wider the interval.

In summary, a 95% confidence interval takes a representative characteristic of a sample (e.g. a relative risk or a mean or proportion) and indicates how closely this will be mirrored in all similar situations in the world at large.[6]

The mathematical calculation of a confidence interval is nowadays performed by statistical software. It involves a fraction that has within its denominator the square root of *n* where, in the case of a relative risk, *n* is essentially the number of people in each of the two groups that that got the disease of interest. There are two ways to make the magnitude of a fraction bigger – either increase the size of the numerator or decrease the size of the denominator. As *n* (in the denominator) gets smaller, so the confidence interval gets wider.

Cohort studies and clinical trials commonly have an exposed or drug-on-trial group and a control group. A comparison of two groups can be expressed as a difference or a ratio, each with its attendant 95% confidence interval. For a difference, the confidence interval is symmetrical – its numerical value is the arithmetic mean of the numerical values of the upper and the lower confidence limits, i.e. ½ × (upper limit + lower limit). On the other hand, the confidence interval of a ratio is *asymmetrical* – its numerical value is the *geometric* mean of the numerical values of the upper and the lower confidence limits, i.e. square root of (upper limit × lower limit).

A relative risk that has the numerical value of 1 means that there is no difference in the proportion of people with the disease of interest, be they exposed or not. Therefore, if the calculated 95% confidence interval includes the numerical value of 1, then the usual interpretation is that any observed difference between the groups – leading to an estimated relative risk that differs from 1 – could be explained simply by random variability between two samples, i.e. that the study has failed to show that being exposed makes any difference to health outcomes.

The 95% confidence interval for the study cited here, as always, is built around the estimated relative risk – in this case 1.67. The lower limit is 0.88 and the upper limit is 3.18. In this case, there is a suggestion that taking the drug on trial is associated with an

---

[6] The notion of *confidence interval* is important. In an attempt to reinforce it with you, I'd like to share one more analogy with you which was offered to me by Dr Arul Mylvaganam, previously a statistician at Monash University, although I believe it originated elsewhere.

I have an idiosyncratic friend, a Collingwood supporter, who favours black attire but who owns a small white dog that goes everywhere with her. I often see my friend in the street with the dog on a old fixed-length lead. Sometimes, I see my friend but not the dog because it is sniffing behind a rubbish bin or whatever – but I know the dog is near because the length of the lead sets the limit of the distance of the dog from its owner. Tonight is dark. I am in the street and I see the white dog but *not* my friend. However, I can tell roughly where my friend will be because I know the length of the dog's lead.

With most research, including epidemiological, we start off 'in the dark'. In epidemiology, we observe the location of some aspect of a *sample* (the dog) then, based on that sample, calculate a *range* (the length of the dog's lead each side of the sample) within which the corresponding feature of the population at large (the person) is likely to be located. Such a range is known as a *confidence interval*.

increased incidence of neuropathy (with a relative risk of 1.67). However, simple sampling variability remains a reasonable explanation of the difference between these two samples, one of 200 with 10% affected and one of 250 with 6% affected. The way that such information is usually expressed is: RR = 1.67, 95% CI = 0.88 to 3.18.

Questions

1. *Does a 95% confidence interval mean that:*

   (a) we are 95% sure that the values fit in the interval; or

   (b) 95% of the values fit within the confidence interval.

ANSWER: Neither. A confidence interval is a range of values generated from a sample (in this case, just 450 people) within which lies the *unknown* value of the relative risk for all people in the world in similar circumstances, i.e. the unknown value of the 'population parameter'. A **95%** confidence interval is worked out by a method that gives the correct answer 95 times out of 100.

2. *Why is a 99% confidence interval wider than a 95% confidence interval?*

ANSWER: A confidence interval is drawn from a sample estimate. A 99% confidence interval includes a greater range of possible values for what is true in the world at large than will a 95% confidence interval.

- The range of a 99% confidence interval is calculated to include 99 out of every 100 of the possible values for the parameter (e.g. relative risk, odds ratio, mean, proportion, rate, count … whatever) whose true value is unknown for all people in the world in similar circumstances – simply because it's impossible to measure everyone in the world so you have to infer what is likely for everyone from the findings from a mere sample that you hope represents all similarly-affected people.
- the range of a 95% confidence interval is calculated to include include 95 out of every 100 (19 out of every 20) of the possible values for the parameter for all people in the world in similar circumstances;
- the range of a 90% confidence interval is calculated to include include 90 out of every 100 (9 out of every 10) of the possible values for the parameter for all people in the world in similar circumstances;
- the range of a 50% confidence interval is calculated to include include half, 50 out of every 100, of the possible values for the parameter for all people in the world in similar circumstances.

Summarising confidence intervals

The calculation of confidence intervals in epidemiology:

- starts by comparing rates of harm (in this case neuropathy) in a group of people taking a drug on trial to rates of harm among those with different therapy;
- obtains a numerical value for this comparison (a relative risk);
- realises that there are others in the world in similar circumstances and that this result may apply to them as well;
- defines all others in similar circumstances as the *population of interest*, and the group studied as a *sample* of that population;
- knows that samples drawn from the same population will vary in their content and may not be perfect miniatures of the population from which they are drawn;
- hence recognises that the numerical value (the relative risk) obtained from the sample may not have exactly that value in the population of interest;
- calculates a range within which that value (the relative risk) for the population of interest is very likely to lie.

P-values – *another way to account for chance variability between samples*

Finding a relative risk, calculating its 95% confidence interval, then seeing whether this interval includes within it the numerical value 1.0 is one way to determine whether an observed difference between the health outcomes is likely to be due to random sampling variability or not.

There is an alternative way to do this. It is called *hypothesis testing*. Here a research investigator starts by proposing that any difference between groups in the proportion of people with neuropathy is purely a coincidence. He or she argues that if many similarly-treated groups of people had been studied then, taken over all, we'd expect to see no consistent association between treatment and neuropathy. In other words, in the case above, he or she would argue that those taking the drug on trial that showed 1.67 times the rate of neuropathy was just a sample variant; it would be be just as likely that another group similarly treated would have, say, two-thirds the rate of neuropathy of a control group.

This negative stance is called *the null hypothesis*. Then, based on the size of the sample and the nature of the sampling distribution, an equation is used to calculate the probability that chance explains the association, i.e. that it would be reasonable to accept the null hypothesis. This probability is called the P-value. This P-value is compared with an arbitrary standard, usually 0.05. This standard is called the level of statistical significance: if the P-value is less than 0.05, then it would be argued that the null hypothesis was untenable, i.e. that there truly is association between use of the drug on trial and increased rate of neuropathy. The use of an arbitrary standard introduces the potential for error – if we set the level too high then we are likely to reject the null hypothesis too often, and if we set the level too low then the opposite applies. Rejecting the null hypothesis when we should have accepted it is called a Type I error; accepting it when we should have rejected it is called a Type II error.

Summary of hypotheses and P-values

In summary, the starting point of any epidemiological study is a research question, e.g. "Does a new drug cause a different rate of neuropathy than conventional therapy?" This definite proposition, put forward in an attempt to ascertain whether use of the drug is associated with a changed rate of neuropathy, is termed a *hypothesis*[7].

This hypothesis is tested by comparing the incidence of neuropathy in the group taking the drug on trial with another group taking conventional therapy. *Seldom* does this comparison show an outstandingly clear association; more likely the *prima facie* answer is *"maybe"* and this is addressed by statistical analysis.

Because the usual end-purpose of enquiry is to make decisions, the statistical analysis is couched to draw decision from uncertainty – it sets rules to ascribe 'black' or 'white' to a shade of grey. Instead of saying "there's probably an association", we say "there *is* an association" or "there's *no* association" and "the probability of our being wrong is ..."

[7] People who want to sound flash will say "*an* hypothesis". This would make sense if the 'h' was silent.

Typically we accept the positive assertion – that there *is* an association – *only* when there's a probability less than 0.05 (one in twenty) that such association is due to chance.

Formally, hypotheses are expressed in the negative, e.g. "that the taking of this drug on trial does **not** change the rate of neuropathy". This negativity-in-expression underlines the status quo, the presumption of non-association. It indicates that we must obtain an out-of-the-ordinary result in order to rightfully declare that an association exists.
A hypothesis expressed in this negative form is called a *null hypothesis*.

Questions:

1. *What is the difference between the use of a P-value and a 95% confidence interval?*

A P-value is the result of a statistical test of significance; the P-value is a measure of the strength of evidence *against* the null hypothesis. The null hypothesis is typically that there is *no* difference between a pair of means or a pair of proportions. If we find a difference we may calculate the probability that a difference *at least as large as this* could have occurred by chance sampling variability. This is the P-value; if it is low – typically less than 0.05 – we reject chance as the explanation.

In using a P-value we start with the null hypothesis, whereas in using a confidence interval we sort of do it the other way around. There we start with an estimate of a difference or a ratio from a study and calculate a range (called a confidence interval) in which the value of the parameter in the population of interest is likely to be. If that interval includes the null value – 0 for a difference, 1 for a ratio – then we generally decide against the alternative hypothesis in favour of the null. An exception may be if a *very wide* confidence interval included the null value; there we'd attempt to repeat the study using more subjects.

To quote *both P-value and confidence interval* is 'belt and braces' but it's reassuring when their outcomes agree with one another.

2. *Why is the arbitrary standard of significance for a P-value set at 0.05?*

There is an article relevant to this by Gerard Dallal called (curiously) "Why P=0.05?" Its web address is http://www.tufts.edu/~gdallal/p05.htm

There are probably two major influences:
(1) it 'feels' about right – as I shall attempt to explain here with a story about tossing a coin several times;
(2) on the normal curve, two (actually 1.96) standard deviations either side of the mean includes 95% of the area or excludes 5% (0.05) of the area.

