TRUTH IN MUSCULOSKELETAL MEDICINE. I: CONFIDENCE INTERVALS

Nikolai Bogduk, Newcastle Bone and Joint Institute, University of Newcastle, Newcastle, NSW, Australia.

Critical reasoning and biostatistics is not something new. The instruments and concepts were developed in the 1950s, and have been elaborated and refined since that time. What is relatively new is their application to medicine at large and to musculoskeletal medicine in particular.

Conspicuously absent in the past, and even to this day, has been an appropriate respect for biostatistics and critical reasoning in undergraduate and postgraduate medical curricula. At best, lip service has been paid to biostatistics but the implications of biostatistics have not been integrated into clinical practice. Instead, medical students are typically taught in a way that implies that the various techniques of physical examination and diagnostic tests in which they are trained are reliable and valid, and that the treatments that they are taught are unquestionably efficacious. This pattern continues into postgraduate training. Yet, ironically when these tests and treatments are subjected to scientific scrutiny they often prove not to be reliable, valid or efficacious. There is, therefore, a mismatch between what is taught and the truth.

This series of articles is designed not to constitute some sort of course of instruction in academic material that makes doctors erudite but which is immaterial to clinical practice. Rather, it is intended to help the reader become a better consumer of clinical information, so that they can recognise the reliable, the valid and efficacious, and thereby distinguish truth from mythology, assertion and speculation. The objective is to equip the reader with devices that allow them to check information for themselves rather than relying on what experts say is right and wrong. If nothing else, the reader will learn what questions to ask, what information to demand, before accepting or believing a speaker or the writer of a journal article.

CONFIDENCE INTERVAL OF A PROPORTION

This first concept is a preface. It does not lead systematically into subsequent topics but recurs in various forms in other areas of biostatistics, and has some immediate applications to day to day practice.

The concept can be introduced by the question: does 3 out of 10 equal 30%?

Mathematicians and philosophers may argue what they please about this question, but in medicine the answer is - no.

The fraction - 3/10, is a proportion, and in medicine will usually reflect the result of some sort of harvest. An investigator will have studied or surveyed 10 cases and found the index condition in three. They are tempted to proclaim a yield of 30%.

The illegitimacy of this temptation stems from the possibility that if the same investigator, or another investigator, repeated the same experiment, they might encounter a slightly different yield - say, 4/10 or 2/10 or even 6/10. What then is the true frequency?

The principle at hand is that there may be a correct or true proportion, that would be evident if every patient or every person in the universe was surveyed, but this proportion will not necessarily be evident if only a small sample of the total, possible population is surveyed. For any small sample a sampling “error” may occur. Just by accident, the investigator might select a group of subjects who happen to exhibit the feature in question somewhat more frequently than the true proportion or somewhat less frequently.

In order to accommodate this possibility a statistical correction applies [1]. The formula is:

$$p^* = p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

where,

p	is the observed proportion,
n	is the number of subjects,
1.96	is a coefficient that generates a 95% probability, and
p^*	is the range within which there is a 95% chance that the true proportion actually lies. Conversely there is a 5% probability that the true value lies outside this range.

If we consider our example,

$p = 3/10 = 0.3$
$(1-p) = (1.0 - 0.3) = 0.7$
$n = 10$

$$p^* = 0.3 \pm 1.96\sqrt{\frac{(0.3)(0.7)}{10}}$$

$= 0.3 \pm 0.28$
$= 0.02$ to 0.58

This result shows that upon sampling 10 subjects and finding an index condition in 3, the prevalence of that condition is not necessarily 30%. The true prevalence could be as low as 2% or as high as 58%. Under these conditions, 30% is not a representative figure. The ambiguity arises because the sample size (n) is small.

Now see the effect of increasing the sample size. Suppose that the investigator studied 100 subjects instead of 10, and found the condition in 30. The prevalence is still not 30%; that figure is still only an estimate because the investigator did not survey every patient in the universe. The confidence interval of the observed proportion must be calculated.

$p = 30/100 = 0.30$
$(1-p) = (1.00 - 0.30) = 0.70$
$n = 100$

$$p^* = 0.30 \pm 1.96\sqrt{\frac{(0.30)(0.70)}{100}}$$

$= 0.30 \pm 0.08$
$= 0.22$ to 0.38

There is still uncertainty but it is less than when the sample size was 10. The true proportion could be as low as 22% or as high as 38%, but for this range, 30% is now a reasonable, indicative figure.

Increasing sample size decreases the size of the 95% confidence interval of the observed proportion. The decrease is exponential. When n is small the confidence interval is wide. As n increases the confidence interval narrows but never gets to zero. The interval would get to zero only if everyone in the universe is sampled, i.e. n = infinity.

