

Onderwijs in wetenschap

dr. R.W.J.G. Ostelo

dr. A.P. Verhagen

prof. dr. ir. H.C.W. de Vet

25

Systematische reviews van effectonderzoek: doel en opzet

H.C.W. de Vet, I. Logghe en A.P. Verhagen

In een tijdperk waarin kosteneffectieve zorg hoog in het politieke vaandel staat, heeft de wetenschappelijke onderbouwing van de paramedische zorg veel aandacht gekregen. Van elke hulpverlener wordt verwacht dat hij informatie bezit over de effectiefste behandeling op zijn vakgebied. Bewijzen voor die effectiviteit worden bij voorkeur geleverd door gerandomiseerde studies ('randomized controlled trial': RCT). Het aantal RCT's is de laatste jaren snel gestegen. Het wordt voor zorgverleners dan ook ondoenlijk alles bij te houden. Literatuuroverzichten, in de vorm van systematische reviews, maken het de zorgverlener gemakkelijker op de hoogte te blijven van de aanwezige kennis op een bepaald terrein in de literatuur. In hoofdstuk 1 zagen we al dat systematische reviews de basis van vormen van 'evidence-based practice' en een belangrijke rol spelen bij het ontwikkelen van behandelstandaarden of -protocollen. In dit hoofdstuk wordt ingegaan op het belang van systematische reviews en worden de opzet en uitvoering beschreven.

De afgelopen jaren is er op de gezondheidszorg steeds meer druk uitgeoefend om de zorg wetenschappelijk te onderbouwen, zowel vanuit de politiek als door de zorgverzekeraars. Vanuit het oogpunt van kosteneffectieve zorg zouden de behandelingen waarvan niet is aangetoond dat ze effectief zijn, op den duur uit het behandelingspakket moeten verdwijnen. Hiermee kreeg de beroepsgroep de taak opgelegd effectiviteitsstudies te verrichten en

aan te tonen dat de toegepaste behandelingen effectief waren. Deze ontwikkelingen weerspiegelen de opkomst van 'evidence-based practice' (zie hoofdstuk 1). Voor de fysiotherapie in Nederland werd dit duidelijk toen de Universiteit Maastricht een rapport publiceerde over de effectiviteit van fysiotherapie.^[1,2] De eindconclusie van dat rapport luidde dat er voor een aantal applicaties binnen de fysiotherapie (bijvoorbeeld electrotherapie; tractie) weinig of geen evidence was (= 'no evidence of effect'). Dat betekent dus dat er geen wetenschappelijk bewijs was ten aanzien van de effectiviteit van deze interventies omdat er op dat moment simpelweg nog geen studies naar waren uitgevoerd. Dat wil dus ook zeggen dat er geen wetenschappelijk bewijs is dat de interventie niet effectief zou zijn. Met andere woorden, de onderzoekers konden deze vraag niet beantwoorden. Deze conclusie van het rapport werd echter al snel en onzorgvuldig vertaald (o.a. in de media, maar zeker ook door de zorgverzekeraar) als 'evidence of no effect'. Om de uitspraak te kunnen onderbouwen dat een interventie niet effectief is moet er dus wel onderzoek zijn uitgevoerd dat aantoont dat een interventie niet effectief is. Dat was hier dus niet het geval. Maar deze 'onzorgvuldige' vertaling leidde er wel toe dat hier 'wetenschappelijke onderbouwing' gevonden leek voor de bezuinigingen op de fysiotherapie.

Het aantal gepubliceerde effectstudies in de paramedische zorg is de laatste decennia ex-

ponentieel gestegen. Wanneer men zich beperkt tot het studiedesign waaraan de meeste bewijskracht wordt toegedicht om de effectiviteit van een behandeling aan te tonen, de 'randomized controlled trial' (RCT) (zie de hoofdstukken 3 en 4), illustreert een blik in het gecomputeriseerde literatuurbestand MEDLINE deze snelle toename: werden er in 1964 nog zestien nieuwe RCT's gepubliceerd op het terrein van de geneeskunde en paramedische beroepen, in 1982 waren dit er al ruim tweeduizend, en in 2000 kwamen er wereldwijd ruim vijfendertigduizend nieuwe RCT's bij. Dit is een verveelvoudiging van het aantal RCT's in een paar jaar tijd. De database van de fysiotherapie (PEDro-database) bevat eind 2011 ruim zestienduizend RCT's, ruim drieduizend systematische reviews en meer dan duizend richtlijnen. Het moge duidelijk zijn dat een behandelend paramedicus deze omvangrijke literatuur nooit kan bijhouden, ook al zou hij de beschikking hebben over alle tijdschriften (meer dan tweehonderd) waarin deze effectstudies worden gepubliceerd.

