*Ton Broeders**

# Rekenen met de psychiater – over de diagnostische waarde van de risicotaxatie

*'Maar het moet voor u als psychiater toch raar zijn om te horen dat uw taxatie ongeveer dezelfde juistheid heeft als kop of munt opgooien.'*
*Paul Witteman in nagesprek bij 'Max, een leven in de tbs', uitgezonden door BNNVARA op 17 juni 2019.*

*De huidige discussie over de tbs-praktijk roept onder andere de vraag op naar de diagnostische waarde van de beschikbare risicotaxatie-instrumenten: in welke mate is de uitslag van een daartoe strekkende test diagnostisch voor de kans op recidive? In de forensische psychiatrie wordt de predictieve validiteit van een risicotaxatietest doorgaans uitgedrukt in een zogenaamde auc-waarde, een getal tussen 0 en 1, waarbij 0,5 staat voor kansniveau. Helaas is de auc-waarde conceptueel nog minder toegankelijk dan de diagnostische waarde of likelihood ratio waarop zij is gebaseerd en die de laatste decennia, behalve in de medische en psychologische wetenschap, met name ook in de forensisch-technische disciplines breed wordt gehanteerd. Voorts blijkt dat kwalificaties als 'veel waarschijnlijker' of 'zeer veel waarschijnlijker', die gangbaar zijn als verbale equivalenten voor de getalsmatige diagnostische waarde van forensisch-technische bevindingen, en 'goed' of 'zeer goed', die gebruikt worden als verbale equivalenten voor de auc-waarde van recidiverisicotaxatietesten, sterk verschillen in de ordegrootte van bewijskracht van de kwantitatieve gegevens waar zij voor staan.*
*Anders dan de in de literatuur vermelde verbale auc-waarden van de onderzochte risicotaxatietesten doen vermoeden blijkt hun diagnostische waarde zeer bescheiden. Om verwarring over de betekenis van de predictiewaarde van een test te voorkomen is vermelding van de toegepaste detectiewaarde, de daarmee samenhangende diagnostische waarde en de gehanteerde a-priorikans essentieel. Het zijn deze waarden, en niet de auc-waarde van de test, die de betekenis van de uitgevoerde test het meest toegankelijk maken.*

## 1. De aanbevelingen van de Onderzoeksraad voor Veiligheid

In maart van dit jaar verscheen een rapport[1] van de Onderzoeksraad voor Veiligheid (OVV) over de forensische zorg naar aanleiding van de zaak Michael P., de man die op 5 juli van dit jaar in hoger beroep werd veroordeeld tot 28 jaar en tbs met dwangverpleging voor het verkrachten en doden van Anne Faber in september 2017.[2] Een van de aanbevelingen die de OVV op grond van zijn onderzoek doet, is het versterken van de aandacht voor het risico op recidive door onder meer:

'(…) een breed gedragen, doelgroepgericht en gevalideerd instrumentarium voor risicotaxatie te ontwikkelen [en] tot die tijd de reeds beschikbare gevalideerde instrumenten voor risicotaxatie te gebruiken.'[3]

Elders in het rapport onderstreept de OVV het belang van de verdere ontwikkeling van valide risicotaxatie-instrumenten, waar hij stelt:

'Onder gedragsdeskundigen in de praktijk en de wetenschap is er discussie over betrouwbaarheid en toepassing van de verschillende soorten instrumenten voor risicotaxatie. Ook verschillen de meningen over de waarde en het gebruik van die instrumenten ten opzichte van het professionele oordeel van gedragsdeskundigen. Deze discussie is nog niet beslecht, waardoor er in Nederland geen algemene overeenstemming bestaat over de toepassing van één gevalideerd instrumentarium voor risicotaxatie.'[4]

## 2. Een gesprek over de waarde van de risicotaxatie

Op 17 juni jl. zond BNNVARA de documentaire 'Max, een leven in de tbs' uit.[5] De uitzending werd gevolgd door een nagesprek. Daarin werd door een panel onder leiding van Paul Witteman (hieronder: PW) bestaande uit mr. Yvo van Kuijck, raadsheer in de penitentiaire kamer in Arnhem (hieronder: YvK), mr. Jan-Jesse Lieftink, advocaat gespecialiseerd in tbs-zaken (hieronder: JJL), en prof. dr. Hjalmar van Marle, psychiater en oud-directeur van het Pieter Baan Centrum (hieronder: HvM), doorgepraat over een aantal vragen rond het tbs-systeem. In de loop van het gesprek ontspon zich de volgende discussie over de waarde van de risicotaxatie:

(…)
PW: Laten we even één stapje verder gaan en het hebben over wat op het ogenblik ook in de maatschappij een belangrijk onderwerp van gesprek is namelijk de risicotaxatie. Er is veel misgegaan de afgelopen tijd over het verlof, namelijk eh …
HvM: … maar niet in de tbs.
PW: Niet allemaal in de tbs. Dat geef ik onmiddellijk toe maar niettemin wel over een verlofsituatie en

* Dr. A.P.A. Broeders is emeritus hoogleraar Criminalistiek aan de Universiteit Leiden en de Universiteit Maastricht en was van 2002 tot 2007 Chief Scientist bij het NFI. Hij is ook redacteur van dit tijdschrift.

1. Onderzoeksraad voor Veiligheid, *Forensische zorg en veiligheid. Lessen uit de casus Michael P.*, Den Haag 2019.
2. Hof Arnhem-Leeuwarden 5 juli 2019, ECLI:NL:GHARL:2019:5542.
3. OVV-rapport, p. 14.
4. OVV-rapport p. 33.
5. Te vinden op www.npostart.nl/2doc-max-een-leven-in-tbs/17-06-2019/BV_101392997.

de risicotaxatie die daaraan verbonden is. Michael P. vermoordde Anne Faber tijdens zijn verlof uit de kliniek in Den Dolder. Peter M. keerde onlangs niet terug van verlof uit diezelfde kliniek en in Lelystad is een man vermoord door een tbs'er op verlof uit de Oostvaarderskliniek. Dat waren, zoals je zegt, niet allemaal tbs'ers.

HvM: Dat is de enige tbs'er van de drie.

PW: Maar is het te doen? Hoe betrouwbaar is de risicotaxatie?

HvM: Nou, de risicotaxatie is betrouwbaar. Maar niet 100% betrouwbaar.

PW: Hoeveel procent wel?

HvM: Er zit een ingewikkelde statistische redenering achter maar je zou kunnen zeggen: in twee op de drie gevallen is risicotaxatie betrouwbaar. Alleen, ik weet niet wie van de drie daarbuiten valt.

PW: Precies. Ja.

YvK: Een op de drie dus niet, hè. Als je nagaat dat je bij kop of munt 50% kans hebt, dan is dat wat de gedragsdeskundige met zijn taxatie toevoegt maar beperkt. Maar we hebben niets beters. Er zijn duidelijk verbeteringen doorgevoerd. Het is niet meer louter de klinische blik van de behandelaar. Het zijn gestructureerde taxaties die gedaan worden op basis van de aanwezige risicofactoren …

JJL: Die klinische blik is er nog steeds.

