

Evaluatie van de werkzaamheid van een behandeling: valkuilen en voorstellen Het voorbeeld van de antidepressiva

Pierre Chevalier, Centre Académique de Médecine Générale, Université Catholique de Louvain
Tekst onder de verantwoordelijkheid van de Franstalige redactie

Een editoriaal in *Minerva* van januari 2013 wees op de valkuilen bij de evaluatie van de werkzaamheid van antidepressiva¹, specifiek als men in meta-analyses de resultaten van studies samenvoegt die niet dezelfde uitkomstmaten gebruiken. Het gebruik van verschillende vragenlijsten voor de evaluatie van depressie is hiervan een voorbeeld: de Hamilton Depression Rating Scale (HDRS) en de Montgomery-Åsberg depression rating scale (MADRS). In dit laatste geval drukten de auteurs van een meta-analyse de resultaten uit als gestandaardiseerd gemiddeld verschil (SMD). Om de effectgrootte van behandelingen onderling te kunnen vergelijken, introduceerden Cohen et al.² drempelwaarden voor de SMD: 0,20 voor weinig effect, 0,50 voor een matig effect en 0,80 voor een groot effect. Nadien gebruikten andere auteurs deze drempelwaarden (waarschijnlijk ten onrechte) om een verschil in klinische werkzaamheid aan te tonen. In hun meta-analyse volgden Kirsch et al.³ de auteurs van de NICE-aanbevelingen in hun keuze voor een afkappunt van 0,50 om te bewijzen dat een behandeling werkzaam is. De keuze voor een drempelwaarde waarvan de klinische relevantie voor discussie vatbaar is, zorgde ervoor dat de betrouwbaarheid van de conclusies van deze meta-analyse in vraag werd gesteld.

In een editoriaal van de *BMJ* gebruiken Turner et al.⁴ een metafoor om aan te tonen dat de werkwijze van NICE⁵ en van Kirsch et al.³ niet adequaat is. Ze drukken de evaluatie van de werkzaamheid van antidepressiva uit in aantal liters 'd-sap' (naar analogie met de 'd' van Cohen, waarbij de 'd' staat voor de gemeten effectgrootte). In hun glas met alle overzichtsartikelen over de relatieve werkzaamheid van antidepressiva, meten Turner et al. 0,41 liter d-sap, terwijl ze in het glas van de FDA 0,31 liter bekomen. Ze besluiten dus dat het glas van de FDA minder vol is dan hun glas. Ze erkennen evenwel dat 0,31 liter een meetbare en belangrijke hoeveelheid is. Als Kirsch et al.³ in hun meta-analyse 0,32 liter d-sap bekomen, is het glas onvoldoende vol (ligt onder het arbitraire afkappunt van $\geq 0,5$) en is hun conclusie dat het glas virtueel geen sap bevat. Turner et al. zijn akkoord dat het 'glas' antidepressiva verre van vol is, maar kunnen niet aannemen dat het volledig leeg is.

Zoals vermeld in het editoriaal van *Minerva*¹, kunnen we het gebruik van de HDRS voor sommige van deze aspecten in vraag stellen. Zweedse onderzoekers stelden vast dat de precisie van de HDRS afneemt naarmate de ernst van de depressie vermindert, dus naarmate de patiënt verbetert⁶. Het lijkt dus weinig betrouwbaar om op die manier de klinische evolutie te vergelijken van patiënten die starten op verschillende niveaus van depressie. Deze onderzoekers besluiten dat de geringe werkzaamheid kan te wijten zijn aan de beperkte precisie van het meetinstrument en aan de lage sensitiviteit om verandering te meten, vooral bij milde tot matige depressie.

Dankzij bovenvermelde publicaties kunnen we in de overzichtsartikelen over de werkzaamheid van antidepressiva 2 problemen identificeren: enerzijds het gebruik van verschillende scorelijsten en/of uitkomstmaten waardoor het moeilijk is om

de resultaten samen te voegen, en anderzijds het gebruik van een evaluatiescore die onvoldoende de gemiddelde verschillen in de evolutie van de behandelde patiënten weergeeft. Gibbons et al.⁷ vonden hiervoor een oplossing... voorwaardelijk! Ze onderzochten de werkzaamheid op korte termijn (6 weken) van fluoxetine en venlafaxine. Daarvoor selecteerden ze alleen de RCT's van de firma's die deze antidepressiva commercialiseren, wat een zeer grote beperking inhoudt, maar hen tegelijkertijd toegang gaf tot de individuele patiëntgegevens. Naargelang de studies gaat het over volwassenen, adolescenten of ouderen. De auteurs includeerden alleen de studies die dezelfde evaluatiescore gebruikten (HDRS voor volwassenen en ouderen, Children's Depression Rating Scale-Revised (CDRS-R) voor adolescenten). Vermits ze beschikten over de individuele patiëntgegevens analyseerden ze de resultaten aan de hand van vrij complexe Bayesiaanse statistische methodes, met een regressiemodel op 3 niveaus (tijdstip van meting, patiënt, studie). Op die manier konden ze bv. voor volwassenen en ouderen een verandering aantonen in HDRS-score van -11,82 met antidepressiva en van -9,26 met placebo, met een schatting van de maximale marginale likelihood van -2,55 (standaardfout 0,20; $p < 0,001$) en een relatieve risicotoename van 27,7%.

Op basis van deze en andere resultaten besluiten de auteurs dat fluoxetine en venlafaxine bij volwassenen en ouderen werkzaam zijn dan placebo op het vlak van respons en remissie. De verschillen in respons (50% vermindering in ernst, NNT 5,41) en in remissie (HDRS-score < 8 , NNT 7,3) lijken gunstig, maar het verschil in HDRS-score van -2,5 blijft onder de door NICE⁵ vastgelegde klinisch relevante drempel van 3 punten (komt overeen met een SMD van 0,50)... We moeten echter voorzichtig zijn bij de interpretatie van de resultaten omwille van de specifieke selectie van de studies, de evaluatieduur van 6 weken en de beperking tot slechts 2 antidepressiva. De werkwijze bij de analyse van de resultaten, vooral de beschikbaarheid van de individuele patiëntgegevens, is evenwel een aspect dat bij andere evaluaties van de werkzaamheid van antidepressiva nuttig kan zijn.

Referenties

- Chevalier P. Evaluatie van een behandeling en betrouwbaarheid van de meetinstrumenten. [Editoriaal] *Minerva* 2013;12(1):1.
- Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. New York: Lawrence Erlbaum Associates, 1988.
- Kirsch I, Deacon BJ, Huedo-Medina TB, et al. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med* 2008;5:260-7.
- Turner EH, Rosenthal R. Efficacy of antidepressants. *BMJ* 2008;336:516-7.
- NICE. Depression: Management of depression in primary and secondary care. Clinical Practice Guideline Number 23. National Institute for Health and Care Excellence. London, 2004.
- Isacsson C, Adler M. Randomized clinical trials underestimate the efficacy of antidepressants in less severe depression. *Acta Psychiatr Scand* 2011;125:453-9.
- Gibbons RD, Hur K, Brown CH, et al. Benefits from antidepressants: synthesis of 6-week patient-level outcomes from double-blind placebo-controlled randomized trials of fluoxetine and venlafaxine. *Arch Gen Psychiatry* 2012;69:572-9.