In de praktijk blijkt dat behandelaars vaak maar van een zeer klein deel van de gepubliceerde studies op de hoogte zijn en op basis daarvan conclusies trekken voor hun handelen. Deze conclusies zouden misschien heel anders zijn als ze toegang hadden tot (bijna) alle informatie die voorhanden is, want de uitkomsten van de verschillende studies over hetzelfde onderwerp kunnen onderling nogal verschillen. Daarom speelt de systematische review een steeds belangrijkere rol bij het samenvatten van de aanwezige kennis op een bepaald vakgebied.

Vormen van systematische reviews

Het doel van een systematische review is een samenvatting te geven van de aanwezige kennis op een bepaald vakgebied. Zo kan de lezer een preciezer inzicht krijgen in de effectiviteit van bepaalde behandelingen en de verschillen daarin bij verschillende patiëntengroepen (bijvoorbeeld acuut versus chronisch) of een beeld krijgen van de invloed van de dosering,

behandelduur en behandelfrequentie. Anderszijds biedt een systematische review zicht op de lacunes in onze kennis en genereert zo vaak nieuwe onderzoeksvragen.

Van oudsher bestond het literatuuronderzoek uit een uitgebreid overzichtsartikel – klassieke of 'narratieve (verhalende) review' – waarin een expert op een bepaald vakgebied zijn mening gaf over de stand van zaken met betrekking tot een bepaalde behandeling of aandoe-ning op basis van (een selectie uit) de literatuur. Hoe deze selectie tot stand kwam en waarop de expert zijn mening baseerde, was niet altijd even duidelijk. Men vertrouwde geheel op de deskundigheid van de auteur.

Vanaf de jaren tachtig kwam daarin verandering en werd het literatuuronderzoek systematischer van opzet. Vanaf die tijd wordt de term 'systematische review' gebruikt voor studies waarin de literatuur op systematische wijze wordt verzameld en beoordeeld. Onderdeel van die beoordeling is dat de methodologische kwaliteit van de in het overzicht betrokken studies wordt bepaald. Het meest kenmerkende van een systematische review is dat duidelijk wordt beschreven op welke manier men de literatuur heeft doorzocht, welke gegevens uit de diverse studies is geëxtraheerd, hoe de kwaliteit van de onderzoeken is beoordeeld en hoe men vervolgens tot een conclusie is gekomen. Dit stelt de lezer in staat de gang van zaken te volgen en zich een mening te vormen over de kracht van het bewijs zelf.

Wanneer een kwantitatieve samenvatting van de resultaten wordt gegeven, wordt een systematische review ook wel meta-analyse genoemd. Dit is niet geheel juist omdat een systematische review de hele methodiek beslaat zoals hierboven beschreven. De meta-analyse is niet meer dan een (belangrijk) onderdeel omdat de resultaten dan ook statistisch samengevat (gepooled) kunnen worden. Het onderscheid is belangrijk omdat men anders zou kunnen denken dat indien er geen meta-analyse is uitgevoerd, het geen systematische review zou zijn. Dit is dus onjuist. In meta-analyses worden de resultaten van alle onderzoeken

ken (grote en kleine) op een gewogen manier (grote studies geven meer gewicht in het gepoolde resultaat) bij elkaar opgeteld. Dit optellen wordt statistische 'pooling' genoemd, omdat met een bepaalde statistische techniek een 'pool' van resultaten wordt gecreëerd, waaruit een algemene conclusie kan worden getrokken. Hierdoor kunnen kleine, maar relevante effecten worden aangetoond, die in de afzonderlijke onderzoeken niet altijd worden opgemerkt vanwege een te kleine omvang van de onderzoekspopulatie.

Opzet van een systematische review

De opzet van een systematische review bestaat uit een aantal onderdelen: de onderzoeksvraag, het formuleren van in- en exclusiecriteria, de zoekstrategie van de literatuur, de selectie van welke artikelen uiteindelijk in de review worden opgenomen, het bepalen van de methodologische kwaliteit van de afzonderlijke artikelen, de analyse (mogelijk een meta-analyse maar dat hoeft niet per se zo te zijn) en het trekken van de conclusies. Deze opzet geldt niet alleen voor reviews waarin de resultaten van RCT's worden samengevat, maar geldt ook voor reviews die resultaten uit observationeel onderzoek samenvatten of reviews over de waarde van een speciale diagnostische test. De uitwerking hierna is gericht op reviews van effectstudies (RCT's). Door de explosieve toename van RCT's is het daarnaast gebruikelijk bestaande systematische reviews te 'updaten'. De oorspronkelijke onderdelen van de review vormen hierbij het uitgangspunt (zie kader 25.1).

