BJP British Journal of Pharmacology

DOI:10.1111/j.1476-5381.2012.02210.x
www.brjpharmacol.org

EDITORIAL

Not different is not the same as the same: how can we tell?

Gordon B Drummond[1] and Sarah L Vowler[2]

[1]*Department of Anaesthesia and Pain Medicine, University of Edinburgh, Royal Infirmary, Edinburgh, UK, and* [2]*Cancer Research UK, Cambridge Research Institute, Cambridge, UK*

Correspondence
Dr Gordon B. Drummond, Department of Anaesthesia and Pain Medicine, University of Edinburgh, Royal Infirmary, Edinburgh, 51 Little France Crescent, Edinburgh EH16 4HA, UK. E-mail: g.b.drummond@ed.ac.uk

This article is being published in *The Journal of Physiology*, *Experimental Physiology*, the *British Journal of Pharmacology*, *Advances in Physiology Education*, *Microcirculation*, and *Clinical and Experimental Pharmacology and Physiology*.

Gordon Drummond is Senior Statistics Editor for *The Journal of Physiology*.

Sarah Vowler is Senior Statistician in the Bioinformatics Core at Cancer Research UK's Cambridge Research Institute.

This article is the 12th in a series of articles on Best Practice in Statistical Reporting. All the articles can be found at http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1476-5381/homepage/statistical_reporting.htm.

Key points

- Testing for the presence of a condition may give an incorrect result
- Sensitivity and specificity of a test depend upon the chosen cut-off value
- 'No difference' usually means 'these data do not suggest a difference'
- The power of a test needs to be known to conclude that there is no difference
- Power calculations should be considered if the null hypothesis is not rejected

Calvin Coolidge was a taciturn president of the United States, nicknamed 'Silent Cal'. When told that he had died, the acerbic writer Dorothy Parker remarked, 'How could they tell?' Telling if something has *not* happened is a perennial problem in science, particularly so in laboratory experiments. Why should this be?

In the case of Cal Coolidge, they probably could tell. But in science, random events make it hard to tell, with certainty. There are also different ways of 'telling'. Mortality is a good categorical measure, an unequivocal endpoint. When diagnostic tests are used to define categories, we meet the concepts of 'false positive' and 'false negative'. Here, we use the expression 'false' to indicate that the conclusion we drew from the test would be incorrect. In the example we use below, we have a test for 'alive'. If we know the true condition, we can classify test results as true or false. Classifying a dead animal as alive would be a false positive. True and false are words usually applied to a diagnosis or similar categorical events but also crop up in other aspects of statistical inference. Categorical analysis and categorical logic (it's either there or it isn't) are not the same as the frequentist logic (the likelihood is small) that is used in common statistical analysis, and this leads to a great deal of argument and misunderstanding. Our previous comment that statistics has an impoverished vocabulary rings true: in this instance, the concepts of false positive and false negative conclusions are applied in different ways to different procedures, causing confusion.

Considering categories, imagine a pool that has become overheated in the sun: some of the frogs in it have died. Our observations suggest that the frogs on the surface of the pool are still alive, and the ones at the bottom of the pool are dead. Is this correct? A random sample produces the data in Figure 1.

Not all the frogs on the surface are alive. If we used 'on the surface' to indicate being alive, we would falsely attribute life to 10 frogs. As 107 frogs are dead, this is a false positive rate of 10/107 (which we only know if we have an unequivocal means of determining death). The positive predictive power

© 2012 The Authors
British Journal of Pharmacology © 2012 The British Pharmacological Society

	Alive	Dead	Total
At the surface	80	10	90
On the bottom	3	97	100
Totals	83	107	190
Test for alive	a	b	a+b
Test for dead	c	d	c+d
	a+c	b+d	a+b+c+d

False positive rate of the test for "alive" = b/(b+d)
False negative rate of the test for "alive"= c/(a+c)
Sensitivity (true positive rate) = a/(a+c)
Specificity (true negative rate) = d/(b+d)

Figure 1
Contingency tables for a diagnostic test.

of the test for life being present, if we were to look at the next frog on the surface, would be 80/90.

