

Het belang van interbeoordelaarsbetrouwbaarheid in psychofarmacologisch onderzoek¹

S. Berendsen, H.L. Van, M.J. van Tricht, L.M.A. Verdegaa, M. Blankers, L. de Haan

- Achtergrond** Er is reden tot zorg over de achteruitgang van de gerapporteerde effectiviteit van antidepressiva en antipsychotica in klinische studies. Een van de mogelijke oorzaken is onvoldoende training om een adequate interbeoordelaarsbetrouwbaarheid (IBB) te bereiken van de gebruikte uitkomstmaten. Het is echter onbekend in hoeverre IBB en trainingsprocedures worden gerapporteerd.
- Doel** Bepalen in welke mate IBB en trainingsprocedures worden gerapporteerd in publicaties van dubbelblinde gerandomiseerde onderzoeken (RCT's) naar antidepressiva of antipsychotica.
- Methode** We extraheerden alle dubbelblinde RCT's uit vijf recente meta-analyses naar antidepressiva of antipsychotica. Verder zochten we op PubMed naar RCT's over antidepressiva in de periode januari 2016–2020, en naar RCT's over antipsychotica in de periode januari 2000–januari 2019.
- Resultaten** In 179 dubbelblinde RCT's naar antidepressiva werd bij 4,5% een IBB-coëfficiënt en bij 27,9% een vorm van training gerapporteerd. In 207 dubbelblinde RCT's naar antipsychotica werd bij 11,2% een IBB-coëfficiënt en werd bij 34,8% over trainingsprocedures gerapporteerd.
- Conclusie** IBB en trainingsprocedures worden zeer weinig gerapporteerd in RCT's naar antidepressiva of antipsychotica. Vanwege de negatieve implicaties van onvoldoende betrouwbaarheid dient men trainingsprocedures en betrouwbaarheidsonderzoek uit te voeren en te rapporteren. Tijdschriftredacties moeten rapportage verplicht stellen.

In de afgelopen decennia is er zorg ontstaan over de afgenomen effectiviteit van de behandeling met antidepressiva of antipsychotica zoals deze gerapporteerd wordt in dubbelblind gerandomiseerd onderzoek (Gelenberg e.a. 2008; Kemp e.a. 2010). Verder leidt ondanks forse investeringen in de ontwikkeling van psychofarmaca slechts 9% van de fase I-studies uiteindelijk tot een klinische toepassing; de helft komt niet tot klinische toepassing vanwege niet-significante verschillen tussen medicatie en placebo in fase II-onderzoeken (Hurko & Ryan 2005).

Een lage betrouwbaarheid van metingen zou bij kunnen dragen aan dit probleem. Juist voor onderzoek waarbij gebruik wordt gemaakt van observationele instrumenten zijn het trainen van onderzoekers en het bereiken van voldoende interbeoordelaarsbetrouwbaarheid (IBB) essentieel. Onvoldoende training leidt dan ook tot een matige IBB bij gebruik van observationele instrumenten (Loevdahl & Friis 1996; Muller & Wetzel 1998).

Gevolgen van onvoldoende betrouwbare metingen

Onvoldoende betrouwbare metingen kunnen een enorme impact hebben op studieresultaten. Ze leiden tot een verlaagde statistische zeggingskracht (studie-power) en daarmee een hoger risico op fout-negatieve bevindingen (Perkins e.a. 2000; Muller & Szegedi 2002). Dit werkt als volgt: analyse van de statistische zeggingskracht is gebaseerd op de geobserveerde variantie in eerdere studies. De geobserveerde variantie bestaat uit ware variantie en de meetfout; hoe hoger de meetfout, des te hoger de variantie en des te lager de betrouwbaarheid. Ter illustratie: een verbetering van de IBB van 0,5 naar 0,8 leidt tot een reductie van de benodigde studiepopulatie van 40% (Perkins e.a. 2000).

Dit effect wordt versterkt door een verwachtingsbias van beoordelaars en de druk om een toereikend aantal patiënten in studies te includeren. Hierdoor ontstaat een risico op artificieel verhoogde symptoomscores bij aan-

AUTEURS

Steven Berendsen, ten tijde van het schrijven van dit artikel: arts in opleiding tot psychiater, Arkin Amsterdam; thans: psychiater, GGz Centraal, en promovendus, afd. Psychiatrie, UMC Amsterdam, locatie AMC.

