STAND VAN ZAKEN

Voor de kritische lezer: vragen bij gerandomiseerd onderzoek

Rolf H.H. Groenwold

Dit artikel geeft een leidraad voor het beoordelen van artikelen over gerandomiseerd onderzoek. Deze leidraad is bedoeld voor diegenen die de gepubliceerde resultaten van gerandomiseerd onderzoek mogelijk willen toepassen in de dagelijkse medische praktijk.

Om de waargenomen verschillen in gezondheidsuitkomsten tussen de onderzochte groepen met zekerheid toe te kunnen schrijven aan de behandeling, is het noodzakelijk dat de groepen vergelijkbaar zijn bij aanvang van het onderzoek, tijdens het onderzoek en aan het einde van het onderzoek, dat wil zeggen: bij het meten van de uitkomstmaat.

Voor het interpreteren van de resultaten van gerandomiseerd onderzoek en de vertaling naar de klinische praktijk is het daarnaast van belang te kijken naar het doel van het onderzoek, de vergelijking die wordt gemaakt, de omvang en richting van het waargenomen effect en op wie dat effect van toepassing is.

Gerandomiseerd onderzoek wordt beschouwd als de gouden standaard om de effecten van medische interventies, bijvoorbeeld een behandeling, te beoordelen.[1] Maar waarom is dat? En waar moet je op letten bij het lezen van een artikel over gerandomiseerd onderzoek? In dit artikel staan deze 2 vragen centraal.

DE KRACHT VAN GERANDOMISEERD ONDERZOEK

Gerandomiseerd onderzoek is een waardevol onderzoeksinstrument, omdat de groepen die worden vergeleken in een onderzoek ook werkelijk vergelijkbaar zijn. Om uitspraken te kunnen doen over de effecten van behandelingen en de causaliteit van die effecten dienen de onderzoeksgroepen die met elkaar worden vergeleken namelijk identiek te zijn en alleen te verschillen ten aanzien van de behandeling die wordt geëvalueerd (zie uitlegkader).[2] Tijdens het onderzoek zijn er 3 momenten waarop, met verschillende instrumenten, vergelijkbaarheid van onderzoeksgroepen kan worden bereikt of behouden: bij aanvang van het onderzoek, gedurende het onderzoek en aan het einde van het onderzoek (tabel).

VERGELIJKBAARHEID BIJ AANVANG VAN HET ONDERZOEK

Gerandomiseerd onderzoek heeft zijn naam te danken aan wat er bij aanvang van het onderzoek gebeurt. Als bijvoorbeeld 2 behandelingen worden vergeleken, splitst men de groep deelnemers aan het begin van het onderzoek op basis van toeval ('at random') in tweeën: het is volkomen willekeurig welke van beide behandelingen een deelnemer krijgt. De 2 groepen die op deze manier ont-

UMC Utrecht, Julius Centrum voor Gezondheidswetenschappen en Eerstelijns Geneeskunde, Utrecht.
Dr. R.H.H. Groenwold, arts-epidemioloog (r.h.h.groenwold@umcutrecht.nl).

UITLEG

Tijdreizen als metafoor van causaal onderzoek

Om de causale effecten van een medische interventie te kunnen onderzoeken, zouden we eigenlijk een tijdreismachine tot onze beschikking moeten hebben. Een groep deelnemers aan een onderzoek wordt dan blootgesteld aan een bepaalde interventie en gevolgd gedurende een vastgestelde periode. Aan het einde van die periode worden gezondheidsuitkomsten gemeten, bijvoorbeeld de kwaliteit van leven, of het aantal nieuwe gevallen van diabetes. Vervolgens gaan alle deelnemers in de tijdreismachine, reizen terug naar het tijdstip dat het onderzoek begon en doorlopen het onderzoek nog een keer op exact dezelfde manier, maar ditmaal zonder dat ze de interventie ondergaan. Gedurende de 2 periodes is alles identiek (de deelnemers, maar ook externe invloeden zoals werkstress, verkeersongevallen, verkiezingsuitslagen, de manier waarop de gezondheidsuitkomsten worden gemeten), behalve de interventie. In dat geval kan een waargenomen verschil in gezondheidsuitkomsten met zekerheid worden toegeschreven aan dat ene waarop de groepen verschillen: de interventie.

staan zullen bij aanvang van het onderzoek – naar verwachting – vergelijkbaar zijn ten aanzien van alle denkbare gemeten én ongemeten kenmerken.