Unfortunately, expressions such as sensitivity, false positive and so on can be used in two different senses. Here, we assume we already have an exact means to determine death, much in the way that some statistics theory will start by defining a population. Thus, we consider that the false positive rate of our test for 'alive' is the proportion of dead frogs on the surface, expressed as a proportion of the frogs that really ARE dead. An alternative view would be if we were to consider the proportion of frogs on the surface that are later found to be dead, we would then be considering the outcome of the test. The diagnostic false positive rate would be 10/90. This is a diagnostic probability, and this form is less commonly used.

Many diagnostic tests do not rely on categories like this: a continuous variable is often used to distinguish categories. Suppose we measure the blood lactate concentration in frogs that are still alive and relate these concentrations to subsequent survival, we might obtain the results shown in Figure 2.

Although the median concentrations are different, some frogs with high lactate concentrations can survive. Is there a particular lactate concentration that will distinguish survival from non-survival? The capacity to predict outcome (sensitivity and specificity) alters as we choose different cut-off values. Using a series of cut-off values, we can plot sensitivity and (1 – specificity) for each chosen value (Figure 3). This is known as a receiver operating characteristic curve, or ROC curve, because one of its first applications was to find how operators of radar receivers distinguish 'real' echoes from background noise. In this case, we are looking for survivors using a lactate concentration that is less than the cut-off value we have chosen.

The large overlap of individual values between the two groups in Figure 2 suggests that lactate concentration is a poor indicator of survival, as there is no value that combines useful specificity and sensitivity. Any cut-off values we could choose would have too many false positive or false negative conclusions. Despite this, there is a statistically significant difference according to the Mann–Whitney test, showing that a significant difference does not mean the measure is the basis of a good diagnostic test.

In other forms of statistical inference, an expression such as 'false positive' is less appropriate and is often not well

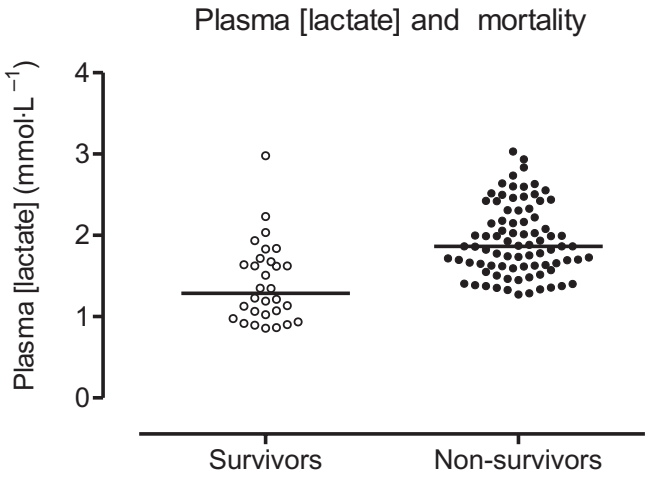

Figure 2
Plasma lactate values in frogs after overheating, related to survival. The groups are different ($P < 0.001$, Mann–Whitney *U*-test). The median value for survivors is 1.29, and for non-survivors, 1.86 (mmol·L^{-1}).

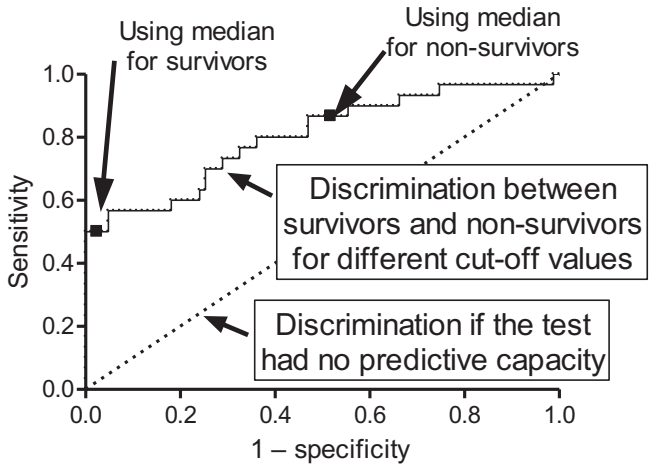

Figure 3
A receiver operating characteristic curve for the capacity of a lactate concentration below a chosen value to predict survival. An effective diagnostic test should have a large area below the curve, which would show a combination of good sensitivity and specificity. As tests vary, so does opinion about the area below the curve that indicates a good test. It should be substantially greater that 50%, which would be no better than chance alone, and the closer to 100%, the better.