Rien Van, opleider psychiatrie (directeur behandelzaken NPI, onderdeel van Arkin, Amsterdam).

Mirjam J. van Tricht, klinisch neuropsycholoog, GGz Centraal, en senior wetenschappelijk onderzoeker, afd. Psychiatrie, UMC locatie AMC, Amsterdam.

Loek M.A. Verdegaal, arts niet in opleiding tot specialist, Groene Hart Ziekenhuis, Gouda.

Matthijs Blankers, senior wetenschappelijk onderzoeker, Trimbos-instituut, Utrecht, en Arkin, Amsterdam.

Lieuwe de Haan, psychiater en hoogleraar Psychotische Stoornissen, UMC locatie AMC, afd. Psychiatrie, Vroege Psychose, en Arkin, Amsterdam.

Correspondentie

Steven Berendsen (s.berendsen@amsterdamumc.nl).

Geen strijdige belangen meegedeeld.

Het artikel werd voor publicatie geaccepteerd op 18-5-2021.

Citeren

Tijdschr Psychiatr. 2021;63(9):838-643

vang van een studie (Kobak e.a. 2010). Na inclusie daalt de gemeten ernst van de symptoomscores dan spontaan omdat ze in werkelijkheid lager zijn.

Onbetrouwbare metingen dragen ook bij aan dit probleem: door metingen die niet volledig betrouwbaar zijn, includeert men altijd een gedeelte patiënten onterecht doordat er foutief een hoge score is gemeten. Dit betekent dat een patiënt geïncludeerd wordt die niet voldoet aan het gestelde ernstcriterium, maar in werkelijkheid een lagere ernstscore heeft (Schatzberg & Kraemer 2000). Deze vertekende ernstscores zijn geassocieerd met een hogere placeborespons, met als gevolg een verlaagde kans op het identificeren van effectverschillen tussen medicatie en placebo. Bovendien kan tijdens de studie beoordelaarsdeviatie ontstaan, dit betekent dat er een verandering in het scoringsgedrag over de tijd optreedt met tot gevolg een toename van onbetrouwbare metingen.

Belang IBB en training

Reeds 20 jaar geleden werd krachtig onderstreept dat het bepalen van IBB cruciaal is voor gerandomiseerd en observationeel wetenschappelijk onderzoek (Kraemer & Thiemann 1989; Kraemer 2000). Echter, de rapportage van IBB-coëfficiënten of trainingsprocedures bleek destijds ondermaats: slechts in 19% van de studies naar depressieve stoornissen gaf men een meting van de betrouwbaarheid en in 9,5% rapporteerde men een trainingsprocedure (Vacha-Haasem 1999; Mulsant e.a. 2002). Deze studies leverden geen precieze informatie over de inclusie van trainingsprocedures of hoogte van IBB-coëfficiënten in gerandomiseerde dubbelblinde behandelstudies naar antidepressiva of antipsychotica.

In de afgelopen twee decennia is de rapportage van onderzoeksmethodologie echter sterk verbeterd, onder andere door de introductie van de CONSORT-richtlijnen voor gerandomiseerd onderzoek. De nog te beantwoorden vraag is daarom of de rapportage van IBB en trainingsprocedures ook is toegenomen. Wij hebben dit bepaald voor publicaties van dubbelblinde gerandomiseerde onderzoeken (RCT's) naar antidepressiva en antipsychotica.

METHODE

We verzamelden de RCT's naar antidepressiva op basis van de meta-analyse van Cipriani e.a. (Cipriani e.a. 2018) waarin RCT's uit de periode 2000-2016 zijn opgenomen. Verder zochten twee auteurs (SB en LV) onafhankelijk van elkaar op PubMed naar RCT's naar antidepressiva in de periode januari 2016-januari 2020. De antipsychoticastudies werden geëxtraheerd uit vier meta-analyses van dubbelblinde RCT's naar antipsychotica die zijn gepubliceerd sinds het jaar 2000 (Leucht e.a. 2012; 2017; Fusar-Poli e.a. 2015; Samara e.a. 2016). Tot slot zocht één auteur (SB) op PubMed naar aanvullende RCT's naar antipsychotica in de periode januari 2000-januari 2019. De inclusiecriteria van onze studie waren: primaire diagnose depressieve stoornis of schizofreniespectrumstoornis, leeftijd > 18 jaar, antidepressiva of antipsychotica, en de uitkomstmaat gemeten met een klinische observatieschaal. Bij de zoektocht naar literatuur in PubMed gebruikten we alle individuele soorten antidepressiva en antipsychotica als trefwoord. Details van de zoekstrategie en het overzicht met alle geïncludeerde studies zijn opvraagbaar bij de eerste auteur.