Het is belangrijk dat het randomiseren eerlijk verloopt. Als bijvoorbeeld op voorhand al bekend is welke behandeling de eerstvolgende deelnemer zal krijgen, kan dat van invloed zijn op hoe of wie wordt gerekruteerd. Als bijvoorbeeld bekend is dat de eerstvolgende deelnemer een placebo zal krijgen, zou de onderzoeker kunnen uitkijken naar een relatief gezonde deelnemer. Dit kan de vergelijkbaarheid van de onderzoeksgroepen beperken.

Zelfs als de randomisatie adequaat gebeurt, kunnen er ogenschijnlijke verschillen bestaan tussen de 2 groepen. Het is echter zinloos om deze verschillen statistisch te toetsen. Wanneer een artikel toch een analyse van die verschillen beschrijft, komt dat neer op het toetsen van de hypothese dat de 2 onderzoeksgroepen voortkomen uit dezelfde populatie. Zonder te toetsen weten we al dat deze hypothese waar is –de groep deelnemers is immers op basis van toeval in tweeën gesplitst – en de uitkomst van een statistische toets zou daar niets aan mogen veranderen.[3] Zolang er maar geen consequenties worden verbonden aan de uitkomsten van dergelijke toetsen, zijn ze nog geen reden het artikel terzijde te schuiven. Gewoon negeren van die toetsen is beter.

TABEL Instrumenten om behandelgroepen in een gerandomiseerd onderzoek vergelijkbaar te maken of te houden

moment	instrument
bij aanvang van het onderzoek	randomisatie
tijdens het onderzoek	blinderen van deelnemers, hun omgeving en hun behandelaars
	randomisatie
aan het einde van het onderzoek (meten van gezondheidsuitkomsten)	standaardisatie van metingen
	blinderen van degene die gezondheidsuitkomsten meet

VERGELIJKBAARHEID GEDURENDE HET ONDERZOEK

Om een waargenomen verschil in gezondheidsuitkomsten tussen behandelgroepen toe te kunnen schrijven aan de behandeling – en niets anders dan de behandeling – dienen de onderzoeksgroepen ook tijdens het onderzoek vergelijkbaar te blijven. Dat betekent bijvoorbeeld dat de deelnemers in de verschillende onderzoeksgroepen zich, gemiddeld genomen, op dezelfde manier moeten gedragen.

Stel dat in een onderzoek naar farmacologische behandeling van obesitas vooral de mensen die de farmacologische behandeling krijgen meer gaan sporten, terwijl dit in de controlegroep niet zo is. Dit kan er mogelijk toe leiden dat er – onterecht – een te grote gewichtsverandering wordt toegeschreven aan de farmacologische behandeling. Maar als de deelnemers in beide groepen meer gaan sporten, en in dezelfde mate, dan is een waargenomen verschil tussen de groepen wél toe te schrijven aan de farmacologische behandeling. Kortom, verandering van gedrag tijdens een gerandomiseerd onderzoek is niet erg, zolang het gedrag maar in beide onderzoeksgroepen in dezelfde mate verandert.

Er zijn 2 manieren om vergelijkbaarheid van groepen tijdens het onderzoek te bereiken. Randomisatie leidt er niet alleen toe dat de onderzoeksgroepen vergelijkbaar zijn bij aanvang van het onderzoek, maar ook dat bijvoorbeeld de aanvankelijke bereidheid om het gezondheidsgedrag te veranderen – bijvoorbeeld meer gaan sporten – gelijk verdeeld is over de onderzoeksgroepen. Of die verandering werkelijk in gelijke mate optreedt, daar heeft de randomisatie geen invloed op.