HvM: Nee, maar kijk. Risicotaxatie dat bekt lekker maar ...

PW: Het is de officiële term hoor.

HvM: Jawel, het bekt lekker maar dan denk je meteen van nou als je risicotaxatie toepast dat doen we, dat is de methode. Maar dat is niet de methode. Het is een onderdeel van de methode. Want het is namelijk dat je lijsten hebt met risicofactoren, dus die neem je af en die kun je aan de persoon zelf vragen, je kunt ze ook invullen vanuit de status, vanuit het dossier. En dan heb je dus een risicoprofiel. Maar of die persoon nu op dit moment aan dat risicoprofiel gaat voldoen als-ie naar buiten gaat, daar heb je gewoon weer het individuele gesprek met de behandelaars voor nodig. Dus risicotaxatie op zichzelf klinkt goed: 'Wij doen risicotaxatie.' Maar je komt niet uit als je alleen maar risicotaxatie doet.

PW: Maar het moet voor u als psychiater toch raar zijn om te horen dat uw taxatie ongeveer dezelfde juistheid heeft als kop of munt opgooien.

HvM: Nee, nee, nee, nee, nee.

YvK: Wel iets meer, heb ik gezegd.

PW: Ongeveer, zei ik.

HvM: Nee, het klinisch oordeel dat is 50%. Dus dat is, dat is niks, hè. Kop of munt. Risicotaxatie draagt ongeveer nog eens een keer een kwart daaraan bij, hè, zullen we zeggen dat is 70% kans. Dus dat is al meer dan toeval.

PW: Ja zeker, maar niet 100%.

HvM: Maar die overige 30% moeten echt de behandelaars zelf met de onderzoekers aan de hand van dat profiel bij elkaar plussen.

(…)

## 3. De diagnostische waarde van een testuitslag

Risicotaxatie levert geen perfecte voorspelling op van de kans op recidive. Dat is op zich niet verwonderlijk. Geen enkele medische of psychologische test geeft 100% zekerheid. Ook een zeer goede test zal, naast correctpositieve (hits) en correctnegatieve (eliminatie) resultaten, een bepaald percentage foutpositieve (loos alarm) en foutnegatieve (missers) uitslagen opleveren maar zal dat minder vaak doen dan een slechte test. Het (relatieve) aantal foute uitslagen zegt dus iets over de kwaliteit van de test. Stel dat het percentage correcte uitslagen van een bepaald risicotaxatie-instrument inderdaad 70% is, zoals Van Marle wellicht bedoelt. Laten we aannemen dat uit retrospectief onderzoek van een bepaalde populatie personen, bijvoorbeeld ex-tbs'ers, is gebleken dat zowel 70% van de recidivisten als 70% van de niet-recidivisten onder hen als zodanig door de test wordt aangemerkt en 30% in beide categorieën ten onrechte niet.[6] In dat geval zijn de twee essentiële maten die ons iets zeggen over de kwaliteit van een test, de *sensitiviteit* en de *specificiteit*, allebei 70%. Met andere woorden, 70% van de recidivisten in de populatie zal als zodanig worden aangemerkt (sensitiviteit) en 70% van de niet-recidivisten (specificiteit) als niet-recidivist. Voorts zal 30% van de recidivisten worden gemist en zal eveneens 30% van de niet-recidivisten ten onrechte als recidivist worden aangemerkt.[7]

Stel dat we 100 recidivisten en 100 niet-recidivisten uit deze populatie testen, dan geeft dat het volgende beeld voor onze taxatietest:

| | **positief testresultaat** | **negatief testresultaat** | **N** | |
|---|---|---|---|---|
| recidivist | 70 | 30 | 100 | sensitiviteit |
| niet-recidivist | 30 | 70 | 100 | specificiteit |

Door nu het (relatieve) aantal correcte uitslagen te delen door het (relatieve) aantal niet-correcte uitslagen kunnen we de *diagnostische waarde* van de testuitslag bepalen. In dit geval is dat, voor zowel een positieve als een negatieve uitslag, 70/30 = 2,3. We kunnen nu zeggen dat voor de onderzochte populatie geldt dat een positieve uitslag 2,3 maal waarschijnlijker is als de geteste persoon een recidivist is dan als het een niet-recidivist betreft. En een negatieve uitslag is eveneens 2,3 maal zo waarschijnlijk wanneer de geteste persoon geen recidivist is als wanneer dat wel het geval is.

6. Bij een retrospectief onderzoek als hier bedoeld wordt voor een groep tbs'ers die in het verleden zijn beoordeeld op recidiverisico een x-aantal jaren later op basis van informatie die beschikbaar was op het moment waarop zij werden beoordeeld een score bepaald op een recidiverisicotaxatie-instrument, die vervolgens voor zowel recidivisten als niet-recidivisten kan worden vergeleken met de eerdere beoordeling.
7. Gelet op de gebruikte formulering: 'in twee op de drie gevallen is risicotaxatie betrouwbaar', doet Van Marle hier strikt genomen feitelijk niet zozeer een uitspraak over de sensitiviteit en de specificiteit van een bepaalde test en daarmee over de diagnostische waarde van de test maar over het omgekeerde, over de positieve en negatieve predictiewaarde, d.w.z. de kans dat iemand die positief test, in feite ook recidiveert en het dus geen loos alarm betreft, respectievelijk de kans dat iemand die negatief test inderdaad niet recidiveert. Die (predictie)waarden hangen echter behalve van de diagnostische waarde van de test ook af van de a-priorikans (ook wel voorafkans genoemd) of prevalentie van de onderzochte conditie in de relevante populatie (zie hieronder).

Zou de test niet slechts 70 maar 90% van alle recidivisten als zodanig aanwijzen dan zou de diagnostische waarde van een positieve testuitslag, bij gelijkblijvende specificiteit van de test, 90/30 = 3 bedragen en de diagnostische waarde van een negatieve testuitslag 70/10 = 7:

| | positief testresultaat | negatief testresultaat | N | |
|---|---|---|---|---|
| recidivist | 90 | 10 | 100 | sensitiviteit |
| niet-recidivist | 30 | 70 | 100 | specificiteit |
| diagnostische waarde | 3 | 7 | | |

## 4. De positieve en negatieve predictiewaarde van een testuitslag

Stel dat deze cijfers gelden voor een medische test, zoals een hiv-test of een borstkankertest. Dan weten we nu dus hoeveel groter de kans is dat we een bepaalde testuitslag krijgen als we te maken hebben met een hiv-geïnfecteerde dan wanneer het gaat om een niet-hiv-geïnfecteerd persoon. Dat is mooi maar in feite willen we natuurlijk altijd iets anders weten, en wel het omgekeerde: hoe groot is de kans dat iemand die positief test daadwerkelijk hiv-geïnfecteerd is, borstkanker heeft of zal recidiveren? Dus niet: hoe groot is de kans op een positief resultaat bij besmetting maar, omgekeerd, hoe groot is de kans op daadwerkelijke besmetting bij een positief resultaat?