Procedure

Twee auteurs (SB en LV) extraheerden onafhankelijk van elkaar de volgende variabelen uit de gepubliceerde manuscripten en supplementen: aanwezigheid van een IBB-coëfficiënt: intraklassecorrelatiecoëfficiënt (ICC), Cohens kappa, Krippendorffs *alpha of agreement coefficient*.

Verder verzamelden we informatie over correlatiecoëfficiënten en het percentage overeenkomst als maat voor IBB, centrale beoordelaars van observationele lijsten en elke gerapporteerde trainingsprocedure voor beoordelaars.

Voor het verkrijgen van de data uit de manuscripten las één van de auteurs (SB of LV) de volledige methodesectie en supplementen. Vervolgens werd dit gecontroleerd door met de zoekfunctie in de pdf na te gaan of in het artikel en eventueel supplement de termen *reliability, ICC, IRR, rater, training, consistency, agreement, Cohen's kappa, correlation, appendix, supplement* voorkwamen. Alle geïncludeerde RCT's naar antipsychotica werden door twee auteurs (SB en LV) gecontroleerd op aanwezigheid van deze variabelen. De RCT's naar antidepressiva werden op een alternatieve wijze gecontroleerd. Bij de start en halverwege de datacollectie bepaalden we de IBB tussen de twee auteurs (SB en LV) van de genoemde extractiemethode. Bij de start van de studie bepaalden we deze voor 25 manuscripten en halverwege de studie voor 20 andere manuscripten.

Bij de start van de studie was Cohens kappa voor aanwezigheid van een IBB-coëfficiënt 1,0, en voor trainingsprocedures 0,915 (95%-betrouwbaarheidsinterval 0,719-1,000). Halverwege de datacollectie bepaalden we de IBB opnieuw tussen de twee auteurs (SB en LV) en was de Cohens kappa voor aanwezigheid van een IBB-coëfficiënt en trainingsprocedures in beide gevallen 1,0. Alle RCT's naar antidepressiva werden vervolgens door een auteur (SB of LV) gecontroleerd. Voor IBB-analyses tussen auteurs SB en LV gebruikten we Statistical Package for the Social Sciences (SPSS) versie 26.

RESULTATEN

Studiekenmerken

In totaal vonden we 179 dubbelblinde RCT's naar antidepressiva, waarvan 168 studies uit de netwerkmeta-analyse van Cipriani e.a. en 11 studies die werden toegevoegd na het zoeken op PubMed. Daarnaast vonden we in totaal 207 dubbelblinde RCT's naar antipsychotica, 108 studies werden geëxtraheerd uit de vier meta-analyses en 99 studies uit het zoeken op PubMed.

In alle geïncludeerde studies gebruikte men observationele instrumenten om de ernstscores van symptomen weer te geven, zoals de *Hamilton Depression Rating Scale* (HDRS; Sharp 2015) of de *Positive and Negative Syndrome Scale* (PANSS; Kay e.a. 1988).

IRR-coëfficiënten en training

Van de 179 RCT's naar antidepressiva gaf men in 4,5% een IBB-coëfficiënt weer en beschreef men in 27,9% dat een trainingsprocedure voor de beoordelaars van de observationele vragenlijst was gevolgd. In totaal rapporteerde men in 11,2% van de 207 RCT's naar antipsychotica een IBB-coëfficiënt en 34,8% gaf informatie over een trainingsprocedure. **Tabel 1** geeft een overzicht. Opvallend was verder dat bij de RCT's naar antidepressiva en antipsychotica in 5,0% (n = 9) en respectievelijk 6,8% (n = 14) werd gerapporteerd dat de betrouwbaarheid was gemeten, zonder weergave van een coëfficiënt of andere score. Bij de RCT's naar antidepressiva vonden we bij 1 RCT gebruik de ICC én percentage overeenkomst, bij de RCT's naar antipsychotica werd in 2,4% (n = 5) alleen een percentage overeenkomst gebruikt. Er werd in geen enkele RCT een betrouwbaarheidsinterval van de IBB-score gerapporteerd.