defined, hidden in the logic of these tests. For example, the *t*-test does not provide black or white conclusions: statistical inference is restricted to judging how large a part chance could play in what we observe. Frequentist statistical tests such as the *t*-tests are measures of *uncertainty*. The usual laboratory experiment question 'do these data show an effect?' has to be worded 'are these samples statistically different?' As we know, the question then is re-worded 'How likely are these results if the samples came from the same population?' Let's say our test result is that $P = 0.01$, which

would usually be taken to indicate there is an 'effect'. This is not so: this *P*-value is the probability that we conclude there is an effect, and we still could be wrong. Although small, a chance still exists that we would be wrong, about 1 in 100 times. This is known as the *alpha error*, the possibility that we might classify our result as showing an effect, when in fact there is none. In the long run, if we were to repeatedly sample two populations that actually *were* the same, then results like the ones we found would be found 1 in 100 times. This is sometimes called the 'false positive' rate. However, it is based on the premise that there is NO difference and is perhaps better called the 'false conclusion' rate.

Reasoning for 'no effect' is even less certain and also often misconstrued. If the *P*-value is 0.12, then we judge that the results are quite possible, assuming the data are from the same population. If we did the same experiment again and again, then we might obtain data like this in almost one out of eight experiments. Is this unlikely? We usually choose to decide this is quite likely and thus do not reject the null hypothesis: if we rejected it, the chance of a false conclusion is too great. This 'chance of a false conclusion' is often misleadingly called a 'false negative rate'. The hypothesis the test uses, which is the NULL hypothesis, is resoundingly negative; there are no positive results lurking here. Positive results come if we are persuaded, by an estimate of probability (the *P*-value), that the null hypothesis is not tenable.

The context of the question 'is there an effect?' is relevant. The consequence of the answer 'no effect' may be unimportant. However, in some cases, the consequence of concluding that there is no difference, because a difference is not apparent, could be devastating. How often do we hear the claim 'there is no evidence that X is toxic?' to assert that X is not toxic? Immediately, the sceptic will ask how toxic and how much evidence? The statistician may add another qualification: how much variation is present?

A reasonable example (not a perfect analogy) is the radar operator: there is a greater certainty about a blip on the screen if the signal is large (accept the answer only if it's very likely), if you can keep looking at the screen (improve the sample size), and if there is not a lot of background noise (Figure 4).

