Hoe studie-armen correct met elkaar vergelijken?

P. Chevalier

In 2009 verscheen een online publicatie van een studie over het effect van een anti-verouderingscrème[1]. De crème bevatte verschillende natuurlijke extracten en peptiden. De auteurs besloten dat de actieve crème in vergelijking met alleen het vehikel een verbetering van gezichtsrimpels veroorzaakte. In hun RCT van zes maanden stelden ze vast dat de actieve behandelingsgroep significant verbeterde (p=0,013) en dat er in de controlegroep geen significant verschil was (p=0,11) in vergelijking met de initiële waarden. Ze besloten hieruit dat het actieve product een klinisch significante verbetering veroorzaakte.
In de gedrukte versie waren de auteurs echter genoodzaakt om hun besluit te wijzigen in 'een niet-significante trend voor klinische verbetering op het vlak van gezichtsrimpels' (p=0,10 voor de vergelijking tussen de twee groepen)[2].

In 2011 legden de statistici Bland en Altman in de BMJ uit wat er fout liep in de online publicatie[3]. Op basis van een significant verschil ten opzichte van de aanvangswaarde in de actieve groep en geen significant verschil ten opzichte van de aanvangswaarde in de controlegroep, hadden de auteurs ten onrechte besloten dat de actieve crème superieur was aan het vehikel alleen.
In een RCT is het essentieel dat de evaluatie gebaseerd is op de vergelijking 'tussen' de groepen en niet op de verschillen tussen eindwaarden en initiële waarden voor elke studie-arm afzonderlijk.
Door middel van een simulatie toonden Bland en Altman aan dat in een studie-opzet zoals de hierboven vermelde RCT, de kans op een significant verschil in één studie-arm en geen significant verschil in de andere studie-arm, 38% bedraagt en niet 5%. We kunnen de p-waarden voor de verschillen tussen aanvangs- en eindwaarden in de afzonderlijke studie-armen dus niet gebruiken om de superioriteit van een behandeling aan te tonen. Bland en Altman bewezen ook dat wanneer de aanvangs- en eindscores in een studie-arm ver uiteen liggen, bijna alle testen significant zullen zijn. Wanneer de aanvangs- en eindwaarden weinig verschillen, zullen bijna alle testen binnen de studie-arm niet-significant zijn.
De gedrukte, gecorrigeerde versie van de RCT over het effect van de anti-verouderingscrème, bracht echter een nieuwe verwarring aan het licht. De auteurs vermeldden dat er bij 43% van de deelnemers in de actieve groep een klinische verbetering van de gezichtsrimpels optrad (p=0,013) en slechts bij 22% in de controlegroep (p-waarde niet-significant). De p-waarde had hier geen betrekking op het aantal responders, maar op een gemiddelde score op de fotonumerieke vragenlijst van Griffith (0 tot 8 met 8 als meest ernstig) voor gezichtsrimpels. In de publicatie van de RCT geven de auteurs geen exacte cijfers en ze vermelden evenmin een drempel voor een klinisch relevant verschil.

De verbetering van de score op zich zegt weinig over hoe de arts en de patiënt de verbetering percipieerden. Indien 'responder' zou betekenen 'een deelnemer wiens score vermindert' (al is het maar met één punt), dan heeft het weinig zin om het aantal responders als uitkomstmaat te nemen. Het verschil weergeven tussen het aantal responders in beide studie-armen is dan evenmin zinvol. In het onderzoeksopzet beschreven de auteurs drie andere te evalueren parameters: verandering in pigmentatie, mate van huidveroudering omwille van blootstelling aan de zon en ruwe huid. Voor deze uitkomstmaten vermeldden de auteurs geen enkel resultaat en ze gebruikten deze eindpunten ook niet voor de evaluatie van het verschil tussen beide studie-armen.
Bij een dergelijk onderzoeksprotocol moet men de resultaten uitdrukken in verschil van de gemiddelde scores met daarbij de vermelding welk verschil in score klinisch relevant is en/of toelaat om te besluiten dat een behandeling werkzaam is.
Een voorbeeld hiervan is de Hamilton-score in studies over antidepressiva voor de behandeling van depressie. Met de Hamilton-score kan men de ernst van de depressie beoordelen. Een vermindering in score van 50% betekent respons op de behandeling en een daling van de score tot 10 of minder betekent remissie. De resultaten van een studie over het effect van antidepressiva moeten dus de evolutie van de Hamilton-score tonen, evenals het aantal deelnemers met respons en met remissie. Al deze gegevens samen laten toe om de resultaten correct te interpreteren, niet alleen op statistische validiteit maar ook op klinische relevantie. Op die manier heeft men kunnen vaststellen dat de effectgrootte van de antidepressiva evenredig is aan de ernst van de depressie bij aanvang.

In de RCT met het anti-verouderingsmiddel liep de studie-opzet reeds mank van bij het begin, waardoor zinvolle conclusies voor de praktijk niet meer mogelijk waren. De fout bij het analyseren van de verschillen (slecht vergeleken binnen de groepen en niet vergeleken tussen de groepen) heeft definitief aangetoond dat een dergelijke publciatie nutteloos is.

Referenties

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.