# Interpretation of research results

Martin Bland

(adapted from work by Trevor Sheldon)

## Why do we need to know about statistics?

We are now in the era of evidence based nursing and the research-led NHS. This means that all healthcare professionals need to be able to read and understand evidence. This evidence often uses statistical methods. There is nothing particularly new about this, of course, and it was at the heart of the earliest development of modern nursing. Florence Nightingale, known as "The Passionate Statistician", was an early member of what is now the Royal Statistical Society and made innovations in statistical methods to use numerical data to promote improvements in healthcare.

Research evidence in usually published in scientific papers and in this lecture we shall look at the basic statistical ideas used in the presentation and interpretation of evidence in such papers. For example, this is the summary of a paper from a nursing journal:

> **Evaluation of an Electrolyte Replacement Protocol in an adult Intensive Care Unit: A retrospective before and after analysis**
>
> Zahra Kanji and Karleen Jung
>
> **Background**
>
> Electrolyte imbalances are frequently encountered in the Intensive Care Unit (ICU) and protocol-driven interventions may facilitate more timely and uniform care.
>
> **Objective**
>
> To compare the effectiveness and timeliness of electrolyte replacement in an adult ICU before and after implementation of an Electrolyte Replacement Protocol (ERP) and to assess nurse and physician satisfaction with the ERP.
>
> **Methods**
>
> Health records of adult patients who experienced hypokalaemia, hypomagnesaemia, or hypophosphataemia in the ICU during the study periods were retrospectively reviewed. Effectiveness of the ERP was assessed by the number of replacement doses indicated but not given and the number of doses and total dose required to normalise the low electrolyte level. Timeliness was evaluated by the time between the laboratory reporting the low electrolyte level and administration of the replacement dose. Nurse and physician satisfaction with the ERP was assessed through a written survey.
>
> **Results**
>
> After implementation of the ERP, the number of replacement doses indicated but not given was reduced for magnesium from 60% to 35% ($p = 0.18$) and for phosphate from 100% to 64% ($p = 0.04$). The time to replacement was reduced for potassium from 79 to 60 min ($p = 0.066$) and for magnesium from 307 to 151 min ($p = 0.15$). Nurses and physicians were satisfied with the ERP.
>
> **Conclusions**
>
> Implementation of an ERP resulted in improvements in the effectiveness and timeliness of electrolyte replacement and nurses and physicians were satisfied with the ERP.
>
> *Intensive and Critical Care Nursing* 2009; **25**: 181-189.

If we want to know whether to implement a similar electrolyte replacement protocol in our unit, we need to understand not just the nursing but also the research methods used in the paper. In this lecture, we shall look first at how we summarise and present data, then at how we interpret them.

## Measures of disease or outcome

When we carry out research in nursing, we usually need to measure either disease or the outcome of an intervention. The way we do this affects how we present and summarise information.

We usually distinguish between:

- qualitative measures, whether something is present or absent, or how things are divided into different categories, and
- quantitative measures (how much of something there is).

Examples of qualitative measures include disease diagnosed, presence of myocardial infarction, death or survival. Whether an indicated dose was given or not is a qualitative measure. Examples of quantitative measures include blood pressure, $PaO_2$, and urine output. Time to replacement of electrolyte is a quantitative measure.

Where a qualitative measure has only two possible outcomes, yes or no, dead or alive, we call it dichotomous. In this lecture we shall look at some ways of dealing with dichotomous and with quantitative measures.

### Dichotomous measures

We usually calculate the risk, the proportion of people in the group that show the outcome of interest (e.g. develops the disease, dies, heals). In this usage, risk can be the chance of good things happening as well as bad.

E.g. If 7 out of 100 patients have a pressure sore during a hospital stay, the risk of in-hospital pressure sore is 7/100 = 0.07 or 7%.

Odds is the number of people with the outcome divided by the number without the outcome.

E.g. If 7 out of 100 patients have a pressure sore during a hospital stay, 93 do not, so odds of in-hospital pressure sore is 7/93 = 0.075.

### Comparing risks

In the study of protocol directed sedation by nurses (Brook *et al.* 1999), patients in the protocol-directed sedation group also had a lower tracheostomy rate compared with patients in the non-protocol-directed sedation group (10 of 162 patients [6.2%] vs., 21 of 159 patients [13.2%]).