Daarnaast worden de deelnemers, hun familie en vrienden en hun behandelaars vaak 'geblindeerd', zodat zij niet weten welke behandeling een deelnemer krijgt. In een onderzoek naar de effectiviteit van een farmacologische behandeling kan men 'blinderen' door bijvoorbeeld gebruik te maken van een placebo; dat is een pil die qua vóórkomen – kleur, geur, smaak, gewicht, et cetera – identiek is aan de pil met de werkzame farmacologische stof, maar die níet de farmacologische stof bevat. In een onderzoek naar de effecten van bijvoorbeeld fysiotherapie of een chirurgische ingreep is het minder duidelijk hoe een behandeling met een identiek vóórkomen, maar

zonder de werkzame component, er uit zou moeten zien. Maar ook in deze setting worden soms placebo's – zogenoemde 'sham'-interventies – gebruikt.

VERGELIJKBAARHEID AAN HET EINDE VAN HET ONDERZOEK

Aan het einde van het onderzoek worden bij alle deelnemers gezondheidsuitkomsten gemeten.

Ongeacht het type onderzoek is het van belang dat dit gestandaardiseerd en bij voorkeur geblindeerd gebeurt. Een gestandaardiseerde meting – dat wil zeggen, volgens een vaste procedure – leidt er enerzijds toe dat meetfouten kleiner worden. Hoe kleiner de invloed van meetfouten, hoe minder deelnemers nodig zijn voor het onderzoek. Anderzijds zal het voor externe beoordelaars, bijvoorbeeld lezers van een wetenschappelijk tijdschrift, ook duidelijker zijn wat de gezondheidsuitkomst precies inhoudt als deze op een gestandaardiseerde manier is gemeten.

Als er al meetfouten optreden, zouden die – idealiter – in beide groepen op dezelfde manier moeten optreden. Stel dat in een onderzoek naar gewichtsverandering het gemeten lichaamsgewicht van deelnemers die de experimentele behandeling krijgen telkens naar beneden wordt afgerond, terwijl de metingen in de controlegroep naar boven worden afgerond. Het gevolg is verre van ideaal, namelijk een vertekening (systematische fout) van het effect van de experimentele behandeling. Als daarentegen degene die de metingen uitvoert niet weet welke behandeling de deelnemer krijgt – met andere woorden: de beoordelaar is geblindeerd – kunnen er nog steeds meetfouten optreden, bijvoorbeeld omdat de gebruikte apparatuur onnauwkeurig meet, maar deze zullen dan naar verwachting gelijk verdeeld zijn over de onderzoeksgroepen.

Overigens is de noodzaak om te blinderen afhankelijk van de gezondheidsuitkomst. In het geval van mortaliteit als uitkomstmaat is er weinig ruimte voor verschillen in interpretatie en zal de uitkomst dus correct worden gemeten, zelfs als de beoordelaar weet welke behandeling een patiënt heeft gehad. Maar als de gezondheidsuitkomst bijvoorbeeld een door de deelnemer gerapporteerde pijnscore is en de deelnemer weet welke behandeling hij of zij heeft ondergaan, dan kan dat – onbewust – leiden tot selectieve 'meetfouten' die verschillen tussen de behandelgroepen.

GERANDOMISEERD ONDERZOEK BEOORDELEN

Bij het beoordelen van een artikel over gerandomiseerd onderzoek zijn er verschillende vragen die de lezer zichzelf moet stellen. Deze komen hieronder aan de orde in de volgorde waarin ze doorgaans worden beschreven in een artikel over gerandomiseerd onderzoek.

ZIJN DE ONDERZOEKSGROEPEN VERGELIJKBAAR?

De belangrijkste vraag die moet worden gesteld bij het beoordelen van een gerandomiseerd onderzoek is of de onderzoeksgroepen inderdaad vergelijkbaar zijn.[4] Zoals hiervoor uiteengezet, heeft een onderzoeker verschillende instrumenten tot zijn beschikking om vergelijkbaarheid te bereiken en te behouden: randomisatie, blindering en standaardisatie van metingen.

WAT IS HET DOEL VAN HET ONDERZOEK?

Gerandomiseerde onderzoeken kunnen worden ingedeeld aan de hand van het doel van het onderzoek: het vaststellen van de werkzaamheid ('efficacy') of de relatieve effectiviteit in de dagelijkse praktijk ('effectiveness').[5] Een andere indeling maakt onderscheid in verklarend ('explanatory') en pragmatisch ('pragmatic') onderzoek.[5,6] In de praktijk worden deze termen nogal eens als synoniemen gebruikt.