ONDERZOEKSVRAAG

Een systematische review heeft als doel een gerichte vraag te beantwoorden. Bijvoorbeeld: hoe effectief is fysiotherapie en manuele therapie voor een behandeling van patiënten met spanningshoofdpijn?^[3] De zoekvraag kan men specificeren door exact aan te geven in welke patiënten (populatie), interventie en/of uitkomstmaat men geïnteresseerd is. Een voorbeeld van een specifiekere vraag is welke

behandeling het effectiefst is bij chronische patiënten, of de vraag of de effectiviteit het meest tot uiting komt in een afname van de pijn, of een vermindering in ziekteverzuim.

Kader 25.1 Samenvatting

Studiedesign. Een systematische review van gerandomiseerde klinische trials (RCT's).

Doelstelling. Het doel van deze systematische review is om de effectiviteit van fysiotherapie en manipulaties bij patiënten met spanningshoofdpijn te evalueren.

Methode. Dit is een update van de bestaande review uit 2004. Er is gezocht naar literatuur in PubMed, CINAHL en Cochrane. Alleen gerandomiseerde klinische trials zijn geïncludeerd die fysiotherapie en manipulaties evalueerden bij een volwassenen populatie (18 jaar en ouder) met spanningshoofdpijn. Twee onderzoekers beoordeelden, onafhankelijk van elkaar, de gevonden literatuur op methodologische kenmerken met behulp van de Delphi-lijst en voerden de data-extractie uit. Analyse vond plaats op basis van een 'best evidence synthesis'.

Resultaten. In totaal zijn twaalf studies geselecteerd. De originele review bestond uit 8 studies en in deze update zijn er vier nieuwe studies toegevoegd. We vonden een grote variatie aan interventies, zoals manipulaties, oefentherapie, fysiotherapeutische applicaties, massage, relaxatietherapie en acupunctuur. Twee studies hadden een laag risico op vertekening van de resultaten (bias). Deze studies evalueerden beide manipulaties in vergelijking met een placebobehandeling (placebolaser) of medicatie. Er werd geen significant verschil gevonden tussen beide behandelingen. Specifieke oefentherapie (cra-niocervicale training) is mogelijk wel

effectief, maar dit betreft slechts één studie.

Conclusie. In deze review is het niet mogelijk om definitieve conclusies te trekken over het effect van manipulaties en fysiotherapie bij volwassenen met spanningshoofdpijn, maar een specifiek oefenprogramma lijkt een veelbelovende interventie.

IN- EN EXCLUSIECRITERIA

Op basis van deze onderzoeksvraag bepaalt men de in- en exclusiecriteria voor de studies die in het overzicht worden opgenomen. In hoofdstuk 4 beschreven we al dat als hulpmiddel voor een goede vraagstelling vaak het PICO-systeem gebruikt wordt: 'Patient – Intervention – Comparison – Outcome'. Deze selectiecriteria kan men in vier hoofdgroepen verdelen. Allereerst wordt bepaald in welk type onderzoeksdesign men is geïnteresseerd. Gouden regel is dat indien er voldoende RCT's gepubliceerd zijn over het onderwerp van keuze, men zich beperkt tot een systematische review van trials. De reden hiervoor is dat RCT's de meeste bewijskracht hebben wat betreft de effectiviteit van een interventie. Bestaan er geen of heel weinig RCT's over het desbetreffende onderwerp – hetgeen op het terrein van de fysiotherapie op dit moment nog maar zelden het geval is – dan kan men ook andere onderzoeksvormen, zoals quasi-experimenten, opnemen. Uiteraard hangt het aantal onderzoeken dat men opneemt ook af van hoe specifiek of hoe breed men de onderzoeksvraag formuleert. Vervolgens bepaalt men in- en exclusiecriteria met betrekking tot de patiëntenpopulatie of aandoening waarin men is geïnteresseerd. Wil men bijvoorbeeld de effectiviteit van een bepaalde behandeling bekijken bij alle patiënten of alleen bij chronische patiënten. In het voorbeeld zijn het patiënten met spanningshoofdpijn, zowel met acute als chronische klachten. Vervolgens bepaalt men in welke interventie men is geïnteresseerd; wil men de effectiviteit weten van

alle soorten fysiotherapie, of alleen oefentherapie? In het voorbeeld is men geïnteresseerd in alle soorten fysiotherapie inclusief manuele therapie. Als laatste wordt bepaald over welke uitkomstmaten in een onderzoek moet zijn gerapporteerd. Als men bijvoorbeeld een uitspraak wil doen over pijnvermindering en 'kwaliteit van leven', dan worden onderzoeken waarin alleen laboratoriumparameters zijn gepresenteerd uitgesloten.