Risk difference = Control risk minus Intervention risk = 13.2% − 6.2% = 7.0 percentage points.

If risk difference = 0 then there is no difference in risk between two groups.

Relative Risk or Risk Ratio = Intervention risk / Control risk = 6.2% / 13.2% = 0.47.

This is less than half the risk.

If relative risk = 1 then there is no difference in risk between the two groups. If relative risk is less than 1 then the risk in intervention group is lower than the risk in comparison group. If relative risk is greater than 1 then the risk in the intervention group is higher than in the comparison group.

Odds ratio = Intervention odds / Control odds =.(10/(162 − 10)) / (21/(159 − 21)) = 0.43
= (6.2/(100.0 − 6.2)) / (13.2/(100.0 − 13.2)) = 0.43.

Odds ratio is also 1.0 if there is no difference, less than 1 if the risk in the intervention group is lower than the risk in the comparison group, and greater than 1 if the risk in the intervention group is higher than in the comparison group.

**Summarising quantitative data**

The first question we usually ask about quantitative data is "how much?". What is the typical or the average value? Two summaries of data which we often use for this are the mean and the median.

To find the mean of a set of measurements, we add all the values together and divide by the number of observations. This is sometimes called the arithmetic mean.

To find the median, we take the value of the middle observation when all the observations are put in order; 50% of the observations lie above the median and 50% lie below.

In "The time to replacement was reduced for potassium from 79 to 60 min", 79 and 60 will be means or medians.

**Measures of effect for continuous data**

Effective interventions should shift the average. For example, an intervention to lower blood pressure in hypertensive patients should result in a lower mean blood pressure.

The average should shift more in people being treated than those not and should shift more in effective treatments than in ineffective treatments. So we examine the difference in mean or median between groups that are treated and those not.

For example, in a study of protocol directed sedation by nurses compared with traditional non-protocol sedation in critically ill patients with acute respiratory failure (Brook *et al.* 1999):

The median duration of mechanical ventilation was 55.9 hours for patients managed with protocol-directed sedation and 117.0 hours for patients receiving non-protocol-directed sedation, a difference of 117.0 − 55.9 = 57.1 hours. Hence the patients receiving protocol directed sedation by nurses had shorter median ventilation time.

For those 132 patients receiving continuous intravenous sedation, those in the protocol-directed sedation group (n = 66) had a shorter mean duration of continuous intravenous sedation (3.5 days) than those in the non-protocol-directed sedation group (5.6 days). The difference in mean stay was 5.6 − 3.5 = 2.1 days.

For another example, consider a trial of a cognitive-behavioural intervention for patients with rheumatoid arthritis (Sharpe 2003). After 18 months, patients in the CBT group were less depressed, as measured by the Hospital Anxiety and Depression Scale (HADS) than at baseline, whereas those in the standard care group became more depressed:

| | Mean HADS depression score: | | |
|---|---|---|---|
| | Before | After 18 months | Reduction |
| CBT treated: | 5.1 | 4.6 | 0.5 (fall) |
| Usual care | 5.3 | 6.7 | −1.4 (rise) |

Difference in reduction in HADS depression = 0.5 − (−1.4) = 1.9

The HADS depression scale has a minimum of 0 and a maximum of 21.

### Variability

The spread of observations is important as well as the average.

- Not all patients are the same
- Values spread out around the average
- Response to treatments vary

Measures of variability include:

- Standard deviation (SD) — average spread — 2/3 of observations are within one standard deviation from mean.
- Range — minimum to maximum observations.
- Interquartile range (IQR) — range containing middle 50% of values.

In the CBT for rheumatoid arthritis study, standard deviations were presented:

| | Mean (SD) HADS depression score | |
|---|---|---|
| | Before | After 18 months |
| CBT treated: | 5.1 (3.9) | 4.6 (3.1) |
| Usual care | 5.3 (3.2) | 6.7 (4.3) |

So we can see that for the CBT group, their scores at the beginning were mostly between 5.1 − 3.9 = 1.2 and 5.1 + 3.9 = 9.0. The HADS scores are actually whole numbers, so between 1 and 9. (We take 8 or more to indicate the presence of depression.) We also know that about 95% of observations are usually between the mean minus two standard deviations and the mean plus two standard deviations. For HADS this is between 5.1 − 2×3.9 = −2.7 and 5.1 + 2×3.9 = 12.9. HADS cannot actually be below 0, so this would be between 0 and 13. All the 5% outside these limits would be above 13. (The maximum score is actually 21.)