Om die laatste vraag te beantwoorden moeten we behalve de diagnostische waarde van de testuitslag nog iets weten en dat is de a-priorikans of voorafkans op de vermoede conditie, bijvoorbeeld hiv-infectie, borstkanker of in dit geval recidive.[8] Immers, als we een populatie testen waarin hiv-infectie bij 1 op de 100 personen voorkomt, dan zal een test, ongeacht zijn diagnostische waarde, meer positieve uitslagen opleveren dan wanneer 1 op de 1.000 personen is geïnfecteerd. Dat is te zien in de volgende kruistabel, waarin we uitgaan van een populatie van 100.000 personen, en een (fictieve hiv-)test met een sensitiviteit van 95% en een specificiteit van 99%.

| | positief testresultaat | negatief testresultaat | N | |
|---|---|---|---|---|
| geïnfecteerd | 95 | 5 | 100 | sensitiviteit |
| niet-geïnfecteerd | 1 | 99 | 100 | specificiteit |

De diagnostische waarde bedraagt 95/1 = 95 voor een positief testresultaat en 99/5 = 19,8 voor een negatief testresultaat. Toegepast op 100.000 personen uit populaties met een prevalentie van respectievelijk 1 op 100 en 1 op 1.000 geeft dit de volgende resultaten:

**Prevalentie 1 op 100**

| | positief | negatief | N |
|---|---|---|---|
| Geïnfecteerd | 950 | 50 | 1.000 |
| Niet-geïnfecteerd | 990 | 98.010 | 99.000 |
| | ppw: 49% | npw: 99,949% | |

*Tabel 1a: Verschillende uitslagen van identieke test bij verschillende a-priorikansen; ppw staat voor positieve predictiewaarde; npw voor negatieve predictiewaarde.*

Als 1 op de 100 is geïnfecteerd, is de positieve predictiewaarde 950/(950 + 990) = 49,0%.[9] De kans dat iemand die positief test daadwerkelijk geïnfecteerd is, bedraagt dus nog geen 50%: er zijn iets meer foutpositieve dan correctpositieve uitslagen. De negatieve predictiewaarde daarentegen is 98.010/(98.010 + 50) = 99,949%, m.a.w. slechts 1 op de ca. 2.000 negatieve testen is dus een 'misser' en in feite positief.

**Prevalentie 1 op 1.000**

| | positief | negatief | N |
|---|---|---|---|
| Geïnfecteerd | 95 | 5 | 100 |
| Niet-geïnfecteerd | 999 | 98.901 | 99.900 |
| | ppw: 8,7% | npw: 99,995% | |

*Tabel 1b: Verschillende uitslagen van identieke test bij verschillende a-priorikansen; ppw staat voor positieve predictiewaarde; npw voor negatieve predictiewaarde.*

Als we dezelfde test toepassen op een populatie waarin de prevalentie 1 op 1.000 is, is de positieve predictiewaarde slechts 95/(95 + 999) = 8,7%: het aantal foutpositieve uitslagen is meer dan tienmaal groter dan het aantal correctpositieve. Er worden dus zeer veel meer mensen ten onrechte als geïnfecteerd aangemerkt dan terecht. Daar staat tegenover dat de negatieve predictiewaarde hier nog hoger is, met 98.901/(98.901 + 5) = 99,995%, oftewel nog niet 1 op 20.000 negatieve testen is een 'misser' en in feite positief. Het laatste voorbeeld illustreert daarmee dat ook een test met een aanzienlijke diagnostische waarde veel meer foutpositieve uitslagen oplevert dan correctpositieve als het gaat om een conditie met een lage frequentie in de populatie, oftewel een lage *base rate* en daarmee een lage voorafkans. Hoe lager de a-priorikans, hoe meer foutpositieve resultaten een test zal opleveren.[10]

## 5. De auc-waarde

Daarmee zijn we beland bij de kern van het probleem. Hoe kunnen we voorkomen dat we grote aantallen foutpositieve (loos-alarm)uitslagen scoren zonder daarbij veel foutnegatieve (missers) te oogsten. Of, omgekeerd, hoe verhogen we de detectie van recidivisten onder bijvoorbeeld tbs'ers, in het belang van de maatschappelijke veiligheid, zonder grote aantallen niet-recidivisten onder hen ten onrechte binnen te houden, waarmee het belang

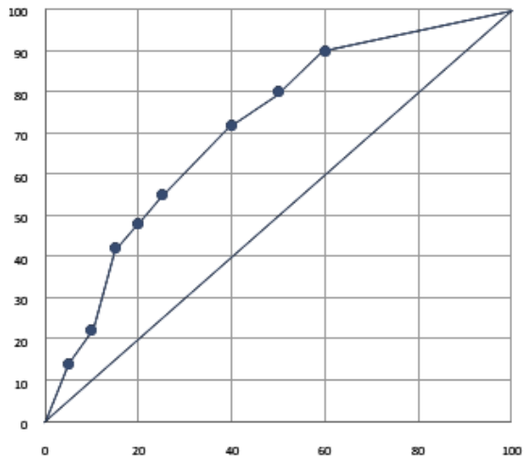
8. In de medische context hangt die voorafkans of a-priorikans samen met de prevalentie in de relevante populatie. Zo zal in het geval van een concrete verdenking op hiv-infectie bij een lid van een risicogroep de voorafkans hoger worden ingeschat omdat de prevalentie in die groep hoger is. In het Engels wordt, naast *prior probability* en *prevalence*, ook de term '*base rate*' gebruikt.
9. De positieve predictiewaarde wordt verkregen door het (relatieve) aantal correctpositieve uitslagen (i.c. 950) te delen door de som van het totale aantal positieve uitslagen (i.c. 950 correctpositief + 990 foutpositief.)
10. Dit is een van de redenen waarom personen in leeftijdscategorieën waarin de prevalentie van een bepaalde ziekte relatief laag is niet worden uitgenodigd voor desbetreffend grootschalig bevolkingsonderzoek, zoals bij het bevolkingsonderzoek naar borstkanker (beperkt tot vrouwen tussen 50 en 75) of darmkanker (beperkt tot personen tussen 55 en 75). Zie hiervoor ook www.rivm.nl/bevolkingsonderzoek-borstkanker en www.rivm.nl/bevolkingsonderzoek-darmkanker.