ZOEKSTRATEGIE

Na het formuleren van de onderzoeksvraag en de in- en exclusiecriteria wordt naar de literatuur gezocht. Dit zoeken moet planmatig gebeuren, want het doel is om alle bestaande studies die aan de in- en exclusiecriteria voldoen ook daadwerkelijk te vinden. Meestal is de zoekstrategie vrij breed, om daarna te beoordelen welke onderzoeken precies aan de in- en exclusiecriteria voldoen. Het meest praktisch is met de zoekactie te beginnen in gecomputeriseerde literatuurbestanden, zoals Cochrane Library, MEDLINE, EMBASE en PEDro. Als het meer psychologisch gerichte vraagstellingen betreft, moet ook in Psycinfo worden gezocht. Het samenstellen van een set trefwoorden ('keywords') is daarbij cruciaal. Men mag geen belangrijke trefwoorden over het hoofd zien. Een manier om dat te controleren is te kijken of alle relevante trefwoorden die bij de gevonden artikelen staan, ook in de zoekstrategie voorkomen. De computerbestanden vormen slechts het begin van een zoekstrategie. Aanvullende zoekstrategieën zijn noodzakelijk. Het ligt voor de hand eerdere reviews op te zoeken en literatuurreferenties in de inmiddels gevonden artikelen na te kijken. Verder wordt aanbevolen contact op te nemen met experts op het desbetreffende terrein, en in databases naar lopende studies te zoeken (om eventuele publicatiebias te voorkomen).

SELECTIE VAN STUDIES

Als het zoeken naar studies is afgerond, moet aan de hand van de in- en exclusiecriteria worden bepaald welke studies in de systema-

tische review worden opgenomen. Soms kan dit op basis van het 'abstract', soms moet daarvoor het gehele artikel worden bestudeerd. Omdat het vaak een subjectieve beslissing is of de studie nu wel of niet aan de in- en exclusiecriteria voldoet, wordt deze procedure vaak door twee reviewers onafhankelijk van elkaar verricht. Het is gebruikelijk om in een flow-diagram aan te geven hoeveel studies men in de diverse selectiestappen vond: d.w.z. aantal titels, volledige artikelen, en het aantal dat aan de inclusiecriteria voldeed. Tevens worden hier redenen vermeld waarom studies uitgesloten werden.

METHODOLOGISCHE KWALITEIT VAN DE AFZONDERLIJKE STUDIES

Voor het beoordelen van de methodologische kwaliteit (tegenwoordig ook wel 'risk of bias' genoemd) van de afzonderlijke studies wordt naar een aantal criteria gekeken, die in criterialijsten worden samengevoegd. Een criterialijst bevat vragen (items) zoals: is er gerandomiseerd? Was de patiënt geblindeerd? Waren

er veel uitvallers? Enzovoort. Voor RCT's bestaan verschillende soorten criterialijsten. De meeste criterialijsten bevatten drie domeinen: interne validiteit (items over randomisatie, blinding, uitval), externe validiteit (items over de patiëntenpopulatie, interventie, effectmaten) en precisie (items over groepsomvang, meetvariatie). Een voorbeeld van een gevalideerde algemene (generieke) criterialijst voor RCT's is de Delphi-criterialijst.^[4] Deze lijst is opgenomen in tabel 25.1. Alle items hebben als antwoordmogelijkheid: 'ja/nee/weet niet'. Deze Delphi-lijst vormde de basis voor twee andere veelgebruikte criterialijsten, namelijk de PEDro-lijst die in de PEDro database veel wordt gebruikt alsook voor de criterialijst van de Cochrane Back Review Group. Soms worden de items van de criterialijst waarop een 'ja' is gescoord, bij elkaar opgeteld tot een somscore. Alle items wegen dan even zwaar. Somscores hebben het voordeel dat ze eenvoudig en inzichtelijk zijn. De score lijkt dan op een rapportcijfer dat elke studie krijgt. Een nadeel van een rapportcijfer is dat

Tabel 25.1 Delphi-criterialijst.^[4]

1	Randomisatie	
	a Is een methode van randomisatie uitgevoerd?	ja / nee / weet niet
	b Was de randomisatiecode onbekend voor degene die patiënten insloot en de behandelingen toeweest?	ja / nee / weet niet
2	Waren de groepen aan het begin van het onderzoek onderling vergelijkbaar wat betreft de belangrijkste prognostische variabelen?	ja / nee / weet niet
3	Zijn de in- en exclusiecriteria genoemd?	ja / nee / weet niet
4	Was de effectbeoordelaar geblindeerd?	ja / nee / weet niet
5	Was de hulpverlener/behandelaar geblindeerd?	ja / nee / weet niet
6	Was de patiënt geblindeerd?	ja / nee / weet niet
7	Zijn de puntschatting en spreidingsmaten gepresenteerd voor de primaire uitkomstmaten?	ja / nee / weet niet
8	Is de analyse volgens het 'intention to treat'-principe* uitgevoerd?	ja / nee / weet niet