Depression has no real units, so we sometimes use standard deviation units instead. We get difference in fall in depression = 1.9 HADS units = 1.9/3.55 = 0.54 standard deviations. (3.55 is the average SD at baseline.)

Differences presented in this way are called standardised mean differences or standardised effect sizes. There are useful to compare data when measurements are in different units or in arbitrary units.

## Drawing conclusions from data

Is the estimate from a single study the ‘true’ answer? If we repeat a study, we will not get exactly the same answer. This is the problem of random variation (sampling error). Even if there really is no treatment effect, the study can show a difference simply by chance. For example, suppose we give dice to two groups of nurses. They roll the dice and each group finds their average score. It is very unlikely that that the two groups will get exactly the same average score; one group will have a higher average than the other. But the dice are random and have no memory. If the nurses roll the dice again and calculate a new average, we cannot be sure that the same group will have the higher average. It could go either way.

How do we show how certain we are that the result is ‘true’? We have two widely-used ways to show how confident we are in the results of our study:

- Confidence intervals
- Significance tests (P values)

**Confidence intervals**

A confidence interval is a plausible range within which we are estimate that the true value lies.

For example, in the study of protocol-directed sedation during mechanical ventilation implemented by nurses: "The median duration of mechanical ventilation was 55.9 hrs (95% confidence interval, 41.0-90.0 hrs) for patients managed with protocol-directed sedation and 117.0 hrs (95% confidence interval, 96.0-155.6 hrs) for patients receiving non-protocol-directed sedation."

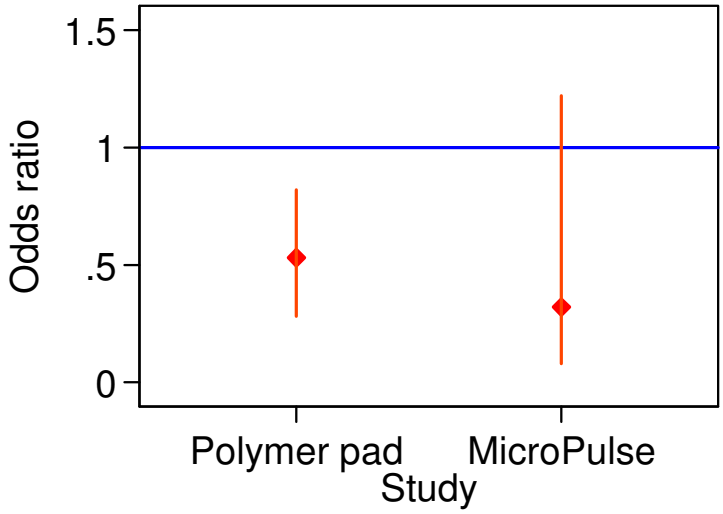
We estimate that, for all possible patients, the median duration with protocol management is between 41 and 90 hours. We do not know where in this range the actual median duration might be and there is a small chance that we might be wrong and it is outside these limits.

In a randomised controlled trial comparing a dry visco-elastic polymer pad and standard operating table mattress in the prevention of post-operative pressure sores (Nixon *et al.* 1998), 222 patients were randomised to the experimental group and 224 to the standard mattress. Pressure sores reported for 11% (22/205) of patients allocated to the dry visco-elastic polymer pad and 20% (43/211) for patients allocated to the standard operating table mattress. There was a significant reduction in the odds of developing a pressure sore on the dry visco-elastic polymer pad as compared to the standard, odds ratio = 0.46 with 95% confidence interval of (0.26, 0.82).

The 95% confidence interval for the odds ratio was 0.28 to 0.82. We estimate that for all possible patients, the odds of a pressure sore when using the dry visco-elastic polymer pad is between 0.28 and 0.82 times the odds of a pressure sore with the standard mattress. We do not know where in this range the actual ratio might be and there is a small chance that we might be wrong and it is outside these limits.