Tabel 1. Inclusie van interbeoordelaarsbetrouwbaarheid (IBB)-coëfficiënten en trainingsprocedures

	Antidepressivastudies (n = 179)	Antipsychoticastudies (n = 207)
IBB-coëfficiënt	4,5% (n = 8)	11,2% (n = 23)
Trainingsprocedures	27,9% (n = 50)	34,8% (n = 72)
Centrale beoordelaars	0,6% (n = 1)	2,9% (n = 6)
Specifieke analyses		
% met ICC	75% (n = 6)	78,3% (n = 18)
% met Cohens kappa	25% (n = 2)	21,7% (n = 5)
% met percentage overeenkomst	12,5% (n = 1)	2,4% (n = 5)
% met correlatiecoëfficiënt	0%	1,9% (n = 4)

DISCUSSIE

Nauwelijks gegevens over IBB en training

Uit ons onderzoek blijkt dat in de meerderheid gegevens ontbreken over IBB en trainingsprocedures in dubbelblinde RCT's naar antidepressiva en antipsychotica. Ongeveer 20 jaar geleden werden vergelijkbare bevindingen gedaan, waarbij men in RCT's voor de behandeling voor depressieve stoornissen in 19% een IBB-coëfficiënt en in 9% een training rapporteerde (Vacha-Haasem e.a. 1999; Mulsant e.a. 2002).

Ondanks duidelijke aanbevelingen zijn er geen aanwijzingen dat er een verbetering is opgetreden in de rapportage van IBB-coëfficiënten in de afgelopen twee decennia (Gelenberg e.a. 2008; Kemp e.a. 2010). Daarenboven werd in ruim 4% van de RCT's naar antipsychotica de IBB met een correlatiecoëfficiënt of percentage overeenkomst weergegeven. Correlatiecoëfficiënten houden echter geen rekening met interindividuele overeenkomst tussen metingen en het percentage overeenkomst corrigeert niet voor overeenkomst op basis van toeval, en zijn daarom niet wenselijk. Betere berekeningen voor de IBB zijn de ICC, Cohens kappa en Krippendorffs alfa (Hayes & Krippendorff 2007; McHugh 2012; Koo & Li 2016).

Lage effectiviteit door lage betrouwbaarheid

Uit onze resultaten concluderen wij tevens dat de verminderde en afnemende effectiviteit van antidepressiva en antipsychotica op theoretisch niveau kunnen worden toegeschreven aan een lage betrouwbaarheid. Dit werkt als volgt: een studie wordt uitgevoerd zonder trainingsprocedures, met een statistische zeggingskracht van 80% leidend tot een steekproefgrootte van 125.

Echter, de ICC zonder trainingsprocedures is gemiddeld rond de 0,3-0,6 (laag tot matig) en hier is geen rekening mee gehouden in de analyse, waardoor de statistische zeggingskracht automatisch lager wordt vanwege een verhoogde variantie in metingen (Berendsen e.a. 2020a, b; 2021). Dit heeft tot gevolg dat de kans op het vinden van significante verschillen tussen medicatie en placebo kleiner is.

Uiteraard zijn hier andere factoren ook van belang, zoals verschillen over de tijd van eigenschappen van geïncludeerde patiënten, of selectie van onderzoekslocaties.

Verklaring gebrekkige rapportage

Een belangrijke vraag is waarom de IBB en trainingsprocedures niet adequaat worden gerapporteerd in dubbelblinde RCT's naar antidepressiva en antipsychotica. Wellicht leeft de gedachte dat voor klinisch ervaren beoordelaars een training voor het betrouwbaar meten van symptomen niet noodzakelijk is. Echter, uit onderzoek naar de PANSS blijkt dat ervaring geen garantie is voor betrouwbare metingen (Muller & Wetzel 1998). Ook is mogelijk dat men trainingen te duur, te tijdsintensief of niet belangrijk vindt, vooral in grote inter-

nationale studies zou men de implementatie te lastig kunnen vinden. Toch kan men substantiële besparingen realiseren door de betrouwbaarheid van metingen te vergroten. Immers, de statistische zeggingskracht wordt daarmee vergroot, waardoor een kleinere steekproef nodig is om effectiviteit aan te kunnen tonen (Perkins e.a. 2000).

