GRADE

Een systeem om niveau van bewijskracht en graad van aanbeveling aan te geven

P. Van Royen

In de aanbeveling 'Gebruik van medicatie bij urgenties' *(zie blz. 472)* wordt voor het eerst met het GRADE-systeem gewerkt. Dit nieuwe systeem, gebaseerd op conclusies van de internationale GRADE-werkgroep, maakt het mogelijk om ook de sterkte (of zwakte) van een aanbeveling aan te duiden. In combinatie met de vroegere niveaus van bewijskracht, gebaseerd op onder meer het type studie en de kwaliteit ervan, krijgen artsen hiermee duidelijkere richtlijnen voor de praktijk.

Het beoordelen van evidentie en resultaten van wetenschappelijk onderzoek is complex en dus moeilijk in de dagelijkse praktijk voor elke huisarts of clinicus. In de klinische praktijk en zeker in de opleiding van huisartsen doet men in toenemende mate een beroep op aanbevelingen of richtlijnen, die zich veelal juist baseren op de resultaten van dit wetenschappelijk onderzoek. Volgens de principes van de 'evidence-based medicine' worden beslissingen in de aanbevelingen genomen op basis van het beste beschikbare bewijsmateriaal [1]. Niet alle 'evidence' is echter in kwaliteit en in vorm identiek.

Men heeft nood aan een transparant en betrouwbaar model waarop het beoordelen van evidentie en het formuleren van de aanbeveling is gebaseerd. Daarom zijn er verschillende methoden ontwikkeld om de evidentie te beoordelen.

Sinds 2002 gebruikt Domus Medica een systeem met drie niveaus van bewijskracht om de kernboodschappen van de aanbevelingen te merken [2]. Op basis van het studietype (klinisch trial, cohortonderzoek, observationele studie) en de methodische kwaliteit van de individuele studies en ook het aantal studies wordt zowel aan de conclusies van wetenschappelijk onderzoek als aan kernboodschappen van een aanbeveling een niveau van bewijskracht of 'level of evidence' toegekend. Men hanteerde drie niveaus van bewijskracht.

Het nadeel van dit systeem met drie niveaus was dat we naast een niveau van bewijskracht niet afzonderlijk konden aangeven hoe sterk we de kernboodschap ook aanbevelen. Het is immers mogelijk dat we een bepaalde behandeling sterk willen aanbevelen, maar dat er slechts beperkt wetenschappelijk onderzoek te vinden is over het nut van die behandeling. Zo beschikken we niet over RCT's in verband met het nut van het toedienen van intramusculair adrenaline in de behandeling van een levensbedreigende anafylaxis met hypotensie, shock en/of respiratoire symptomen. We blijven slechts bij observationele of casestudies om hiervan het nut aan te tonen; de 'level of evidence' zal dus eerder niveau 3 zijn. Het is ook weinig waarschijnlijk dat een geneesmiddel betere of snellere resultaten zal geven dan adrenaline. Maar met de kennis over de voor- en nadelen zijn we er wél van overtuigd dat men adrenaline op een correcte wijze en zo snel mogelijk moet toedienen in een situatie van levensbedreigende anafylaxis, ondanks het gebrek aan harde bewijzen uit de literatuur. Daarom willen we dit voor de klinische praktijk sterk aanbevelen.

In de aanbeveling *'Gebruik van medicatie bij urgenties'* kennen we voor het eerst ook een graad van aanbeveling toe, naast het niveau van bewijskracht, aan de belangrijkste besluiten of kernboodschappen.

Tot nu toe konden we naast een niveau van bewijskracht niet afzonderlijk aangeven hoe sterk we de kernboodschap ook aanbevelen

Van niveaus naar GRADE

Een systematische en expliciete benadering om een oordeel te vellen over het niveau van bewijskracht en de sterkte van aanbeveling helpt om fouten te vermijden bij de toekenning van niveaus en bevordert een kritische beoordeling van de informatie en open communicatie tussen ontwikkelaars en gebruikers van aanbevelingen. Sinds ruim dertig jaar hebben heel veel organisaties elk hun eigen systeem van beoordeling uitgewerkt met als gevolg dat eenzelfde boodschap kan worden gerangschikt als niveau 2, C+, B of sterke evidentie, naargelang het systeem dat wordt gebruikt. Dit schept veel verwarring en belet een effectieve communicatie hierover.