A trial compared an alternating pressure overlay intra-and post-operatively (the MicroPulse system) compared to a gel overlay (Russell and Lichtenstein 2000). 7/100 patients developed a pressure sore in the gel overlay compared to 2/98 in the MicroPulse system: odds ratio = 0.32. The 95% confidence interval is (0.08 to 1.22). Although for MicroPulse the odds ratio is further from 1 than is the odds ratio for the polymer pad study, the sample size is smaller and so is the risk of a pressure ulcer. This makes the confidence interval wider. The interval includes the value 1.0, so we could get an odds ratio as small as 0.32 with a sample of this size if there were no difference between the two systems.

Compare the two studies graphically:




The polymer pad study shows clearly that the polymer pad is superior to the standard mattress. The MicroPulse study does not show clearly that the MicroPulse system is superior to the gel overlay, although it suggests that it might be.

**Statistical significance**

If the effect size found is so big as to be unlikely to have occurred by chance if there really were no effect, we say it is statistically significant. If effect is small then we cannot exclude chance.

In the polymer pad study "There was a significant reduction in the odds of developing a pressure sore on the dry visco-elastic polymer pad as compared to the standard, odds ratio = 0.46 with 95% confidence interval of (0.26, 0.82), P = 0.010."

What is a P value? P is the proportion of possible samples which would have a difference as big or bigger IF there were really no difference in all possible trial participants.

In the polymer pad study, P = 0.010. Only 1 in 100 trials would produce a difference as big as this. This is good evidence that the polymer pad works.

What about the MicroPulse system? P = 0.17. The difference observed, large though it is, could happen in 17 out of every hundred trials. We do not have good evidence that MicroPulse works.

If the P value is small then we say that the result is statistically significant.

The usual decision points for P values are:

- P > 0.05 — no evidence or poor evidence for an effect, not statistically significant.
- P < 0.05 or P = 0.05 — reasonable evidence for an effect, statistically significant.
- P < 0.01 — good evidence for an effect, highly statistically significant.
- P < 0.001 — very strong evidence for an effect, very highly statistically significant.

**Confidence intervals and P values**

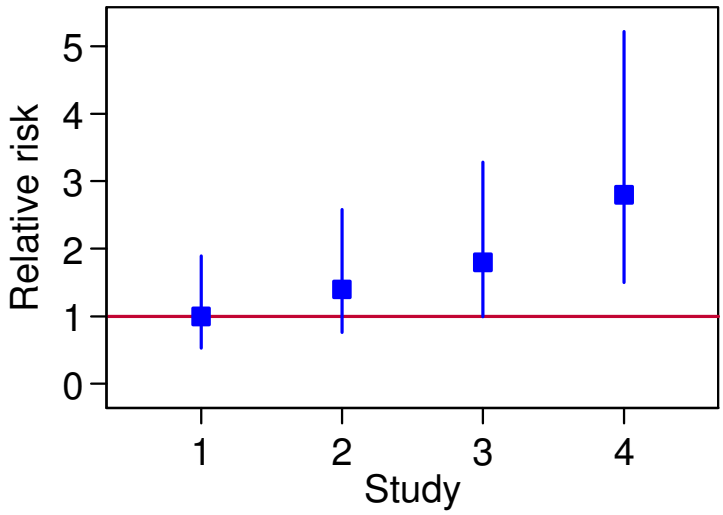
If P < 0.05, the "no effect" value lies outside the confidence interval. For a difference in means, the "no effect" value = 0. For odds ratios or relative risks, the "no effect" value = 1.

Example: four relative risks for pairs of samples of size 100. The risk in the control group is 0.25. For each hypothetical study, the relative risk found is given, as 1, 1.4, 1.8, and 2.8, and the confidence interval and the P value have been calculated.

The results are:

| | RR | 95% CI | P |
|---|---|---|---|
| Study 1 | 1.0 | 0.53 to 1.90 | 1.0 |
| Study 2 | 1.4 | 0.76 to 2.58 | 0.3 |
| Study 3 | 1.8 | 1.00 to 3.28 | 0.05 |
| Study 4 | 2.8 | 1.50 to 5.21 | 0.001 |

We can also show this graphically:




For Study 1, the estimate is 1.0 and the test is not significant, we have no evidence that risk ratio is different from 1.0. For Study 2, the estimate is 1.4, but the confidence interval includes 1.0, the P value is greater than 0.05, the test is not significant, and we do not have good evidence that the raised risk ratio we see in the sample would be present in the population of all the people like this. For Study 3, the estimate is 1.8, the confidence interval ends at 1.0, the P value is 0.05, the test is just statistically significant, and we do have evidence that the raised risk ratio we see in the sample would be present in the population of all the people like this. For Study 4, the estimate is 2.8, the confidence interval does not include 1, the P value is much less than 0.05, the test is highly statistically significant, and we have strong evidence that the raised risk ratio we see in the sample would be present in the population of all the people like this.

