FORMULIER Vb

voor het beoordelen van een

SYSTEMATISCHE REVIEW
VAN DIAGNOSTISCH ONDERZOEK

Evidence-Based RichtlijnOntwikkeling

Formulier Vb: beoordeling systematische review van onderzoek naar de waarde van een diagnostische test

Dit formulier is bestemd voor het beoordelen van systematische reviews van onderzoek naar de waarde van een diagnostische test waarin één of meer indextesten (de testen waarvan de diagnostische waarde geëvalueerd wordt) worden vergeleken met een referentietest (een zogenoemde gouden standaard).

NB: gelet op de complexiteit van deze materie is het gewenst dat iemand met expertise op het gebied van systematische reviews van diagnostisch onderzoek bij de beoordeling betrokken is.

Dit formulier is ontwikkeld door een werkgroep bestaande uit vertegenwoordigers van het Dutch Cochrane Centre, het Kwaliteitsinstituut voor de Gezondheidszorg CBO, het Nederlands Huisartsen Genootschap, het institute for Medical Technology Assessment, de Werkgroep Onderzoek Kwaliteit, het College voor Zorgverzekeringen, Zorgonderzoek Nederland (ZonMw) en de Orde van Medisch Specialisten en wordt ondersteund door het Nederlands Paramedisch Instituut, de Vereniging voor Integrale Kankercentra en de Werkgroep Infectieziektenpreventie.

Voor het beoordelen van de kwaliteit van andere typen onderzoek zijn eveneens formulieren ontwikkeld. Deze staan samengevat in onderstaande tabel.

Type onderzoek	Formulier
Dwarsdoorsnedeonderzoek (waarde diagnostische test)	I
Randomised controlled trial (RCT)	II
Cohortonderzoek	III
Patiënt-controleonderzoek	IV
Systematische review van	
RCT's (therapie en preventie)	Va
diagnostisch onderzoek	Vb
observationeel onderzoek (etiologie/"harm"/prognose)	Vc
Kosten-effectiviteitsanalyse	VI
Richtlijn	AGREE

Instructie beoordeling

- De bruikbaarheid van een publicatie voor een richtlijn wordt in de formulieren op drie facetten beoordeeld: validiteit, toepasbaarheid in de praktijk en toepasbaarheid in de Nederlandse gezondheidszorg
- Daarnaast wordt gevraagd om de belangrijkste kwantitatieve gegevens te extraheren en op een uniforme wijze te presenteren.
- De opmaak van de beoordelingsformulieren maakt het u makkelijk:
  a) op diverse plaatsen is een beslismoment ingebouwd: indien een publicatie op dat moment niet aan de vereisten van validiteit of toepasbaarheid voldoet hoeft u met de beoordeling niet verder te gaan.
  b) de criteria en manier van data-extractie worden telkens op de tegenoverliggende pagina kort toegelicht.

Zend opmerkingen en suggesties aangaande dit formulier naar cochrane@amc.uva.nl.

Vraag 1. *Vraagstelling.* Een systematische review dient uit te gaan van een expliciete klinische vraagstelling. De vraag dient bij voorkeur de volgende aspecten te bevatten: de doelpopulatie, de indextest, de referentietest en het beoogde resultaat (bijvoorbeeld het aantonen of uitsluiten van een aandoening, een hoge voorspellende waarde van een testuitslag of een laag percentage foutnegatieven). Een leidraad kan het Engelse acroniem “PICO” zijn: **P**atient, **I**ndextest(en), **C**omparison (gouden standaard of referentietest, waarvoor soms ook het beloop van de aandoening genomen wordt) en **O**utcome.