Lage IBB als breder probleem

Het probleem van een afnemende effectiviteit van behandelingen over de tijd wordt in de psychiatrie niet alleen in studies over psychofarmaca gerapporteerd, het is ook gesignaleerd bij psychotherapeutische interventies, bijvoorbeeld bij cognitieve gedragstherapie (Johnsen & Friberg 2015). De auteurs geven daarvoor echter andere verklaringen zoals minder goede scholing in de uitvoering van de therapie, toename van medicatiegebruik tijdens studies of veranderingen in de diagnostiek van depressies over de tijd. Maar gezien het veelvuldig gebruik van observationele meetinstrumenten zoals de HDRS zou onvoldoende IBB ook hier een rol kunnen spelen. Het bereiken van voldoende betrouwbaarheid van observationele metingen in de psychiatrie is zeker mogelijk, maar training is noodzakelijk (Muller e.a. 1998; Kobak e.a. 2004).

In andere medische vakgebieden spelen ook problemen rond het bereiken van adequate IBB. Bijvoorbeeld in de dermatologie, waarbij de IBB voor het bepalen van de ernst van atopisch eczeem of hidradenitis suppurativa zeer variabel blijkt (ICC: 0,07-0,72) (Thorlacius e.a. 2019). De auteurs adviseren om trainingen uit te breiden en meer bedside teaching toe te passen voordat de onderzochte instrumenten worden toegepast in de dermatologische praktijk.

Ook in de chirurgie blijkt het behoud over de tijd van adequate IBB zeer beperkt (ICC: -0,17) zonder trainingsprocedures van laparoscopische vaardigheden door chirurgen (Gawad e.a. 2019). De onderzoekers concluderen dat er meer aandacht moet zijn voor onderhoudstraining om voldoende betrouwbaarheid te behouden.

Het is ons onbekend of er in deze vakgebieden van de geneeskunde aandacht is voor trainingsprocedures en bepaling van IBB in gerandomiseerd onderzoeksverband.

Beperkingen en sterke punten

Onze studie heeft een aantal beperkingen. Ten eerste hebben wij geen contact gelegd met individuele auteurs van RCT's voor aanvullende informatie betreffende training en bepaling van IBB. Daarom is het mogelijk dat trainingsprocedures of IBB-metingen zijn uitgevoerd zonder dat deze vermeld zijn. We achten dit echter niet zo waarschijnlijk omdat auteurs meestal geneigd zijn om sterke aspecten van hun onderzoek onder de aandacht te brengen.

Ten tweede geven de geïncludeerde artikelen wisselende en vaak slechts beperkte informatie over de trainingsprocedures. Om deze reden kunnen wij niet nagaan wat

de rol is van bijvoorbeeld het aantal beoordelaars per studie, hun professionele achtergrond, de aard en duur van de training of de methode van het bepalen van IBB. De kracht van ons onderzoek is de inclusie van een groot aantal dubbelblinde RCT's betreffende zowel antidepressiva en antipsychotica, waarbij wij denken dat we alle belangrijke Engelstalige psychofarmacologische publicaties van de afgelopen 19 jaar voor RCT's betreffende antipsychotica en 20 jaar voor RCT's betreffende antidepressiva hebben opgenomen.

Implicaties

Op basis van onze bevindingen doen we een aantal dringende aanbevelingen.

In klinisch onderzoek is het noodzakelijk trainingsprocedures toe te passen om voldoende IBB te bereiken, te meten en deze vervolgens ook te rapporteren. Bepaling van de IBB kan met een statistische analyse zoals de ICC, Cohens kappa of Krippendorffs alpha of agreement coefficient-1 (McHugh 2012; Koo & Li 2016; Zec e.a. 2017).