My gut feel is that, if two standard deviations either side of the mean of a Normal distribution had instead excluded 4% or 6% of the area, then 0.04 or 0.06 would have been instead chosen as the level of significance.

However, let's say that for the next Federal election, you decide to newly establish a party called "The Wee Pee Point O-Three Party" whose platform was to reduce the level of significance of P-values from 0.05 to 0.03. Now, strange as it may seem, you get the deciding vote in a close Senate result and introduce a private member's bill to achieve your party platform. Some less honourably-behaving members may call you a P-nut or P-leaf and heckle "P's-on-you", but they would *not* be able to show that you were wrong. The 0.05 level is indeed arbitrary but most people think that it's around about right, rather like the idea of three meals a day.

I borrowed a coin-tossing idea from Prof Andrew Forbes, Head of the Statistics Unit in the Monash University School of Public Health and Preventive Medicine. I tossed a coin several times in front of a group of students. I said that I'd call tails each time and see how often I was right.

I tossed it once. Result: heads. The probability of one head by chance in one throw of a fair coin is 0.5. *Students' reaction*: "OK".

I tossed it again. Result: heads. Probability of two heads by chance in two throws of a fair coin is 0.25. *Students' reaction*: "OK".

I tossed it again. Result: heads. Probability of three heads by chance in three throws of a fair coin is 0.125. *Students' reaction*: "Bit sad, eh!"

I tossed it again. Result: heads. Probability of four heads by chance in four throws of a fair coin is 0.0625. *Students' reaction*: "Hey, I reckon there's something suspicious!"

I tossed it again. Result: heads. Probability of five heads by chance in five throws of a fair coin = 0.03125. *Students' reaction*: "There's DEFINITELY something skew about that coin!"

This is, of course, a single example – but, on the face of it, it seems like a realistic human reaction.

So, somewhere between a probability of 0.06 and 0.03, this little group of people made up their mind that the probability of this event was so small that pure chance was no longer accepted by them as the explanation. On that (perhaps flimsy) basis, it seems that a figure of 0.05 is roughly in the 'right ballpark'.

Comment: *P-values can be hard because everything is expressed in the negative!!*

When we are attempting to understand an elusive item of knowledge or to master a skill, it is easier to do this when the words we use are positive rather than negative. For example it's better to start a novice in a factory with "Always push the blue button" rather than "Never press the red button". In general, health messages expressed in terms of a gain are more imperative than those expressed as a loss.

So it is with P-values. If we haven't fully grasped P-values, then to try to talk ourselves through it using negatives, e.g. reject, or even technical negatives, e.g. null-hypothesis, gives us a pretty slippery path. Of course, once we gain an understanding, then we can use the negatives quite fluently – and so we should – but I suggest that they can be mean and awkward guides for a person who is just feeling his or her way in an unfamiliar arena.

So let's escape for a moment from health, and take the situation of Alexis Smart, newly-appointed town planner at the City of Sunny Rises. She notices that of the city's 2000 streets, eight of them have names starting with X as compared with the average rate in cities which is, say, 1 in 5000. She wonders whether such a finding suggests X-ist tendencies among the city's planners or could it be simply a chance variation.

The first and most fundamental thing that Alexis realises is that if you take a sample from any larger body of items, then its characteristics are likely to be a bit different from any other sample that you take. Therefore, she recognises that having 8 in 2000 street names that start with X could simply represent the sorts of ups and downs in numbers that tend to happen across different municipalities.

Alexis then decides to work out how likely it is that you would get eight (or more) street names in her city starting with X just by chance, when you would expect fewer than one. She gets out her statistical software and calculates a probability of 0.00001 that this would occur by chance.

She then faces a decision: "Does this indicate an X-ist tendency or not?" The probability that 8 street names starting with X occurred by chance is very low indeed and so she decides that there is indeed an X-ist tendency in her municipality. Be aware though, that there *is* a very small probability that this actually represents random variation. If that were indeed so (and I cannot tell you whether it is), she has made what is called a Type I error.

I could have used a more technical explanation of this story. It would go like this. Alexis obtained a P-value of 0.00001 and so decided to reject the null hypothesis.

A further question: *What do you mean by bias and why is it important in epidemiology?*

Epidemiology is largely about seeking information on what agent causes what health effects (good or bad). If we had infinite wealth, we could study everyone in the world that comes in contact with an agent of interest (e.g. a nominated drug) and find out what happens to their health as a result of that contact. However, our resources are always limited and so we must make do with just a sample of people who come in contact with that agent. We then find out what happens to the health of people in that sample and then use those results to predict what is likely to happen to all others in the world who may come in contact with that agent – the target population or population of interest. This process of projecting the results of a sample on to the target population in the world-at-large is called extrapolation or generalisation.

We can only fairly generalise when the sample has essentially the same important features as the 'target population'. If the people in the sample differ in any more than minor ways from those in the target population, then the sample cannot be fairly said to represent the target population. It is then said to be biased.

Text books on epidemiology talk about several types of bias. Essentially, it happens in two ways. There is bias in selection of the people for the sample, i.e. they have a different age range or state of health from the target population, or there is a bias in measurement so that information about health or the level of exposure to the agent is missing or recorded inaccurately.

Bias is about whether a sample is fit to make a generalisation from. If you cannot generalise, then the things that you found out about the people you studied apply only to them. It is a story about them but is not particularly useful for the world at large.

Association and causation.

We have referred to measuring the health outcomes in two groups with different exposures. If an observed difference appears very *unlikely to be due to* chance variation, (the sort of random variation that happens between samples drawn from the same population) then the research investigator will argue that the taking of a drug and the side effect of interest are *associated*.

Thus, a study will start with a research question that goes something like: "Will the taking of a particular drug adversely affect the health of people in a particular way?" The research question asks about cause, however the study can provide merely an *association*, i.e. it can tell that particular health effects occur consistently when the drug is taken, but it cannot formally account for that. Epidemiology indicates what happens but it does not (on its own) *account* for that, i.e. it *cannot alone prove* causation.

In the English language the terms *cause* and *effect* correlate. Their mutual relationship is a cliché. We often hear people say: "cause and effect". The Oxford English Dictionary defines *cause* in terms of *effect* and vice versa. Thus it is difficult to disarticulate these linked words in order to interpose the intermediary concept of *association*. Our mindset doesn't easily allow it hence, in all but the most disciplined minds, *association* becomes tainted with the smell of *causation* whether that is appropriate or not.

So, over centuries our language has developed a vocabulary of terms which go with *cause and effect*. There is no such vocabulary that fits with the more abstract notion of *association*. Accordingly we borrow terminology from *cause* and apply it to *association*. Hence the term used to describe the *strength of association* is **relative risk;** yet the very use of the word "*risk*" tends to imply *causation*.

Thus the important distinction between debutante *association* and well-savoured and justified *causation* may easily blur for all but careful and discerning readers.

I have a birthday card that says:

> "Birthdays are good for you. Statistics show that people who have most birthdays live the longest."

Now, I ask you: "Does this indicate that having birthdays truly causes the effect of longevity?!"

Components of causation

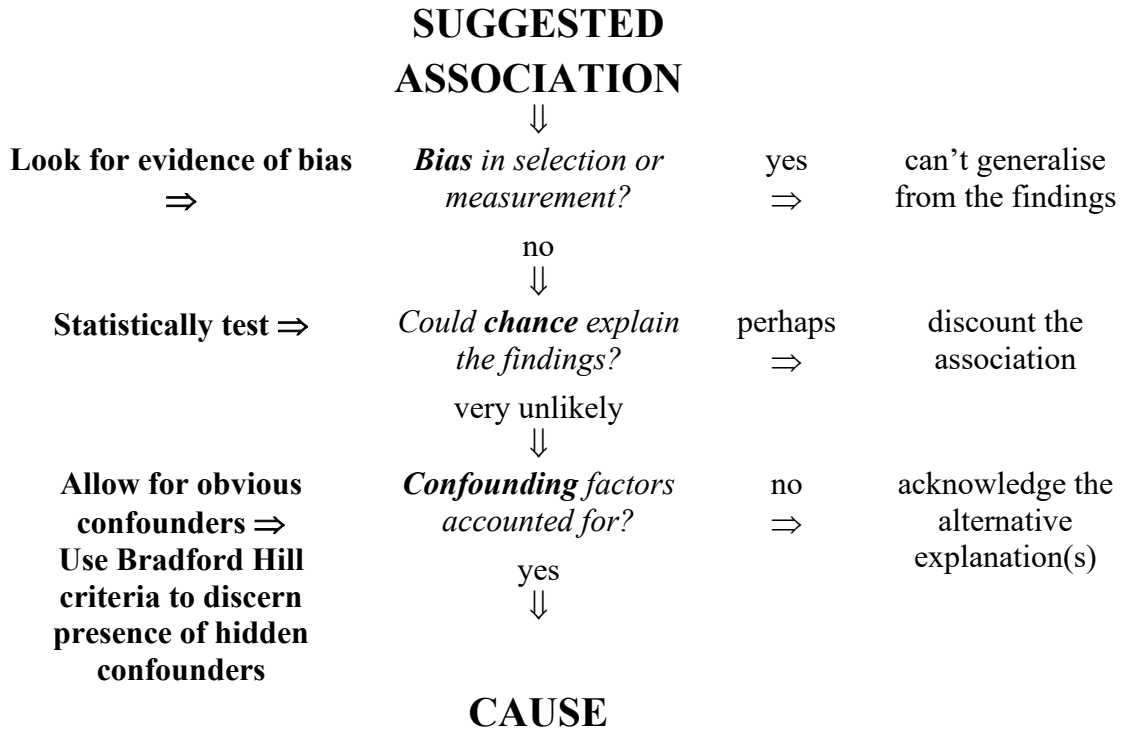
A cause is the factor that brings an effect, a change. Some change may occur only when or after a *particular* factor is present. More often, *no* such specific factor is identified and *enough of several component factors must join* to bring a change.