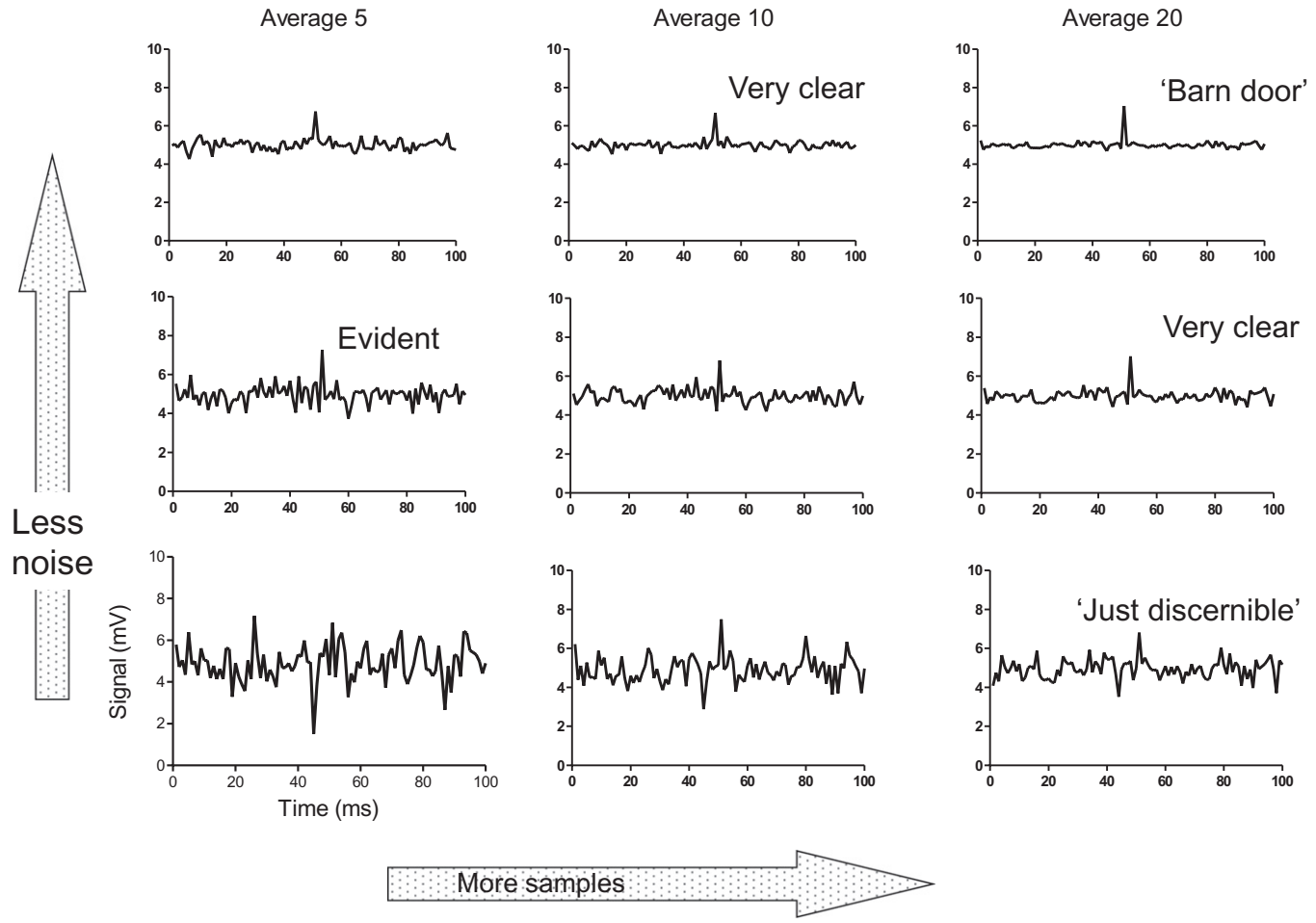

Figure 4

In each of these examples of signal averaging, there is an identical signal with added random noise. The signal is more easily seen if the noise is reduced (vertical arrow). Signal detection is improved by taking more samples to allow more averaging (horizontal arrow). The final feature of signal detection is how certain we wish to be (from 'just discernible' to 'barn door') that the event is not random (a concept analogous to the *P*-value).

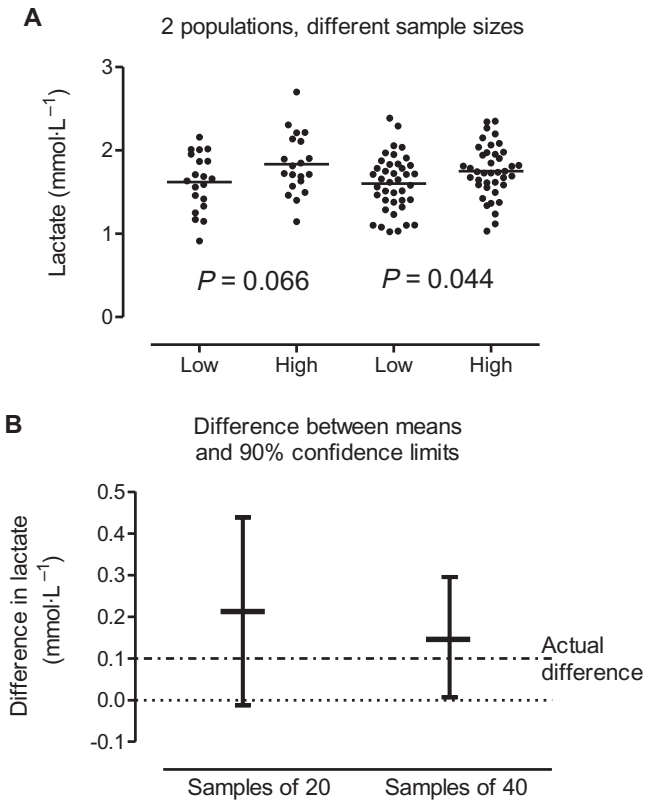

Figure 5

(A) Random samples from two different populations, low and high, where the population means differ by 0.1 mmol·L^{-1}. A *t*-test finds the *P*-values shown. (B) A valuable further analysis that should be applied before erroneously assuming 'no difference' is to inspect the 90% confidence limits of the difference between the means. This indicates the size of the possible differences.