Training kan bijvoorbeeld bestaan uit een cursus interviewvaardigheden, met daarbij live-interviews van patiënten gevolgd door directe feedback van ervaren beoordelaars. Nog beter is om te werken met centrale, adequaat getrainde beoordelaars, in plaats van met ad-hocbeoordelaars ter plaatse (Kobak e.a. 2004). Een eerdere studie toonde aan dat bij gebruik van dergelijke centrale beoordelaars een lagere meetfout en verminderde placeborespons optreden (Kobak e.a. 2010). Door het maken van video-opnames van metingen tijdens de studie en deze opnieuw te laten scoren door andere beoordelaars, kan beoordelaarsdeviatie worden voorkomen. Vervolgens kan men de oorspronkelijke metingen zo nodig corrigeren en kan men aanvullende training aanbieden aan matig presterende beoordelaars. Tot slot achten wij het noodzakelijk dat rapportage over trainingsprocedures en IBB in de CONSORT-richtlijnen en bijbehorende aanvullende documenten en in de auteursinstructies van wetenschappelijke tijdschriften wordt opgenomen.

CONCLUSIE

Onderzoekers en redacties van wetenschappelijke tijdschriften onderkennen het belang van adequate trainingsprocedures en IBB nog steeds te weinig. De gevolgen van onbetrouwbare metingen voor de interpretatie van studie-uitkomst is immers al geruime tijd bekend, maar desondanks aanwezig. In ieder geval worden trainingsprocedures en IBB nog steeds te weinig beschreven in publicaties betreffende RCT's naar antidepressiva of antipsychotica. Dat baart zorgen want onbetrouwbare metingen en onvoldoende training hebben potentieel een grote impact op studieresultaten en kunnen een rol spelen bij het niet vinden van verschillen tussen placebo en antidepressiva en antipsychotica.

Wij onderstrepen daarom het belang van standaardisering van het uitvoeren en rapporteren van IBB en trainingsprocedures in behandelstudies – bijvoorbeeld door deze in de CONSORT-richtlijnen op te nemen.

NOOT

- 1 Dit artikel is deels gebaseerd op twee eerdere publicaties:
 - Berendsen S, Van HL, Verdegaa LMA, van Tricht MJ, Blankers M, de Haan L. Burying our heads in the sand: the neglected importance of reporting inter-rater reliability in antipsychotic medication trials. *Schizophr Bull* 2020a; 46: 1027-9.
 - Berendsen S, Verdegaa LMA, van Tricht MJ, Blankers M, Van HL, de Haan L. An old but still burning problem: inter-rater reliability in antidepressant trials. *J Affect Disord* 2020b; 276: 748-51.

LITERATUUR

- Berendsen S, Kapitein P, Schirmbeck F, e.a. Pre-training inter-rater reliability of clinical instruments in an international psychosis research project. *Schizophr Res* 2021; 230: 104-7.
- Cipriani A, Furukawa TA, Salanti G, e.a. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: A systematic review and network meta-analysis. *Lancet* 2018; 391: 1357-66.
- Fusar-Poli P, Papanastasiou E, Stahl D, e.a. Treatments of negative symptoms in schizophrenia: Meta-analysis of 168 randomized placebo-controlled trials. *Schizophr Bull* 2015; 41: 892-9.
- Gawad N, Fowler A, Mimeault R, e.a. The inter-rater reliability of technical skills assessment and retention of rater training. *J Surg Educ* 2019; 76: 1088-93.
- Gelenberg AJ, Thase ME, Meyer RE, e.a. The history and current state of antidepressant clinical trial design: A call to action for proof-of-concept studies. *J Clin Psychiatry* 2008; 69: 1513-28.
- Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* 2007; doi: 10.1080/19312450709336664.
- Hurko O, Ryan JL. Translational research in central nervous system drug discovery. *NeuroRx* 2005; 2: 671-82.
- Johnsen TJ, Friberg O. The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: A meta-analysis. *Psychol Bull* 2015; 141: 747-68.
- Kay SR, Opler LA, Lindenmayer JP. Reliability and validity of the positive and negative syndrome scale for schizophrenics. *Psychiatry Res* 1988; 23: 99-110.
- Kemp AS, Schooler NR, Kalali AH, e.a. What is causing the reduced drug-placebo difference in recent schizophrenia clinical trials and what can be done about it? *Schizophr Bull* 2010; 36: 504-9.
- Kobak KA, Engelhardt N, Williams JB, e.a. Rater training in multicenter clinical trials: Issues and recommendations. *J Clin Psychopharmacol* 2004; 24: 113-7.
- Kobak KA, Leuchter A, DeBrot D, e.a. Site versus centralized raters in a clinical depression trial: Impact on patient selection and placebo response. *J Clin Psychopharmacol* 2010; 30: 193-7.
- Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016; 15: 155-63.
- Kraemer HC. Pitfalls of multisite randomized clinical trials of efficacy and effectiveness. *Schizophr Bull* 2000; 26: 533-41.
- Kraemer HC, Thieman S. A strategy to use soft data effectively in randomized controlled clinical trials. *J Consult Clin Psychol* 1989; 57: 148-54.
- Leucht S, Leucht C, Huhn M, e.a. Sixty years of placebo-controlled antipsychotic drug trials in acute schizophrenia: Systematic review, bayesian meta-analysis, and meta-regression of efficacy predictors. *Am J Psychiatry* 2017; 174: 927-42.