Let's take an example from toxicology. Absorption of the metal lead into a person's body is *necessary* in order for lead poisoning to occur; lead poisoning *cannot* happen without absorption of lead. However, some little absorption of lead will *not* be enough to bring lead poisoning; *sufficient* dose is required to bring poisoning. *Sufficient* dose inevitably brings poisoning.

Where a cause *inevitably* produces its effect then it is designated a *sufficient cause*. However, most causes that are of interest to population health studies are mere components of sufficient cause; they are *not* sufficient in themselves. If some (not necessarily all) of these component causes are identified and removed, then the sum of the remaining components may well become *insufficient* to bring the disease – or, at least, cause its appearance to be delayed.

Epidemiological studies may identify *sufficient causes* and the *components of sufficient causes,* be these several components acting in parallel (such as cardiac risk factors) or in series (such as the several mutations that occur over a period of time before some cancers start). Repeated studies may reveal several sufficient causes. Any component common to all sufficient causes is designated a *necessary cause*.

The following figure summarises the three steps from *association* to *cause*.



An association between an exposure and a harmful health effect may be explained as:

- exposure *causing* the health effect;
- the health effect *causing* exposure;
- exposure and health effect being related through some third factor (e.g. lung cancer and heavy alcohol consumption are related through smoking rather than through cause and effect). This artefact of association is termed *confounding*.

The middle one of these dot points may, at first sight, seem almost silly. However, in retrospective studies, i.e. case-control studies, it is sometimes difficult to be sure which came first – the exposure itself or the early, sub-clinical stage of a disease of slow onset. For example, does increased intake of lead cause children to have a low IQ or vice versa[8]?

A confounder is something that provides an *alternative explanation* for what seems to be a relationship between exposure and disease. Confounding happens where both exposure and disease are linked by their association with a third factor. It has the potential to bring an error in the *interpretation* of the results of a study.

When a confounding factor is known or likely, then a study may be *designed* to cope with it by:

- randomising - good, but suitable only for intervention studies;
- restricting entry to the study - e.g. eliminating or separately categorising smokers from a study of occupational lung disease;
- matching controls so that potential confounding factors, e.g. age, gender, smoking, are distributed similarly to the cases or exposure group.

However, if study design did not take sufficient account of confounding, then it may be coped with in the analysis stage of a study by:

- stratifying the subjects, e.g. into age-strata, then using multipliers to align the age mix of one group with that of the other – standardisation[9];
- using sophistocated statistical methods such as multivariable analysis.

Sometimes, though, confounders are *not* easily recognised. In that case, we use criteria of causation to help take fair account of their hidden presence.

---

[8] de Silva PE & Christophers AJ. Lead exposure and children's intelligence: do low levels of lead in blood cause mental deficit? *J Paediat Child Health* 1997; 33: 12-17.

[9] Standardisation is a change made to findings that makes them more readily comparable with other facts that refer to similar things. Standardisation may be *direct* or *indirect*. The maths of this is *not* directly relevant to you except to say that:

- *direct standardisation* starts with **rates** of disease (or death) obtained from a study and arranged in five-year or ten-year intervals of age. The age-mix of the population (from a recent census or other relevant, reliable record) is then used to calculate an overall predicted disease rate. This is compared with the overall disease rate obtained in the study;
- *indirect standardisation* starts instead with **numbers** of cases of disease (or death) obtained from a study and arranged in intervals of age. Knowledge of the relevant age-related incidence of that disease (e.g. from national health statistics) is used to calculate an overall number of 'expected' cases. This is compared with the number of cases obtained in the study.

Bradford Hill criteria of causation

We use criteria to try to sort *causal* associations from those which are *not*. The most widely quoted criteria were enunciated by Bradford Hill in 1965[10] and are paraphrased here as eight questions.

- Did the exposure occur before the disease happened?
- How strong is the association, i.e. is the relative risk or odds ratio large?
- Are larger exposures associated with higher rates of disease?
- Is reduction in exposure associated with lower rates of disease?
- Is a plausible mechanism known from, say, animal experiments that is consistent with the natural history and biology of the disease?
- Is the result consistent with other studies, especially different populations using different study designs?
- Is there similarity to other cause-effect relationships?
- Are the findings specific, i.e. does exposure give a narrow, identifiable range of health outcomes?

These various criteria of causation are mutually supportive, overlapping and complementary. By themselves, the individual criteria are *only sometimes* decisive.

An association may be confidently *rejected as causal* only if:

- disease precedes exposure, or
- there is *in*coherence with known facts or
- where enough other well-conducted and relevant studies have found *no* basis for causation.

Strong evidence of a *causal relationship* between an exposure and a disease exists when the association is strong, there is a standard dose-response relationship, or the supposed relationship has been successfully used to *predict* other related causal links.

Sometimes we lean from association toward causation when, after reflection and asking a range of informed others (perhaps in a special closed meeting), we can think of *no other way to explain* the existence of an association.

Ultimately, for two events long separated in time such as some past chemical exposure and the development of cancer, a decision to assert that these are cause and effect is a belief, a judgment. We should ever remain open to continuing good information and be prepared to change our preventive policies and legal precedents accordingly, even when this means stopping action that we once diligently and energetically pursued in the name of prevention.

[10] Bradford Hill A. The environment and disease: association or causation. *Proc Roy Soc Med* 1965; 58: 295-300.

Question

*Why do I say that the decision to assert that 'this exposure **causes** that chronic disease' is a matter of **judgment** rather than fact? Doesn't the application of the Bradford Hill criteria exclude confounders?*

Application of the Bradford Hill criteria assists detection of hidden confounders but does not reliably exclude them. Even if, say, three of the criteria are met, it is still possible that confounding factors explain the association. Thus, in the face of this uncertainty, it often becomes a social (even political) process to accept an association as causal. This particularly relates to causes of cancer where the time-lag between early cell change and diagnosis is so long – years, even decades. If you are too ready to accept an association as causal you'll be chastened as a "Chicken Little" (who thought the sky was falling when an acorn fell on his head). If you're over-cautious to accept an association as causal, you'll be accused of "sitting on your hands while people die". Particularly if the health effect is severe and the preventive remedy is carries a large cost, then it may be a very tough call as to how much evidence is enough to regard an association as causal. Inevitably, a too-hasty decision will drag preventive dollars from useful projects into something that will do little to aid community health. And there are always special-interest groups in the arena that will arm themselves with whatever selected information promotes their particular preventive or compensation interests.

Interpretation of findings – two types of error

In using an arbitrary standard to determine whether an association found between an exposure and disease in a sample of people is statistically significant, a statistical test may be correct in two ways and in error in two ways. This is shown in the Table below. A type I error is saying there's an association when there really isn't. A type II error is saying there's no association when there really is.

Relationship between the results of a statistical test on a sample of people and the actual (usually unknown) situation for the population

	ASSOCIATION TRULY EXISTS (which we never really know for sure)	
	YES	**NO**
STATISTICAL TEST FINDS ASSOCIATION	Positive test result is correct	Type I error
STATISTICAL TEST FINDS *NO* ASSOCIATION	Type II error	Negative test result is correct

There is a trade-off between type I and type II errors. Type I errors are like when a smoke alarm near the kitchen is set off by normal cooking. To continue the analogy, a type II error occurs when a smoke alarm fails to respond to a house fire. The probability of type I errors may be brought to zero by removing the battery from the smoke alarm, but this maximises the risk of type II errors. The appropriate solution is inevitably a compromise between the two types of error[11]. The use of a *95% confidence interval* or *a 'level of statistical significance' for a P-value of 0.05* means that the probability of a Type I error has been set at 0.05. A common setting for the probability of a Type II error is 0.2.

[11] The smoke-alarm analogy was drawn from Gonick L, Smith W. The cartoon guide to statistics. New York: HarperCollins, 1993.

The P-value is compared with the set probability of a Type I error. On the other hand, the power of a study, i.e. the ability to detect an association, difference or real effect when one in fact exists is one minus the probability of a Type II error, i.e. commonly $1 - 0.2 = 0.8$.

Lack of power: when a negative result may be wrong

Conventionally, the probability of getting a particular array of numbers on a table has to be small before we reject chance as their explanation. The stringency of this practice means that a weak association sometimes passes unnoticed, particularly if the number of subjects studied was small. That negative result belies the association; it means that the study lacks sufficient *statistical power*. The form of a power statement is something like:

> "This study is designed to detect a difference of 15% with a probability of 0.8."

Standard textbooks of statistics discuss *power* and ways to calculate the number of subjects required for a study to be of sufficient power. Particularly for a study to assess the effects of an intervention, it would be unethical to set out to perform an under-powered study because you are subjecting individuals to risk without likelihood of a worthwhile result.

QUESTION: How do you work out the sample size, i.e. the number of patients needed for a study?

The sample size depends on four critical factors:

1) the type I error rate (α) that the investigator sets usually at 5%;
2) the type II error rate (β) that the investigator sets usually at 20%;
3) the variability of the data (this variance may be estimated from previous studies or a pilot study);
4) the size of the effect that the investigator is seeking.

It seems *intuitively* easier to work out the sample size for a clinical trial than for a case-control study. In a case-control study where cases are few, then up to four times the number of controls as cases are used in order to boost the power of the study. [It is rare for a worthwhile case-control study to have fewer than 100 subjects.]