In the evaluation of an electrolyte replacement protocol in an adult intensive care unit, after implementation of the ERP, the number of replacement doses indicated but not given was reduced for magnesium from 60% to 35% (P = 0.18) and for phosphate from 100% to 64% (P = 0.04). The time to replacement was reduced for potassium from 79 to 60 min (P = 0.066) and for magnesium from 307 to 151 min (P = 0.15). Although things look better, the only thing for which we have reasonable evidence of a real improvement is phosphate. Even that evidence is not strong.

In the trial of protocol-directed sedation during mechanical ventilation implemented by nurses, the protocol-directed sedation group had statistically significantly shorter durations of mechanical ventilation than patients in the non-protocol-directed sedation group (P = 0.008). Lengths of stay in the intensive care unit (P = 0.013) and hospital (P < 0.001) were also significantly shorter among patients in the protocol-directed sedation group. Hence we have good evidence that the system improves patient outcome.

**Statistical Misconceptions**

It is not true that a very low P-value signifies a very strong clinical effect. Low P-values only tell us that there is good evidence that an effect exists. We should also look at the size of the effect and its confidence interval.

For example, the UK Prospective Diabetes Study Group randomised 1,148 hypertensive diabetic patients to atenolol or captopril. At the end of the trial, the atenolol group had significantly lower mean diastolic blood pressure than did the captopril group, P = 0.02. The authors did not recommend giving everybody atenolol, however. They reported that captopril and atenolol were equally effective in reducing blood pressure, with a difference, captopril minus atenolol, in both mean systolic and mean diastolic pressure of 1 mm Hg. The difference of 1 mm Hg is tiny and would not influence our choice of treatment. The 95% confidence interval for the difference in diastolic pressure was 0 to 2 mm Hg and for the non-significant difference in systolic pressure it was −1 to 3 mm Hg. Hence we cannot say that there is no difference in mean systolic pressure, but we can estimate that it is at most 3 mm Hg.

It is not true that a large P-value means that there is no clinical effect. There may be a effect but the sample may be too small to detect it reliably. It would be very rash to claim that the MicroPulse system did not work because the difference was not significant. The effect could also be quite dramatic.

## References


Brook AD, Ahrens TS, Schaiff R, Prentice D, Sherman G, Shannon W, Kollef MH. (1999) Effect of a nursing-implemented sedation protocol on the duration of mechanical ventilation *Critical Care Medicine* **27**: 2609-2615.

Kanji Z, Jung K. Evaluation of an Electrolyte Replacement Protocol in an adult Intensive Care Unit: A retrospective before and after analysis. *Intensive and Critical Care Nursing* 2009; **25**: 181-189.

Nixon J, McElvenny D, Mason S, Brown J, Bond S. A sequential randomised controlled trial comparing a dry visco-elastic polymer pad and standard operating table mattress in the prevention of post-operative pressure sores. *International Journal of Nursing Studies* 1998; **35**: 193-203

Russell JA, Lichtenstein SL. Randomised controlled trial to determine the safety and efficacy of a multi-cell pulsating dynamic mattress system in the prevention of pressure ulcers in patients undergoing cardiovascular surgery. *Ostomy/Wound Management* 2000; **46**(2):46-51, 54-5.

Sharpe L, Sensky T, Timberlake N, Ryan B, Allard S. Long-term efficacy of a cognitive behavioural treatment from a randomized controlled trial for patients recently diagnosed with rheumatoid arthritis. *Rheumatology* 2003; **42**: 435-441.

UKPDS Group. Efficacy of atenolol and captopril in reducing risk of macrovascular and microvascular complications in type 2 diabetes. *British Medical Journal* 1998; **317**: 713-720.