Verklarend onderzoek richt zich op de werkzaamheiden wordt daarom uitgevoerd onder sterk gecontroleerde omstandigheden om het maximale effect van de werkzame component te meten. Blindering van deelnemers en behandelaars is een belangrijk element van dergelijk onderzoek, om met zekerheid een waargenomen verschil in gezondheidsuitkomsten te kunnen toeschrijven aan de werkzame component.

Gerandomiseerd onderzoek dat zich richt op behandeleffecten in de dagelijkse praktijk (relatieve effectiviteit) wordt ook wel pragmatisch gerandomiseerd onderzoek genoemd. De in- en exclusiecriteria voor deelname zijn in een pragmatisch onderzoek vaak minder strikt dan in een verklarend onderzoek, deelnemers en behandelaars worden niet altijd geblindeerd – conform de dagelijkse praktijk – en in plaats van de weinig realistische placebo wordt de experimentele behandeling vergeleken met een realistisch behandelalternatief, bijvoorbeeld geen behandeling, een andere behandeling, of dezelfde behandeling, maar in een andere dosering.

Juist wanneer het doel van een onderzoek is om het gehele effect van een behandelstrategie te evalueren – en niet slechts de werkzaamheid van bepaalde component, bijvoorbeeld de farmacologische stof – valt te overwegen om niet te blinderen. In de praktijk weten patiënten immers ook welke behandeling zij krijgen wanneer een bepaalde strategie wordt aangeboden en zijn zij dus ook niet geblindeerd.

WELKE BEHANDELINGEN WORDEN VERGELEKEN?

Welke behandelingen worden vergeleken is aan de onderzoeker: om farmacologische werkzaamheid vast te stellen ligt een vergelijking met 'geen behandeling' of placebo voor de hand, terwijl een vergelijking tussen 2 reële behandelopties, bijvoorbeeld 2 verschillende geneesmid-

delen, de vraag beantwoordt wat de meest aangewezen behandeling in de praktijk is. Let in het laatste geval vooral ook op doseringen: is het een eerlijke vergelijking of is de controlebehandeling suboptimaal gedoseerd zodat de experimentele behandeling gemakkelijk kan winnen?

WORDEN DE GEGEVENS VAN ALLE DEELNEMERS GEBRUIKT IN DE ANALYSE?

Zoals al opgemerkt, leidt randomisatie – naar verwachting – tot onderzoeksgroepen die bij aanvang van het onderzoek vergelijkbaar zijn. Maar om die vergelijkbaarheid te behouden moeten dan wel de gegevens van alle deelnemers ook worden meegenomen in de analyse van het onderzoek. Als een deel van hen tijdens het onderzoek stopt en er daarom geen gezondheidsuitkomsten bij hen zijn gemeten, is niet eenvoudig te zeggen hoe daar mee omgegaan moet worden in de analyse. Maar als de deelnemers die gestopt zijn eenvoudigweg buiten beschouwing worden gelaten, moet dat vragen oproepen: is de vergelijkbaarheid van de groepen die worden geanalyseerd nog wel intact?[7] Of is er sprake van selectieve uitval van deelnemers waardoor de onderzoeksgroepen misschien niet meer vergelijkbaar zijn?

WAT IS DE KLINISCHE RELEVANTIE EN DE OMVANG VAN HET EFFECT?

Om het geschatte behandeleffect klinisch te duiden is het in de eerste plaats van belang kritisch te zijn over de relevantie van de gemeten gezondheidsuitkomsten. Gezondheidsuitkomsten zoals sterfte en ziekte, maar ook kwaliteit van leven, zijn voor toekomstige gebruikers van een behandeling relevanter dan bijvoorbeeld intermediaire uitkomstmaten zoals biomarkers, bijvoorbeeld de serumcholesterolwaarde.

Daarnaast zijn de richting en de omvang van het effect van belang. Is het waargenomen effect klinisch relevant? Hoe groot is bijvoorbeeld het gemiddelde verschil in kwaliteit van leven als patiënten behandeling A of behandeling B krijgen? En hoe snel treedt dat effect eigenlijk op en blijft het ook na langere tijd bestaan?