Multiple hypotheses: when a positive result isn't necessarily true [also called multiple comparisons]

There is an almost universal practice in epidemiology that makes *statistically significant* findings really less 'significant' than is claimed. This is the practice of making dozens or even hundreds of comparisons between test and control groups on the one set of data.

Because epidemiological studies are costly, investigators compare their groups in regard to various disease outcomes and various types, intensities and periodicity of exposure. In the midst of a plethora of hypothesis tests on the same set of data, chance positive results can be expected.

You should expect to find the misleading effects of multiple hypotheses in the following three common situations:

- where it is clear that an investigator has performed many hypothesis tests on the same set of data; or
- where there is a cluster of cases at a particular location or during a particular period; or
- where an investigator has undertaken a study without a clear purpose but, on the basis of some of their findings, has formulated a hypothesis during analysis of the data rather than at the commencement of the study.[12]

Question: *The probability of a test giving the right answer each time it is used is 0.95.*

(a) *If three such tests were performed, each independent of the other, what is the probability of getting the ALL 3 answers correct?* Answer: $(0.95)^3 = 0.86$

(b) If fifteen such tests were performed, each independent of the other, what is the probability of getting the ALL 15 answers correct? Answer: $(0.95)^{15} = 0.46$

(c) If thirty such tests were performed, each independent of the other, what is the probability of getting the ALL 30 answers correct? Answer: $(0.95)^{30} = 0.21$

(d) If fifty such tests were performed, each independent of the other, what is the probability of getting the ALL 50 answers correct? Answer: $(0.95)^{50} = 0.08$

(e) So, if fifty such tests were performed, each independent of the other, what is the probability of getting at least one wrong answer?. Answer: 1 – 0.08 = 0.92, i.e. a very high probability of at least one wrong answer.

Three common situations where multiple hypotheses occur

(i) Where many questions are asked

Multiplicity happens when the investigator (or group of investigators) starts with a quite broad research question (e.g. that exposure to chemicals brings adverse outcomes to pregnancy). This splinters into a bundle of subordinate hypotheses based, say, on occupation, type of chemical, nature of health outcome, each of which the investigator tests individually. In this situation, the investigator should rightfully signal to the reader the possibility of some associations occurring by chance. A naïve or artful investigator will fail to do this.

(ii) Clusters (= multiple hypotheses by stealth)

Let's take an example: three children in a particular suburban street get leukaemia in a particular year.

By knowing the rates of leukaemia and the number of children in the street, it seems at first glance that three cases of leukaemia in one year in this population is a very unlikely event. Residents then hypothesise: "Could this occurrence be due to a local environmental toxin?"

However, leukaemia is just one cancer of many, and this is just one street among many in one year among many. In a large city, there is a moderate probability of getting a cluster

[12] This is akin to an archer entering a forest, shooting an arrow without specific aim, hitting a tree and then asserting "that's the tree I meant to hit". Any of the nearby trees could have been hit by the arrow - the late nomination of one convenient tree cannot get away from the large clutch of other possibilities.

of three cases of cancer among children in one of its ten thousand streets in any year simply by chance.

One may liken the chances of the situation to an investigator observing that one Saturday night all cars in serious accidents in Melbourne had a double 7 in their number plate, and then proceeding to postulate that the road toll could be reduced by removing from the road all cars with such number plates. A person seriously interested in accident prevention would recognise that this was just one of many Saturday nights and that 'double 7' was just one of many 'special' number combinations; e.g. it could have been a double 6, a 123, or whatever. Any hypothesis that is made up ***after*** *an event* is *inevitably* one among many possibles. On the other hand, if a cluster of leukaemias or double 7s was *predicted and then occurred*, the finding is far more arresting.

The occurrence of clusters alert us to possibilities in the same way as the boy that cried "Wolf!" in the children's tale[13]. In fact, most clusters are coincidences. Unnervingly, however, the people who raise the alert will usually remind us that on one occasion the wolf in the children's tale really did come.

(iii) Starting a study without a clear initial purpose

Charles Dickens' classic, *David Copperfield*, featured a character, Mr Wilkins Micawber, whose schemes were ill-conceived yet who remained ever-optimistic that something would "turn up".

In parallel vein, some investigators do *not* plan their statistical analyses before their data are collected. Instead of formulating precise hypotheses, they collect data then use computer packages to comprehensively analyse it in the hope of turning up associations. This exploratory type of analysis is legitimate ***only*** if the authors report that this is what they've done [which too often they don't]. It is, in a sense, a means of artificially generating clusters. Further data should then be collected to test the hypothesis generated by the exploratory analysis. On the rare occasions that a lot of data are available, then one part may be used to explore and the rest used to test the hypothesis generated by the exploration.

Whenever a hypothesis is formulated *after* the data are collected, then other hypotheses will inevitably lurk although the investigator may choose to ignore them.

The problem with multiple hypotheses

When multiple hypotheses are applied to a single set of data, their very multiplicity degrades the strength of the conclusions that can be drawn from the statistical analysis unless special allowance is made. P-values (and levels of confidence) lose their true meaning in an *absolute* sense[14] although *not* in a *relative* sense; P-values drawn from the same set of data may still be usefully compared with *each other*.

---

[13] *The boy that cried wolf* is one of the fables credited to Aesop, a story-teller who lived in ancient Greece.

[14] The definition of a *P-value* is the probability *due to chance* of observing a result as extreme or more extreme than the one actually observed. When *more than one* hypothesis is made on the data then the probability due to chance of some extreme result becomes greater than any of the P-values calculated individually. When very many hypotheses are made, it becomes quite likely that a few extreme results will occur by chance.

The existence of multiple hypotheses can be a *major* source of misinformation in epidemiology because repeated generation of P-values will bring one or more that falls below the arbitrary standard of 0.05 simply by chance, so generating false associations between exposures and diseases. If such falsehood is something scary but plausible – like an association between an occupational exposure and an increase in cancer – then the impact of its revelation leads to anxieties, enters compensation courtrooms and pushes policy-makers to spend in one area and so lose opportunity elsewhere. Particularly if the association is scary or attractive to litigants, social forces cause the moral onus to shift to those who would argue *against* the association to prove it is *not* so. Of course such rebuttal is impossible – because no-one can prove such a negative – so the 'smell' lingers on. Health policy-makers seem inevitably pushed to act – especially when the findings are roundly and noisily touted as "significant'; and probably, the more of these that there are, the more likely is such wasteful action.

Coping with multiple hypotheses

There are mathematical ways of coping with multiple hypotheses. The problem is that these ways all serve to diminish the statistical *power* of a study, i.e. the ability of a study to find a difference or real effect when one in fact exists. So it becomes harder for an investigator to obtain a positive result unless a larger sample size is used. To do this adds cost and delay.

The prominent American epidemiologist, Rothman,[15] in 1990 stated his *opposition to mathematical adjustment for multiple comparisons* (multiple hypotheses). He said that such adjustment suppresses the opportunity to accept and later explore some 'odd' finding that may lead to exciting new knowledge of causation. He considers that science could be the poorer for this. He styles adjustment for multiple hypotheses as a 'penalty for peeking' – a discouragement to question too widely because investigators will be penalised for so-doing by having to put their findings through a statistical sieve that is tightened in rough proportion to the number of questions asked. In other words, Rothman fears that mathematical adjustment for multiple hypotheses will thin out the 'byplay'. However, the problem is now that many authors seize the byplay and pretend it's the main game. They want to both:

- explore widely;
- proclaim *without humility* the significance of *any* positive finding.

Some investigators will test, say, eighty hypotheses, get two low P-values, and then exclaim "Gee whiz!" when, instead, a realistic response would be "Ho Hum". They will promenade "It's significant!" with nigh as much fervour as a new father proclaims, "It's a girl!" or "It's a boy!" And, for most of us, the term *statistically significant* holds power that is augmented by the very fact that there are *numbers* proffered by the authors to shore it up.[16] Also, with positive results, authors typically recommend that "further research is needed".

[15]Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology 1990; 1: 43-46.

[16] Many would-be critics are nervous to argue about the quality of numbers - there is *no* escape from embarrassment if one makes a critical comment that proves foolish. With words or ideas one may at least feign misunderstanding; with numbers, you are usually either right or wrong.

The issue then has potential to become a '*tar baby*'[17] because, although some of the false positives will be later shown up for what they are, new ones will be generated. If you keep asking questions, you'll get stuck with the answers!

QUESTION

Why is it a problem to make lots of hypotheses on a single set of research data?

A hypothesis test using a 95% confidence interval or setting a level of significance for a P-value of 0.05 is using a method that gives the correct answer 95 times out of 100. So an incorrect answer will appear, on rough average, about 5 times in 100 hypotheses. That's OK – it's inevitable. However, where very many hypotheses are tested, an honest investigator will signal the risk of false findings to the reader or perhaps attempt to make some adjustment to the level of significance, e.g. a P-value of 0.01 (≡ 99% confidence interval).

Limits of epidemiology

Very low exposures

Airborne asbestos fibres are the best-known of occupational carcinogens. One issue is how low an exposure must be before the added risk of mesothelioma becomes negligible.

Compensation has been received by workers with mesothelioma whose past asbestos exposure has appeared to be minor. Of course, it is impossible to tell in retrospect whether the minor asbestos exposure was the culprit and equally impossible to prove that it wasn't. The courts, faced with this evidential vacuum, tend to decide on the basis of "when in doubt, favour the injured". This is a humanitarian expedient; the reality is we simply don't know.

Epidemiology can never clarify the lower limits of risk because, in trying to tell the difference between health outcomes with *minimal* exposure and health outcomes with *no* exposure, the epidemiologist always faces an unforgiving mathematical operation – dividing by the square root of *n*. Where *n*, the number of cases is small, as inevitably it is when exposures are low, it becomes impossible to show that minimal exposure differs in any health-related way from no exposure.

