Artigo original

Testes estatísticos no tratamento de feridas

Palavras-chave testes estatísticos inferenciais, valor de p, intervalos de confiança, rácio de probabilidade
Como referência 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
Submetido 6 Novembro 2022, Aceite 1 Dezembro 2022

Os estudos de tratamento de feridas geralmente produzem dados quantitativos; e esses dados requerem alguma forma de análise. Quase todos os estudos incluem um resumo descritivo dos participantes, tal como foi discutido no meu artigo anterior[1]. No entanto, muitos estudos também requerem alguma forma de testes estatísticos inferenciais; geralmente se o objetivo é o de generalizar os resultados da amostra a uma população.

Os diferentes desenhos de estudo requerem métodos de ensaio diferentes, embora o objetivo básico na maioria dos casos seja idêntico. Isto serve para avaliar a *significância* do *efeito* em avaliação; isto é, para estabelecer se qualquer efeito que vemos na nossa amostra de dados é um reflexo do estado das coisas na população-mãe, ou se é provável que este não seja mais do que uma variação aleatória nos nossos dados. Um efeito poderia ser uma diferença observada entre grupos de estudo (como a quantidade de biofilme presente num grupo de controlo e num grupo em que os pacientes são tratados com um penso antimicrobiano), a diferença entre uma medida tomada na linha de base e no pós-intervenção de um único grupo de pacientes (tais como os níveis de dor durante e após a cirurgia), uma relação observada entre duas variáveis (como o grau de utilização da tecnologia de saúde móvel e o conhecimento sobre cuidados com feridas) ou muitas outras quantidades.

O estabelecimento da significância estatística requer um teste de uma hipótese. Normalmente testamos uma *hipótese nula* (ou seja, sem efeito); por exemplo, que a diferença de meios na população é zero. Miller et al[2] conduziram um ensaio piloto cego aleatório controlado (RCT) para examinar a concordância e a aceitabilidade da terapia de estimulação elétrica (EST) em pacientes com úlceras venosas de perna (VLU) e que não tinham tolerado compressão moderada a alta. O tratamento foi testado contra a hipótese nula de nenhum efeito; ou seja, que a concordância com o tempo total recomendado para o tratamento sob controlo ou sob EST era a mesma. Num estudo de coorte sobre a eficácia de um penso biocelulósico contendo poli-hexanida em 28 pacientes, Lenselink e Andriessen[3] mediram (entre outras quantidades) a percentagem de tecido de granulação e de tecido amarelo, testando várias hipóteses relacionadas com diferenças nos resultados dos pacientes entre a linha de base e as 24 semanas. Gunes et al[4], num estudo de pacientes diabéticos e não diabéticos, analisaram vários biomarcadores, principalmente a galectina-3 (a qual promove a angiogénese e a formação de novos vasos) e testaram várias hipóteses relacionadas com as relações entre a galectina-3 e vários outros biomarcadores.

A significância é quantificada utilizando o valor *p* familiar (se não for amplamente compreendido), o qual é uma *probabilidade condicional*: a probabilidade de que os resultados observados, ou algo mais extremo, terem sido obtidos, dado que a hipótese nula é verdadeira. Na linguagem quotidiana é referida vagamente como "a oportunidade de uma oportunidade": ou seja, a probabilidade de os nossos dados terem surgido como resultado de uma variação natural e não porque exista realmente um efeito. Convencionalmente, um valor *p* inferior a 0,05 (5%) é considerado como que indicando um resultado de significância estatística (ao nível de 5% de significância) e uma consequente rejeição da hipótese nula de nenhum efeito. O estudo de Miller et al produziu um valor *p* de 0,671 para o seu resultado primário; o que indica provas insuficientes para uma diferença entre os grupos de tratamento. O estudo de Lenselink e Andriessen produziu um valor *p* inferior a 0,04 para um teste que compara a percentagem média de tecido de granulação da linha de base com 24 semanas de pós-tratamento; indicando provas de um efeito de tratamento. O estudo de Gunes et al produziu um valor *p* inferior a 0,001 num teste de correlação entre os biomarcadores de proteína galectina-3 e C-reativa, indicando fortes evidências de uma relação significativa entre estes dois biomarcadores.