* Alle patiënten worden geanalyseerd in de groep waaraan ze zijn toegewezen, ongeacht welke interventie ze uiteindelijk hebben ontvangen.

een 'nee' op het ene item kan worden gecompenseerd met een 'ja' op een ander item. Een onderzoek kan dan toch een 6 halen, terwijl op items die als heel belangrijk worden gezien een 'nee' is gescoord. Het is ook mogelijk de items individueel te bekijken. Hoe een criterialijst moet worden gebruikt, is niet altijd duidelijk aangegeven. In het geval van de Delphi-criterialijst is het aan de onderzoeker te bepalen of er een somscore wordt berekend of niet. In de publicatie dient te worden vermeld voor welke optie men heeft gekozen. Ook wordt geadviseerd per afzonderlijke studie de scores per item weer te geven in een tabel, zodat de lezer inzicht krijgt hoe de scores ten aanzien van de risk of bias van de afzonderlijke studies tot stand is gekomen. Het scoren van de risk of bias van de studies wordt meestal door twee reviewers onafhankelijk van elkaar gedaan. Daarna worden de resultaten naast elkaar gelegd en worden de discrepanties besproken. Door middel van discussie of het oordeel van een derde persoon probeert men tot een eindoordeel te komen.

ANALYSE

De volgende stap in een systematische review is de analyse. Daarbij moet een besluit worden genomen of de resultaten van de afzonderlijke onderzoeken bij elkaar worden opgeteld, ofwel statistisch gepoold, of niet. Indien men statistisch kan poolen wordt dat ook wel meta-analyse genoemd. Of een meta-analyse wel of niet kan worden uitgevoerd hangt af van een aantal factoren. Een noodzakelijke voorwaarde hiervoor is dat de individuele studies minimaal een puntschatting (gemiddelde, modus, mediaan) en een spreidingsmaat (standaarddeviatie, betrouwbaarheidsinterval) presenteren voor de uitkomstmaat waarin men is geïnteresseerd, anders kan er niet worden gepoold. Verder dient afgewogen te worden of de individuele studies inhoudelijk voldoende op elkaar lijken wat betreft patiëntenpopulatie, interventies en uitkomstmaten om de gegevens bij elkaar te kunnen optellen. Dit wordt ook wel homogeniteit genoemd. Of anders gesteld: als studies te heterogeen zijn,

dan is een meta-analyse niet verstandig omdat men dan appels met peren gaat vergelijken. In het voorbeeld (zie kader 25.1) bleken de interventies te divers (te heterogeen) om de resultaten statistisch te poolen.^[2] Waar de grens ligt bij wat men mag poolen en wat niet, is meer een kwestie van klinische en methodologische afwegingen dan van statistische overwegingen.

PRESENTATIE VAN DE RESULTATEN

In een systematische review worden meestal veel en grote tabellen gepresenteerd. Vaak vindt men een tabel met een overzicht van de geïncludeerde studies, met kenmerken van de patiënten, aantal patiënten, en kenmerken van de interventie en soms uitkomstmaten. Zoals eerder is aangegeven is ook gebruikelijk om iets te laten zien van de risk of bias van de afzonderlijke studies. Per uitkomstmaat worden meestal de grootte van het effect (effect-schatting ofwel effect sizes) gepresenteerd, eventueel samen met het gepoolde effect. Een apart overzicht van alle uitkomstmaten vindt men vervolgens in een aparte tabel, ook wel de Summary of Findings (SoF)-tabel genoemd. Hierin staan alle belangrijke uitkomstmaten, ook als er geen enkele studie (of slechts enkele) die uitkomstmaat presenteerde(n). In deze SoF-tabel staan voor alle relevante uitkomstmaten de gepoolde effectschatting (of een andere indicatie van het effect); het aantal studies en aantallen patiënten dat daaraan bijgedragen heeft.

GRADE: HOE STERK IS HET BEWIJS?

Als men een idee heeft gekregen van de grootte van het effect, is de volgende vraag: hoe sterk is het bewijs (de evidence) daarvoor? Guyatt et al. ontwikkelden het GRADE-systeem om de sterkte van het bewijs uit te drukken.^[5] GRADE staat voor Grading the Recommendations, Assessment, Development and Evaluation. Uitgaande van de gepoolde effectschatting, of van het overzicht van de effecten die in de verschillende studies zijn gevonden, wordt gekeken naar een aantal domeinen die de sterkte/kracht van het bewijs

aantasten. Het gaat om de volgende domeinen: a) de studieopzet (design) en risk of bias van de studie; b) de betrouwbaarheid van het resultaat (precision); c) de consistentie van de resultaten; d) of het bewijs direct of indirect is; en e) of er mogelijk sprake is van publicatiebias. GRADE wordt in principe toegepast op reviewniveau en niet op studieniveau. De risk of bias wordt natuurlijk wel op studieniveau vastgesteld. GRADE kent uiteindelijk vier niveaus van bewijskracht: 1) sterk bewijs; 2) beperkt bewijs; 3) weinig bewijs en 4) zeer weinig bewijs.

