Article original

Tests statistiques dans le traitement des plaies

Mots clés tests statistiques inférentiels, valeur p , intervalles de confiance, rapport de cotes

Pour les références 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

Soumis le 6 novembre 2022, Accepté le 1er décembre 2022

Les études sur les soins des plaies produisent généralement des données quantitatives, et ces données nécessitent une certaine forme d'analyse. Presque toutes les études comprennent un résumé descriptif des participants, comme je l'ai expliqué dans mon précédent article[1]. Cependant, de nombreuses études nécessitent également une certaine forme de tests statistiques inférentiels, particulièrement si l'intention est de généraliser les résultats de l'échantillon à une population.

Les différents modèles d'étude nécessitent des méthodes de test différentes, bien que l'objectif de base soit le même dans la plupart des cas. Il s'agit d'évaluer *l'importance* de *l'effet* en question, c'est-à-dire d'établir si l'effet que nous observons dans notre échantillon de données reflète l'état de la situation dans la population mère ou s'il s'agit simplement d'une variation aléatoire dans nos données. Un effet peut être une différence observée entre des groupes d'étude (comme la quantité de biofilm présente dans un groupe témoin et un groupe où les patients sont traités avec un pansement antimicrobien), la différence entre une mesure prise au départ et après l'intervention d'un seul groupe de patients (comme les niveaux de douleur pendant et après l'opération), une relation observée entre deux variables (comme l'étendue de l'utilisation de la technologie de santé mobile et la connaissance du soin des plaies) ou de nombreuses autres quantités.

L'établissement de la signification statistique nécessite de tester une hypothèse. Nous testons généralement une *hypothèse nulle* (d'absence d'effet); par exemple, que la différence de moyennes dans la population est nulle. Miller et al.2 ont mené un essai pilote contrôlé et randomisé (ECR) en simple aveugle pour examiner la concordance et l'acceptabilité de la thérapie par stimulation électrique (TSE) chez les patients qui n'avaient pas toléré une compression modérée à élevée et souffrant d'ulcères de jambe veineux (UJV). Le traitement a été testé par rapport à l'hypothèse nulle d'absence d'effet, c'est-à-dire que la concordance avec la durée totale de traitement recommandée sous traitement témoin ou sous TSE était la même. Lenselink et Andriessen[3] ont mesuré (entre autres) le pourcentage de tissu de granulation et de tissu jaune dans une étude de cohorte sur l'efficacité d'un pansement de biocellulose contenant du polyhexanide chez 28 patients, en testant plusieurs hypothèses relatives aux différences de résultats pour les patients entre le début de l'étude et 24 semaines. Dans une étude portant sur des patients diabétiques et non diabétiques, Gunes et a.l[4] ont analysé plusieurs biomarqueurs, principalement la galectine-3 (qui favorise l'angiogenèse et la formation de nouveaux vaisseaux) et ont testé plusieurs hypothèses relatives aux relations entre la galectine-3 et divers autres biomarqueurs.

La signification est quantifiée à l'aide de la valeur *p* familière (bien que peu comprise), qui est une *probabilité conditionnelle*: la probabilité que les résultats observés, ou quelque chose de plus extrême, auraient été obtenus, étant donné que l'hypothèse nulle est vraie. Dans le langage courant, on parle de "hasard du hasard", c'est-à-dire de la probabilité que nos données soient tombées comme elles sont tombées à cause d'une variation naturelle et non parce qu'un effet existe réellement. Par convention, une valeur *p* inférieure à 0,05 (5%) est considérée comme indiquant un résultat de signification statistique (au niveau de signification de 5%) et un rejet consécutif de l'hypothèse nulle d'absence d'effet. L'étude de Miller et al. a donné une valeur *p* de 0,671 pour son résultat primaire, indiquant une insuffisance de preuve d'une différence entre les groupes de traitement. L'étude de Lenselink et Andriessen a donné une valeur *p* inférieure à 0,04 pour un test comparant le pourcentage moyen de tissu de granulation entre le début de l'étude et 24 semaines après le traitement, prouvant un effet du traitement. L'étude de Gunes et al. a donné une valeur *p* inférieure à 0,001 dans un test de corrélation entre les biomarqueurs de la galectine-3 et de la protéine C-réactive, montrant des preuves solides d'une relation significative entre ces biomarqueurs.