van de betrokkene wordt geschaad. De verhouding tussen die twee grootheden hangt in de eerste plaats af van de *base rate* oftewel de frequentie van recidive in de relevante populatie van bijvoorbeeld tbs'ers. Daaraan kunnen we echter, zeker op korte termijn, niets veranderen. In de tweede plaats hangt de verhouding tussen die twee grootheden samen met de sensitiviteit en de specificiteit en dus met de diagnostische waarde van de risicotest. En die waarden hangen weer af van de manier waarop we de test scoren. In die gevallen waarin de test een continue variabele betreft die een getalsmatige score oplevert en geen binaire uitslag (ja of nee), kunnen we die verhouding variëren door de score die we als positief aanmerken, de zogenaamde grenswaarde of het afkappunt, te verhogen. Dit heeft tot gevolg dat we minder correctpositieve resultaten krijgen en dus meer missers. Maar we kunnen de afkapwaarde ook verlagen. In dat geval krijgen we meer correctpositieve resultaten maar daarmee ook meer foutpositieve. Naarmate we meer recidivisten terecht binnen de poort houden, zal dat ertoe leiden dat we ook meer niet-recidivisten ten onrechte binnen houden.
Hoe het verband tussen deze twee variabelen ligt, kan worden geïllustreerd met de cijfers in Tabel 2 die zijn overgenomen uit een hoofdstuk over risicotaxatie van Lammers (2018).[11]

| afkapwaarde | % detectie | % vals alarm |
|---|---|---|
| hoog | 14 | 5 |
| | 22 | 10 |
| | 42 | 15 |
| | 48 | 20 |
| | 55 | 25 |
| | 72 | 40 |
| | 80 | 50 |
| laag | 90 | 60 |

*Tabel 2: detectiepercentage en vals-alarmscore van een test met auc-waarde 0,72.*

Zoals de tabel illustreert, neemt door verlaging van de grenswaarde of afkapwaarde de detectie weliswaar spectaculair toe (tot 90%) maar stijgt daarmee ook de vals-alarmscore tot grote hoogte (60%). In figuur 1 zijn de waarden uit Tabel 2 in een grafiek geplaatst, met de detectiescore of sensitiviteit op de y-as en de vals-alarmscore (of 1-specificiteit) op de x-as.




*Figuur 1: ROC-curve; x-as: percentage vals alarm (1 – specificiteit) ; y-as: percentage correctpositief (detectie)*

Daardoor ontstaat een curve, de ROC-curve of *receiver operating characteristic*-curve, die zich verheft boven de diagonaal uit de oorsprong, die de toevalsscore markeert. De omvang van het gebied onder de curve, de *area under the curve*, kan worden uitgedrukt in een score, de auc-score, in dit geval 0,72, waar de theoretisch maximaal haalbare waarde 1 bedraagt.[12]
De waarden in Tabel 2 illustreren de consequenties van het schuiven met de afkapwaarde. De keuze daarvan is uiteraard geen wetenschappelijke maar een politieke.

Zonder hier nader in te gaan op de wijze waarop de auc-waarde wordt berekend, is het instructief om te bezien welke diagnostische waarden en positieve en negatieve predictiewaarden de cijfers in Tabel 2 representeren. Daarbij gaan we uit van een *base rate* of a-priorikans van 24%, zoals vermeld in Verkes (2018).[13]

| afkap-waar-de | % detec-tie | % vals alarm | d.w. pos. | d.w. neg. | ppw | npw |
|---|---|---|---|---|---|---|
| hoog | 14 | 5 | 2,8 | 1,10 | 46,9% | 77,7% |
| | 22 | 10 | 2,2 | 1,15 | 40,1% | 78,5% |
| | 42 | 15 | 2,8 | 1,47 | 46,9% | 82,3% |
| | 48 | 20 | 2,4 | 1,54 | 43,1% | 83,0% |
| | 55 | 25 | 2,2 | 1,67 | 40,1% | 84,1% |
| | 72 | 40 | 1,8 | 2,14 | 36,2% | 87,1% |
| | 80 | 50 | 1,6 | 2,50 | 33,6% | 88,8% |
| laag | 90 | 60 | 1,5 | 4,00 | 32,1% | 92,7% |

*Tabel 3: d.w. staat voor diagnostische waarde of likelihood ratio; pos. voor positief; neg. voor negatief; ppw voor positieve predictiewaarde; npw voor negatieve predictiewaarde. A-priorikans of voorafkans op recidive 24%.*

11. Lammers 2018 (p. 298), baseert haar tabel op Brand 2005 (p. 445), die de gegevens weer baseert op een studie van Grann e.a. 1998 en 1999. Nadere bibliografische informatie over deze teksten van Grann ontbreekt overigens zowel bij Brand 2005 als bij Lammers 2018.
12. Zie voor nadere illustratie van het begrip auc-waarde Brand 2005, Brand & Van Emmerik 2015, Lammers 2018 of Verkes 2018. Lammers 2018 (p. 298) geeft overigens 0,75 als auc-waarde voor de data die zij aan Brand 2005 (p. 440) ontleent. Brand & Van Emmerik 2015 (p. 355) benadrukken dat, hoewel dit vaak gebeurt, de auc-waarde niet kan worden opgevat als een indicatie van het percentage correcte uitslagen dat de test in kwestie gemiddeld oplevert, i.c. circa 72%.
13. Het cijfer is afkomstig van een score op de HKT-R-test bij een cohort van tbs-gestelden in de jaren van 2004 tot 2008 zoals gemeld door Verkes 2018 (p. 311), die het weer ontleent aan Spreen e.a. 2014. Dit cijfer betreft uiteraard slechts de geregistreerde recidive, die per definitie een onderschatting vormt van de werkelijke recidive. Hierbij rijst de vraag welke *base rate* gehanteerd kan worden bij recidiverisicotaxaties in het kader van de pro-justitia-rapportage en, in het verlengde daarvan, de vraag of de risicotaxatie bij het ontbreken van recidivedata voor deze populatie niet gestaakt moet worden omdat de diagnostische waarde ervan dan niet kan worden bepaald.

Wat opvalt is dat de diagnostische waarden in alle gevallen zeer bescheiden zijn. In combinatie met de relatief lage voorafkans of a-prioriwaarschijnlijkheid van 24% leidt dit tot positieve predictiewaarden die nooit boven de 50% uitkomen. De implicatie daarvan is zonder meer alarmerend: ongeacht de hoogte van de afkapwaarde zal een positieve testuitslag in meer dan de helft van de gevallen foutpositief of loos alarm zijn.
Waar een lage afkapwaarde (hoge sensitiviteit, relatief weinig missers maar veel vals-alarmscores) van belang is voor de veiligheid van de samenleving en een hoge afkapwaarde (lage sensitiviteit, weinig vals alarm maar veel missers) vooral het belang dient van de geteste persoon die niet onterecht wil worden vastgehouden, illustreert de tabel dat maximalisatie van het eerste belang, de bescherming van potentiële slachtoffers, leidt tot een positieve predictiewaarde van slechts 32,1%, wat betekent dat twee van de drie positieve uitslagen loos alarm zijn. Daar staat wel tegenover dat de negatieve predictiewaarde in dat geval ruim 92% bedraagt, waarmee minder dan 8% van de negatieve testresultaten missers, d.w.z. recidivisten betreft die onterecht negatief scoren.
Maar ook als de grenswaarde voor detectie hoog wordt gelegd, met een detectiescore van maar 14%, is de positieve predictiewaarde nog geen 50%, is dus meer dan 50% van de positieve resultaten foutpositief en komt de negatieve predictiewaarde nog maar net uit boven de driekwart, wat betekent dat bijna een kwart van de personen die negatief testen in feite recidiveren.
Stel dat we niet uitgaan van een recidiverisico van 24 maar van 50%. Leidt dit tot betere resultaten?
Het korte antwoord is nee. Immers, zolang de diagnostische waarde van de test gelijk blijft, blijft ook de mate waarin de test onzekerheid reduceert, gelijk.