Vraag 2. *Zoekactie.* Uitgaande van de vraagstelling dient op gestructureerde wijze met behulp van duidelijk omschreven zoektermen in de internationale literatuur te worden gezocht. De zoektermen moeten zodanig gekozen zijn, dat naar verwachting nagenoeg alle relevante artikelen op het omschreven gebied gevonden zullen worden. Een minimum vereiste omtrent de databases is een zoekactie in MEDLINE en controle van de referenties van de verzamelde artikelen. Indien niet in MEDLINE gezocht is, hoeft u met de beoordeling niet verder te gaan. Of een aanvullende zoekactie in EMBASE of specialistische databases zoals Cinahl en Psyclit meerwaarde heeft bij het zoeken naar onderzoek betreffende de waarde van een diagnostische test, is niet bekend. Optionele aanvullende zoekacties zijn gedrukte indexen, het aanschrijven van experts en fabrikanten en het handmatig doorzoeken van tijdschriften. Eveneens zullen de eventuele restricties in de zoekactie beschreven moeten zijn. Hierbij kan gedacht worden aan restrictie in jaartal van publicatie, publicatiestatus (wel of niet gepubliceerd) of in taal. Hoewel geen empirisch onderzoek gedaan is naar de invloed van taalrestricties op het resultaat van een diagnostische review (bijvoorbeeld alleen inclusie van Engelstalige onderzoeken) kan een beperking tot één of meerdere talen leiden tot minder precieze, en mogelijk ook vertekende, schattingen.
Vanzelfsprekend moet de zoekactie aansluiten op de vraagstelling zoals beoordeeld in vraag 1.

Vraag 3. *Selectie*. Een uitputtende zoekactie (vraag 2) zal resulteren in een groot aantal treffers. Veel aldus gevonden artikelen zullen niet relevant zijn voor de systematische review en dus buiten de systematische review gelaten moeten worden. De selectie van artikelen die geïncludeerd moeten worden, dient bij voorkeur door twee reviewers onafhankelijk van elkaar te geschieden aan de hand van expliciet beschreven in- en exclusiecriteria. Een minimum van twee reviewers voor de definitieve selectie is bedoeld om de kans op selectiebias in te perken.
Onder in- en exclusiecriteria vallen bijvoorbeeld: een specifieke onderzoekspopulatie (bijvoorbeeld patiënten in de eerste lijn), gebruik van een bepaalde referentietest, etc. De in- en exclusiecriteria dienen aan te sluiten op de vraagstelling en dienen expliciet beschreven te zijn om de reproduceerbaarheid van de systematische review te verhogen. Tevens moet de selectie vanaf de initiële database zoekactie tot de uiteindelijke inclusie in de systematische review duidelijk beschreven zijn.

Vraag 4. *Kwaliteitsbeoordeling*. De kwaliteit van de opgenomen onderzoeken dient beoordeeld te zijn. Indien geen kwaliteitsbeoordeling heeft plaatsgevonden, hoeft u niet met de beoordeling verder te gaan. Idealiter dient de kwaliteitsbeoordeling door twee reviewers onafhankelijk van elkaar aan de hand van een expliciete lijst van criteria beoordeeld te zijn. De onafhankelijke beoordeling verkleint de kans op informatiebias.
De kwaliteit van de opgenomen onderzoeken is mede bepalend voor de validiteit van de conclusies van en aanbevelingen in de systematische review. Er zijn verschillende criterialijsten om de kwaliteit van diagnostisch onderzoek te beoordelen, maar gevalideerde checklists zijn niet voorhanden. Van een aantal criteria is door empirisch onderzoek de vertekenende invloed (overschatting van testkwaliteiten) aangetoond. Het betreft de criteria *case-control* design (een design waarin de niet-zieken niet representatief zijn voor de doelgroep, bijvoorbeeld gezonde vrijwilligers in plaats van de gangbare patiëntenpopulatie), het toepassen van verschillende referentietesten voor verschillende deelnemers, onvoldoende beschrijving van de indextest en onvoldoende beschrijving van de onderzochte populatie. De gebruikte criterialijst zal derhalve tenminste deze criteria moeten bevatten. Van de overige items, zoals bijvoorbeeld het ontbreken van onafhankelijke (blinde) beoordeling van de indextest en referentietest (review bias) en toepassen van de referentietest afhankelijk van het resultaat van de indextest (work-up bias), is niet overtuigend empirisch aangetoond dat de eventuele vertekening systematisch is. Het resultaat van de kwaliteitsbeoordeling per individueel onderzoek zal in het artikel terug te vinden moeten zijn.
De wijze waarop tot consensus gekomen wordt en de mate van overeenstemming tussen de reviewers dient beschreven zijn.

Beoordeling van de kwaliteit van een systematische review van onderzoek naar de waarde van een diagnostische test

Naam beoordelaar:.....Datum:.....

Titel: .....

Auteurs: .....

Bron: .....

Beoordeling van de validiteit

Korte beschrijving van de onderzochte indextest(en): .....