Asbestos is a 'celebrity' hazard. Years of adverse publicity has given it horror symbolism, so that many frightened people consider that disaster starts with one fibre – as if having a tiny bit of asbestos is like having a tiny bit of mesothelioma. Because we have no good evidence for what happens at low levels of exposure, asbestos tends to be treated in a qualitative way, i.e. "better out than in", rather than the quantitative approach of sufficiently lowering the extent of exposure. Measurements of asbestos in air are done but results are essentially treated on a dichotomous basis – below the level of detection or above it.

[17] The tar baby features in the late 19th century 'Uncle Remus' children's stories by Joel Chandler Harris where animals were personified and named "Brer Rabbit", "Brer Fox" etc. ('Brer' or 'Br'er' is a contraction of 'Brother'). Brer Rabbit encountered a child-shaped sculpture made of sticky tar and hit it with his paw. This stuck in the tar. As he tried to free his paw, his leg stuck. As he tried to free his leg, the other paw stuck

When the health outcome is ill-defined

Disease (or *injury*) is about pathological change whereas *illness* is about the reporting of symptoms and a change to behaviour. There is substantial but not complete overlap between disease and illness. *Disease happens; falling ill can be a choice.* There is pathology without illness (e.g. degenerative changes to joints) and sometimes illness without recognisable or comparable pathology (e.g. some arm or spine pains).

Sometimes, a condition brought by a person to a health practitioner cannot be diagnosed, even after a careful history, physical examination and a handful of special investigations. The person is clearly distressed but, with symptoms only, the practitioner finds it very difficult to define his or her illness. In the context of occupational health, it takes an edge to it when the person asks the practitioner for support in their making of a compensation claim. If the practitioner has no proper diagnosis then he or she can neither predict the natural course of the patient's illness nor choose the most effective intervention. Nor can the practitioner advise on appropriate prevention, nor feel confident about return to work. If you don't know what's there, you cannot tell when it's gone. It's like feeling that you are being stalked whilst you are walking at night or in a fog.

Counting occupational cancers

Reliably establishing the causes of any one cancer-type is very difficult because cancer proceeds from a combination of events affecting a cell in the body. This series of events occurs over a period of years or decades, and causal factors seldom lay their fingerprint on the cancer histology. So disputes are commonplace about what cancers are occupational and what are not. Protagonists bring their individual values to address this uncertainty; and key decision-makers, such as courts or standards-setting authorities, have different perspectives depending whether their main preoccupation is a humanitarian care for the afflicted or practical prevention of injury and disease. Without a way of classing cancers as occupational or not, it is difficult to make a reliable count.

There is only limited information on the type and intensity of occupational *exposures* to known and suspected human carcinogens in Australia. Some overseas countries do better, e.g. Finland. However, like medical practice itself, constant and rapid technological changes to workplaces make nigh impossible the task of finding out in broad-scale who is exposed to what. This lack of exposure data adds uncertainty to the tally of cancers characterised as occupational. And, as previously stated, one area of enduring dispute is the degree to which exposure to a carcinogen must be lowered before the risk of exposure becomes too small to be of concern.

In May 2007, the Australian media made an initial suggestion that a cluster of breast cancers in a radio studio in Brisbane may have been due to non-ionising radiation. The truth of such an allegation is extremely difficult to decide in retrospect, although courts are forced to attempt it because compensation is sourced on cause.

Conclusion

Epidemiology is a powerful method for finding out about whether exposures can increase the risk of harm to health. Its terminology, and particularly the mathematical aspects, makes it hard for many to understand. It also has its limitations and potential to mislead.

APPENDIX

This appendix refers to two further issues.

- Randomisation and 'intention to treat' analysis;
- Absolute risk reduction.
- Screening and diagnostic tests.

Randomisation and intention to treat analysis

The terms 'intention to treat analysis' and 'per protocol analysis' refer to randomised trials. Let's imagine a trial where the test group receives a new drug and the control group receives conventional therapy. The term *randomised* means that anyone who is enrolled in the trial has a 50:50 chance of being in either the test group or the control group. This is done in an attempt to make the general demographic features of each group very similar. If the groups are indeed similar then, if the outcome for the test group differs from the outcome for the control group, we can reasonably argue that this difference has resulted from a difference in efficacy between the new drug and the conventional therapy. Randomisation is of *central importance* to being able to sustain this line of argument.

Intention to treat analysis holds that randomisation is so important that it should not be violated. Accordingly, at the end of the trial, the outcome for all those originally designated as members of the test group is compared with the outcome for all those originally designated as members of the control group. This is regardless of whether or not the members of each group completed the course of treatment as planned. People who advocate intention to treat analysis argue that this method of analysis gives a better indication of how the new drug will perform when commercially released because there will always be patients who do not complete their course of treatment. The disadvantage is that, by including patients who don't give the new drug a 'fair go', it tends to narrow the gap measured between the effect of the new drug and the effect of conventional therapy. In effect, use of intention to treat analysis reduces selection bias but increases measurement bias.

Some people disagree with the use of intention to treat analysis. They argue that it is silly to count patients who have not adhered to the treatment protocol among those that have. The argument is that you get a better comparison of the actual effect of the new drug compared with its conventional counterpart if you count only those patients who properly adhered to the treatment protocol. The others should be removed from consideration. If only a small proportion are removed from consideration, this will not greatly affect randomisation. This is called *per protocol* analysis. The people who argue for per protocol analysis are prepared to accept some selection bias in order to reduce measurement bias.

Top quality journals insist on intention to treat analysis. But there is no reason why a per protocol analysis could not also be included.

Absolute risk reduction (compared with relative risk)

Imagine a medical condition that, if left untreated, causes complications. An example would be high blood pressure that, if left untreated, will increase a person's risk of stroke or kidney disease.

There is already treatment for this medical condition but we have a new drug that we hope will decrease the risk of complications of this condition. We put the new drug on trial by comparing the rate of complications among a group taking this drug with the rate of complications among a group taking standard (or conventional) treatment. Let's specify a period of follow-up of one year.

Now, let's put in some figures. Say that a group of 250 people were treated with the drug on trial and, among these, there were 15 who suffered complications. This is a rate of 15/250 or 6%/year, i.e. the incidence of complications was 6 per hundred people per year.

A comparison group of 240 people were given conventional treatment and, among these, there were 30 who suffered complications. This is a rate of 30/240 = 12.5%/year, the incidence of complications was 12.5 per hundred people per year.

A *relative risk* simply compares these two numbers as a ratio, i.e. 6/12.5 = 0.48. This is interpreted as "the rate of complications in the group that took the drug on trial is just below half the rate of complications of those taking conventional treatment". It may also be expressed as a *relative risk reduction*, 1 – relative risk, i.e. 1 – 0.48 = 0.52. This may be interpreted as "taking the drug on trial more than halves the rate of complications compared with conventional treatment". Because relative risk (or relative risk reduction) is simply a ratio, it has no units.

The **absolute risk reduction** is obtained by *subtracting* the incidence of complications for those taking the drug on trial with the incidence of complications for those taking conventional treatment, thus: 12.5%/year – 6%/year = 6.5%/year. This may be interpreted as "the difference in incidence rate between those taking the drug on trial and those taking conventional treatment is 6.5% per year". Expressed another way, you could say: "If 100 people took the drug on trial, complications would be expected in 6 of those people in a year. If 100 people took conventional therapy, complications would be expected in 12.5 of those people in a year. The difference in complication rate is 6.5 people in every 100 – in favour of the drug on trial". (Of course, it's hard to imagine half a person, so you could say instead: "If 200 people took the drug on trial, complications would be expected in 12 of those people in a year. If 200 people took conventional therapy, complications would be expected in 25 of those people in a year. The difference in complication rate is 13 people in every 200 – in favour of the drug on trial".)

The absolute risk reduction retains the units of 'cases per hundred people per year'. In other words, the risk reduction is expressed in terms of the total number of people involved. The reason for expressing incidence in terms of 'per hundred' is because the test group (of 250) and the control group (of 240) differed a little in size. Expressing it as 'per hundred' makes for a fairer comparison.

Now let's imagine a much lower rate of complications. Let's say that for the drug on trial the incidence of complications was 0.60% and for conventional treatment it was 1.25%.

The relative risk remains at 0.60/1.25 = 0.48 (with the relative risk reduction again equal to 0.52).

However, the absolute risk reduction is 1.25% – 0.6% = 0.65%/year. The *relative risk reduction* remains impressive, but the *absolute risk reduction* is minor. In this case, because the complication rate is so low, the quite dramatic effect of the drug brings benefit to far fewer people. Thus, if the drug on trial was a whole lot more expensive, it would probably not be favoured over conventional treatment.

Number needed to treat

Another way to think the absolute risk reduction, the *ARR*, is:

$$\frac{\text{the number of complications saved by use of the new drug}}{\text{100 people}}$$

In the case of a 6.5% absolute risk reduction:

$$\frac{\text{6.5 complications saved by use of the new drug}}{\text{100 people}}$$

If we turn this fraction on its head, i.e. $\frac{1}{ARR}$, we get:

$$\frac{\text{100 people}}{\text{6.5 complications saved by use of the new drug}}$$

Since $\frac{100}{6.5} = 15$, this inverted fraction can be expressed:

$$\frac{\text{15 people}}{\text{each complication saved by use of the new drug}}$$

So, on average, *for every 15 patients treated with the new drug*, there will be one fewer complication than had these patients been treated with conventional therapy.

This is known as the **number needed to treat**. It is $\frac{1}{\text{absolute risk reduction}}$ when the absolute risk reduction is expressed as a decimal fraction (*not* as a percentage).

What happens if the rate of complications is low?