| afkapwaarde | % detectie | % vals alarm | d.w. pos. | d.w. neg. | ppw | npw |
|---|---|---|---|---|---|---|
| hoog | 14 | 5 | 2,8 | 1,10 | 73,7% | 52,4% |
| | 22 | 10 | 2,2 | 1,15 | 68,7% | 53,5% |
| | 42 | 15 | 2,8 | 1,47 | 73,7% | 59,5% |
| | 48 | 20 | 2,4 | 1,54 | 70,1% | 60,6% |
| | 55 | 25 | 2,2 | 1,67 | 68,7% | 62,5% |
| | 72 | 40 | 1,8 | 2,14 | 64,3% | 68,2% |
| | 80 | 50 | 1,6 | 2,50 | 61,5% | 71,4% |
| laag | 90 | 60 | 1,5 | 4,00 | 60,0% | 80,0% |

*Tabel 4: d.w. staat voor diagnostische waarde of likelihood ratio; pos. voor positief; neg. voor negatief; ppw voor positieve predictiewaarde; npw voor negatieve predictiewaarde. A-priorikans of voorafkans op recidive 50%.*

De percentages van de positieve predictiewaarde liggen weliswaar steeds hoger dan in Tabel 3 maar dat is enkel het gevolg van het hanteren van een hogere a-priorikans. Bovendien ligt de negatieve predictiewaarde steeds lager dan in Tabel 3 en worden er dus meer recidivisten gemist. De relatieve onzekerheidsreductie die de test oplevert, blijft gelijk.

## 6. Verbale equivalenten voor de kwantitatieve scores bij de risicotaxatie

Hoewel het alleen al voor de onderlinge vergelijkbaarheid wenselijk is de diagnostische waarde of de auc-waarde van een test in cijfers of getallen uit te drukken, is daarmee op zichzelf nog niet duidelijk bij welke waarde we te maken hebben met een goede test of een met een slechte. Een antwoord op die vraag lijkt te worden gegeven door Harte & Breukink (2010) wanneer zij schrijven:

'Het is gebruikelijk om de predictieve validiteit van risicotaxaties uit te drukken in zogenoemde AUC-waarden (Brand, 2005). Een AUC-waarde van .50 betekent dat de voorspellende waarde gelijk is aan toeval en een AUC-waarde van 1.0 betekent een perfecte voorspelling. Een AUC-waarde tussen .50 en .60 wordt beschouwd als onvoldoende, tussen .60 en .70 als matig, tussen .70 en .80 als redelijk, tussen .80 en .90 als goed en boven de .90 als zeer goed.'[14]

Brand (2005) zelf is van oordeel dat 'een AUC-waarde van rond de .75 ongeveer het maximum van nauwkeurigheid is wat men vandaag de dag kan bereiken op het gebied van het voorspellen van gewelddadig of delinquent gedrag'. Hij acht het 'niet ethisch om bewust voor een risicotaxatielijst te kiezen welke minder goed voorspelt dan haalbaar is'.[15] Tien jaar later schrijven Brand & Van Emmerik (2015): 'Als minimale AUC voor een risicotaxatielijst wordt momenteel een waarde tussen de .70 en .75 gebruikt.'[16]

Lammers (2018) schrijft:

'Onderzoekers denken nogal verschillend over de vraag hoe hoog een AUC-waarde moet zijn. Enige tijd beschouwde men een AUC-waarde vanaf 0,70 als "redelijk" en een AUC vanaf 0,75 als "goed". De laatste tijd lijken onderzoekers wat strenger en worden AUC-waarden boven de 0,80 pas als "goed" beschouwd.'[17]

Combinatie van deze gegevens levert de volgende tabel op:

| auc-waarde | verbaal equivalent |
|---|---|
| 0,50 | toeval |
| 0,50-0,60 | onvoldoende |
| 0,60-0,70 | matig |
| 0,70-0,80 | redelijk tot goed |
| 0,80-0,90 | goed |
| > 0,90 | zeer goed |

*Tabel 5. Auc-waarden en de daarmee geassocieerde verbale equivalenten.*

14. Harte & Breukink 2010 (p. 64) voegen daaraan toe dat anno 2010 van tien instrumenten de auc-waarden bekend zijn.
15. Brand 2005, p. 453.
16. Brand & Van Emmerik 2015, p. 367.
17. Lammers 2018 (p. 296) verwijst hierbij naar Bogaerts e.a. 2017 (p. 2260), die, ondanks dat zij de auc-waarden die zij verkrijgen voor een door hen onderzocht taxatie-instrument, de HKT-R, omschrijven als '*modest*' en '*marginal*', niettemin concluderen '(...) *that this risk assessment instrument appears to be a satisfactory instrument for risk assessment*'.

## 7. Verbale equivalenten voor de kwantitatieve waarschijnlijkheidstermen in het forensisch-technisch onderzoek

In het forensisch-technisch onderzoek wordt de laatste decennia het gewicht van het bewijs veelal eveneens bij voorkeur tot uitdrukking gebracht in getalsmatige of kwantitatieve termen, en wel in de vorm van de *likelihood ratio* of de diagnostische waarde van het bewijs bij twee elkaar uitsluitende hypothesen.[18] Zo wordt de bewijswaarde van een DNA-match berekend als de verhouding van de kans op de match onder twee elkaar uitsluitende aannamen of hypothesen: bijvoorbeeld de kans op een match onder de aanname dat de verdachte de bron is van het biologisch materiaal en de kans op een match onder de aanname dat een willekeurige, niet-verwant de donor is van het celmateriaal. Voor een matchende verdachte is die kans 100% of 1, voor een willekeurige niet-verwant wordt die kans doorgaans gerapporteerd als kleiner dan 1 op 1 miljard, een getal dat correspondeert met de geschatte frequentie van het profiel in de relevante populatie. De kansverhouding of *likelihood ratio* en de diagnostische waarde van een match met een volledig profiel is daarmee groter dan 1 miljard.[19] Voor een partieel profiel kan het gaan om aanzienlijk lagere waarden, afhankelijk van het aantal willekeurige niet-verwanten van wie het profiel zou matchen met dat van het spoor en daarmee van de frequentie van het betrokken profiel in de relevante populatie.

In veel gevallen zal de forensisch-technisch deskundige echter geen harde kwantitatieve uitspraak kunnen doen over de *likelihood ratio*, met name voor typen bewijs waarvoor de kans op een match niet in cijfers kan worden uitgedrukt, bijvoorbeeld omdat kwantitatieve empirische gegevens over de frequentie van bepaalde relevante sporenkenmerken ontbreken. Zo kan voor veel soorten sporenbewijs, zoals handschrift, glas, verf, vezels, vuurwapens, schoensporen of werktuigsporen, de frequentie van de relevante kenmerken in de populatie niet altijd worden gekwantificeerd doordat er, anders dan bij DNA, geen of onvoldoende bruikbare populatiegegevens in de vorm van referentieverzamelingen zijn.[20] Ook dan kan de *likelihood ratio* of diagnostische waarde worden aangegeven, zij het uitgedrukt in verbale termen. Een voorbeeld van zo'n reeks verbale waarschijnlijkheidstermen is de reeks die is ontwikkeld door het Nederlands Forensisch Instituut (NFI).[21] De verbale waarschijnlijkheden die worden gerapporteerd zijn in dat geval niet gebaseerd op objectieve, kwantitatieve empirische data, zoals in het DNA-voorbeeld hierboven, maar vinden hun oorsprong in de ervaring van de onderzoeker en kunnen worden opgevat als deels gebaseerd op geïnternaliseerde frequentieschattingen.