Het betrouwbaarheidsinterval kwantificeert de onzekerheid van het geschatte effect. Het 95%-betrouwbaarheidsinterval is het interval dat, als het onderzoek eindeloos zou worden herhaald, in 95% van de gevallen het ware behandeleffect zal bevatten.[8] In 5% van de gevallen zal de waarheid dus niet in het betrouwbaarheidsinterval liggen. In de praktijk wordt het betrouwbaarheidsinterval gezien als een maat voor de precisie van het geschatte effect.

OP WIE ZIJN DE RESULTATEN VAN TOEPASSING?

Een gerandomiseerd onderzoek zegt feitelijk slechts iets over de effecten van de behandeling in de groep mensen die meededen aan het onderzoek.[9] Toch is de intentie niet om uitspraken te doen over de deelnemers aan het onderzoek, maar juist over een grotere groep, in het bijzonder potentiële gebruikers. Met andere woorden: we willen de resultaten generaliseren.

Generaliseerbaarheid is de mate waarin de resultaten van het onderzoek van toepassing zijn op mensen die niet deelnamen aan het onderzoek.[10] Soms is het evident dat resultaten niet kunnen worden gegeneraliseerd naar bepaalde groepen. Het is bijvoorbeeld bespottelijk om het waargenomen effect van orale anticonceptiva op het voorkómen van zwangerschap te generaliseren naar mannelijke gebruikers; die zijn er niet. Maar hoe ver kunnen we normaal gesproken gaan bij het generaliseren van de resultaten van een gerandomiseerd onderzoek?

Wanneer de onderzoeksgroep representatief is – in termen van bijvoorbeeld ernst of progressie van de ziekte – voor de groep toekomstige gebruikers, zijn de resultaten van het onderzoek doorgaans eenvoudig te generaliseren. De in- en exclusiecriteria van een gerandomiseerd onderzoek bieden hier inzicht in. Wanneer de groep patiënten die meedeed aan een onderzoek echter wezenlijk verschilt van een toekomstige patiënt bij wie zich een therapeutische vraag voordoet, dan zullen er extra aannames moeten worden gedaan, namelijk dat de effecten van de behandeling dezelfde zullen zijn bij de nieuwe patiënt.[10,11] Eigenlijk draait alles om de vraag in hoeverre de effecten van de behandeling zullen verschillen tussen individuen. Als bijvoorbeeld in een onderzoek naar de effecten van antibiotica bij een middenoorontsteking alleen meisjes deelnemen, zijn de resultaten dan van toepassing op jongens? Ander onderzoek of kennis van de pathofysiologie kan dan helpen: als een middenoorontsteking net zo vaak dezelfde bacteriële origine heeft bij jongens als bij meisjes, zullen de antibiotica waarschijnlijk in beide groepen even effectief zijn. De resultaten van het onderzoek waaraan alleen meisjes deelnemen zijn dan ook van toepassing op jongens. Dus ook als een onderzoekspopulatie niet representatief is voor de toekomstige gebruikers, bijvoorbeeld in termen van de verhouding jongens/meisjes. kunnen we de resultaten generaliseren als we bereid zijn aanvullende aannames te doen.

WAT WETEN WE NOG NIET?

Na het lezen van elk artikel zullen er nog vragen onbeantwoord zijn. Het is waardevol om te benoemen welke antwoorden het onderzoek heeft gegeven, maar ook wat we nog niet weten na de afronding van het onderzoek. Daarbij valt te denken aan bijwerkingen van de experimentele behandeling of langetermijneffecten.