Now let's imagine a much lower rate of complications. Let's say that for the drug on trial the incidence of complications was 0.60% and for conventional treatment it was 1.25%. The relative risk remains at 0.60/1.25 = 0.48 (with the relative risk reduction again equal to 0.52).

However, the **absolute risk reduction** is 1.25% – 0.6% = 0.65%, i.e. 0.0065. The *relative risk reduction* remains impressive, but the *absolute risk reduction* is minor.

Here, the **number needed to treat** would be $\frac{1}{0.0065}$ i.e. 154, i.e., on average, you'd need 154 patients treated to be treated with the new drug in order to get one fewer with a complication than had these patients been treated with conventional therapy. Because the complication rate is so low, the drug brings benefit to far fewer people.

Screening and diagnostic tests

This topic is best introduced by presenting examples of questions. The two questions were provided to me and I don't know their source. Question 1 refers to screening. After this, I explain likelihood ratio. Then Question 2 refers to a diagnostic test.

First off, let's define two terms that commonly appear in these tests.

Sensitivity is the probability that a person who actually has the disease of interest will have a positive (abnormal) test result.

Specificity is the probability that a person who does *not* have the disease will have a negative (normal) test result. A test is considered highly specific if it is positive for only a very small proportion of those *without* the disease.

Question 1

You are trying to decide whether to institute a screening program for type 2 diabetes. You find data indicating that a HbA_{1c} of 6.4% or greater has a sensitivity of 40% and a specificity of 80%. You estimate the prevalence of undiagnosed type 2 diabetes in your community is 20%.

What will be the ratio of true positives to false positives if you institute a screening program?

Response to Question 1

I shall respond to this question by providing a worked response. In responding to any question about screening or diagnostic tests, I recommend that you give yourself some practice at quickly scribbling down a 2 × 2 table.

Imagine you have 100 people. From the question, 20%, i.e. 20 will have undiagnosed type 2 diabetes. Of these the test with its sensitivity of 40% will pick up 40% of 20 = 8.

Again from the question, 80%, i.e. 80 will *not* have type 2 diabetes. Of these, the test with its specificity of 80% will show 80% of 80 = 64 as truly negative. Therefore, there will be 80 – 64 = 16 as falsely positive.

You have 8 true positives and 16 false positives, so the **answer to the question is 1:2.**

On the next page, this is laid out on a 2 × 2 table. I have labelled the four boxes of the 2 × 2 table in a standard way – *a*, *b*, *c* and *d*. For calculations about a screening or diagnostic test, I have added boxes for sub-totals on the bottom and right-hand edge, and a box for the grand total in the bottom right-hand corner.

	HAVE DIABETES	DON'T HAVE DIABETES	TOTAL
TEST POSITIVE	*a* 40% of 20 = 8	*b* 80 – 64 = 16	*a* + *b*
TEST NEGATIVE	*c* 20 – 8 = 12	*d* 80% of 80 = 64	*c* + *d*
	a + *c* 20	*b* + *d* 80	*a* + *b* + *c* + *d* 100

The contents of the boxes on my table may be talked about in the following way:

Box ***a***: True positive test results, i.e. people with the disease and a positive test.

Box ***b***: False positive test results, i.e. people without the disease but with a positive test.

Box ***c***: False negative test results, i.e. people with the disease but with a negative test.

Box ***d***: True negative test results, i.e. people without the disease and with a negative test.

The contents of the boxes containing *sub-totals* may be talked about in the following way:

Sub-total box ***a*** + ***c***: All people with the disease.

Sub-total box ***b*** + ***d***: All people that do *not* have the disease.

Sub-total box ***a*** + ***b***: All people with a positive test result.

Sub-total box ***c*** + ***d***: All people with a negative test result.

Grand total box ***a*** + ***b*** + ***c*** + ***d***: Everyone involved.

Estimated prevalence of the disease in the population to be tested, i.e. *pre-test probability*:

$$\text{i.e. } \textit{pre-test probability} = \frac{\text{all those with the disease}}{\text{everyone involved}} = \frac{a+c}{a+b+c+d}$$

Sensitivity, i.e. proportion of those with disease that show a positive test result:

$$\text{i.e. } \textit{sensitivity} = \frac{\text{those with the disease and positive test results}}{\text{all those with the disease}} = \frac{a}{a+c}$$

Specificity, i.e. proportion of those *without* disease that show a negative test result:

$$\text{i.e. } \textit{specificity} = \frac{\text{those without the disease and with negative test results}}{\text{all those that do not have the disease}} = \frac{d}{b+d}$$

Sensitivity and *specificity* are properties of *the test itself*.

Positive predictive value, i.e. the predictive value of a positive test result, i.e. proportion of those with a positive test result that actually have the disease:

$$\text{i.e. } \textit{positive predictive value} = \frac{\text{those with the disease and positive test results}}{\text{all those with positive test results}} = \frac{a}{a+b}$$

A *negative predictive value*, i.e. the predictive value of a negative test result, may also be calculated.

Positive predictive value (or post-test probability) is determined by the properties of the test **together with** the likely prevalence of the disease in the group that is tested, or your estimation of the probability that your patient has the disease before you conduct the test.

There is a way to combine sensitivity and specificity by using a term called *likelihood ratio.* There are two likelihood ratios – *positive* and *negative*.

The *positive likelihood ratio answers* the following question: "How many times more likely is a *positive* test in a person *with* the disease than in a person *without* the disease?" So, if the test has the property of being three times more likely to show positive in a person with the disease than in a person without the disease, then the positive likelihood ratio of this test for this disease is 3. A test with no ability to discern a person with the disease from one who doesn't would have a likelihood ratio of 1.

So, going back to this 2 × 2 table, the positive likelihood ratio of this test is:

$$\frac{\text{the proportion of people } \textit{with} \text{ the disease that have a } \textit{positive} \text{ test result (i.e.the sensitivity)}}{\text{the proportion of people } \textit{without} \text{ the disease that have a } \textit{positive} \text{ test result } (1-\text{specificity})}$$

$$= \frac{a}{a+c} \text{ divided by } \frac{b}{b+d} = \frac{8}{20} \text{ divided by } \frac{16}{80} = 2.$$

A positive likelihood ratio of 2 is not strong, i.e. a positive test result is a fairly weak indicator of whether the person has the disease or not. This is largely because the specificity is quite a bit less than 1, i.e. a specificity of only 0.8 (80%) is of limited value for diagnosis (although sometimes it's the best you can get).

On the other hand, the negative likelihood ratio answers the following question: "How many times more likely is a *negative* test in a person *with* the disease than in a person *without* the disease?" So, if the test has the property of being half as likely to show positive in a person with the disease than in a person without the disease, then the negative likelihood ratio of this test for this disease is ½ or 0.5.

Going back to this 2 × 2 table, the *negative* likelihood ratio of this test is:

$$\frac{\text{the proportion of people } \textit{with} \text{ the disease that have a } \textit{negative} \text{ test result } (1- \text{ sensitivity})}{\text{the proportion of people } \textit{without} \text{ the disease that have a } \textit{negative} \text{ test result (specificity)}}$$

$$= \frac{c}{a+c} \text{ divided by } \frac{d}{b+d} = \frac{12}{20} \text{ divided by } \frac{64}{80} = 0.75.$$

A negative likelihood ratio of 0.75 is weak indeed, i.e. gaining a negative result with *this* test is *not* a helpful way to tell whether a person is clear of the disease. This is largely because the sensitivity is a bare 40%. In general, for a negative test result to be useful in *excluding* disease, the negative likelihood ratio of the test must be closer to 0 than to 1.

Likelihood ratio and using odds

One essential life skill is to build new understandings upon what we already know. A medical consultation is exactly like that. Seeing a patient enter the room – his or her age, gender, ethnicity, facies, colour, mobility, breathing – will immediately make some diagnoses more likely than others. After history-taking, a shorter list of probable diagnoses will appear, and physical examination will refine this further.

The process of medical reasoning is partly about reducing uncertainty – adjusting earlier probability based on what has been newly ascertained by observing a patient, by history, by physical examination. Doctors do this informally – without applying strictly numerical methods. However, for the purposes of explanation, the thought-process may be demonstrated more explicitly by using numbers.

In the diagnostic situation, the manipulation of numbers becomes easier if, rather than probability, we use a related term called *odds*. Most people are familiar with use of odds at a racetrack which are the odds *against* a horse winning. Thus, in the view of a bookmaker, a horse at 25**:**1 is quite *un*likely to win. Yet, in the process of diagnosis, we use odds *in favour* rather than odds against. So the odds of "25**:**1 against" transforms into $\frac{1}{25}$**:**1 or "0.04**:**1 in favour". This reversal introduces fractions which makes the concept a little less vivid. Also, when diagnosticians refer to odds, the "**:**1" is usually omitted. Hence, odds of 0.04**:**1 are spoken of simply as 0.04. The absence of the "**:**1" gives the odds a similar appearance to probability, creating potential for confusion.

Odds are related to probability, p, in the following way: odds = $\frac{p}{1-p}$.

The table offers equivalences:

probability of diagnosis:	1 in 100 or 0.01	1 in 20 or 0.05	1 in 10 or 0.1	1 in 5 or 0.2	1 in 3 or 0.33	1 in 2 or 0.5
equivalent odds in favour of diagnosis	0.01:1 or simply 0.01	0.05:1 or simply 0.05	0.11:1 or simply 0.11	0.25:1 or simply 0.25	0.50:1 or simply 0.50	1:1 or simply 1

You can see that, *below 10%*, the odds are virtually the same as the probability so, effectively, when the probability is *low*, probability and odds can be used interchangeably.