Only you can answer how big the sample should be. The answer translates into a question to you - how close do you want the estimated proportion to be to the true proportion? If you are satisfied with a 20% range you might settle for a small sample, but if you want to be within 5% of the true proportion, you will

demand a larger sample. The equation is available to you in order to calculate these answers. This equation is one that you should either memorise or carry in your wallet; you never know when you will need to use it as a consumer of new information.

EXAMPLES

The following are a series of examples that illustration the application of confidence intervals in situations that may befall practitioners in musculoskeletal medicine. They are not esoteric or academic applications but ones that should be of day to day concern or interest.

Example 1: success rate of a new treatment

A speaker announces a success rate of 70% for a new treatment. Do you believe him? You should not. You should first ask - what was *n*?

If n = 10, the success rate is 7/10. You should calculate the confidence interval of this proportion. It amounts to 42% to 98%. In real-life terms this means that if *you* were to repeat the experiment, i.e. adopt the treatment, you might encounter a result as good as 98% or as bad as 42%. You cannot expect 70%; you should be prepared for 42%.

If n = 100, the confidence interval changes to 61% to 79%. In this case, 70% is a more indicative figure of what *you* might expect to encounter; but be prepared for a success rate as low as 61% instead of 70%.

This example shows that confidence intervals are not a tool used only by research scholars. They are relevant to you as a therapist. Every time someone advises you to adopt a new therapy, they are effectively inviting you to become an investigator and to repeat their experiment. Therefore, you should have no illusion that reported proportions are absolute. Your experience may be different from that of the first investigator, and the confidence interval formula indicates to you how different your experience might be.

You can also use the equation in reverse to calculate the appropriate n, if you have a certain confidence interval in mind. Let's say you want to ensure that the success *you* are prepared to accept is anywhere within 15% of the speaker's reported success rate. On what sample size should the speaker's results be based? Under these conditions,

$$p^* = 0.70 \pm 0.15$$

$$1.96\sqrt{\frac{(0.70)(0.30)}{n}} = 0.15$$

$$\sqrt{\frac{(0.70)(0.30)}{n}} = \frac{0.15}{1.96}$$

$$\sqrt{\frac{0.21}{n}} = 0.0765$$

$$\frac{0.21}{n} = 0.0059$$

$$\frac{0.21}{0.0059} = n$$

$$n = 35.59$$

Thus, for you to expect a success rate in your hands of 70% $\pm$ 15%, the investigator should provide you with a success rate of 70% based on 35.59 subjects, i.e. at least 36 subjects. If the investigator's study is smaller than this, you cannot rely on achieving a result in your hands that falls within 15% of 70%; your result might be considerably worse. The same calculation could be repeated if you wanted a tighter range, say 5%, and if the reported success rate were any other figure, say 80% or 60%. The general formula is:

p = the reported success rate

p^* = the range which you would accept, i.e.

$p^* = p \pm z$

where

$$z = 1.96\sqrt{\frac{p(1-p)}{n}}$$

from which n can be calculated.

Example 2: could it be placebo?

An investigator audits a new treatment. He finds a success rate of 8/10. Is this impressive, or might it be a placebo response? In order to answers these questions, calculate the confidence intervals.

The confidence interval of 8/10 is 55% to 100%. Prima facie this looks like the result is not a placebo, for 55% is well above the conventional estimate of 30% for a placebo response rate. However, one must also calculate the confidence interval of the placebo response rate.

In a sample of only 10 patients, the confidence interval of 3/10 is 2% to 58%. This means that in a sample of only 10 patients, a series of investigators could encounter placebo rates as low 2% or as high as 58%.

The figure - 58%, is higher than the figure - 55%. Thus, the confidence intervals of 8/10 and 3/10 overlap. This means that statistically, it possible for a success rate of 8/10 to overlap the possible placebo range. Thus, prima facie, 8/10 could well be a placebo response rate. The investigator would have to provide you with a larger study, with a narrower confidence interval of the success rate, before you can accept that the result is not placebo.

This sort of calculation does not prove that a reported result is or is not a placebo response; that can only be shown directly in a controlled study. But it does show that statistically 8/10 is not necessarily an impressive result. Calculating the confidence intervals prevents you from being seduced into believing that such an impressive could not possibly be a placebo.

Investigators wanting to plan a study can use these sorts of calculations in reverse in order to determine what size of study is required to provide prima facie evidence that the observed success rate is unlikely to be due to a placebo effect. Such calculations can be used to determine if the treatment in question is worthy of a controlled trial. There is no point expending effort if the prima facie evidence is consistent with a placebo effect.

What size study should be conducted if the observed success rate is 75%, and the placebo rate is assumed to be 30%?