De GRADE-werkgroep *(Grading of Recommendations, Assessment, Development and Evaluation)*, samengesteld uit meer dan zestig vertegenwoordigers van belangrijke

organisaties die ook richtlijnen ontwikkelen zoals het National Institute of Clinical Excellence in Groot-Brittannië (NICE) en de Wereldgezondheidsorganisatie (WHO), heeft een nieuw systeem ontwikkeld die al deze uitdagingen en tekortkomingen tracht op te vangen. Dit systeem maakt een duidelijk onderscheid tussen het niveau van bewijskracht en de sterkte van de aanbeveling.

Onder niveau van bewijskracht verstaat men de mate waarin men op basis van de kwaliteit van de evidentie er zeker van kan zijn dat de aanbeveling valide/correct is. De sterkte van aanbeveling wijst eerder op de mate waarin men erop kan vertrouwen dat het volgen van de richtlijn meer goed dan kwaad zal doen. De eerste publicatie van deze GRADE-werkgroep dateert van 2004 [3]. Een nog eenvoudiger en meer klinisch georiënteerde beschrijving van het systeem werd gebruikt door het *American College of Chest Physicians (ACCP)* [4]. Het is op basis van dit laatste schema dat we de aanbevelingen ontwikkeld door de Domus Medica, vanaf nu en ook in de toekomst zullen '*graden*'.

Voor elke uitkomstmaat in een studie of meta-analyse bepaalt men de kwaliteit van de evidentie volgens deze elementen: studiedesign, studiekwaliteit, consistentie en directheid van bewijs

Niveaus van bewijskracht of methodologische kwaliteit van studies

Studiedesign

Het niveau van bewijskracht wordt in de eerste plaats bepaald door het *type studie of studiedesign*. Het krachtigste bewijs in de geneeskunde wordt geleverd door de klinische studies, waarbij de patiënten aselect een actieve of een placebobehandeling krijgen, de zogenaamde *RCT (Randomised Controlled Trial)*. Wanneer de deelnemende patiënten, de behandelaars en de onderzoekers daarbij niet op de hoogte zijn van de toegewezen behandeling (een dubbelblinde onderzoeksopzet), kunnen we aannemen dat alles in het werk werd gesteld om beïnvloeding of vertekening van de resultaten tegen te gaan. De resultaten van een correct uitgevoerde RCT met duidelijk klinische relevante uitkomsten (zoals mortaliteit, genezing enzovoort) wegen daarom zwaar door op de bewijsbalans. Ook *systematische reviews of meta-analyses* van verschillende RCT's rangschikken we bij aanvang hoog; het samenvoegen van de resultaten in een meta-analyse geeft daarbij ook een nauwkeuriger beeld van de grootte van het effect.

Het is echter niet altijd praktisch of ethisch mogelijk om een interventieonderzoek uit te voeren. Bij het optreden van een levensbedreigende anafylaxis kunnen we niet dubbelblind aan de ene groep adrenaline en aan de andere groep fysiologisch water toedienen. Dan is enkel *observationeel onderzoek* mogelijk, zonder het toetsen van een interventie of experimentele behandeling. Alle observationeel en beschrijvend onderzoek, met inbegrip van case-controleonderzoek en cohortonderzoek, scoren we in principe lager dan RCT-onderzoek.

Studiekwaliteit

Naast de hiërarchie van studietypes is het ook nodig om de *kwaliteit* van de bestudeerde artikels onder de loep te nemen. Er bestaan allerlei checklists of scorelijsten om artikels op hun waarde of validiteit te beoordelen wat diagnose, prognose en therapie betreft [5]. Daarbij staan aspecten zoals klinische relevantie, weergave van de resultaten, de onderzoeksmethode en statistische analyse centraal. Het onderscheiden van methodologische kernelementen binnen een bepaald onderzoekstype is van belang om te kunnen bepalen welke onderzoeken de hoogste validiteit vertegenwoordigen. Zo is het voor het evalueren van een diagnostische test belangrijk dat deze onafhankelijk en blind wordt vergeleken met de referentietest bij een voldoende grote groep van opeenvolgende patiënten. Wordt aan één of meerdere van deze vereisten niet voldaan, dan worden de conclusies als minder krachtig beschouwd [5]. Dit kan betekenen dat een RCT niet meer de hoogste maar slechts een matige graad van bewijskracht meekrijgt.

Consistentie van de resultaten en directheid van bewijs

Als verschillende RCT's erg verschillende schattingen van het effect van een behandeling als resultaat geven (heterogeniteit of variabiliteit van de resultaten), dan spreken we van een lage *consistentie* van de onderzoeksresultaten en is het belangrijk naar een verklaring te zoeken voor deze heterogeniteit. Behandelingen kunnen bijvoorbeeld een groter effect hebben bij erg zieke patiënten dan bij minder zieke populaties. Als men geen verklaring vindt voor het verschil, dan is dit eveneens een reden om een lager niveau van bewijskracht toe te kennen (*zie tabel 1*). De mate waarin de populatie, interventie en uitkomstmaten van het wetenschappelijk onderzoek of artikel overeenkomt met deze waarin men geïnteresseerd is voor de aanbeveling, noemt men de *directheid van bewijs*.