.....

Korte beschrijving van de referentietest(en): .....

.....

METHODEN

1. Is de vraagstelling adequaat geformuleerd?
   [ ] Ja
   [ ] Nee
   [ ] Te weinig informatie in het artikel om dit te beantwoorden

2. Is de zoekactie adequaat uitgevoerd?
   [ ] Ja
   [ ] Nee
   [ ] Te weinig informatie in het artikel om dit te beantwoorden

3. Is de selectieprocedure van artikelen adequaat uitgevoerd?
   [ ] Ja
   [ ] Nee
   [ ] Te weinig informatie in het artikel om dit te beantwoorden

4. Is de kwaliteitsbeoordeling adequaat uitgevoerd?
   [ ] Ja
   [ ] Nee
   [ ] Te weinig informatie in het artikel om dit te beantwoorden

Vraag 5. *Data-extractie.* Data-extractie is het verzamelen van de resultaten uit de oorspronkelijke onderzoeken. In geval van diagnostisch onderzoek zal dit doorgaans de samenvattende 2x2 tabellen betreffen. Soms dienen de reviewers deze tabellen zelf te reconstrueren. Daarom dient het data-extractieproces duidelijk beschreven te zijn en weer bij voorkeur door twee reviewers onafhankelijk van elkaar plaatsgevonden te hebben.

Vraag 6. *Beschrijving oorspronkelijke onderzoeken.* De belangrijkste kenmerken van de oorspronkelijke onderzoeken moeten in de systematische review beschreven zijn. Wat belangrijk is, hangt af van het onderwerp. Tenminste beschreven moeten zijn:

- *onderzochte populatie*: per onderzoek tenminste de leeftijds- en geslachtsverdeling, aard, ernst en duur van de klachten, spectrum van de zieken en niet-zieken
- *index- en referentietest*: aard, wijze van uitvoering en criteria voor duiden testresultaat; indien van toepassing dienen de gehanteerde afkappunten voor het benoemen van een negatieve en positieve test te zijn vermeld.
- *indextestkarakteristieken*: per indextest samenvattende 2x2 tabel (aan de hand waarvan alle gewenste parameters berekend kunnen worden) of aantal onderzochte personen, prevalentie van de onderzochte aandoening, sensitiviteit, specificiteit, likelihoodratio van positieve en negatieve indextestuitslag.

Vraag 7. *Meta-analyse.* Omdat verschillen in resultaten tussen de afzonderlijke onderzoeken vaak voorkomen, is het niet eenvoudig om bij diagnostische reviews tot samenvattende, meer precieze schattingen van de verschillende parameters te komen. Een diagnostische review zal dan ook in de eerste plaats de bronnen van eventuele variabiliteit moeten onderzoeken. In een aantal gevallen is het wel mogelijk om, bij voldoende homogeniteit, tot samenvattende schattingen te komen, hoewel nog geen consensus bestaat over de ideale wijze om een diagnostische meta-analyse uit te voeren.
De volgende strategie (zie ook de figuur op de laatste pagina) kan een leidraad zijn bij het lezen en beoordelen.
1. Beoordeel de homogeniteit van de resultaten van de onderzoeken door deze weer te geven in een spreidingsdiagram met sensitiviteit op de Y-as en specificiteit (of 1-specificiteit) op de X-as.
2. Indien zowel de sensitiviteit als de specificiteit van alle onderzoeken onderling homogeen zijn (als voldoende vergelijkbaar beschouwd kunnen worden), kunnen beide parameters gepoold worden. Dat geldt ook voor de likelihoodratio van een positieve test (LR+) en de likelihoodratio van een negatieve test (LR–) die rechtstreeks voortkomen uit sensitiviteit en specificiteit.
3. In een diagnostische meta-analyse kan heterogeniteit van sensitiviteit en/of specificiteit veroorzaakt worden door verschillen tussen de onderzoeken in de keuze van het afkappunt voor het positief duiden van de indextest. Een diagnostische meta-analyse dient dan ook aan dit aspect aandacht te besteden. Indien sprake is van afhankelijkheid van een afkappunt, zijn sensitiviteit en specificiteit negatief gecorreleerd en kunnen de resultaten worden weergegeven in een samenvattende receiver operating characteristic (ROC) curve (*Summary ROC curve,* SROC-curve) (zie vraag 9 voor een interpretatie hiervan).
4. Indien de resultaten heterogeen zijn en er geen aanwijzingen zijn voor afhankelijkheid van een afkappunt, dient meta-analyse achterwege te blijven of dient onderzocht te worden of de heterogeniteit het gevolg is van klinische of methodologische verschillen tussen de onderzoeken (klinische of methodologische heterogeniteit). Opdelen van de onderzoeken in vergelijkbare subgroepen kan de heterogeniteit eventueel weer doen verdwijnen. Aan dat laatste zit echter het nadeel verbonden van een doelbewuste reconstructie: door het achteraf schuiven met groepen is het namelijk vrijwel altijd mogelijk om *post hoc* (kleine) groepen samen te stellen waarin homogeniteit statistisch niet kan worden verworpen.