Daarbij kunnen de volgende uitspraken worden gedaan:

De bevindingen van het onderzoek zijn:
*ongeveer even waarschijnlijk/*
*iets waarschijnlijker/*
*waarschijnlijker/*
*veel waarschijnlijker/*
*zeer veel waarschijnlijker/*
*extreem veel waarschijnlijker*
onder hypothese (1) (van het Openbaar Ministerie), i.c. dat de verdachte de bron is van het celmateriaal, als/dan onder hypothese (2) (van de verdediging), i.c. dat een willekeurig lid van de relevante populatie de bron is van het spoor.

Ook hier gaat het weer om uitspraken over de waarschijnlijkheid van een bevinding (i.c. de vaststelling van een bepaalde mate van overeenkomst van bepaalde kenmerken van een spoor met die van het referentiemateriaal) onder twee elkaar uitsluitende hypothesen.
Ook als harde kwantitatieve uitspraken niet mogelijk zijn, kan de deskundige vaak wel een getalsmatige indicatie geven van de orde van grootte van de waarschijnlijkheid van zijn bevindingen onder de verschillende relevante hypothesen en daarmee van de *likelihood ratio*. Uiteraard is het dan van belang dat rapporteurs uit verschillende deskundigheidsgebieden aan de verschillende verbale termen vergelijkbare ordegroottes toekennen. Om de uniformiteit in de formulering van de conclusies voor de diverse deskundigheidsgebieden te bevorderen heeft het NFI de verbale termen numeriek gedefinieerd volgens het onderstaande schema, waarin de getalsmatige bewijskracht van de verbale termen in de linkerkolom en de ordegrootte van de *likelihood ratio* in de rechterkolom worden weergegeven.
Daarbij worden de volgende uitspraken gedaan:

| **verbale term** | **ordegrootte bewijskracht (LR)** |
|---|---|
| *ongeveer even waarschijnlijk* | *1-2* |
| *iets waarschijnlijker* | *2-10* |
| *waarschijnlijker* | *10-100* |
| *veel waarschijnlijker* | *100-10.000* |
| *zeer veel waarschijnlijker* | *10.000-1.000.000* |
| *extreem veel waarschijnlijker* | *> 1.000.000* |

*Tabel 6. Relatie numerieke en verbale conclusies volgens Vakbijlage NFI (2017: 6).*[22]

## 8. Vergelijking van de verbale equivalenten gebruikt door gedragskundigen en forensisch-technisch deskundigen

Vergelijking van de ordegrootte van de verbale equivalenten die het NFI hanteert voor het forensisch-technisch onderzoek en de waarden die Lammers noemt voor de auc-waarden van risicotaxatietesten laat zien dat de

18. Zie hiervoor bijvoorbeeld Broeders 2016.
19. Een 'volledig' profiel wordt verkregen wanneer men alle DNA-kenmerken die men wil bepalen ook heeft kunnen bepalen. Men spreekt van een partieel of deelprofiel als dat niet in alle gevallen mogelijk is. De kans op een match met een willekeurige niet-verwant zal in dat laatste geval altijd groter zijn, bijvoorbeeld 1 op 1000, wat zou neerkomen op een LR van 1000.
20. Dit speelt vooral bij het zogenaamde bron- of herkomstonderzoek, waarbij het gaat om het bepalen van de unieke bron van een spoor.
21. NFI 2017, p. 2.
22. NFI 2017. De ENFSI, het *European Network of Forensic Science Institutes*, waarbij 71 Europese laboratoria uit 38 landen zijn aangesloten, adviseert eveneens het gebruik van kwantitatieve *likelihood ratios* met vermelding van de daarmee geassocieerde verbale equivalenten, die qua ordegrootte overeenkomen met de door het NFI gehanteerde termen.

laatste waarden zeer veel lager liggen dan de eerste. Zo geldt als verbaal equivalent van een auc-waarde van 0,75 volgens Lammers (2018) de term 'redelijk tot goed'.[23] Kijken we nu in Tabel 3 of 4 naar de diagnostische waarden die daarmee corresponderen dan blijken acht van de zestien diagnostische waarden lager dan 2. De overige acht diagnostische waarden liggen tussen de 2 en 5. Diagnostische waarden van een dergelijke (zeer) bescheiden ordegrootte vallen in de NFI-tabel in de categorieën 'ongeveer even waarschijnlijk' en 'iets waarschijnlijker': de testuitslag is in die gevallen binnen de NFI-schaal op te vatten als 'ongeveer even waarschijnlijk' bij niet-recidive als bij recidive, respectievelijk 'iets waarschijnlijker' bij recidive dan bij niet-recidive. Het gewicht van het bewijs is dus verwaarloosbaar tot buitengewoon gering. Dat een door de beroepsgroep als 'redelijk tot goed' gewaardeerde test geen tot zeer weinig bewijswaarde heeft is op zijn minst genomen verwarrend voor de gebruiker die met deze beide typen bewijs wordt geconfronteerd.

In een welkom pleidooi voor het gebruik van *likelihood ratios* of diagnostische waarden in het psychologisch onderzoek verwijst Rassin elders in dit nummer van *EeR* naar een artikel van Jarosz & Wiley (2014), waarin zij wijzen op verbale equivalenten voorgesteld door respectievelijk Raftery (1995) en Jeffreys (1961). Ook deze volgens Rassin voor 'regulier' wetenschappelijk onderzoek gebruikte waarden verschillen qua ordegrootte weliswaar eveneens zeer aanzienlijk van de door het NFI gebruikte termen maar liggen anderzijds ook weer niet zeer dicht bij de verbale equivalenten die Lammers (2018) geeft voor auc-waarden.

| **diagnostische waarde** | **verbaal equivalent (Raftery 1995)** | **verbaal equivalent (Jeffreys 1961)** |
|---|---|---|
| 1-3 | *weak* | *anecdotal*[24] |
| 3-10 | *positive* | *substantial* |
| 10-20 | *positive* | *strong* |
| 20-30 | *strong* | *strong* |
| 30-100 | *strong* | *very strong* |
| 100-150 | *strong* | *decisive* |
| > 150 | *very strong* | *decisive* |

*Tabel 7. Kwantitatieve en verbale termen voor diagnostische waarde of likelihood ratio naar Jarosz & Wiley 2014.*