O RCT realizado por Miller et al e o teste de alterações a partir da linha de base conduzido por Lenselink e Andriessen são exemplos de desenhos de estudo agrupados: um *estudo não emparelhado*, no qual são comparados dois grupos de estudo não relacionados, no caso de Miller et al e um desenho de *estudo emparelhado*, no qual as medidas são tomadas em duas ocasiões a partir do mesmo grupo de pacientes, no caso de Lenselink e Andriessen (a palavra "emparelhado" refere-se ao facto de cada paciente contribui com um par de leituras, não que esteja envolvido um par de grupos). Uma outra variante deste estudo surge quando são tomadas medidas de pacientes que correspondem a variáveis demográficas ou a variáveis relacionadas com a saúde. O estudo de Gunes et al é um exemplo de um desenho correlacional utilizando dados não agrupados. Estes três desenhos são, provavelmente, as escolhas mais comuns de desenho de estudo no tratamento de feridas, embora possam ser encontrados vários outros tipos de estudo, que também visam inferir resultados a partir de dados de uma amostra para uma população, incluindo, por exemplo, testes de concordância, estudos de rastreio/diagnóstico, análise de sobrevivência e meta-análise.

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

Os resultados numéricos de estudos não pareados podem ser comparados utilizando o teste t de amostras independentes, ou, se estiverem envolvidos mais de dois grupos, um teste como a análise de variância (ANOVA). Os resultados categóricos são geralmente analisados utilizando o teste de qui-quadrado para a associação. Para o desenho de um estudo emparelhado, a técnica de análise padrão é a do teste t de amostras emparelhadas ou as medidas repetidas ANOVA. Os estudos correlacionais são geralmente quantificados através do coeficiente de correlação de Pearson e podem ser estendidos para uma definição de regressão linear. Todos estes procedimentos são procedimentos *paramétricos* que fazem certas suposições distributivas sobre os dados; se estas suposições não forem cumpridas, as técnicas não paramétricas correspondentes, tais como o teste de rank-sum, o teste de rankings assinado, a avaliação do coeficiente de correlação de rank de Spearman ou os procedimentos de bootstrapping podem com frequência funcionar bastante bem. Todos os procedimentos podem ser facilmente implementados utilizando software estatístico padrão e todos produzirão uma avaliação da significância estatística, medida pelo valor *p* dado pelo software.

Embora o valor *p* facilite uma dedução da significância estatística ou não, este não nos dá uma medida de precisão nos nossos resultados. Este é outro lado dos testes inferenciais. Podemos encontrar na nossa amostra um efeito de uma determinada dimensão, mas não seria realista esperar que existisse um efeito de tal magnitude na população em geral. Embora nunca saberemos qual é realmente a dimensão do efeito populacional, é possível derivar para ele um *intervalo de confiança* (IC), também conhecido como *intervalo de incerteza*. Lsto é frequentemente interpretado vagamente como a gama de valores dentro da qual podemos afirmar com um certo grau de confiança (convencionalmente de 95%) que um valor populacional se encontra. Mais formalmente, se repetirmos o procedimento muitas vezes, então o intervalo de valores determinado em cada vez iria conter o verdadeiro parâmetro populacional em 95% das ocasiões.

Os IC não quantificam a força da prova contra a hipótese nula, como o valor *p* faz, mas em vez disso dão uma medida da precisão de uma estimativa (por exemplo, a diferença entre, ou a razão dos valores médios nos grupos de tratamento). Atualmente, os IC (e a estimativa do efeito em torno do qual se encaixam) são cada vez mais relatadas conjuntamente com os valores *p* nos testes estatísticos e parecem ser, para muitos, mais informativas e mais fáceis de compreender. Embora não façam parte de um teste de hipóteses, a maioria do software estatístico reportará automaticamente um IC como um subproduto do resultado do teste de hipóteses.