L'ECR mené par Miller et al. et le test des changements par rapport au départ mené par Lenselink et Andriessen sont des exemples de modèles d'étude groupés: une *étude non appariée*, dans laquelle deux groupes d'étude non reliés sont comparés, dans le cas de Miller et al. ainsi qu' un modèle d' *étude apparié*, dans lequel les mesures sont prises à deux occasions sur le même groupe de patients, dans le cas de Lenselink et Andriessen (le mot "apparié" fait référence au fait que chaque patient fournit une paire de lectures, et non qu'une paire de groupes est impliquée). Une autre variante de cette étude se présente lorsque les mesures sont prises auprès de patients appariés sur des variables démographiques ou liées à la santé. L'étude de Gunes et al. est un exemple de conception corrélationnelle utilisant des données non groupées. Ces trois modèles sont probablement les choix les plus courants en matière de conception d'études dans le domaine du traitement des plaies, bien que l'on puisse trouver divers autres types d'études, qui visent également à déduire les résultats d'un échantillon de données pour les appliquer à une population, y compris par exemple des tests de concordance, des études de dépistage/diagnostic, des analyses de survie et des méta-analyses.

John Stephenson
PHD FRSS (GradStat) CMath(MIMA)
Maître de conférence en statistiques biomédicales
Université de Huddersfield, Royaume-Uni
Courriel: J.Stephenson@hud.ac.uk

Les résultats numériques des études non appariées peuvent être comparés à l'aide du test t des échantillons indépendants ou, si plus de deux groupes sont concernés, d'un test tel que l'analyse de la variance (ANOVA). Les résultats catégoriels sont généralement analysés à l'aide du test d'association du Khi deux. Pour le modèle d'étude apparié, la technique d'analyse standard est le test t des échantillons appariés ou l'ANOVA à mesures répétées. Les études corrélationnelles sont généralement quantifiées par le coefficient de corrélation de Pearson, et peuvent être étendues à un cadre de régression linéaire. Toutes ces procédures sont des procédures *paramétriques* qui peuvent redistribuer les hypothèses concernant les données. Si ces hypothèses ne sont pas satisfaites, les techniques non paramétriques correspondantes, telles que le test de la somme des rangs, le test des rangs marqués, l'évaluation du coefficient de corrélation des rangs de Spearman ou les procédures de bootstrapping, peuvent souvent donner d'assez bons résultats. Toutes les procédures peuvent être facilement mises en œuvre à l'aide d'un logiciel statistique standard et toutes donneront lieu à une évaluation de la signification statistique, mesurée par la valeur *p* fournie par le logiciel.

Si la valeur *p* facilite l'inférence de la signification statistique ou non, elle ne nous donne pas une mesure de la précision de nos résultats. C'est un autre aspect des tests inférentiels. Nous pouvons trouver dans notre échantillon un effet d'une taille donnée, mais il ne serait pas réaliste de s'attendre à ce qu'un effet de cette ampleur existe dans la population générale. Bien que nous ne sachions jamais quelle est la taille réelle de l'effet de population, il est possible d'en déduire un *intervalle de confiance* (IC), également appelé *intervalle d'incertitude*. Ce terme est souvent interprété de manière large comme la plage de valeurs dans laquelle nous pouvons affirmer avec un certain degré de confiance (conventionnellement 95%) qu'une valeur de la population se situe. Plus formellement, si nous devions répéter la procédure de nombreuses fois, la plage de valeurs déterminée à chaque fois contiendrait le véritable paramètre de la population dans 95% des cas.

Les IC ne quantifient pas la robustesse des données contre l'hypothèse nulle, comme le fait la valeur *p*, mais donnent plutôt une mesure de la précision d'une estimation (par exemple, la différence, ou le rapport entre des valeurs moyennes dans les groupes de traitement). Aujourd'hui, les IC (et l'estimation de l'effet autour de laquelle ils sont ajustés) sont de plus en plus souvent indiqués au côté des valeurs *p* dans les tests statistiques et semblent pour beaucoup plus informatifs et plus faciles à comprendre. Bien qu'ils ne fassent pas partie d'un test d'hypothèse, la plupart des logiciels statistiques signalent automatiquement un IC comme un produit dérivé dans les résultats du test d'hypothèse.

