Podosophia (2018) 26:130–132
https://doi.org/10.1007/s12481-018-0199-4

# Diagnostische waarden en beoordelaarsbetrouwbaarheid

**Sil Aarts · Anja de Koning**

Published online: 11 July 2018


**Samenvatting** Testuitslagen spelen een grote rol in de medische wereld. Zo worden bijvoorbeeld medicijnen of interventies voorgeschreven op basis van testuitslagen en onderzoeksresultaten. Diagnostische waarden en beoordelaarsbetrouwbaarheid spelen in dergelijke tests een belangrijke rol.

**Trefwoorden** sensitiviteit · specificiteit · kappa · ICC

## Inleiding

Als één na laatste analysemethode van deze statistiekreeks komen twee verschillende zaken aan bod: diagnostische waarden en de beoordelaarsbetrouwbaarheid. Deze methoden worden dikwijls gebruikt in podotherapeutisch onderzoek. Dit artikel zal zich in het eerste deel richten op diagnostische waarden. Het tweede deel gaat over betrouwbaarheid.

## Diagnostische tests

Testuitslagen spelen een grote rol in de medische wereld. Medicatie is gebaseerd op testuitslagen en ook interventies worden op basis van testuitslagen ingezet. In de podotherapeutische praktijk leveren diverse testinstrumenten en onderzoeksmethoden een bijdrage aan de diagnostiek: voetdrukmetingen, echografie en het bepalen van de enkel-armindex zijn hier voorbeelden van. De uitslagen van dergelijke onderzoeksmethodieken kunnen leiden tot het al dan niet inzetten van interventies. Het is daarom van belang om na te gaan welke toegevoegde waarde een test heeft.

Bij de keuze van een diagnostische test speelt de voorspellende waarde van die test een belangrijke rol. Een *voorspellende waarde* is de kans dat een uitslag, bijvoorbeeld een testuitslag 'diabetes mellitus type 2' (DM2) juist is. We onderscheiden een positief en een negatief voorspellende waarde. Een *positief voorspellende waarde* is het deel van de onderzochte patiënten met een positieve testuitslag die de ziekte, in dit geval DM2, daadwerkelijk heeft. De *negatief voorspellende waarde* is het deel van de onderzochte patiënten met een negatieve testuitslag die, in dit geval, geen DM2 heeft.

## Diagnostische waarde

*Sensitiviteit* en *specificiteit* (zie kader) worden ook gebruikt in de podotherapie. Idealiter zijn zowel de sensitiviteit als de specificiteit van een test 100%. In de praktijk is dit echter een utopie: er zijn geen instrumenten of testmethoden die dit percentage halen. Meestal wordt er afgewogen wat belangrijker is: bij een patiënt een negatieve diagnose stellen terwijl de ziekte wel aanwezig is (een ziekte 'missen') of een patiënt een diagnose meegeven terwijl hij de ziekte niet heeft ('loos alarm'). Dit laatste lijkt de voorkeur te hebben; liever een diagnose krijgen en de ziekte niet hebben dan andersom. Vooral bij aandoeningen zoals kanker of aids speelt de balans tussen specificiteit en sensitiviteit echter een grote rol. Een hoge sensitiviteit is nodig omdat je geen ziekte 'over het hoofd wil zien', maar ook de specificiteit moet hoog zijn; een fout-positieve diagnose, 'loos alarm', kan na-

In deze rubriek dragen de auteurs een steentje bij aan het vergroten van de kennis over wetenschappelijk onderzoek en de toepasbaarheid ervan in de podotherapeutische praktijk.

dr. S. Aarts (✉)
Fontys Paramedische Hogeschool, Eindhoven, Nederland
s.aarts@fontys.nl

A. de Koning
Vrije Universiteit Amsterdam, Amsterdam, Nederland

**Tabel 1** Samenhang tussen ziekte en testuitslag

| | DM2 aanwezig | DM2 afwezig | Totaal | Formule |
|---|---|---|---|---|
| Test positief | cel A: echt-positieven (TP[a]) | cel B: fout-positieven (FP[a]) | TP + FP | PVW = TP / (TP + FP) |
| Test negatief | cel C: fout-negatieven (FN[a]) | cel D: echt-negatieven (TN[a]) | TN + FN | NVW = TN / (TN + FN) |
| Totaal | TP + FN | FP + TN | | |

*TN* true negatives, *TP* true positives, *FN* false negatives, *FP* false positives, *NVW* negatief voorspellende waarde, *PVW* positief voorspellende waarde
[a]Afkortingen van de Engelse termen die ook in Nederland worden gebruikt

melijk bij dergelijke ziektes ook grote consequenties hebben voor de patiënt en zijn naasten.

