# Interpretation of research results

Martin Bland

Professor of Health Statistics
University of York

(adapted from work by Trevor Sheldon)

http://martinbland.co.uk/

## Why do we need to know about statistics?

We are now in the era of evidence based nursing and the research-led NHS.

Healthcare professionals need to be able to read and understand evidence.

This evidence often uses statistical methods.

Research evidence in usually published in scientific papers.

We shall look at the basic statistical ideas used in the presentation and interpretation of evidence.

## Why do we need to know about statistics?

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.

*Intensive and Critical Care Nursing* 2009; **25**: 181-189.

## Why do we need to know about statistics?

This is the summary of a paper from a nursing journal:

**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.

## Why do we need to know about statistics?

This is the summary of a paper from a nursing journal:

**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.

## Why do we need to know about statistics?

This is the summary of a paper from a nursing journal:

**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.

**Why do we need to know about statistics?**

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.

**Why do we need to know about statistics?**

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.

- How do we summarise and present data?
- How do we interpret data?

**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,
- quantitative measures (how much of something there is).

## Measures of disease or outcome

Examples of qualitative measures:

disease diagnosed,
presence of myocardial infarction,
dead or alive,
whether an indicated dose was given or not.

Only two possible outcomes, we call it dichotomous.

Examples of quantitative measures:

blood pressure,
PaO2,
urine output,
time to replacement of electrolyte.

In this lecture we shall look at some ways of dealing with quantitative and dichotomous measures.

## Dichotomous measures

Risk = proportion of people in the group that show the outcome of interest (e.g. develops the disease, dies, heals).

"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 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 hospital stay, 93 do not, so odds of in-hospital pressure sore is 7/93 = 0.075.

## Comparing risks

Example: 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).

Patients in the protocol-directed sedation group 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.

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.

## Comparing risks

Protocol directed sedation by nurses:

Patients in the protocol-directed sedation group 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%]).

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

This is less than half the risk.

Relative risk = 1.0 ➔ no difference in risk.

Relative risk < 1.0 ➔ risk in intervention group lower.

Relative risk > 1.0 ➔ risk in the intervention group is higher.

## Comparing risks

Protocol directed sedation by nurses:

Patients in the protocol-directed sedation group 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%]).

Odds ratio = Intervention odds / Control odds

= (10 / (162 − 10)) / (21/(159 − 21)) = 0.43.

= (6.2 / (100 − 6.2)) / (13.2/(100 − 13.2)) = 0.43.

Odds ratio = 1.0 ➔ no difference in risk.

Odds ratio < 1.0 ➔ risk in intervention group lower.

Odds ratio > 1.0 ➔ risk in the intervention group is higher.

## Summarising quantitative data

First question: “how much?”.

What is the typical or the average value?

Two summaries of data which we often use are

- mean,
- median.

## Summarising quantitative data

Mean:

> add all the values together and divide by the number of observations.

Sometimes called the arithmetic mean.

Median:

> 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.

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.

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.

## Measures of effect for continuous data

Example: 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.

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.

## Measures of effect for continuous data

Example: 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).

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.

## Measures of effect for continuous data

Example: trial of a cognitive-behavioural intervention for patients with rheumatoid arthritis (Sharpe 2003).

Hospital Anxiety and Depression Scale (HADS) depression score (0 to 21) after 18 months:

| | 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

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.

## 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 observations.

## Variability

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) |

Approximately 2/3 of observations are within one standard deviation from mean.

For the CBT group, 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. (8 or more ➔ depression.)

## Variability

About 95% of observations are usually between the mean minus two standard deviations and the mean plus two standard deviations.

For CBT, between 5.1 − 2×3.9 = −2.7
and 5.1 + 2×3.9 = 12.9.

HADS cannot 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.)

## Variability

Depression has no real units.

Sometimes use standard deviation units instead.

Difference in fall in depression = 1.9 HADS units
= 1.9/3.55
= 0.54 standard deviations.

(3.55 = average SD at baseline.)

Called standardised mean differences or standardised effect sizes.

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.

**Drawing conclusions from data**

Even if there really is no treatment effect, the study can show a difference simply by chance.

Example: give two groups of nurses dice.

They roll the dice, each group finds their average score.

One group will have a higher average than the other.

Dice are random and have no memory.

Roll the dice again and calculate a new average.

Cannot be sure the same group will have the higher average.

**Drawing conclusions from data**

How do we show how certain we are that the result is ‘true’?

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.

Example, protocol-directed sedation 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."

Estimate that, for all possible patients, the median duration with protocol management is between 41 and 90 hours.

Do not know where in this range the actual median might be.

Small chance that it is outside these limits.

## Confidence intervals

Example: randomised controlled trial comparing a dry visco-elastic polymer pad and standard operating table mattress in the prevention of post-operative pressure sores.

222 patients randomised to the experimental group
224 patients randomised to the standard mattress.