Il existe une correspondance exacte entre les IC et la valeur p correspondante: un IC à 95% qui exclut la valeur clé 0 (pour une différence entre les groupes d'étude) ou 1 (pour un rapport entre les groupes d'étude) correspond à un effet significatif au niveau de signification de 5% (c'est-à-dire une valeur *p* inférieure à 0,05). Inversement, un IC à 95% qui comprend une valeur clé correspond à une valeur *p* qui est statistiquement non significative au niveau de signification de 5% (c'est-à-dire qu'elle est égale ou supérieure à 0,05). Atkinson et al.5 ont étudié l'effet de divers facteurs sur le risque d'infection du site chirurgical lors d'une opération de la colonne vertébrale et, dans une présentation courante de résultats tabulés (ci-dessous), ont exposé les statistiques d'un modèle comprenant à la fois les valeurs *p* et les IC. Il est à noter que le facteur «niveaux rachidiens», qui est significatif selon la valeur *p* (0,019), présente un ratio de cotes avec un IC associé de 1,04 à 1,54, ce qui exclut la valeur clé (pour un ratio) de 1; tandis que le facteur «région rachidienne», qui n'est pas significatif selon la valeur *p* (0,103), présente un ratio de cotes avec un IC associé de 0,71 à 44,3, ce qui inclut la valeur clé. Ce tableau est également un bon exemple de la façon dont un effet d'une ampleur relativement faible (chaque niveau rachidien supplémentaire est associé à une augmentation de 26% des chances d'infection) peut être significatif, alors qu'un effet d'une grande ampleur (une chirurgie pratiquée dans la région thoracique plutôt que non thoracique est associée à une augmentation d'environ cinq fois les chances d'infection) peut être non significatif.

La sélection d'un test statistique approprié pour un plan d'étude donné n'est pas toujours simple et nécessite un examen attentif des paramètres de l'étude. Aucun test unique ne convient à tous les types d'études. Pour les études groupées, comme les modèles non appariés et appariés discutés ci-dessus, nous pouvons avoir besoin de considérer, par exemple, le nombre de groupes, la taille des groupes, la distribution des données, l'indépendance des unités et la présence ou l'absence de facteurs de confusion dans le choix d'un test. Pour les études non groupées, telles que les études corrélationnelles, nous pouvons nous demander si nos données se présentent sous la forme de données brutes ou de classements; et pour les études non groupées évaluant de multiples facteurs, nous pouvons nous demander si nous ne sommes pas en train de surajuster nos données (en modélisant le bruit plutôt que le signal) en essayant d'analyser trop de facteurs pour un échantillon d'une taille donnée.

Des mesures spéciales sont nécessaires pour les modèles complexes: par exemple lorsque les données sont regroupées (comme les patients dans les services d'un hôpital, ou les sites anatomiques chez les patients); lorsque des tests multiples sont effectués (comme cela peut se produire dans les études avec des mesures de résultats multiples, lorsque la variable de traitement clé a plusieurs niveaux, ou lorsque des analyses séparées sont effectuées sur des sous-groupes d'individus et/ou à des points de temps multiples); lorsque les résultats ne sont pas connus avec précision (par exemple, lorsqu'un résultat est le délai d'apparition d'un événement indésirable chez des patients qui ne sont pas suivis fréquemment); lorsqu'une série d'observations est effectuée sur les mêmes patients; ou lorsqu'il est nécessaire de déterminer une synthèse de l'estimation de l'effet à partir de plusieurs études.

Tableau 1: extrait du tableau de données exposées par Atkinson et al.5

Facteur/covariable	valeur p	Rapport de cotes	IC à 95% pour le rapport de cotes
Nombre de niveaux rachidiens	0,019	1,26	(1,04, 1,54)
Région spinale primaire - non thoracique (référence) Thoracique	0,103	5,59	(0,71, 44,3)

Certaines caractéristiques des données, telles que la présence d'un grand nombre de valeurs manquantes ou invalides ou de valeurs aberrantes, peuvent également nécessiter des processus statistiques supplémentaires. Dans toutes ces situations, il est recommandé de demander l'avis d'un statisticien biomédical.

RÉFÉRENCES

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