Vraag 8. *Algemeen oordeel*. Hier wordt een inschatting gevraagd van de validiteit (Is dit een goede systematische review?) en toepasbaarheid (Gaat deze review over mijn PICO?). Let hierbij ook op eventuele fouten in de review onderzoek die funest zijn voor de validiteit ervan (*red flags, fatal flaws*). Er zijn geen regels te geven voor welke items positief gescoord moeten worden of hoeveel items tenminste positief gescoord moeten worden. Dit is deels afhankelijk van de “state-of-the-art” met betrekking tot het betreffende onderwerp. Het gaat er hier om het samenvattend oordeel van wat de beoordelaar de werkgroep zou willen mededelen over de bruikbaarheid van het artikel voor de besluitvorming.

5. Is adequaat beschreven hoe data-extractie heeft plaatsgevonden?
   [ ] Ja
   [ ] Nee
   [ ] Te weinig informatie in het artikel om dit te beantwoorden

6. Zijn de belangrijkste kenmerken van de oorspronkelijke onderzoeken beschreven?
   [ ] Ja
   [ ] Nee
   [ ] Te weinig informatie in het artikel om dit te beantwoorden

7. Is meta-analyse op een correcte manier uitgevoerd?
   [ ] Ja / niet van toepassing (terecht geen meta-analyse uitgevoerd)
   [ ] Nee
   [ ] Te weinig informatie in het artikel om dit te beantwoorden

ALGEMEEN OORDEEL

8. Zijn de resultaten van de systematische review valide en toepasbaar?

[ ] Voldoende valide en toepasbaar	⇐ ga verder bij 9
[ ] Twijfelachtig	⇐ ga verder bij 9; voorzichtig oordeel bij 12
[ ] Onvoldoende valide en toepasbaar	U kunt stoppen met het invullen van de checklist, tenzij er geen betere artikelen op dit gebied zijn (terugkoppelen naar de werkgroep)

Vraag 9*. Resultaten*.

NB: zie ook de toelichting bij vraag 7!

Indien sprake is van homogene resultaten (maar dit is meer uitzondering dan regel!), kunnen de berekende parameters (sensitiviteit, specificiteit, LR+, LR–) uit de systematische review worden overgenomen. Noteer de gepoolde schatters en hun 95%-betrouwbaarheidsinterval. (NB: in formulier I wordt de betekenis van de verschillende diagnostische parameters toegelicht)

Gepoolde likelihoodratio's kunnen ook berekend worden aan de hand van gepoolde sensitiviteit en specificiteit:

Likelihoodratio positieve test (LR+) = Sens / (1–Spec)
Likelihoodratio negatieve test (LR–) = (1–Sens) / Spec

Indien sprake is van heterogene resultaten die het gevolg zijn van het toepassen van verschillende afkappunten, dient de SROC-curve zelf geïnterpreteerd te worden. Dit vergt kennis van zaken en beoordelingsvermogen. Een grove leidraad is de volgende.

1. Hoe groter het oppervlak tussen de curve en de diagonaal, des te beter het onderscheidingsvermogen van de test. (Deze benadering is vooral nuttig indien in een systematische review meerdere testen onderling vergeleken worden.)
2. Als het doel van de test het uitsluiten van de onderzochte aandoening is, moet de sensitiviteit van de indextest hoog zijn. De SROC-curve bevindt zich dan in het bovengedeelte van de grafiek. Bij hoge sensitiviteit is de voorspellende waarde van een negatieve testuitslag hoog: als de indextest negatief is, is de kans dat de ziekte inderdaad afwezig is, groot.
3. Als het doel van de test het insluiten van de onderzochte aandoening is, moet de specificiteit van de indextest hoog zijn. De SROC-curve bevindt zich dan in het linker gedeelte van de grafiek. Bij hoge specificiteit is de voorspellende waarde van een positieve testuitslag hoog: als de indextest positief is, is de kans dat de ziekte inderdaad aanwezig is, groot.