TOT SLOT

Een gerandomiseerd onderzoek grijpt in op de autonomie van patiënt en arts: niet de patiënt en zijn arts bepalen of en zo ja, welke behandeling er wordt gestart. Dat wordt bepaald door het toeval. De belasting voor de deelnemers aan het onderzoek dient mede daarom zo beperkt mogelijk te zijn. Daarnaast is het includeren en volgen van deelnemers en het meten van gezondheidsuitkomsten een kostbaar proces. Het gevolg is dat het aantal deelnemers en de looptijd van het onderzoek meestal worden beperkt. Er worden juist voldoende deelnemers ingesloten om de primaire onderzoeksvraag te beantwoorden, maar niet meer dan dat. Zeldzame bijwerkingen of bijvoorbeeld subgroepeffecten – is de behandeling meer of minder effectief in bepaalde subgroepen? – kunnen daardoor onvoldoende worden onderzocht. Bovendien wordt de looptijd van gerandomiseerd onderzoek zo kort mogelijk gehouden: als de effecten van een behandeling zichtbaar zijn na 6 weken, wordt de uitkomst niet pas na 6 maanden gemeten. Langetermijneffecten en de wat zeldzamere bijwerkingen blijven zodoende onopgemerkt. Gerandomiseerd onderzoek is er in vele soorten en maten, met uiteenlopende kwaliteit en relevantie. Voor het beoordelen van de effecten van medische behandelingen is het essentieel dat de onderzochte groepen bij aanvang van een onderzoek vergelijkbaar zijn. Randomisatie is het krachtigste instrument om dit te bereiken. Maar om het onderzoek op waarde te kunnen schatten is het van belang om naar meer te kijken dan alleen die randomisatie. De leidraad in dit artikel voor het beoordelen van gerandomiseerd onderzoek is niet uitputtend, maar kan richting geven bij het duiden van de onderzoeksresultaten.

LEERPUNTEN

De volgende vragen vormen een leidraad voor lezers die artikelen over gerandomiseerd onderzoek op waarde willen schatten:

- **Zijn de onderzoeksgroepen vergelijkbaar?**
- **Wat is het doel van het onderzoek: het vaststellen van de werkzaamheid, of van de relatieve effectiviteit in de dagelijkse praktijk?**
- **Welke behandelingen worden vergeleken?**
- **Worden de gegevens van alle deelnemers gebruikt in de analyse?**
- **Wat is de klinische relevantie, de omvang en de precisie van het effect van de behandeling?**
- **Op wie zijn de resultaten van toepassing?**
- **Welke vragen blijven onbeantwoord? Denk hierbij aan bijvoorbeeld langetermijneffecten en bijwerkingen.**

In de rubriek Stand van zaken verschijnen regelmatig bijdragen over methoden die gebruikt worden bij het opzetten van wetenschappelijk onderzoek. De artikelen in deze serie illustreren op begrijpelijke wijze wat een bepaalde methode behelst, zonder dat hier uitvoerige methodologische kennis voor nodig is. Zowel oude als nieuwe methodologische principes worden zo inzichtelijk gemaakt voor artsen die klinische onderzoeken goed willen interpreteren.

Belangenconflict en financiële ondersteuning: geen gemeld.

Aanvaard op 1 april 2015

Citeer als: Ned Tijdschr Geneeskd. 2015;159:A9018

KIJK OOK OP WWW.NTVG.NL/A9018

LITERATUUR

1 Vandenbroucke JP. Observational research, randomised trials, and two views of medical science. PLoS Med. 2008;5:e67.
2 Grobbee DE, Hoes AW. Clinical Epidemiology: Principles, Methods and Applications for Clinical Research. 2nd ed. Londen: Jones and Bartlett Publishers; 2014.
3 Senn S. Testing for baseline balance in clinical trials. Stat Med. 1994;13:1715-26.
4 Higgins JP, Altman DG, Gøtzsche PC, et al; Cochrane Bias Methods Group; Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928.
5 Thorpe KE, Zwarenstein M, Oxman AD, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. J Clin Epidemiol. 2009;62:464-75.
6 Sedgwick P. Explanatory trials versus pragmatic trials. BMJ. 2014;349(nov13 3):g6694.
7 Groenwold RH, Moons KG, Vandenbroucke JP. Randomized trials with missing outcome data: how to analyze and what to report. CMAJ. 2014;186:1153-7.
8 Kirkwood BR, Sterne JAC. Essential Medical Statistics. 2nd ed. Oxford: Wiley-Blackwell; 2003.
9 Senn S. Statistical issues in drug development (Statistics in Practice). 2nd ed. Chichester, Engeland: Wiley-Interscience; 2007.
10 Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. Int J Epidemiol. 2013;42:1012-4.
11 Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet. 2005;365:82-93.