In tab. 1 staat een voorbeeld. Aan de ontwikkeling van nieuwe technologische toepassingen in de zorg wordt veel tijd en geld besteed. Stel er wordt een nieuwe app ontwikkeld, waarmee je, met behulp van een teststrip, een druppel bloed kunt analyseren en in een paar minuten kan vaststellen of iemand DM2 heeft. Cel A betreft dan de mensen die écht DM2 hebben en die door de testuitslag van de app ook als zodanig worden gediagnosticeerd. De positief voorspellende waarde wordt dan berekend als de verhouding: *true positives* / positieve testen = TP / (TP + FP).

De sensitiviteit en specificiteit worden dan als volgt berekend:

- sensitiviteit = TP / (TP + FN)
- specificiteit = TN / (FP + TN)

### Sensitiviteit en specificiteit

Sensitiviteit en specificiteit gebruik je om de uitslag van een nieuwe test of meetmethode te vergelijken met een 'gouden standaard' (een instrument dat dient als ijkpunt). Let wel, dikwijls moet de (para)medische praktijk zich tevreden stellen met een 'gouden standaard', terwijl die vaak niet berust op de absolute 'waarheid'. Denk aan de ziekte van Alzheimer, een aandoening die pas postmortem daadwerkelijk kan worden vastgesteld, terwijl de beschikbare klinisch-diagnostische criteria als 'gouden standaard' gelden.

*Sensitiviteit*

Sensitiviteit heeft betrekking op personen die door de gouden standaard als positief worden bestempeld (de ziekte is aanwezig). Sensitiviteit is dan het percentage van die groep dat ook door de nieuwe test als positief wordt bestempeld: de correct geïdentificeerde positieven. Om met grote zekerheid te kunnen vaststellen dat een persoon een ziekte heeft, is een hoge sensitiviteit nodig. Een hoge sensitiviteit is vooral belangrijk als je fout-negatieve uitslagen wilt voorkomen: patiënten die een negatieve uitslag of diagnose krijgen, bij wie de ziekte wel aanwezig is.

*Specificiteit*

Specificiteit heeft betrekking op personen die door de gouden standaard als negatief worden bestempeld (de ziekte is afwezig). Specificiteit is dan het percentage van die groep dat ook door de nieuwe test als negatief wordt bestempeld: de correct geïdentificeerde negatieven. Een hoge specificiteit is vooral belangrijk als je doel is om weinig fout-positieve uitslagen te krijgen, dus: er zijn weinig patiënten die een positieve testuitslag krijgen, terwijl de ziekte afwezig is ('loos alarm').

## Beoordelaarsbetrouwbaarheid

Voor wetenschappelijk onderzoek is het belangrijk om te weten hoe groot de overeenstemming is tussen bijvoorbeeld beoordelaars of bij een test-hertest (twee keer een meting uitvoeren, met dezelfde test, bij dezelfde deelnemers). Zowel bij subjectieve metingen (bijv. kwaliteit van leven) als objectieve metingen (bijv. bloeddruk) is het belangrijk om te weten of en in welke mate uitkomsten consistent zijn over beoordelaars of tests. Er worden twee betrouwbaarheden onderscheiden: intrabeoordelaarsbetrouwbaarheid (de mate van overeenstemming tussen twee of meer metingen die worden uitgevoerd door één beoordelaar) en de interbeoordelaarsbetrouwbaarheid (de mate van overeenstemming tussen metingen die wordt uitgevoerd door twee of meer beoordelaars).

Een veel gebruikte maat in podotherapeutisch (wetenschappelijk) onderzoek is *kappa*. Kappa is de overeenstemming tussen kwalitatieve beoordelingen. De kappa wordt gebruikt voor het bepalen van de betrouwbaarheid van dichotome schalen (bijv. slecht *vs.* goed) of ordinale schalen (bijv. slecht, voldoende, goed) [1]. In tab. 2 zijn de gegevens van twee beoordelaars te zien. Stel dat deze tabel gaat over twee podotherapeuten die echografie gebruiken om te beoordelen of er bij hun patiënten sprake is van weefselschade aan de achillespees. Er werden in totaal 100 patiënten gezien. Elke beoordelaar heeft de achillespees van 100 patiënten bekeken en beoordeeld of er sprake was van weefselschade (weefselschade aanwezig) of niet (weefselschade afwezig). De geobserveerde overeenkomst tussen beide beoordelingen is dan ook 60 / 100 = 60 %. Dit lijkt hoog, maar de kappa is slechts 0,2. De kappa heeft een waarde tussen 0 en 1. Een lage kappawaarde geeft aan dat de overeenstemming tussen beoordelaars op toeval berust; een hogere waarde duidt erop dat de beoordelingen vaker

