Original article

Statistical testing in wound care

Keywords inferential statistical tests, *p* value, confidence intervals, odds ratio

For referencing Stephenson J. Statistical testing in wound care. WCET® Journal 2022;42(4):38-40

DOI https://doi.org/10.33235/wcet.42.4.38-40

Submitted 6 November 2022, Accepted 1 December 2022

Wound care studies generally yield quantitative data; and that data require some form of analysis. Almost all studies include a descriptive summary of participants, as discussed in my previous article[1]. However, many studies also require some form of inferential statistical testing; usually if the intention is to generalise findings from the sample to a population.

Different study designs require different testing methods, although the basic aim in most cases is the same. This is to assess the *significance* of the *effect* of interest; that is, to establish whether any effect we see in our sample data is a reflection of the state of affairs in the parent population, or likely to be nothing more than random variation in our data. An effect could be an observed difference between study groups (such as amount of biofilm present in a control group and a group where patients are treated with an anti-microbial dressing), the difference between a measure taken at baseline and post-intervention from a single patient group (such as pain levels during and after surgery), an observed relationship between two variables (such as the extent of mobile health technology use and wound care knowledge) or many other quantities.

Establishment of statistical significance requires a test of a hypothesis. We usually test a *null hypothesis* (of no effect); for example, that the difference in means in the population is zero. Miller et al[2] conducted a pilot single-blinded randomised controlled trial (RCT) to examine concordance with and acceptability of electric stimulation therapy (EST) in patients with venous leg ulcers (VLUs) who had not tolerated moderate to high compression. The treatment was tested against the null hypothesis of no effect; i.e. that concordance with the total recommended treatment time under control treatment or under EST was the same. Lenselink and Andriessen[3] measured (among other quantities) the percentage of granulation tissue and yellow tissue in a cohort study on the efficacy of a polyhexanide-containing biocellulose dressing in 28 patients, testing several hypotheses relating to differences in patient outcomes between baseline and 24 weeks. In a study of diabetic and non-diabetic patients, Gunes et al[4] analysed several biomarkers, primarily galectin-3 (which promotes angiogenesis and new vessel formation) and tested several hypotheses relating to relationships between galectin-3 and various other biomarkers.

Significance is quantified using the familiar (if not widely understood) *p*-value, which is a *conditional probability*: the probability that observed results, or something more extreme, would have been obtained, given that the null hypothesis is true. In everyday parlance it is referred to loosely as 'the chance of a chance': that is, the probability that our data has fallen the way it has just as a result of natural variation and not because an effect actually exists. Conventionally, a *p*-value below 0.05 (5%) is taken as indicating an outcome of statistical significance (at the 5% significance level) and a consequential rejection of the null hypothesis of no effect. The study of Miller et al yielded a *p*-value of 0.671 for its primary outcome; indicating insufficient evidence for a difference between treatment groups. The study of Lenselink and Andriessen yielded a *p*-value reported to be less than 0.04 for a test comparing the mean percentage of granulation tissue from baseline to 24 weeks post-treatment; indicating evidence for a treatment effect. The study of Gunes et al yielded a *p*-value of less than 0.001 in a test for correlation between the galectin-3 and C-reactive protein biomarkers, indicating strong evidence for a significant relationship between these biomarkers.

The RCT conducted by Miller et al and the test of changes from baseline conducted by Lenselink and Andriessen are examples of grouped study designs: an *unpaired study*, in which two unrelated study groups are compared, in the case of Miller et al and a *paired study* design, in which the measures are taken on two occasions from the same group of patients, in the case of Lenselink and Andriessen (the word 'paired' refers to the fact that each patient contributes a pair of readings, not that a pair of groups is involved). Another variant of this study arises when measures are taken from patients matched on demographic or health-related variables. The study of Gunes et al is an example of a correlational design using ungrouped data. These three designs are probably the most common choices of study design in wound care, although various other study types, which also aim to infer results from sample data

John Stephenson
PHD FRSS(GradStat) CMath(MIMA)
Senior Lecturer in Biomedical Statistics
University of Huddersfield, United Kingdom
Email J.Stephenson@hud.ac.uk

to a population may be found, including for example tests of agreement, screening/diagnostic studies, survival analysis and meta-analysis.

Numerical outcomes in unpaired studies can be compared using the independent samples t-test, or, if more than two groups are involved, a test such as analysis of variance (ANOVA). Categorical outcomes are usually analysed using the chi-squared test for association. For the paired study design, the standard analysis technique is the paired samples t-test or the repeated measures ANOVA. Correlational studies are usually quantified via Pearson's correlation coefficient, and may be extended into a linear regression setting. All these procedures are *parametric* procedures which make certain distributional assumptions about the data; if these assumptions are not fulfilled, corresponding non-parametric techniques, such as the rank-sum test, signed ranks test, evaluation of Spearman's rank correlation coefficient or bootstrapping procedures can often work quite well. All procedures may be easily implemented using standard statistical software and all will yield an assessment of statistical significance as measured by the *p*-value and given by the software.