Risk of bias

Als een systematische review alléén RCT's includeert heeft deze een grotere bewijskracht dan wanneer niet-gerandomiseerde experimenten of observationele studies zijn geïnccludeerd. In dat geval begint met een 'sterk niveau van bewijs'. Maar ook RCT's moeten volgens de regels der methodologische kunst worden uitgevoerd en dat wordt bekeken aan de hand van de risk of bias per studie. Als de systematische review vooral studies met een hoog risico op vertekening van de resultaten bevat (high risk of bias), dan doet dit afbreuk aan de bewijskracht. Men spreekt binnen de GRADE-systematiek dan van het 'downgraden' van de bewijskracht waardoor deze bijvoorbeeld nog maar beschreven wordt als 'beperkt niveau van bewijs.'

Betrouwbaarheid

De betrouwbaarheid van de (gepoolde effect) schatting wordt aangegeven door het betrouwbaarheidsinterval. Daarbij wordt natuurlijk eerst gekeken of het effect statistisch significant is. Maar ook als dat het geval is geeft een smal betrouwbaarheidsinterval een preciezere schatting van de grootte van het effect, waardoor de bewijskracht steviger wordt dat de geschatte grootte van het gevonden effect klopt, dan wanneer het betrouwbaarheidsinterval erg breed is.

Consistentie van het effect

Dit gaat om de vraag of steeds ongeveer hetzelfde effect wordt gevonden in een groep studies. Als dat het geval is heeft men meer vertrouwen in die uitkomst dan wanneer er sommige onderzoeken een groot effect laten zien en anderen helemaal geen. Vaak is dat ook al een reden om het effect niet statistisch te poolen over de studies. We zeggen dan dat de uitkomsten te heterogeen zijn. Deze heterogeniteit kan mogelijk veroorzaakt worden door verschillen in de patiëntenpopulatie, interventie of uitkomstmaten. De vraag of er sprake is van heterogeniteit kan het best worden beantwoord op basis van gezond verstand. Naarmate men strengere in- en exclusiecriteria hanteert, wordt de kans op heterogeniteit kleiner, maar de kans dat men geen onderzoeken overhoudt, wordt des te groter. Soms worden er ook verschillende effecten gevonden in studies die erg op elkaar lijken. We spreken dan van statistische heterogeniteit. Deze kan worden berekend, maar aan de interpretatie van die berekening zitten nogal wat haken en ogen. Indien de resultaten van studies niet consistent is wordt het niveau van de bewijskracht binnen GRADE ook weer een niveau lager, ofwel 'downgraded'.

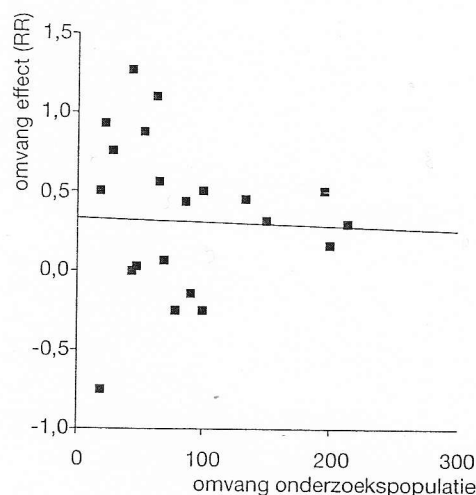
Direct of indirect effect

Indirect bewijs wil zeggen dat er een conclusie wordt getrokken die niet direct over de onderzoeksvraag gaat. Een voorbeeld is onderzoek naar het effect van een cholesterolverlagend middel op het aantal gevallen van hart- en vaatziekten. Om dit vast te stellen kijkt men bijvoorbeeld meestal naar het effect op de hoogte van het cholesterol in het bloed, terwijl men eigenlijk wil weten of dit middel ook het aantal gevallen van hart- en vaatziekten vermindert. Dit soort 'afgeleide' maten worden ook wel intermediaire maten of proxy-uitkomsten genoemd. De mate waarin zo een intermediaire uitkomst nu ook daadwerkelijk iets zegt over de echte uitkomst bepaalt in zulke gevallen de waarde van het onderzoek. In dit geval kun je het effect op het cholesterolgehalte dan een vorm van indirect bewijs

noemen voor de eigenlijke uitkomst: hart- en vaat ziekten. Het is helder dat indirect bewijs minder sterk is dan direct bewijs. In de GRADE-systematiek zal de bewijskracht ook weer een niveau lager uitvallen, ofwel worden 'downgraded' indien er sprake is van indirect bewijs.