Er kan twijfel zijn over de directheid van het bewijs, als de patiënten in uw populatie veel ouder of meer comorbiditeit hebben dan deze in de studies. Zo is de gemiddelde leeftijd van patiënten in meeste therapiestudies bij hartfalen

rond de 60 jaar, wat veel lager is dan de gemiddelde leeftijd van patiënten met hartfalen in de huidige medische praktijk, veelal rond 74 jaar [6].

Andere factoren die het niveau van bewijskracht kunnen verlagen, zijn onnauwkeurige of gebrekkige data of publicatiebias. Als studies erg weinig patiënten of weinig events bevatten, dan zullen de uitkomsten ook minder precies zijn. Men verkrijgt dan grote betrouwbaarheidsintervallen, die iets zeggen over de nauwkeurigheid van de in het onderzoek gevonden waarden. Men spreekt van publicatiebias indien positieve onderzoeksresultaten een grotere kans op publicatie hebben dan onderzoek met een negatief of 'niet-significant' resultaat. Dan kan men onterecht gaan denken dat er toch een effect bestaat. Deze vorm van vertekening is belangrijk bij meta-analyses.

Relatief risico in observationele studies

Resultaten uit observationele studies kunnen toch een matige of zelfs hoge graad van evidentie worden toegekend, als een *belangrijke associatie* werd gevonden. Het relatief risico (RR) is daarbij een belangrijke maat, want het geeft de schatting aan van het aantal keer dat de kans op een bepaalde gebeurtenis, bijvoorbeeld om ziek te worden bij blootstelling aan een bepaalde risicofactor zoals roken, groter (of kleiner) is dan in de niet-blootgestelde groep, in casu de niet-rokers. Als dit risico twee keer zo groot is (of juist de helft), dan wel vijf keer zo groot (of slechts één vijfde), dan kunnen we uitgaan van een sterke of zeer sterke associatie. Eenzelfde bedenking kan men maken als stijgende dosissen van een bepaald geneesmiddel meer effect gaan geven (*dosis-responsgradiënt*); ook dan is er reden om een hoger niveau van bewijskracht toe te kennen.

Niveaus van bewijskracht: A, B en C

Voor elke uitkomstmaat in een studie of meta-analyse bepaalt men de kwaliteit van de evidentie volgens de bovenstaande elementen: studiedesign, studiekwaliteit, consistentie en directheid. Bij aanvang klasseert men een RCT bij een hoge kwaliteit van evidentie en elke observationele studie bij een lage kwaliteit van evidentie.

De GRADE-werkgroep heeft nog geen schema ontwikkeld voor diagnostisch onderzoek, maar men kan ervan uitgaan dat een goed opgezet diagnostisch onderzoek, met name een onafhankelijke blinde vergelijking van een diagnostische test met een referentietest, ook als hoge kwaliteit van evidentie kan worden beschouwd.

Het is dan op basis van een aantal factoren dat men de kwaliteit van de evidentie kan doen dalen of stijgen, wat uiteindelijk resulteert in drie niveaus van bewijskracht of kwaliteit van evidentie (*zie tabel 1*):

- Hoog niveau van bewijskracht, aangeduid met de letter A, waarbij verder onderzoek ons vertrouwen in de schatting van het effect zeer waarschijnlijk niet zal veranderen.
- Matig niveau van bewijskracht, aangeduid met de letter B, waarbij verder onderzoek waarschijnlijk een belangrijke invloed zal hebben op ons vertrouwen in de schatting van het effect en deze schatting zou kunnen veranderen.
- Laag niveau van bewijskracht, aangeduid met de letter C, waarbij verder onderzoek zeer waarschijnlijk een belangrijke invloed zal hebben op ons vertrouwen in de schatting van het effect en waarschijnlijk deze schatting zal veranderen OF eender welke schatting van het effect is zeer onzeker.

Tabel 1: Criteria voor het toewijzen van niveaus van bewijskracht.

Studietype:
- RCT's zonder beperkingen of sterk overtuigende evidentie van observationele studies = HOOG (A);
- RCT's met beperkingen of sterke evidentie van observationele studies = MATIG (B);
- observationele studies/ casestudies en RCT's met majeure beperkingen = LAAG (C).