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

## Confidence intervals

Example: randomised controlled trial comparing a dry visco-elastic polymer pad and standard operating table mattress in the prevention of post-operative pressure sores.

222 patients randomised to the experimental group
224 patients randomised to the standard mattress.

Pressure sores reported for:

Dry visco-elastic polymer pad, 11%.

Standard mattress, 20%.

Odds ratio = 0.46 with 95% confidence interval of (0.26 to 0.82).

## Confidence intervals

Example: randomised controlled trial comparing a dry visco-elastic polymer pad and standard operating table mattress in the prevention of post-operative pressure sores.

Odds ratio = 0.46 with 95% confidence interval of (0.26 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.

Do not know where in this range the actual ratio might be.

Small chance that it is outside these limits.

## Confidence intervals

Example: trial compared an alternating pressure overlay intra-and post-operatively (the MicroPulse system) compared to a gel overlay.

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.

## Confidence intervals

Example: trial compared an alternating pressure overlay intra-and post-operatively (the MicroPulse system) compared to a gel overlay.

Pressure sores:

Gel overlay, 7/100 patients.

MicroPulse system, 2/98 patients.

Odds ratio = 0.32.

The 95% confidence interval is 0.08 to 1.22.

## Confidence intervals

Polymer pad vs. standard mattress:

odds ratio = 0.46, 95% CI is 0.26 to 0.82.

MicroPulse systems vs. gel overlay:

odds ratio = 0.32, 95% CI is 0.08 to 1.22.

For MicroPulse the odds ratio is further from 1.0 than is the odds ratio for the polymer pad.

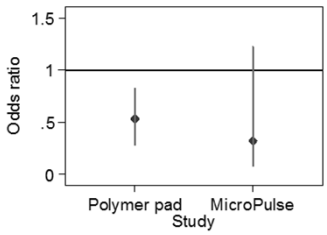
Sample size is smaller and so is the risk of a pressure ulcer.

This makes the confidence interval wider.

For MicroPulse, confidence interval includes 1.0.

Could get odds ratio as small as 0.32 with a sample of this size if there were no difference between the two systems.

## Confidence intervals




Clear that polymer pad is superior to the standard mattress.

Not clear that the MicroPulse system is superior to the gel overlay, although study 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.

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.”

**Statistical significance**

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.

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.

**Statistical significance**

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.

MicroPulse study: 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.

**Statistical significance**

What is a P value?

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.

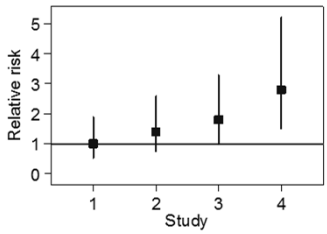
For odds ratios or relative risks, the no effect value = 1.

Example: four relative risks for pairs of samples of size 100.

The risk in the control group is 0.25.

| | 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 |

## Confidence intervals and P values




P = 1.0 0.3 0.05 0.001

Study 1, no evidence that risk ratio is different from 1.0.

Study 2, do not have good evidence.

Study 3, do have some evidence.

Study 4, strong evidence.

## Confidence intervals and P values

Example, 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.

## Confidence intervals and P values

Example, trial of protocol-directed sedation 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."

We have good evidence that the system improves patient outcome.

## Statistical Misconceptions

"A very low P-value signifies a very strong clinical effect."

**Not true!**

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.

## Statistical Misconceptions

Example: the UK Prospective Diabetes Study Group randomised 1,148 hypertensive diabetic patients to atenolol or captopril.

The atenolol group had significantly lower mean diastolic blood pressure than did the captopril group, P = 0.02.

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.

## Statistical Misconceptions

Example: the UK Prospective Diabetes Study Group randomised 1,148 hypertensive diabetic patients to atenolol or captopril.

The atenolol group had significantly lower mean diastolic blood pressure than did the captopril group, P = 0.02.

Did not recommend giving everybody atenolol.

Reported that captopril and atenolol were equally effective in reducing blood pressure.

Difference, captopril minus atenolol, in both mean systolic and mean diastolic pressure = 1 mm Hg.

Difference = 1 mm Hg is tiny and would not influence choice of treatment.

## Statistical Misconceptions

Example: the UK Prospective Diabetes Study Group randomised 1,148 hypertensive diabetic patients to atenolol or captopril.

Diastolic pressure: 95% confidence interval for the difference was 0 to 2 mm Hg.

Systolic pressure: 95% confidence interval for the difference was −1 to 3 mm Hg (difference not significant).

We cannot say that there is no difference in mean systolic pressure, but we can estimate that it is at most 3 mm Hg.

## Statistical Misconceptions

"A large P-value means that there is no clinical effect."

**Not true!**

There may be an 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.

Pressure sores: gel overlay, 7/100 patients,
MicroPulse system, 2/98 patients.

Odds ratio = 0.32, 95% confidence interval is (0.08 to 1.22).

The effect could also be quite dramatic.