Publicatiebias

Publicatiebias houdt in dat er vertekening (bias) in de review optreedt doordat niet alle onderzoeken die bestaan zijn gepubliceerd, of dat niet alle publicaties zijn gevonden omdat ze in minder goed toegankelijke tijdschriften zijn gepubliceerd. Men gaat ervan uit dat het risico van publicatiebias het grootst is bij kleine studies waarin men geen of zelfs negatieve effecten vindt. Als een grafische presentatie (plot) wordt gemaakt van de relatie tussen de effectschatting en de omvang van de onderzoekspopulatie, krijgt men bij voldoende studies een soort trechtervorm van puntjes te zien: een 'funnelplot' (figuur 25.1). Als er erg veel puntjes ontbreken in het gebied van de kleine studies met geen of een negatieve uitkomst, kan er sprake zijn van publicatiebias. In de hier gepresenteerde figuur is dat niet of nauwelijks het geval. In dit geval zal de bewijskracht dus niet een niveau worden verlaagd voor het domein 'publicatiebias'. Kortom, op basis van deze vijf domeinen bepaalt men hoe sterk het bewijs is van de conclusie in een systematische review. Deze methode is nog vrij nieuw en levert tijdens de uitvoering vaak nog onderlinge discussie op bij de auteurs. Het is echter wel een expliciete methode waarbij de auteurs zelf mogen aangeven hoe sterk het bewijs is en daar transparant in moeten zijn, wat een groot voordeel is. Voordat we de GRADE-systematiek gebruiken werd ook de methode van de 'levels of evidence' veel gebruikt om de sterkte van het bewijs weer te geven. Deze indeling wordt nog steeds vaak in richtlijnen gebruikt. Soms ook is het aan de lezers of gebruikers van de systematische review om een beoordeling te geven aan de sterkte van het bewijs. Deze lieten zich nog meer dan de auteurs leiden door



Figuur 25.1 Funnelplot voor de schatting van publicatiebias.

het feit of de uitkomst van de review hen beviel of niet.

Noodzaak van onderzoeksprotocol

Het is zeer belangrijk van tevoren een onderzoeksprotocol te maken, waarin bovenbeschreven stappen en de manier waarop men die gaat uitvoeren worden vastgelegd. Dit protocol beschrijft onder andere wie zich bezighouden met de informatieverzameling (reviewers, al dan niet inhoudsdeskundig), hoe de risk of bias van de onderzoeken wordt vastgesteld, en hoe dit in de eindconclusie wordt verwerkt. Binnen de Cochrane Collaboration is het zelfs verplicht en worden de protocollen voorafgaande aan de uitvoering van de review ook gepubliceerd in de Cochrane Library.

Mode of noodzaak?

Het is bekend dat het enige tijd duurt voordat resultaten van studies doordringen tot de kennis en praktijk van de behandelaars. In hoofdstuk 1 beschreven we al dat dit van alle tijden is, zelfs als het om nijpende problemen

gaat. Systematische reviews en meta-analysen kunnen een belangrijke rol spelen om die tijd te bekorten. Antman c.s. hebben in 1992 onderzocht hoe snel bevindingen uit wetenschappelijk onderzoek in medische tekstboeken terechtkomen.^[6] Zij hebben gekeken naar de behandelingen en secundaire preventie van hart- en vaatziekten. Van streptokinase (een antistollingsmiddel) was in 1973 de effectiviteit al bekend, maar het middel werd pas in 1985 voor het eerst in de tekstboeken opgenomen als zijnde een adequate behandeling bij hart- en vaatziekten. Een ander voorbeeld betreft het middel lidocaïne (een middel tegen ventrikelfibrilleren). In vijftien onderzoeken tussen 1970 en 1987 (met in totaal 8745 gerandomiseerde patiënten) bleek lidocaïne niet effectief te zijn. Toch komt het middel als aanbevolen profylaxe tot 1990 (en misschien nu nog) voor in medische tekstboeken. Uit het onderzoek van Antman c.s. bleek ook dat, waarschijnlijk doordat men niet op de hoogte was van al bestaande onderzoeken, er onnodig veel studies op een bepaald vakgebied zijn uitgevoerd.^[6] Zo zijn er tussen 1959 en 1985 33 RCT's uitgevoerd naar het effect van streptokinase als trombolytische (antistolling)therapie. Als na acht RCT's een meta-analyse zou zijn uitgevoerd, zou een significante vermindering van het aantal overleden patiënten zijn gevonden door het gebruik van streptokinase (totaal 2432 patiënten gerandomiseerd). De 25 daarop volgende RCT's (totaal nog eens 34.542 patiënten gerandomiseerd) gaven in de meta-analyse geen verandering van het effect te zien. Al deze 34.542 patiënten zijn voor niets in trials opgenomen en aan de helft daarvan is een gunstige/effektieve behandeling onthouden. Eén opmerking moet hierbij wel worden gemaakt, namelijk dat bij deze meta-analysen niet is gekeken naar de risk of bias van de desbetreffende onderzoeken.