Factoren die de methodologische kwaliteit van studies doen dalen:
- beperkingen van de studiekwaliteit,
- inconsistentie van de resultaten,
- indirectheid van evidentie (indirecte populatie, interventie, uitkomstmaten),
- onnauwkeurige of gebrekkige data (grote betrouwbaarheidsintervallen),
- kans op publicatiebias.

Factoren die de methodologische kwaliteit van studies doen stijgen:
- grootte van het effect of sterk bewijs van associatie (direct bewijs, RR>2 of RR<0,5) zonder mogelijke confounders of zeer sterk bewijs van associatie (direct bewijs, RR>5 of RR<0,2, geen bedreiging van validiteit);
- alle mogelijke confounders zouden het effect verminderd hebben;
- bewijs van een dosis-responsgradiënt.

Graad van aanbeveling

Sterke of zwakke graad: 1 of 2

Men kan een bepaalde kernboodschap uit de aanbeveling sterk (aangeduid met het cijfer 1) of zwak (aangeduid met het cijfer 2) aanbevelen.

Een keuze hiertussen zal worden gemaakt op basis van een afweging tussen de voor- en nadelen van een bepaalde interventie of actie. Van een sterke aanbeveling spreken we als de voordelen duidelijk de nadelen of risico's overtreffen. Bij een zwakke aanbeveling is er eerder een evenwicht tussen de voor- en nadelen of risico's. Soms is zelfs de balans tussen voor- en nadelen onzeker of zijn er zeker geen nettovoordelen. Als er voor- en nadelen met elkaar in balans zijn, is het tevens nuttig om de kosten af te wegen tegenover de nettowinst.

Wat bepaalt verder de sterkte van een aanbeveling?

- De methodologische kwaliteit van de studies en het resulterend niveau van bewijskracht;
- Het belang van de uitkomstmaat die de actie of behandeling voorkomt, waarbij harde eindpunten zoals mortaliteit of ernstige morbiditeit uiteraard belangrijker zijn dan de symptoombehandeling en dus de interventies die hierop werken, een sterkere aanbevelingsgraad krijgen.
- De grootte van het effect van de behandeling, bijvoorbeeld hoeveel risicoreductie levert een behandeling met statines op bij patiënten met diabetes mellitus type 2 in de primaire preventie van cardiovasculaire events.
- Het risico dat iets ook gebeurt, wat samenhangt met het basisrisico in bepaalde groepen, bijvoorbeeld het risico op cardiovasculaire pathologie bij diabetespatiënten ligt twee- tot viermaal hoger dan in de gewone bevolking, zodat kernboodschappen die hierop ingrijpen, sterker aan te bevelen zijn.
- Verschillen in waarden of voorkeuren van de doelgroep van de aanbeveling, bijvoorbeeld werkende populatie versus anderen bij de behandeling van acute diarree.
- De toepasbaarheid van de evidentie in een specifieke setting, waarbij men rekening houdt met de beschikbaarheid van de interventie of het geneesmiddel, de beschikbare expertise bij de huisartsen of gezondheidswerkers, organisatorische of financiële barrières. Hoe beter toepasbaar, des te sterker de aanbeveling.

De graad van aanbeveling heeft uiteraard directe implicaties voor de praktijk. Bij een sterke graad van aanbeveling zullen alle goed geïnformeerde patiënten dezelfde keuze maken. Bij een dergelijke kernboodschap gebruiken we een bewoording als *"we bevelen aan om iets al of niet te doen"* of *"doe dit of doe dit niet"*. Bij een zwakke aanbeveling is het aan de patiënt en zijn huisarts om tussen de verschillende mogelijkheden een keuze te maken. Dan gaan we bewoordingen gebruiken zoals *"we adviseren om iets al of niet te doen"* of *"doe dit of doe dit mogelijk niet"*.

Tabel 2: Graden van aanbeveling.