**Tabel 2** Kruistabel van beoordelingen tussen twee onafhankelijke beoordelaars

| | | Beoordelaar I | | |
|---|---|---|---|---|
| | | Weefselschade aanwezig | Weefselschade afwezig | Totaal |
| Beoordelaar II | Weefselschade aanwezig | 20 | 30 | 50 |
| | Weefselschade afwezig | 10 | 40 | 50 |
| | Totaal | 30 | 70 | 100 |

dan op basis van toeval, met elkaar overeenstemmen. Sporadisch komt een negatieve kappa voor. Dit duidt op het ontbreken van overstemming in de beoordelingen. Als leidraad geldt dat een waarde ten minste 0,60/0,70 moet aannemen om te kunnen spreken van een goede betrouwbaarheid. Zie het kader voor de manier waarop kappa wordt berekend.

Als je in kaart wil brengen wat de overeenstemming is tussen twee beoordelingen die vaak numerieke waarden bevatten, wordt de intraclasscorrelatiecoëfficiënt (ICC) gebruikt [2]. Stel dat twee beoordelaars punten geven voor een toets waarop een student minimaal een 1 kan score en maximaal een 10. Net als de kappa kan de ICC een waarde aannemen tussen de 0 en de 1. Over het algemeen wordt gesteld dat er bij een ICC van 0,70 een redelijke overeenstemming is. Een hoge ICC duidt op kleine verschillen tussen de beoordelingen en een lage ICC op grotere verschillen in de beoordeling [2].

## Kanttekening bij kappa en ICC

De indeling van kappa en ICC is arbitrair: 'slecht' (< 0), 'gering' (0–0,20), 'matig' (0,21–0,40), 'redelijk' (0,41–0,60), 'voldoende tot goed' (0,61–0,80) en 'bijna perfect' (0,81–1,00) [1]. Het is dan ook zaak om bij de interpretatie van dergelijke maten altijd de klinische relevantie in het achterhoofd te houden: "Wat betekent dit resultaat voor de maatschappij of de klinische praktijk?" en "Wat betekent dit resultaat voor mij als podotherapeut en voor mijn patiënten"?

Alle informatie is te vinden in diverse handboeken, zoals die van Field [3] en Howell [4].

### Berekening kappa

Kappa = (Pr(o) – Pr(e)) / (1-Pr(e))*
PR(o) = (20 + 40) / 100 = 0,60
PR(e) = kans op 'aanwezig' + kans op 'afwezig'
Aanwezig: (30 × 50) / 100 = 15; 15 / 100 = 0,15
Afwezig: (70 × 50) / 100 = 35; 35 / 100 = 0,35
Aanwezig + afwezig = 0,15 + 0,35 = 0,5

*Kappa* = (Pr(o) – Pr(e)) / (1-Pr(e)) = (0,6 – 0,50) / (1 – 0,5) = 0,2

* Pr(o) = *P* van *proportion, o* van *observed* = de waargenomen overeenstemming.
* Pr(e) = *P* van *proportion,* e van *expected* (*by chance*) = de verwachte overeenstemming bij onafhankelijkheid (gegeven de randtotalen van een tabel).

## Take home message

In de medische wereld is het van belang dat testinstrumenten of methoden, wanneer er twee gebruikt worden, dezelfde resultaten laten zien. Dat geldt ook voor twee beoordelaars: bij een en dezelfde patiënt moet de diagnose van podotherapeut A gelijk zijn aan de diagnose van podotherapeut B. Om overeenstemming tussen toetsinstrumenten en beoordelingen te onderzoeken, kunnen maten zoals sensitiviteit, specificiteit, kappa en ICC worden gebruikt.

## Vooruitblik

Het volgende en laatste artikel in deze reeks over statistische toetsen zal zich focussen op correlatieanalyses.

## Literatuur


1. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.
2. Koch G. Intraclass correlation coefficient. In: Kotz S, Johnson NL, redactie. Encyclopedia of statistical sciences. Deel 4. New York: John Wiley & Sons; 1982.
3. Field A. Discovering statistics using SPSS, deel 9. 3e druk. Londen: SAGE; 2009.
4. Howell DC. Statistical methods for psychology. 8e druk. Boston: Cengage Learning; 2012.



**dr. Sil Aarts,** docent/onderzoeker

**Anja de Koning,** beleidsmedewerker Faculteit der Gedrags- en Bewegingswetenschappen