Concluderend kan worden gesteld dat de resultaten van systematische reviews van belang zijn voor zowel behandelaars als patiënten, zodat niet onnodig lang niet-effectieve behandelingen worden voorgeschreven of effec-

tieve behandelingen worden onthouden aan patiënten. Een andere les die kan worden getrokken uit dit verhaal, is dat men altijd een systematische review moet uitvoeren alvorens men aan een nieuwe effectstudie begint. Deze review heeft niet alleen als functie te kijken of de onderzoeksvraag misschien al beantwoord is, maar kan ook goede sturing geven aan de keuze van onderzoekspopulatie, interventie en uitkomstmaten.

Cochrane Collaboration en evidence-based medicine

Steeds meer systematische reviews worden uitgevoerd onder de vlag van de Cochrane Collaboration (CC). De in 1992 opgerichte CC is genoemd naar Archie Cochrane, een Britse arts-epidemioloog die schreef dat het een schande was dat hulpverleners niet op de hoogte waren van de onderzoeksbevindingen in hun eigen vakgebied. De CC is een internationaal samenwerkingsverband van een sterk groeiend aantal wetenschappers. Het doel van deze organisatie is het systematisch verzamelen van informatie uit (effect)studies en vervolgens het in kaart brengen van de effectiviteit van behandelingen in alle takken van de gezondheidszorg in een actuele systematische review. Deze systematische reviews en alle RCT's worden verzameld in de Cochrane Library (<http://www.cochrane.org>). Het uitvoeren van systematische reviews wordt gestimuleerd en gecoördineerd, zodat zo weinig mogelijk studies dubbel worden gedaan. Bovendien inventariseert de CC de zogenaamde 'witte vlekken', gebieden waar nog weinig RCT's zijn gedaan en niet of nauwelijks systematische reviews voorhanden zijn. Volgens de CC dient voornamelijk die zorg te worden aangeboden waarvan de effectiviteit is aangetoond ('evidence-based medicine'). Evidence-based medicine of evidence-based practice betekent letterlijk het geven van die zorg die is gebaseerd op wetenschappelijk bewezen effectiviteit. Systematische reviews en meta-analysen leveren hiervoor de basis ofwel de 'evidence'. Deze kennis over de ef-

fectiviteit wordt vervolgens in richtlijnen en behandelstandaarden opgenomen.

Beschouwing

Met de niet-aflatende vraag om kosteneffectieve zorg is er behoefte om evidence-based practice door te voeren in alle aspecten van zorg. Vooral de overheid en zorgverzekeraars spelen hierin een grote rol. Alleen het publiceren van resultaten is niet voldoende om de praktijk te veranderen of te beïnvloeden. Het samenvatten van de aanwezige onderzoeksgegevens in systematische reviews en deze gebruiken bij het opstellen van behandelstandaarden en richtlijnen zijn methoden om de paramedische zorg optimaal gebruik te laten maken van de aanwezige kennis op het gebied van zowel effectstudies als prognostische en diagnostische studies. De paramedische zorg is wat dit betreft op de goede weg.

Literatuur

- 1 Beckerman H, Bouter LM, editors. Effectiviteit van fysiotherapie: een literatuuronderzoek. Maastricht: Rijksuniversiteit Limburg, Vakgroep Epidemiologie; 1991.
- 2 Beckerman H, Bie RA de, Bouter LM, Oostendorp RAB. De effectiviteit van lasertherapie bij aandoeningen van het bewegingsapparaat en de huid: een meta-analyse van patiëntgebonden onderzoek. *Ned Tijdschr Fysiother.* 1990;100: 306-6.
- 3 Verhagen AP, van Heest M, Have D van der, Lenssinck M-L.B. Berger MY, Verkerk K, Paschier J, Koes BW. Het effect van fysiotherapie en manipulaties bij volwassenen met spanningshoofdpijn; een update van een systematische review. *Ned Tijdschr Fysiother.* 2009;119:85-92.
- 4 Verhagen AP, Vet HCW de, Bie RA de, Kessels AGH, Boers M, Bouter LM, et al. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. *J Clin Epidemiol.* 1998;51:1235-41.
- 5 Guyatt G, Oxman ADD, Akl EA et al. Introduction – GRADE evidence profiles and summary of finding tables. *J Clin Epidemiol.* 2011;64: 383-94.
- 6 Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. *JAMA.* 1992;268:240-8.