Graden van aanbeveling		Voordelen versus nadelen en risico's	Methodologische kwaliteit van de studies	Implicaties
1 A	Sterke aanbeveling, hoge graad van evidentie	Voordelen overtreffen duidelijk de nadelen of risico's.	RCT's zonder beperkingen of sterk overtuigende evidentie van observationele studies.	Sterke aanbeveling, kan worden toegepast bij de meeste patiënten en in de meeste omstandigheden.
1 B	Sterke aanbeveling, matige graad van evidentie	Voordelen overtreffen duidelijk de nadelen of risico's.	RCT's met beperkingen of sterke evidentie vanuit observationele studies.	Sterke aanbeveling, kan worden toegepast bij de meeste patiënten en in de meeste omstandigheden.
1 C	Sterke aanbeveling, lage of zeer lage graad van evidentie	Voordelen overtreffen duidelijk de nadelen of risico's.	Observationele studies of casestudies.	Sterke aanbeveling, maar dit kan veranderen als er hogere evidentie beschikbaar komt.
2 A	Zwakke aanbeveling, hoge graad van evidentie	Evenwicht tussen voor- en nadelen of risico's.	RCT's zonder beperkingen of sterk overtuigende evidentie van observationele studies.	Zwakke aanbeveling, de beste actie kan verschillen naargelang de omstandigheden, patiënten of maatschappelijke waarden.
2 B	Zwakke aanbeveling, matige graad van evidentie	Evenwicht tussen voor- en nadelen of risico's.	RCT's met beperkingen of sterke evidentie vanuit observationele studies.	Zwakke aanbeveling, de beste actie kan verschillen naargelang de omstandigheden, patiënten of maatschappelijke waarden.
2 C	Zwakke aanbeveling, lage of zeer lage graad van evidentie	Onzekerheid over voor- of nadelen – evenwicht tussen beide is mogelijk.	Observationele studies of casestudies of RCT's met majeure beperkingen.	Erg zwakke aanbeveling, alternatieven kunnen evengoed te verantwoorden zijn.

Het GRADE-systeem in de praktijk

De combinatie van een niveau van bewijskracht met een graad van aanbeveling leidt tot een cijfer-lettercombinatie van het GRADE-systeem. In *tabel 2* vind je alle mogelijkheden van combinatie tussen graden van aanbeveling en niveaus van evidentie of bewijskracht.

Laten we dit illustreren met enkele voorbeelden uit de aanbeveling *'Gebruik van medicatie bij urgenties'*:

- Het toedienen van aspirine bij een acuut myocardinfarct is ruimschoots onderbouwd met goed uitgevoerde RCT's en krijgt aldus een niveau van bewijskracht A toegekend. De studies tonen geen beperkingen. Omdat de voordelen van deze behandeling de nadelen of risico's duidelijk overstijgen en omdat de behandeling ook toepasbaar is bij de meeste patiënten in de meeste omstandigheden, bevelen we dit sterk aan. Dit resulteert dus in een GRADE 1A voor deze aanbeveling.
- Bij een hypertensieve crisis met symptomen kan men in afwachting van de MUG sublinguale nitraten toedienen. Outcomestudies in de huisartsenpraktijk over de antihypertensieve therapie bij hypertensieve crisis zijn er niet. Men kan enkel terugvallen op expertadviezen en studies in het ziekenhuismilieu. De studies, die vaak slechts observationele studies zijn, bieden slechts een indirect bewijs. We kunnen dus enkel een laag niveau van bewijskracht (niveau C) toekennen. Het is evenmin heel duidelijk of de voordelen wel duidelijk opwegen tegen mogelijke risico's zoals een verlaagde bloeddruk. We zullen dit dus zwak aanbevelen, wat uiteindelijk resulteert in een GRADE 2C.

Besluit

GRADE is een internationaal veel gebruikt systeem om zowel de kwaliteit van evidentie als de sterkte van aanbeveling op een heldere wijze te beoordelen. Deze aanpak kan richtlijnontwikkelaars helpen om weloverwogen beslissingen te nemen en kan leiden tot meer reflectie over de waarde van aanbevelingen bij de gebruikers.

Auteur

P. Van Royen is huisarts te Antwerpen, diensthoofd van de vakgroep Huisartsgeneeskunde van de Universiteit Antwerpen en voorzitter van de commissie Aanbevelingen van Domus Medica vzw.

Literatuur

1 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology, a basic science for clinical medicine. Boston: Little, Brown and Company, 1991.

2 Van Royen P. Niveaus van bewijskracht: levels of evidence. *Huisarts Nu* 2002; 31:54-7.

3 Atkins D, Best D, Briss P, at al. Grading quality of evidence and strength of recommendations. *BMJ* 2004;328:1490-500.

4 Guyatt G, Gutterman D, Baumann MH, et al. Grading strength of recommendations and quality of evidence in clinical duidelines: report from an American College of Chest Physicians Task Force. *Chest* 2006;129:174-81.

5 Offringa M, Assendelft W, Scholten R. Inleiding in evidence-based medicine. Klinisch handelen gebaseerd op bewijsmateriaal. Houten/ Diegem: Bohn Stafleu van Loghum, 2008.

6 Van Royen P, De Keulenaer G, et al. Systematisch literatuuronderzoek ter voorbereiding van de consensusvergadering rond het doelmatig gebruik van geneesmiddelen bij hartfalen in de ambulante behandeling. Brussel: Riziv, 2008.