Considérer les bonnes différences entre bras d'étude

P. Chevalier

En 2009, la publication en ligne[1] de l'évaluation de l'efficacité d'une émulsion antivieillissement de la peau, émulsion contenant différents ingrédients d'extraits naturels et de peptides, montrait versus véhicule seul un bénéfice du produit en termes d'amélioration des rides faciales. Les auteurs de cette étude, Watson et coll., notaient pour leur RCT de 6 mois de traitement, une amélioration significative (p=0,013) dans le groupe traitement actif et l'absence de différence significative (p=0,11) dans le groupe contrôle versus examen initial. Ils concluaient à une amélioration clinique significative avec le produit actif.

Dans la version finalement publiée dans la présentation imprimée de la revue[2], cette conclusion était corrigée en : tendance mais non significative d'amélioration clinique en termes de rides faciales, avec un p=0,10 pour la comparaison entre les 2 groupes. Dans un article récent paru dans le BMJ[3], Bland et Altman expliquent l'erreur faite initialement par Watson et coll. Ceux-ci avaient considéré à tort que l'absence d'une différence significative dans le groupe contrôle versus valeur initiale et que la présence d'une différence significative dans le groupe produit actif versus valeur initiale permettait de conclure à une supériorité du produit actif sur le comparateur.

Dans une RCT, l'évaluation doit porter non sur les différences obtenues respectivement dans les bras d'étude versus valeurs basales mais bien sur la différence entre ces groupes sur l'ampleur de la différence respectivement observée.

Bland et Altman montrent, exemple à l'appui dans une simulation, que la probabilité d'une différence significative dans un des deux groupes mais pas dans l'autre dans une étude au protocole semblable à celui de l'étude de Watson, est de 38% et non de 5% ; les valeurs p des différences dans les groupes versus valeurs initiales ne peuvent pas être utilisées comme points de comparaison pour montrer la supériorité d'un traitement. Ils montrent que si la différence est importante entre valeur finale et valeur initiale dans un bras, presque tous les tests seront significatifs. Si la différence est faible par contre, presque tous les tests seront non significatifs.

Dans la dernière version publiée, celle imprimée, Watson et coll. introduisent une nouvelle confusion. Ils mentionnent 43% de sujets avec amélioration clinique des rides faciales dans le groupe produit actif (p=0,013) et seulement 22% dans le groupe contrôle (p=NS). La valeur p ne porte pas sur le nombre de répondeurs mais sur un score moyen dans le groupe basé sur le Griffiths photonumeric scale for photoaged skin (score de 0 à 8, 8 étant le plus sévère) pour les rides du visage. Les chiffres précis ne sont pas mentionnés dans la publication et les auteurs ne définissent également pas ce qu'ils estiment être une différence cliniquement pertinente. L'amélioration d'un score, en elle-même, ne dit pas grand-chose de l'amélioration perçue par le praticien ni par le patient. Si répondeur signifie personne diminuant son score ne fût-ce que d'un point, donner un nombre de répondeurs n'est d'aucune utilité ; mentionner une différence pour le nombre de répondeurs entre les deux bras n'est alors également d'aucune utilité. Dans le protocole d'étude, 3 autres paramètres sont décrits comme devant être scorés également : dyspigmentation, grade de vieillissement lié à l'exposition solaire, rugosité cutanée au palper ; aucune mention de résultats pour ces critères n'est faite et les résultats pour ces critères importants ne sont donc pas repris pour juger de la différence entre les 2 bras d'étude.

Des résultats doivent être exprimés, dans ce type d'étude, en différence de score moyen, mentionner la différence au score qui a une pertinence clinique et/ou permet d'affirmer qu'il y a une réponse au traitement. Un exemple en est le score de Hamilton dans les études évaluant les antidépresseurs dans le traitement de la dépression. Le score de Hamilton évalue la sévérité de la dépression et une diminution de ce score de 50% est considérée comme une réponse au traitement et une diminution du score ≤10 comme une rémission. Les résultats d'une étude avec un antidépresseur doivent donc comprendre les résultats d'évolution au score d'Hamilton et les taux de réponse et de rémission. L'ensemble de ces résultats permettra alors une interprétation pour la validité statistique mais aussi pour la pertinence clinique. Dans ce domaine des antidépresseurs, cette interprétation a permis de déterminer, entre autres, que l'ampleur de l'efficacité des antidépresseurs est proportionnelle à la sévérité initiale.

Dans l'étude sur l'émulsion antivieillissement cutané, le protocole était initialement défectueux, ne permettant pas, au départ déjà, de tirer des conclusions utiles pour la pratique et les erreurs d'analyse des différences (intragroupes mal comparées, intergroupes non faites) ont fait sombrer définitivement l'intérêt d'une telle publication.

Références

1. Watson RE, Ogden S, Cotterell LF, et al. A cosmetic 'anti-ageing' product improves photoaged skin: a double-blind, randomized controlled trial. Br J Dermatol 2009;doi10.1111/j.1365-2133.-2009.09216.x.
2. Watson RE, Ogden S, Cotterell LF, et al. A cosmetic 'anti-ageing' product improves photoaged skin: a double-blind, randomized controlled trial. Br J Dermatol 2009;161:419-26.
3. Bland JM, Altman DG. Comparisons within randomised groups can be very misleading. BMJ 2011;342:d561.