Rassin (2019) wijst er terecht op dat de waarden die het NFI hanteert primair betrekking hebben op het zogenaamde herkomstonderzoek van materiële (fysische, chemische, biologische of digitale) forensische sporen. Daarbij gaat het om een vraag op bronniveau: wat is de unieke bron van het vingerspoor, het handschrift, het celmateriaal, de huls enz.? Voor de vraag hoe en wanneer een bepaald spoor is achtergelaten en daarmee voor de vraag, op handelings- of activiteitenniveau, naar de activiteit of handeling waarbij het spoor is vrijgekomen spreekt hij zijn twijfel uit over de mogelijkheid hier *likelihood ratios* te gebruiken en acht hij de verbale termen in de NFI-tabel minder bruikbaar. In plaats daarvan stelt hij onder verwijzing naar eerdergenoemde Jarosz & Wiley (2014) andere bandbreedten voor, waarbij bewijs met een *likelihood* score tussen 30 en 100 als 'zeer sterk', en bewijs met een score groter dan 100 als 'definitief' zou worden aangemerkt.[25] Dat komt neer op een zeer aanzienlijke overwaardering van het bewijs die verhult dat de bewijskracht van het onderzoek van dit type veelal minder sterk zal zijn en de binnen de forensisch-technische wereld nagestreefde uniformiteit bij de waardering van het gewicht van bewijs ernstig ondermijnt.

## 9. Eindconclusie

Zoals Van Marle in het eerder gememoreerde nagesprek over de betrouwbaarheid van de risicotaxatie binnen de tbs-praktijk opmerkt – 'er zit een ingewikkelde statistische redenering achter' – is het onderliggende statistische model betrekkelijk complex. Zijn poging om de betrouwbaarheid van de risicotaxatie te concretiseren – 'je zou kunnen zeggen: in twee op de drie gevallen is risicotaxatie betrouwbaar' – suggereert dat de uitslag gemiddeld in bijna 70% van de gevallen juist is. Waarvoor dit cijfer precies staat en waarop het is gebaseerd is onduidelijk.

Zoals we hierboven hebben gezien zeggen de positieve en negatieve predictiewaarde van een uitgevoerde test niet zeer veel over de kwaliteit van de test of de mate van onzekerheidsreductie die de test oplevert. Ze worden immers medebepaald door (1) de toegepaste detectiegrens of afkapwaarde, (2) de daarmee samenhangende diagnostische waarde, en (3) de gehanteerde a-priorikans of voorafkans.

Harte (2017) verwijst naar een metastudie van Fazel e.a. (2012) waaruit bleek dat de onderzochte taxatie-instrumenten weliswaar weinig foutnegatieve (missers) opleverden maar meer dan 50% foutpositieve resultaten:

'Deze studie[26] toonde aan dat wanneer instrumenten personen aanwijzen als hebbende een lage kans op recidive, dit meestal klopt. Echter, van de mensen die door het instrument als recidivegevaarlijk waren aangewezen, recidiveerde de meerderheid niet. Dit impliceert dat toepassing van de instrumenten leidt tot opleggingen en verlengingen van maatregelen terwijl dat feitelijk niet nodig is.'[27]

Dat laatste is juist maar is een onvermijdelijk gevolg van het feit dat de voorafkans op recidive (aanzienlijk) onder de 50% ligt. Naarmate die kans lager is neemt het absolute aantal positieve uitslagen af maar neemt het aandeel foutpositieve uitslagen binnen die groep toe. Daarnaast is de positieve predictiewaarde, zoals we hebben gezien,

23. Diagnostische waarden voor hogere auc-waarden dan 0,75 zullen uiteraard hoger uitvallen, met dien verstande dat zelfs een auc-waarde als 0,90, die door de beroepsgroep als 'zeer goed' wordt gekwalificeerd, bescheiden diagnostische waarden oplevert.
24. De term '*anecdotal*' zou daarmee van toepassing zijn op 8 van de 16 hierboven vermelde diagnostische waarden.
25. Termen als 'definitief' of '*decisive*' (Nederlands: beslissend, doorslaggevend) lijken ook vanwege hun categorisch karakter minder gelukkig gekozen.
26. Fazel e.a. 2012. Harte 2016 (p. 203) vermeldt dat deze meta-analyse van 73 databestanden uit 13 landen gebaseerd op scores van 24 827 personen een positieve predictiewaarde van 41% en een negatieve predictiewaarde van 91% te zien gaf, bij een *base rate* van 23,7%. De diagnostische waarde van een matig tot hoge testuitslag bedraagt daarmee 2,24, die van een lage testuitslag 3,14.
27. Harte 2017, p. 2388.

mede afhankelijk van het gekozen detectieniveau. Zo zal het hanteren van een lagere afkapwaarde bij hantering van hetzelfde risicotaxatie-instrument met dezelfde, bescheiden auc-waarde, weliswaar leiden tot minder opleggingen en verlengingen van maatregelen maar tevens tot meer missers en daarmee tot meer recidive.

De auc-waarde van een risicotaxatietest is een veelgebruikte maat om de predictiewaarde van verschillende risicotaxatie-instrumenten te vergelijken. Zij is echter om een aantal redenen ongeschikt om de predictiewaarde van concrete toetsuitslagen te beoordelen.[28] Bovendien blijkt dat verbale equivalenten van de auc-waarde die worden gebruikt om de validiteit van risicotaxatie-instrumenten tot uitdrukking te brengen, sterk tot zeer sterk verschillen van de verbale termen die in het reguliere wetenschappelijk, respectievelijk het forensisch-technisch onderzoek worden gehanteerd. Om verwarring te voorkomen zou daarom bij de rapportage van de risicotaxatie niet slechts de positieve of negatieve predictiewaarde, maar ook de toegepaste detectiegrens of afkapwaarde moeten worden gespecificeerd, alsmede de daarmee samenhangende diagnostische waarde en de gehanteerde a-priorikans. Zeker wanneer in de toekomst risicotaxatie-instrumenten niet langer alleen door de gedragsdeskundige maar ook door de rechter of officier worden toegepast, zoals Harte verwacht, is het consequente gebruik van standaarden van essentieel belang.[29]

Brand & Van Emmerik (2015) wijden een uitvoerige bespreking aan de voor- en nadelen van de auc-waarde en introduceren een aantal nieuwe maten voor de vergelijking van predictiewaarden van verschillende risicotaxatie-instrumenten, waaronder de D80 en D90. Deze nieuwe maten staan voor de predictiewaarden die de risicotaxatietest in kwestie oplevert als de detectiegrens (D) of afkapwaarde op 80 respectievelijk 90% wordt gesteld. De reden die zij aanvoeren om juist deze afkapwaarden te hanteren is dat ze het meest relevant zijn aangezien in de praktijk 'in veel landen meer gewicht wordt toegekend aan het voorkomen van recidive dan aan een (onnodig) lang verblijf in een forensisch psychiatrisch ziekenhuis'.[30] Hoewel voorstellen als deze zeker tot beter begrip van de predictiewaarde van de risicotaxatie kunnen leiden, lijkt gezien de zeer bescheiden diagnostische waarden van de huidige generatie risicotaxatie-instrumenten, de meeste winst te behalen met een verbetering van de kwaliteit van deze instrumenten, zoals ook de OVV adviseert. Voor die stelling is ook steun te vinden bij Harte, die concludeert: '(…) wanneer we de studies naar predictieve validiteit overzien, dan moeten we helaas concluderen dat het met de voorspellende waarde van de instrumenten matig is gesteld.'[31]