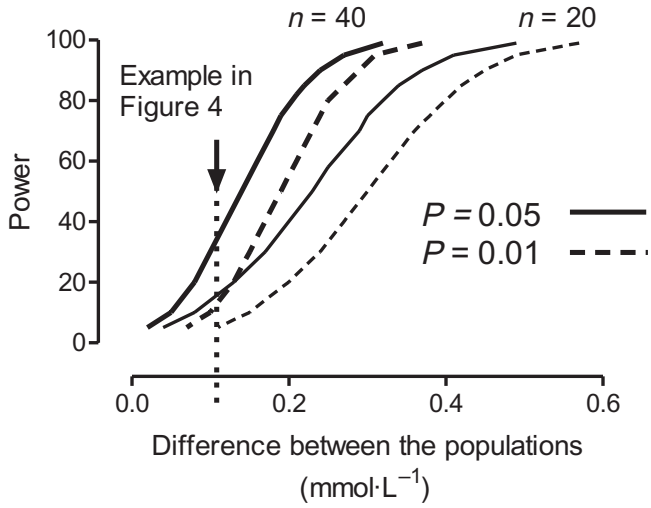

Figure 6

Using the samples in Figure 4, the power of the test (the chance of finding a specified difference between samples, given that it exists) is shown, where $n = 20$ and $n = 40$, for two different *P*-values. The populations that provided the example data in Figure 4 were different, by 0.1 mmol·L^{-1}, indicated by the vertical dotted line. If *P* is accepted at 0.05 and $n = 40$, the power of the test is only about 40%.

In this example, a signal is always present, but we just can't be sure in all the cases. This is equally so when statistics are applied to the null hypothesis. We have to consider the *beta error*: the possibility that we might classify our result as showing no effect where in fact there could be a difference. To demonstrate, consider two populations that are known to be different, as shown in Figure 5.

Here, a small sample of 20 per group, and a small true difference between the groups, does not reject the null hypothesis, with a *P*-value of 0.05. In this particular case, there is indeed a true difference in the populations from which these samples have been taken, but the power of the test applied has been inadequate. The power of a test is the chance of finding a *defined* difference: a power of 90% would mean that there was only a 10% chance of failing to find a difference of the defined magnitude. Thus, power is $(1 - \beta)$. The 10% false negative conclusion rate or beta error is also known as the type II error. Some call this a 'false negative', but this could be misleading as there is never a 'true' negative here.

The frequentist definition is that if the experiment were repeated over and over, the conclusion would be incorrect occasionally, and the rate is the beta error. For the example we have chosen, the relationships between power, sample size and the difference we seek to detect are shown in Figure 6.

The retrospective use of power calculations has been often criticized, but this is usually because the authors were seeking to *reject* a null hypothesis (they would have liked to find a difference). If a null hypothesis cannot be rejected, *this is not sufficient evidence that there is NO difference*. Claiming 'no difference' can be a serious error. If we wish to convince ourselves, and others, that there is no difference, we must be sure that the test was capable of rejecting a specifically defined difference, if it were present. If a small difference is considered important, and variation in the population is substantial, then a small sample will be inadequate. Put simply, the study is underpowered. Although there are software packages to calculate power, a simple approach for a simple two group comparison is to plot the 90% confidence limits of the difference between the means (Figure 5B). This will show the likely range of differences that could be detected.

The logical error (concluding that no significant difference is equivalent to the same) is frequently made worse when test results are compared; for example, the effect of substance A causes a significant difference, but substance B does not; thus, A has more effect than B. As we explain above, several factors, not explicitly stated, could affect this result. It could well be that B has a greater, but more varied, effect.

Because most tests are looking for a positive, many scientists fail to consider negatives adequately. In the old song 'Accentuate the positive', the accomplished song writer Johnny Mercer also cautioned to 'eliminate the negative, don't mess with Mr In-Between'.