- Leucht S, Tardy M, Komossa K, e.a. Antipsychotic drugs versus placebo for relapse prevention in schizophrenia: A systematic review and meta-analysis. *Lancet* 2012; 379: 2063-71.
- Loevdahl H, Friis S. Routine evaluation of mental health: Reliable information or worthless 'guesstimates'? *Acta Psychiatr Scand* 1996; 93: 125-8.
- McHugh ML. Interrater reliability: The kappa statistic. *Biochem Med (Zagreb)* 2012; 22: 276-82.
- Muller MJ, Rossbach W, Dannigkeit P, e.a. Evaluation of standardized rater training for the Positive And Negative Syndrome Scale (PANSS). *Schizophr Res* 1998; 32: 151-60.
- Muller MJ, Szegeedi A. Effects of interrater reliability of psychopathologic assessment on power and sample size calculations in clinical trials. *J Clin Psychopharmacol* 2002; 22: 318-25.
- Muller MJ, Wetzel H. Improvement of inter-rater reliability of panss items and subscales by a standardized rater training. *Acta Psychiatr Scand* 1998; 98: 135-9.
- Mulsant BH, Kastango KB, Rosen J, e.a. Interrater reliability in clinical trials of depressive disorders. *Am J Psychiatry* 2002; 159: 1598-600.
- Perkins DO, Wyatt RJ, Bartko JJ. Penny-wise and pound-foolish: The impact of measurement error on sample size requirements in clinical trials. *Biol Psychiatry* 2000; 47: 762-6.
- Samara MT, Dold M, Gianatsi M, e.a. Efficacy, acceptability, and tolerability of antipsychotics in treatment-resistant schizophrenia: A network meta-analysis. *JAMA Psychiatry* 2016; 73: 199-210.
- Sharp R. The Hamilton Rating Scale For Depression. *Occup Med (Lond)* 2015; 65: 340.
- Thorlacius L, Garg A, Riis PT, e.a. Inter-rater agreement and reliability of outcome measurement instruments and staging systems used in hidradenitis suppurativa. *Br J Dermatol* 2019; 181: 483-91.
- Vacha-Haasem T, Ness C, Nilsson J, e.a. Practices regarding reporting of reliability coefficients: A review of three journals. *The Journal of Experimental Education* 1999; 67: 335-41

SUMMARY

The importance of inter-rater reliability in psychopharmacological trials

S. Berendsen, H.L. Van, M.J. van Tricht, L.M.A. Verdegaal, M. Blankers, L. de Haan

- Background** There are concerns about the declining efficacy of antidepressants and antipsychotics in clinical trials. A potential cause may be found in poor training practices to achieve sufficient inter-rater reliability (IRR). However, it is unknown whether IRR and training procedures are currently reported.
- Aim** To determine the proportion of publications concerning double-blind randomized controlled trials (RCTs) investigating antipsychotics or antidepressants that report IRR and training procedures.
- Method** We extracted all double-blind RCTs from five large meta-analyses concerning antidepressants and antipsychotics. Further, we conducted a Medline-search for double-blind RCTs investigating antidepressants from January 2016 - January 2020, and antipsychotics from January 2000 - January 2019.
- Results** In 179 double-blind RCTs with antidepressants, only 4.5% reported an IRR coefficient whereas 27.9% reported on training procedures. Further, in 207 double-blind RCTs with antipsychotics, 11.2% reported an IRR coefficient and 34.8% reported training procedures.
- Conclusion** There is a substantial lack of reporting IRR and training procedures in RCTs with antidepressants and antipsychotics. Considering the implications of insufficient IRR, it is necessary to conduct and report training procedures and IRR. Reporting IRR and training procedures should be made mandatory by editorial boards of scientific journals.