For there to be prima facie evidence of the treatment not being due to placebo, the upper confidence limit of the placebo rate should be less than the lower confidence limit of the success rate, i.e.

$$0.3 + 1.96\sqrt{\frac{(0.3)(0.7)}{n}} < 0.75 - 1.96\sqrt{\frac{(0.75)(0.25)}{n}}$$

$$0.3 + 1.96\sqrt{\frac{0.21}{n}} < 0.75 - 1.96\sqrt{\frac{0.1875}{n}}$$

$$1.96\sqrt{\frac{0.21}{n}} + 1.96\sqrt{\frac{0.1875}{n}} < 0.45$$

$$\sqrt{\frac{0.21}{n}} + \sqrt{\frac{0.1875}{n}} < 0.23$$

Upon squaring both sides,

$$\frac{0.21}{n} + \frac{0.1875}{n} + \frac{2\sqrt{(0.21)(0.1875)}}{n} < 0.053$$

$$\frac{0.21}{n} + \frac{0.19}{n} + \frac{0.40}{n} < 0.053$$

$$\frac{0.80}{0.053} < n$$

$$n > 15.09$$

Thus, for this success rate and expected placebo rate, the study must be based on at least 16 subjects. A smaller study would not show a success rate greater than the possible placebo rate. Conversely, if the placebo rate was greater, or the success rate smaller, the sample size would need to be appropriately larger. Remember, however, that such calculations do not prove that the success is not due to placebo; they provide only prima facie evidence that it is *unlikely* to be due to placebo. The utility of the calculations is not to substitute for a controlled trial, but to prevent controlled trials being wasted on success rates that are possibly within the placebo range. They also protect the consumer from being seduced by figures that numerically look impressive but which are based on too small a study.

Example 3: population studies

In an epidemiological study, an investigator samples 50 individuals with a history of neck pain after a motor vehicle accident and finds that none developed chronic neck pain. He concludes that chronic neck pain after whiplash does not occur. Is this deduction correct?

Assume that the true prevalence of chronic neck pain after whiplash is 6%. Did the author have a large enough sample to exclude this prevalence?

For a sample of 50, the confidence interval of 6% is

$$p^* = 0.06 \pm 1.96\sqrt{\frac{(0.06)(0.94)}{50}}$$

$$= 0.06 \pm 0.065$$

$$= 0.0 \text{ to } 0.125$$

The figure - 0.0, indicates that with a sample of only 50 the investigator could well be studying a population in which the true prevalence was 6% but would find zero cases. Another investigator, using the same sample size might find 12.5% of 50 = 6.25, i.e. 6 cases. Thus, the study does not exclude a 6% prevalence.

For interest, calculate what prevalence does a sample of 50 reasonably exclude.

$$p - 1.96\sqrt{\frac{p(1-p)}{50}} > 0$$

$$p > 0.071$$

Thus, a sample of 50 might only exclude a prevalence of more than 7%.

However, remember that this is not an absolute result. The confidence interval expresses only a 95% chance. Thus, a sample of 50 has only a 95% chance of excluding a prevalence of 7%. There remains a 5% chance that a prevalence of 7% would not be excluded by a sample of 50.

Another warning is that towards the extremes, confidence interval calculations come to grief. When the proportions approach 0% or 100%, the conventional formula does not apply and certain mathematical adjustments need to be applied [2].

Armed with this example, the reader is invited to analyse for themselves, as an assignment exercise, the data and conclusions provided by a recent study in Lithuania on the prevalence of whiplash [3]. [Hint: find how many patients suffered neck pain immediately after the accident and how many of these went on to develop chronic symptoms. Using these figures, calculate the prevalence that would be excluded by this sample size, of chronic neck pain arising in patients who suffer neck pain immediately after an accident. Compare that finding with the prevalence of chronic neck pain in individuals who simply are involved in an accident without suffering neck pain.]

CONCLUSION

This is the first step towards incorporating biostatistics into everyday practice. The ability to calculate the confidence interval of a proportion equips the reader with a survival technique in the world of medical consumerism. It protects the reader against being hoodwinked by figures that look good but which are based on too small a sample. The confidence interval is one of the devices that helps answer the question - was the study big enough. Instead of appealing to experts, readers can now use the confidence interval formula to answer this question for themselves, whenever the occasion arises. In future articles, we will address truth in diagnosis and truth in therapy.

REFERENCES

1. Armitage P, Berry G. Statistical Methods in Medical Research, 3rd edn. Oxford: Blackwell, 1994 pp93-125.

2. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical Epidemiology. A Basic Science for Clinical Medicine, 2nd edn. Boston: Little, Brown & Co. 1991; pp175-176.

3. Schrader H, Obelieniene D, Bovim G, Surkiene D, Mickeviciene I, Sand T. Natural evolution of late whiplash syndrome outside the medicolegal context. Lancet 1996;347:1207-1211.