While the *p*-value facilitates an inference of statistical significance or otherwise, it does not give us a measure of precision in our results. This is another side to inferential testing. We may find in our sample an effect of a given size, but it would not be realistic to expect that an effect of precisely that magnitude exists in the wider population. While we will never know what the size of the population effect actually is, it is possible to derive a *confidence interval* (CI), also known as an *uncertainty interval*, for it. This is often interpreted loosely as the range of values within which we can state to a certain degree of confidence (conventionally 95%) that a population value lies. More formally, if we were to repeat the procedure many times, then the range of values determined each time would contain the true population parameter on 95% of occasions.

CIs do not quantify the strength of evidence against the null hypothesis, as the *p*-value does, but instead give a measure of the precision of an estimate (for example, the difference between, or ratio of, the mean values in treatment groups). Nowadays CIs (and the estimate of effect around which they are fitted) are increasingly reported alongside *p*-values in statistical testing and appear to many to be more informative and easier to understand. While they do not form part of a hypothesis test, most statistical software will automatically report a CI as a by-product of the hypothesis test output.

There is an exact correspondence between CIs and the corresponding *p*-value: a 95% CI that excludes the key value 0 (for a difference between study groups) or 1 (for a ratio between study groups) corresponds to a significant effect at the 5% significance level (i.e. a *p*-value that is less than 0.05). Conversely, a 95% CI that includes a key value corresponds to a *p*-value that is statistically non-significant at the 5% significance level (i.e. is 0.05 or greater). Atkinson et al[5] investigated the effect of various factors on risk of surgical site infection during spinal surgery and in a typical presentation of tabulated results (below), reported statistics from a model including both *p*-values and CIs. Note that the spinal levels factor, which is significant according to the *p*-value (0.019), has an odds ratio with an associated CI of 1.04 to 1.54, which excludes the key value (for a ratio) of 1; while the spinal region factor, which is non-significant according to the *p*-value (0.103), has an odds ratio with an associated CI of 0.71 to 44.3, which includes the key value. This table is also a good example of how an effect of relatively small magnitude (each additional spinal level is associated with a 26% increase in odds of infection) may be significant; whereas an effect of large magnitude (surgery performed in the thoracic, rather than non-thoracic region is associated with about a fivefold increase in odds of infection) may be non-significant.

Selection of an appropriate statistical test for a given study design is not always straightforward and requires careful consideration of study parameters. No single test is suitable for all types of studies. For grouped studies, such as the unpaired and paired designs discussed above, we may need to consider, for example, the number of groups, the size of the groups, the distribution of data, the independence of units and the presence or absence of confounding factors in selecting a test. For ungrouped studies, such as correlational studies, we may wish to consider whether our data is in the form of raw data or rank orderings; and for ungrouped studies assessing multiple factors, we may wish to consider whether we are potentially overfitting our data (modelling noise rather than signal) by attempting to analyse too many factors for a sample of a given size.

Special measures are needed for complex designs: for example when data is clustered (such as patients within hospital wards, or anatomical sites within patients); when multiple tests are being conducted (such as may arise in studies with multiple outcome measures, where the key treatment variable has multiple levels, or where separate analyses are conducted on sub-groups of individuals and/or at multiple time points); when outcomes are not known exactly (such as when an outcome is the time to an adverse event in patients who are monitored infrequently); when a series of observations are made on the same patients; or when it is required to determine a synthesised estimate of effect from multiple studies. Certain

Table 1: extract from data table reported by Atkinson et al[5]

Factor/covariate	p value	Odds ratio	95% CI for odds ratio
Number of spinal levels	0.019	1.26	(1.04, 1.54)
Primary spinal region-non-thoracic (reference) Thoracic	0.103	5.59	(0.71, 44.3)

data features, such as the presence of extensive missing or invalid values or outliers may also lead to a requirement for additional statistical processes. In all such situations it is recommended that the advice of a biomedical statistician is sought.

REFERENCES

1. Stephenson, J. (2022). Descriptive presentation of wound care data. World Council of Enterostomal Therapists Journal. 42, 3, p. 30-33
2. Miller C, McGuiness W, Wilson S, Cooper K, Swanson T, Rooney D, Piller N, Woodward M (2017). Concordance and acceptability of electric stimulation therapy: a randomised controlled trial. Journal of Wound Care Vol. 26, No. 8. https://doi-org.libaccess.hud.ac.uk/10.12968/jowc.2017.26.8.508
3. Lenselink E, Andriessen A (2011). A cohort study on the efficacy of a polyhexanide-containing biocellulose dressing in the treatment of biofilms in wounds. Journal of Wound Care 20 (534)
4. Gunes EA, Eren MA, Koyuncu I, Taskin A, Sabuncu T (2018). Investigation of galectin-3 levels in diabetic foot ulcers. Journal of Wound Care (27); 12
5. Atkinson R, Stephenson J, Jones A, Ousey K. An assessment of key risk factors for surgical site infection in patients undergoing surgery for spinal metastases J Wound Care 2016; 25(S9); S30-S34