Let's say that, based on your patient's appearance and history, you estimate the probability of one diagnosis-of-interest to be about 20% (0.2), i.e. odds of 0.25. You then conduct a physical examination. You observe a particular set of physical signs.

From a respected source, you ascertain that this set of physical signs is about three times more likely to occur in any person with your diagnosis-of-interest than in a healthy person that does not have that diagnosis. This is a property of *this* set of physical signs in relation to *that* diagnosis. They are independent of you and your patient, i.e. they are, in general, the same regardless of whichever patient they are applied to. As stated on the previous page, this characteristic is given a special name – *positive likelihood ratio*.

So how does the occurrence of this set of physical signs in your patient alter your odds of 0.25 in favour of your diagnosis-of-interest – the odds you had estimated prior to conducting a physical examination of your patient. Well, you simply multiply the prior

odds by the likelihood ratio to calculate your updated odds, thus: 0.25 × 3 = 0.75. So now, your updated odds in favour of the diagnosis-of-interest are 0.75**:**1.

Should you wish, these updated odds may be converted back to a probability:

updated odds	0.25	0.33	0.50	0.75	1	2	3	4	5	6
equivalent probability	0.2 20%	0.25 25%	0.33 33%	0.43 43%	0.50 50%	0.67 67%	0.75 75%	0.80 80%	0.83 83%	0.86 86%

Probability = $\frac{\text{odds}}{1+\text{odds}}$. Odds are able to have a value greater than 1 whereas probability, of course, has values only between 0 and 1.

In this case, odds of 0.75 equate to a probability of 0.43.

You then take another step. You arrange a laboratory test for your patient. From your medical website, you are aware that this test has the following properties:

- The test has the property that a *positive* test result is 8× more likely in people *with* the diagnosis than in a same-sized group of people *without* the diagnosis. Put another way, people with the disease are 8× as likely to show a positive test result as people without the disease. This figure of 8 is the "likelihood ratio of a positive test result" or *positive likelihood ratio*, for short (abbreviated *L*+);
- The test also has the property that a *negative* test result is 3× more likely in people *without* the diagnosis than in a same-sized group of people *with* the diagnosis. Put another way, people with the disease are one-third (0.33) as likely to show a negative test result as people without the disease. This figure of 0.33 is the "likelihood ratio of a negative test result" or *negative likelihood ratio* for short (abbreviated *L*–).

Let's take two scenarios in turn. Firstly, let's say that your patient has a *positive* test result. In this case, your prior odds (given the presence of the set of signs on physical examination) are 0.75. The likelihood ratio is 8. Multiplying these prior odds by 8, i.e. 0.75 × 8 = 6. Having odds in favour of a diagnosis of 6 (i.e. 6**:**1) equates to a probability of around 0.86 or 86% that your patient has this condition.

On the other hand, let's say that your patient has a *negative* test result. Your prior odds are 0.75 and the likelihood ratio is now 0.33. Multiplying these prior odds by 0.33, i.e. 0.75 × 0.33 = 0.25, i.e. you are back to where you were before you conducted the physical examination. Having odds in favour of a diagnosis of 0.25 (i.e. 0.25**:**1) equates to a probability of around 0.2 or 20% that your patient has this condition.

A whole string of likelihood ratios may be multiplied together to give the odds of a particular diagnosis. Some of these may be positive, some negative. For example, let's say that with another patient, you estimate from her age, nationality and presenting symptom that the odds of a particular diagnosis is around 0.20. Following this, you proceed further with history, examination and test results and discover the following which I shall number in the order that you find them:

1 presence of symptoms with a positive likelihood ratio of 2, i.e. $L_1+ = 2$
2 absence of *other* symptoms with a negative likelihood ratio of 0.75, i.e. $L_2– = 0.75$
3 presence of physical examination findings with a positive likelihood ratio of 2.5, i.e. $L_3+ = 2.5$
4 a positive test result with a positive likelihood ratio of 5.0, i.e. $L_4+ = 3.5$
5 a negative test result with a negative likelihood ratio of 0.65, i.e. $L_5– = 0.65$
6 a positive test result with a positive likelihood ratio of 4.3, i.e. $L_6+ = 4.3$.

From this, the odds that your patient has this diagnosis $= 0.20 \times L_1 \times L_2 \times L_3 \times L_4 \times L_5 \times L_6$
$= 0.20 \times 2 \times 0.75 \times 2.5 \times 3.5 \times 0.65 \times 4.3 = 7.3$

which, using the formula, $\text{probability} = \frac{\text{odds}}{1 + \text{odds}}$, is a probability of 0.88 (88%)

There are few published figures for the likelihood ratio of *symptom*s so the figure that a doctor estimates for their positive or negative likelihood ratios will be based on his or her own experience. What I've attempted to show is a simulated example of the reasoning behind reaching a diagnosis.

Cautions

Here, numbers have been used to illustrate the thought-process behind diagnosis. However, one great problem of using numbers like this is that it gives a false sense of precision. Please be aware of this, *be very aware*!

To start with, your prior estimation of the probability that your patient has this diagnosis is inevitably an educated guess – an approximation to the truth. You may put it at 20%, but you would be unlikely to contradict stridently a colleague that chose instead to make it 15% or 25%.

Secondly, the likelihood ratio for many physical signs is not known. Perhaps no good-enough sized study has been done or perhaps there is dispute about the 'gold standard' for diagnosis. For example, what constitutes the 'gold standard' of diagnosis of carpal tunnel syndrome? Is it the constellation of examination findings plus nerve conduction studies, or is it relief of symptoms following relevant surgery?

Many quoted likelihood ratios have wide 95% confidence intervals so, in cases like that, it is misleading to quote *precise* odds or probabilities of diagnosis for your patient after having used such an imprecise likelihood ratio in a multiplication sum.

What I am saying here is simply *be cautious*. Recognise the imprecision of using prior probabilities and likelihood ratios. Use the numbers as an aid to reasoning rather than as a stentorian driver to action.

Also, have a feel for the threshold of odds or probability above which you will decide to treat. If, say, your clinical findings have already delivered odds of around 6 (a probability of 0.86) that your patient has a particular diagnosis, then ask yourself whether you really need a confirmatory test with a positive likelihood ratio of, say, 5? This will increase the odds in favour of the disease from 6 up to 6×5, i.e. 30, (a probability of 0.97). In most circumstances, it is unlikely that changing odds from 'very high' to 'extremely high' will make a difference to your clinical decision to treat.

And realise that a test with a likelihood ratio of 1 (or close to 1) for the situation at hand will afford ***no*** diagnostic value. The further away from 1 is the likelihood ratio – whether

up or down – the better value a test will offer to assist you to make or, respectively, refute a provisional diagnosis.

When a test may be used in several different diagnostic quests, be aware that the test will have a different likelihood ratio for each diagnosis for which the test is used. For example, the likelihood ratio of a positive test result for serum bilirubin will differ depending whether the test is used in relation to haemolysis or liver cell disease. And, as a clinician, you will need to decide what constitutes "positive" – it is not necessarily what falls to one side of the arbitrary line that delimits the edge of the normal range. That is well-known to occupational physicians who commonly accept a higher blood level of, say, a metal in a work-exposed person than is recognised as the upper limit of the normal range for the population at large.

Finally, you may happen to be more familiar with the terms *sensitivity* and *specificity* than with *likelihood ratio*. Recognise that likelihood ratio is a way of combining these two terms. Using the abbreviation *Sn* for sensitivity and *Sp* for sensitivity, the likelihood ratio of a positive test result ($L+$) is $\frac{Sn}{1-Sp}$, and the likelihood ratio of a negative test result ($L-$) is $\frac{1-Sn}{Sp}$.

What happens when you use a test for screening?

When a test is used for diagnosis, the probability of that diagnosis is often reasonably high, e.g. 0.1 or more. However, when a population is screened, the likely prevalence of the disease in the population being tested is often very low.

Let's take a situation where the prevalence of the disease in the population to be screened happens to be around 1%, i.e. the probability is 0.01 that a person selected at random from that population will have that disease. Let's say that the sensitivity of the test is 0.99 and its specificity is 0.95. In this situation, the likelihood ratio of a positive test result is $\frac{0.99}{1-0.95}$ = 19.8, let's say 20, which is a strong likelihood ratio.

At 0.01, both probability and odds have the same value, so the post-test odds would be:

$$\text{post-test odds} = \text{pre-test odds} \times \text{likelihood ratio}$$
$$= 0.01 \times 20 = 0.20$$

Post-test odds of 0.20 translate into a post-test probability of 0.17. *This is the probability that the person **has** the disease, given a positive test result.* Putting it the other way around, there is a probability of 0.83 (83%) that the person does *not* have the disease, despite the positive test result.

The post-test odds can be improved by increasing either the pre-test probability or the likelihood ratio. Increasing the pre-test probability can be done by limiting the screening to a narrower population with a higher vulnerability to the disease in question. If pre-test probability were increased to 0.05, the pre-test odds would be 0.05. If the sensitivity and specificity of the test remained the same, then:

$$\text{post-test odds} = \text{pre-test odds} \times \text{likelihood ratio}$$
$$= 0.05 \times 20 = 1.0$$

Post-test odds of 1.0 translate into a post-test probability of 0.50, i.e. a 50:50 chance that the person has the disease if the test is positive.

Similarly, the post-test odds can be increased by increasing the *specificity* of the test. If, say, the specificity of the test were increased to, say, 0.99 (with the same sensitivity), the likelihood ratio would be:

$$\frac{0.99}{1-0.99} = 99$$

If the pre-test odds were 0.01, then:

$$\begin{aligned}\text{post-test odds} &= \text{pre-test odds} \times \text{likelihood ratio} \\ &= 0.01 \times 99 = 0.99\end{aligned}$$

Again, this translates into a post-test probability of 0.50.