Existe uma correspondência exata entre o IC e o valor *p* correspondente: um IC 95% que exclui o valor chave 0 (para uma diferença entre grupos de estudo) ou 1 (para um rácio entre grupos de estudo) corresponde a um efeito significativo ao nível de significância de 5% (ou seja, um valor *p* inferior a 0,05). Por outro lado, um IC de 95% que inclua um valor chave corresponde a um valor *p* estatisticamente não significativo ao nível de 5% de significância (ou seja, é 0,05 ou superior). Atkinson et al[5] investigaram o efeito de vários fatores no risco de infeção do local cirúrgico durante a cirurgia da coluna vertebral e numa apresentação típica de resultados tabelados (abaixo), relataram estatísticas de um modelo que incluía tanto valores *p* como IC. Note-se que o fator dos níveis da coluna vertebral, que é significativo de acordo com o valor de *p* (0,019), tem um rácio de probabilidade com um IC associado de 1,04 a 1,54, que exclui o valor chave (para um ratio) de 1; enquanto o fator da região espinal, que não é significativo de acordo com o valor de *p* (0,103), tem um rácio de probabilidade com um IC associado de 0,71 a 44,3, que inclui o valor chave. Esta tabela é também um bom exemplo de como um efeito de magnitude relativamente pequena pode ser significativo (cada nível adicional da coluna vertebral está associado a um aumento de 26% nas probabilidades de infeção); enquanto um efeito de grande magnitude (a cirurgia realizada na região torácica, em vez de não torácica, está associado a um aumento de cerca de cinco vezes nas probabilidades de infeção) pode não ser significativo.

A seleção de um teste estatístico apropriado para um determinado desenho de estudo nem sempre é linear e direto e requer uma cuidadosa consideração dos parâmetros do estudo. Nenhum teste único é adequado para todos os tipos de estudos. Para estudos agrupados, tais como os desenhos não emparelhados e emparelhados discutidos anteriormente, podemos ter de considerar, por exemplo, o número de grupos, o tamanho dos grupos, a distribuição dos dados, a independência das unidades e a presença ou a ausência de fatores de confusão na seleção de um teste. Para estudos não agrupados, tais como estudos correlacionais, podemos querer considerar se os nossos dados se encontram sob a forma de dados em bruto ou de ordenações de classificação; e para estudos não agrupados que avaliam múltiplos fatores, podemos considerar se estamos potencialmente a sobreajustar os nossos dados (modelando o ruído em vez do sinal), tentando analisar uma excessiva quantidade de fatores para uma amostra de um determinado tamanho.

Medidas especiais são necessárias para desenhos complexos: por exemplo, quando os dados são agrupados (tais como pacientes dentro de enfermarias hospitalares, ou locais anatómicos em pacientes); quando são realizados múltiplos testes (tais como podem surgir em estudos com medidas de resultados múltiplos, onde a variável-chave de tratamento tem múltiplos níveis, ou onde análises separadas são conduzidas em subgrupos de indivíduos e/ou em múltiplos momentos no tempo); quando os resultados não são conhecidos exatamente (por exemplo, quando um resultado é o momento de um acontecimento adverso em pacientes que são monitorizados com pouca frequência); quando é efetuada uma série de observações sobre os mesmos pacientes; ou quando é necessário determinar uma estimativa sintetizada do efeito a partir de estudos múltiplos. Certas características dos dados, tais

Quadro 1: extrato da tabela de dados comunicados por Atkinson et al[5]

Fator/covariável	valor p	Rácio de probabilidade	95% CI para rácio de probabilidade
Número de níveis da coluna vertebral	0,019	1,26	(1,04, 1,54)
Região espinal primária - não torácica (referência) Torácica	0,103	5,59	(0,71, 44,3)

como a presença de extensos valores em falta ou inválidos ou discrepantes, podem também levar a uma exigência de processos estatísticos adicionais. Em todas estas situações é recomendado que se procure o conselho de um estatístico biomédico.

REFERÊNCIAS

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