Voor het inschatten van de waarde van een test kan ook nog gekeken worden naar voorspellende waarden (achterafkansen) bij verschillende voorafkansen (prevalentie) op de aandoening. Hiertoe dienen eerst likelihoodratio's berekend te worden waarna de voorspellende waarden afgeleid kunnen worden aan de hand van het nomogram (zie laatste pagina).

1. Als gepoolde schatters voorhanden zijn, kunnen de gepoolde LR+ of LR– als uitgangspunt genomen worden.
2. Indien uitgegaan moet worden van een SROC-curve: kies op de curve (= combinatie van sensitiviteit en (1–specificiteit) de laagste, gemiddelde en hoogste sensitiviteit en lees de bijbehorende specificiteiten af. Bereken dan voor iedere combinatie de LR+ en LR–.
3. Bepaal aan de hand van de likelihoodratio's voor verschillende voorafkansen en indextestresultaten (positief of negatief) de voorspellende waarden van de test.

Voorbeeld: LR+ = 10 en de voorafkans ("pretest probability") op de aandoening = 20%.
Trek een rechte lijn door het punt 20 op de linker as en het punt 10 op de middelste as. Waar de lijn de rechter as ("post-test probability") snijdt wordt de achterafkans afgelezen: een positieve indextestuitslag bij een voorafkans van 20% is ongeveer 75%. De kans op ziekte neemt derhalve toe van 20% naar 75% als de indextest positief is.

9. Resultaten:

Indextest: .....

Referentietest: .....

Aantal onderzoeken: ..

Indien sprake is van homogene resultaten:

	Waarde	95%-BI
Sensitiviteit		
Specificiteit		
Likelihoodratio van een positieve test (LR+)		
Likelihoodratio van een negatieve test (LR–)		

Voorspellende waarden bij voorafkans	Achterafkans bij een	
	positieve testuitslag (VW+)	negatieve testuitslag (VW–)
%		
%		
%		

In geval van een SROC-curve:

	Bijbehorende specificiteit	LR+	LR–
1. Laagste sensitiviteit			
2. Gemiddelde sensitiviteit			
3. Hoogste sensitiviteit			

Voorspellende waarde van een positieve testuitslag (VW+):

	Achterafkans bij		
Voorafkans	1. laagste sensitiviteit	2. gemiddelde sensitiviteit	3. hoogste sensitiviteit
%	%	%	%
%	%	%	%
%	%	%	%

Voorspellende waarde van een negatieve testuitslag (VW–):

	Achterafkans bij		
Voorafkans	1. laagste sensitiviteit	2. gemiddelde sensitiviteit	3. hoogste sensitiviteit
%	%	%	%
%	%	%	%
%	%	%	%

Vraag 10 en 11. *Toepasbaarheid in de Nederlandse gezondheidszorg.* Beide vragen zijn een belangrijk onderdeel van richtlijnontwikkeling en dienen daarom in de werkgroep bediscussieerd te worden.

Vraag 12. *Conclusie met betrekking tot de systematische review en de waarde van de diagnostische test*

Geef hier een globale samenvatting van het eindoordeel over de systematische review. Probeer, indien aanwijzingen bestaan voor vertekening van de resultaten, tenminste een inschatting te maken van de richting van de vertekening (overschatting of onderschatting van de testkarakteristieken) en zo mogelijk ook over de grootte van de vertekening. Eventuele aanwijzingen voor mogelijke belangenverstrengeling van de auteurs met belanghebbende opdrachtgevers kunt u hier ook rapporteren. Ook is het verstandig ingezonden brieven en/of redactionele commentaren op het hier door u beoordeelde onderzoek te raadplegen bij het formuleren van uw conclusie.