## 10. Ten slotte

Vergelijking van de termen die worden gebruikt als verbale equivalenten voor de auc-waarde van risicotaxatie-instrumenten met de termen die als zodanig worden gebruikt voor de diagnostische waarde van testen in andere wetenschapsgebieden geeft grote verschillen te zien. Daardoor worden de onderliggende reële verschillen in diagnostische waarde in belangrijke mate verhuld. Vermelding in de rapportage van de in het risico-onderzoek gehanteerde detectiegrens en de daarmee samenhangende diagnostische waarde kan hier voor meer duidelijkheid zorgen en overspannen verwachtingen over de waarde van risicotaxatie-instrumenten wegnemen.

## 11. Literatuur

**Blaauw, Bogaerts & Spreen 2019**
E. Blaauw, S. Bogaerts & M. Spreen, 'Risicotaxatie in de Nederlandse rechtspraktijk: op naar een best practice', *EeR* 2019, afl. 2, p. 71-77.

**Bogaerts e.a. 2017**
S. Bogaerts, M. Spreen, P. ter Horst & C. Gerlsma, 'Predictive validity of the HKT-R risk assessment tool: two and 5-year violent recidivism in a nationwide sample of Dutch forensic psychiatric patients', *International Journal of Offender Therapy and Comparative Criminology* 2017, p. 1-12.

**Brand 2005**
E.F.J.M. Brand, 'Een maat voor de kwaliteit van instrumenten voor risicotaxatie', in: M.J. Sjerps & J.A. Coster van Voorhout (red.), *Het onzekere bewijs: gebruik van statistiek en kansrekening in het strafrecht*, Deventer: Kluwer 2005, p. 429-456.

**Brand & Van Emmerik 2015**
E.F.J.M. Brand & J.L. van Emmerink, 'Een nieuwe maat voor de voorspellende waarde van risicotaxatie-instrumenten in de forenische psychiatrie', in: P.A.M. Mevis. J.H.M. Thulen, B.C.M. Raes, E.A. Mulder, M.J.F. van der Wolff & S.R. Bakker (red.), *Omzwervingen tussen psychiatrie en recht* (Liber Amicorum prof. dr. H.J.C. van Marle), Deventer: Wolters Kluwer 2015, p. 351-368.

**Broeders 2016**
A.P.A. Broeders, 'Forensisch bewijs', in: M. Boone, C. Brants & R. Koole (red.), *Criminologie en strafrecht*, Den Haag: Boom 2016 (2e druk), p. 117-161.

**ENFSI 2015**
ENFSI, 'ENFSI Guideline for evaluative reporting in Forensic Science', 2015, www.enfsi.eu.

28. Zie Brand & Van Emmerik 2015 voor een uitvoerige discussie van de beperkingen van de auc-waarde als maat voor predictiewaarde van een concrete toets.
29. Harte 2017 (p. 2388): 'Het is te verwachten dat de Nederlandse rechtspraktijk binnen afzienbare tijd wordt geconfronteerd met modellen die officieren van justitie en rechters zelf kunnen toepassen.' Onduidelijk blijft waarop deze verwachting is gebaseerd en hoe reëel deze is.
30. Brand & Van Emmerik 2015, p. 357.
31. Harte 2016, p. 208. Ook Hummelen 2019 (p. 1), die schrijft over de toepassing van risicotaxatie-instrumenten in de forensische behandeling, waarschuwt voor overspannen verwachtingen van het gebruik van deze instrumenten: 'De aanbeveling [van de OVV] om systematisch risicotaxaties uit te voeren is zinvol, maar kan tegelijkertijd leiden tot de misvatting dat daarmee de gesignaleerde problemen grotendeels kunnen worden opgelost.'

**Fazel e.a. 2012**
S. Fazel, J.P. Singh, H. Doll & M. Grann, 'Use of risk assessment instruments to predict violence and antisocial behavior in 73 samples involving 24 827 people: Systematic review and meta-analysis', *British Medical Journal* 2012, 345: e4692.

**Harte 2016**
J.M. Harte, 'Predictie van criminaliteit', in: M. Boone, C. Brants & R. Koole (red.), *Criminologie en strafrecht*, Den Haag: Boom 2016 (2e druk), p. 187-213.

**Harte 2017**
J.M. Harte, 'Recidive inschatten met behulp van een empirisch model. Kansen voor de strafrechtspraktijk', *NJB* 2017/33, p. 2386-2389.

**Harte & Breukink 2010**
J.M. Harte & M.D. Breukink, 'Objectiviteit of schijnzekerheid? Kwaliteit, mogelijkheden en beperkingen van instrumenten voor risicotaxatie', *Tijdschrift voor Criminologie* 2010, 52 (1), p. 52-72.

**Hummelen 2019**
J.W. Hummelen, 'Geen risicotaxatie zonder risicobewustzijn', *EeR* 2019, afl. 4, p. 129-131.

**Jarosz & Wiley 2014**
A.F. Jarosz & J. Wiley, 'What are the odds? A practical guide to computing and reporting Bayes factors', *Journal of Problem Solving* 2014, afl. 7, p. 2-9.

**Jeffreys 1961**
H. Jeffreys, *Theory of probability*, Oxford (UK): Oxford University Press 1961.

**Lammers 2018**
S.M.M. Lammers, 'Risicotaxatie', in: J.W. Hummelen, R.J. Verkes & M.J.F. van der Wolf (red.), *Forensische psychiatrie en de rechtspraktijk*, Utrecht: De Tijdstroom 2018, p. 287-306.

**NFI 2017**
NFI, 'De reeks waarschijnlijkheidstermen van het NFI en het Bayesiaanse model voor interpretatie van bewijs', *Vakbijlage NFI*, mei 2017, versie 2.2.

**Raftery 1995**
A.E. Raftery, 'Bayesian model selection in social research', in: P.V. Marsden (red.), *Sociological methodology*, Cambridge (MA): Blackwell 1995, p. 111-196.

**Rassin 2019**
E. Rassin, 'Likelihood ratios in rechtspsychologische rapporten', *EeR* 2019, afl. 5, p. 188-194.

**Spreen e.a. 2014**
M. Spreen, E. Brand, P. ter Horst & S. Bogaerts, *Handleiding HKT-R, Historische, Klinische en Toekomstige – Revisie*, Groningen: Stichting FPC Dr. S. van Mesdag 2014.

**Verkes 2018**
R.J. Verkes, 'Technische begrippen bij risicotaxatie-instrumenten', in: J.W. Hummelen, R.J. Verkes & M.J.F. van der Wolf (red.), *Forensische psychiatrie en de rechtspraktijk*, Utrecht: De Tijdstroom 2018, p. 307-312.