Now, let's consider doing both – increasing the pre-test probability to 0.05 *and* the specificity of the test to 0.99. Then:

$$\begin{aligned}\text{post-test odds} &= \text{pre-test odds} \times \text{likelihood ratio} \\ &= 0.05 \times 99 = 1.0 = 5.0 \text{ (to two significant figures)}\end{aligned}$$

Post-test odds of 5.0 are equivalent to a post-test probability of 0.83. Most would agree that this level of probability requires action. Putting it the other way around, there is a probability of a mere 0.17 that the person does *not* have the disease, given this positive test result.

Essential point: For a screening test to minimise the risk of false positive results, its specificity must be *very high* and the population screened should be limited to those with *more than a trivial probability* of having the disease.

RELEVANCE TO YOUR EXAMINATION

For your examination in occupational and environmental medicine, you would ***not*** be expected to quote the numerical value of a likelihood ratio. However, the concepts of:

- "more likely" *versus* "less likely", or
- "strongly supportive" *versus* 'tending to dissuade"
- 'requiring action" *versus* "giving pause"

should, of course, colour your path to diagnosis and patient care, even if raw probabilities and odds are not written in your thoughts.

Question 2

A new blood test for ulcerative colitis has been designed. Colonoscopy [the gold standard] demonstrated disease in 50% of a study sample. The blood test has a true positive rate of 89% and a false positive rate of 21%. You see a child and estimate the child's probability of having ulcerative colitis is 10%. The blood test shows positive. What is the probability that the child has ulcerative colitis?

A	10%	D	70%
B	30%	E	72%
C	50%		

Response to Question 2

Let's first put aside the child's test and look at the *sensitivity* and *specificity* of the test itself. In the study that determined the *sensitivity* and *specificity* of the test, 50% of those in the study had ulcerative colitis and so 50% didn't have it. Let's for argument sake say there were 200 people in the study. Putting this on a 2 × 2 table would give the following numbers in the bottom row of sub-totals.

	HAS U.C.	**DOESN'T HAVE U.C.**	**TOTAL**
TEST POSITIVE	a	b	$a + b$
TEST NEGATIVE	c	d	$c + d$
	$a + c$ 50% of 200 = 100	$b + d$ also 100	$a + b + c + d$ 200

The question refers to a true positive rate of 89%. This means that of those that have the disease, 89% will show a positive test result. This is another way of saying that the sensitivity, $\frac{a}{a+c}$, is 89%, i.e $\frac{89}{100}$.

The question refers to a false positive rate of 21%. This means that of those that do *not* have the disease, 21% will show a positive test result. So $\frac{b}{b+d}$ is 21%, i.e. $\frac{21}{100}$. So, the completed table will look like this:

	HAS U.C.	**DOESN'T HAVE U.C.**	**TOTAL**
TEST POSITIVE	a [true +ve] 89	c [false +ve] 21	$a + b$
TEST NEGATIVE	c 100 – 89 = 11	d 100 – 21 = 79	$c + d$
	$a + c$ 50% of 200 = 100	$b + d$ also 100	$a + b + c + d$ 200

This brings us to the positive likelihood ratio which is:

$$\frac{\text{the proportion of people } \textit{with} \text{ the disease that have a } \textit{positive} \text{ test result (i.e.the sensitivity)}}{\text{the proportion of people } \textit{without} \text{ the disease that have a } \textit{positive} \text{ test result (1−specificity)}}$$

$$= \frac{a}{a+c} \text{ divided by } \frac{b}{b+d} = \frac{89}{100} \text{ divided by } \frac{21}{100} = 4.24.$$

So the positive likelihood ratio of this test is 4.24.

Now we go back to the child who you estimate [based on history and examination] has a probability of 10% or 0.1 of having ulcerative colitis. You perform this test on the child and it returns a positive result.

Based on this test result, you want to know the probability that the child has ulcerative colitis, i.e. the post-test probability.

A direct formula to convert the pre-test probability to the post-test probability using the likelihood ratio was shown on page 7. However, a simpler formula uses *odds* rather than probability. The formula using odds is:

$$\text{post-test odds} = \text{pre-test odds} \times \text{likelihood ratio}$$

The formula to convert probability to odds is $\text{odds} = \frac{\text{probability}}{1-\text{probability}}$. For a probability of 0.1, this converts to odds of $\frac{0.1}{1-0.1} = 0.11$.

So, post-test odds = $0.11 \times 4.24 = 0.47$.

Convert back to probability: $\text{probability} = \frac{\text{odds}}{1+\text{odds}} = \frac{0.47}{1+0.47} = 0.32$.

Given the positive test result, this is the probability that the child has ulcerative colitis.

This decimal fraction can be converted to the percentage of 32%.

The nearest response option for the question is 30% which is the answer.

And intuitively, 30% is the correct answer, even if you don't have much facility with numbers. For the answer to be 10%, means that the test offers no benefit whatsoever, i.e. a positive likelihood ratio of 1. That is *not* the case here. And given a pre-test probability of just 10% (0.1), you'd need a test with a BIG likelihood ratio, i.e. 10, in order to get to a post-test probability of 50%. Few new lab tests are that powerful.

Some definitions

Confidence interval: a *range* of values calculated from a *sample* within which we believe the *true value* for the population lies. A 95% confidence interval is calculated by a method that gives the *right answer 95% of the time.*

estimate: a number which is inferred to be a plausible value for some parameter of interest.

inference: process of drawing conclusions about a population on the basis of measurement or observations made on a sample of individuals from that population.

odds: probability that an event will occur *divided by* the probability that the event will *not* occur.[18]

odds ratio: the statistic used in *case-control* studies. It is the *odds of a particular exposure* among people with a disease of interest divided by the corresponding odds of exposure among persons *without* that disease.

parameter: a number that describes some characteristic of a whole population.

probability: an assessment, based on experience or theory, of the *proportion* of *outcomes of interest* that are likely to occur.

proportion: the number of people (or observations) with the characteristic of interest divided by the total number of people (or observations). Thus, a proportion is always a value between 0 and 1 although it may be expressed as a percentage. An example of proportion is *prevalence.*

P-value: the *probability* due to chance of observing a result *as extreme or more extreme* than the one actually observed.

rate: a proportion that takes the additional dimension of time e.g. 6 new cases per thousand of population per year. Over one year, this could also be expressed 6 cases per thousand person-years. An example of rate is *incidence.*

ratio: a comparison of two values – the number of observations with a characteristic of interest (e.g. exposure) divided by the number *without* that characteristic. *Odds* is an example of a ratio (so an 'odds ratio' is, in fact, a ratio of two ratios).

The ratio 8**:**5 may be expressed $\frac{8}{5}$**:** 1, i.e. 1.6 **:** 1 or, by leaving off the ' **:** 1', as simply 1.6.

relative risk: a comparison of the *incidence* of a disease among persons with *exposure* to an agent with the incidence of the disease among those *without* exposure.[19] Formally, in mathematical terms, this comparison is expressed as a ratio. Relative risk is also called *risk ratio*.

statistic: a value that *summarises* one characteristic of a *sample.* Examples of statistics include *mean*, *proportion*, *count*. A statistic may also *compare two parts* of a sample, e.g. *relative risk*, *odds ratio*, or a *difference* between means or proportions

type I error: rejection of the null hypothesis when it is actually correct.

type II error: failure to reject the null hypothesis when it is actually *in*correct.

---

[18] Because the results of an individual study may be used to predict the probability of future events, odds may also be defined using data from a single study as 'the number of times that an event has occurred *divided by* the number of times that the event has *not* occurred'. So, if there were 25 people exposed out of 100 with disease, then the proportion exposed would be 25/100 = 0.25. If this proportion were used to predict what may be the case among people similarly afflicted in the future, then we'd say the probability of exposure was 0.25. We could also say that the probability of *non-exposure* was 0.75 and so the odds of exposure was 0.25/0.75 = 0.33.

[19] In reality, relative risk is a comparison of the incidence of a disease among those with *more* exposure with the incidence of disease among those with *less* exposure.

Challenging for *both* learners and teachers

With epidemiology, there are challenges for the learner but also for the teacher. Here is a poem drawn from *Head KJ, Blessinger P. Teaching as a human experience. An anthology of contemporary poems. Newcastle upon Tyne: Cambridge Scholars Publishing, 2015*. p 70. It speaks about teaching *standard error*, the basis of confidence intervals, and how many students struggle but one 'got it' and used it.

Standard error

An inference that's very often made –
'a population tends to closely share
some feature that a smaller group displayed' –
is why my students need to be aware
of standard error. Yet, it's hard to learn,
and students' drive to listen fades with each
attempt of mine to strive for ways to earn
attention till my message has its reach.

The teaching cycle brings around today
my yearly chance to make this topic clear.
I'll aid my students, shorten what I say,
and they can later go and persevere …
… ask how it differs from the things it's like,
and why we need it, what we couldn't do
if it weren't here. But can my students strike
the hours in busy lives to see this through …
… to think and tell themselves of things they know …
related things … then try to make the link
with standard error, find their gaps and so
seek remedy where knowledge meets its brink?

It's four months on. A student whom I taught
then calls to see me and in measured way
explains she liked the insight I had brought
to 'estimation', then goes on to say:
"My research gathered data from a group,
defined a mob of which the group's just part;
a feature in the group at six per cent,
would be, I thought, like echoed in the mob.
I viewed the group as sample of the mob;
but samples vary some from whence they're drawn.
Your standard error helped me calculate
how far from six my estimate might stray."

She smiles and says she'd thought I'd like to know
that standard error served to underlie
her thinking. Yes, I'm happy that is so –
but more, she'd thought it worth enough to try.

David Goddard
June 2014