Voorbeeld: “Goed opgezette review. Er is een uitgebreide zoektocht zonder restricties uitgevoerd aan de hand van adequaat gekozen zoektermen. De opgenomen onderzoeken zijn voor het merendeel van lage kwaliteit. De VW+ en VW- van de onderzochte test zijn zodanig laag voor een aantal voorafkansen dat het weinig zinvol is voor een huisarts deze test toe te passen.”

TOEPASBAARHEID IN DE NEDERLANDSE GEZONDHEIDSZORG

10. Kan het gevonden resultaat worden toegepast op de Nederlandse situatie?
    (hierbij valt bijvoorbeeld te denken aan de beschikbare faciliteiten)

    [ ] Ja
    [ ] Nee
    [ ] Te weinig informatie in het artikel om dit te beantwoorden

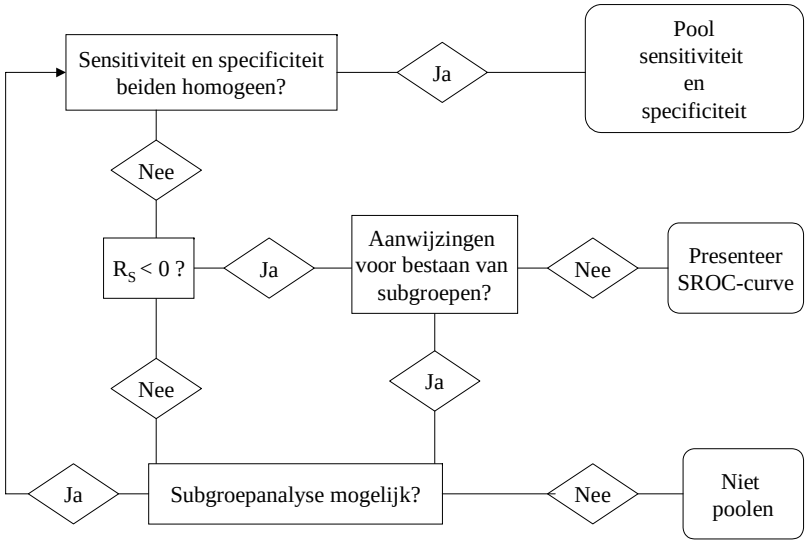
11. Op welk(e) echelon(s) kan het resultaat worden toegepast?
    (meerdere opties tegelijk mogelijk)

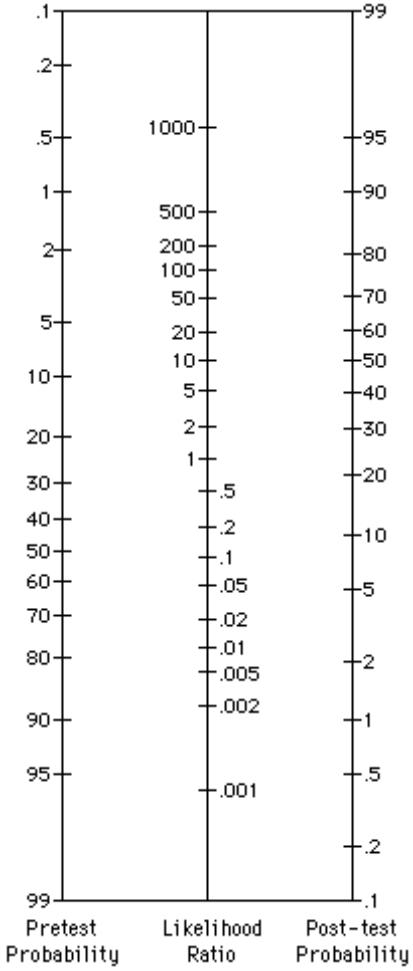
    [ ] algemene bevolking
    [ ] eerste lijn
    [ ] tweede lijn
        [ ] academische ziekenhuizen
        [ ] perifere ziekenhuizen
    [ ] derde lijn

CONCLUSIE

12. Conclusie met betrekking tot de systematische review en de waarde van de diagnostische test(s)

SCHEMA VOOR HET UITVOEREN VAN EEN DIAGNOSTISCHE META-ANALYSE



R_S = correlatiecoëfficiënt van Spearman; SROC = summary receiver operating characterictic

NOMOGRAM VOOR HET BEPALEN VAN DE ACHTERAFKANS AAN DEHAND VAN DE VOORAFKANS EN LIKELIHOODRATIO