Tijdschrift voor psychiatrie 25, 1983/3

De Nederlandse vertaling van de Present State Examination 9e editie

door C. J. Slooff, W. R. E. H. Mulder-Hajonides van der Meulen en R. H. van den Hoofdakker

I Betrouwbaarheidsaspecten*

Inleiding

Studies over de tussenbeoordelaarsovereenkomst van psychiatrische diagnosen en classificaties hebben in het verleden sterk uiteenlopende, veelal teleurstellende, resultaten opgeleverd. De verschillen in uitkomsten zijn toe te schrijven aan de verschillen in definities van de symptomen, het ontbreken van beslissingsregels voor het rangschikken ervan ten behoeve van een classificatie en aan verschillen in classificatiesystemen. Andere belangrijke oorzaken hiervan zijn de verschillen in onderzoeksdesigns, in patiëntenpopulaties en in de eigenschappen van de beoordelaars, zoals interviewstijl en opleiding (Beck, 1962; Frank, 1975; Helzer, 1977; Kendell, 1975; Spitzer, 1978; Ward, 1962).
Controle van deze variabelen leidde tot de ontwikkeling van gestandaardiseerde psychiatrische interviews, waarvan de Present State Examination (PSE) een bekend voorbeeld is (Pen, 1977; Wing e.a., 1974). In dit interview worden 140 klassieke psychiatrische symptomen beoordeeld aan de hand van een semi-gestructureerd interviewprotocol. De symptomen zijn verdeeld in drie groepen, waarvan er twee worden vastgesteld door middel van vragen aan de patiënt. Dit zijn de neurotische en psychotische symptomen en ze betreffen afwijkende ervaringen van de patiënt in de maand vóór het interview. De derde groep wordt vastgesteld door observatie. Hier betreft het afwijkend gedrag, affect en taal tijdens het interview. Definities van symptomen, beslissingsregels voor de beoordeling van de aanwezigheid en ernst van symptomen alsmede regels voor de scoring zijn uitgebreid beschreven in een begeleidende handleiding. Bij de PSE behoort een computerprogramma (CATEGO), dat de symptoomscores verwerkt op een manier, die ontleend is aan de beslissingsprocedure uit de klinisch psychiatri-

* Deel II zal gepubliceerd worden in het volgende nummer van dit tijdschrift.

Schrijvers zijn respectievelijk als psychiater, als psychologe en als psychiater-hoogleraar verbonden aan de afdeling Biologische Psychiatrie Rijksuniversiteit Groningen; hoofd prof. dr. R. H. van den Hoofdakker.

sche praktijk. Deze verwerking leidt tot CATEGO-syndromen en CATEGO-klassen, die, eventueel aangevuld met gegevens uit de ziektegeschiedenis van patiënten, een classificatie volgens de International Classification of Diseases (ICD) opleveren. Het CATEGO-programma is beschreven door Wing (1974) en Verhey (1981). De achtste en negende editie van de PSE zijn uitvoerig op betrouwbaarheid onderzocht in klinisch en in internationaal epidemiologisch onderzoek bij opgenomen en bij ambulante patiënten. Deze studies hebben aangetoond, dat de betrouwbaarheid van psychiatrische symptomen en classificaties tot een redelijk hoog niveau kan worden gebracht (Cooper, 1977; Duckworth, 1978, Luria, 1974; Wing, 1967, 1977). In Nederland rapporteerden Pen en anderen (1977) echter, overeenkomstig Kendell (1968), dat de observatieitems tijdens het interview minder betrouwbaar werden vastgesteld dan symptomen, die beoordeeld werden aan de hand van de informatie verstrekt door de patiënt. Tevens meldden deze auteurs, dat de symptomatologie bij psychotische patiënten minder betrouwbaar werd vastgesteld dan die bij andere ziektebeelden. De tussenbeoordelaarovereenkomst van de symptomen bleek bovendien in twee opeenvolgende interviews, gevoerd door verschillende beoordelaars, beduidend lager dan de overeenkomst die bereikt werd met twee onafhankelijke beoordelingen op eenzelfde tijdstip. Dit geheel in overeenstemming met de onderzoeksresultaten uit de literatuur (Cooper, 1977; Helzer, 1977; Kreitman, 1961; Wing, 1974, 1977). De implicaties hiervan voor de CATEGO- en ICD-classificaties werden onvoldoende onderzocht.
Gegevens betreffende de inhoudelijke validiteit van CATEGO werden verzameld in het US-UK-project (Cooper e.a., 1972) en de International Pilot Study of Schizophrenia (WHO, 1973, 1979). In deze studies werden de CATEGO-classificaties vergeleken met de ICD-classificaties van de onderzoekers. De overeenkomst tussen deze twee onafhankelijk van elkaar verkregen ICD-classificaties bleek hoog te zijn voor de categorieën functionele psychosen en neurosen (ICD 295-298 en 300), alhoewel het aantal patiënten met een affectieve stoornis of een neurose ons inziens vergelijkenderwijs te klein was voor een duidelijke conclusie.
De predictieve validiteit van de PSE-CATEGO-klassen is tot nu toe slechts bestudeerd voor patiënten met een functionele psychose (Brockington, 1978; Kendell, 1979; WHO, 1979). In een vergelijking met andere classificatiesystemen voor schizofrene psychosen bleek de CATEGO-classificatie van beperkte betekenis bij de predictie van het sociale functioneren bij een vervolgbeoordeling na 5 jaar. In de vervolgstudie van de IPSS (WHO, 1979) bleken depressieve patiënten vergeleken met schizofrene patiënten, geclassificeerd volgens CATEGO, in het algemeen weliswaar een betere prognose te hebben, doch de voorspellende waarde van de klasse-indeling voor de prognose bleek gering. De verschillen in symptomatologie van de ziektebeelden kwamen niet voort uit pathognomonische symptomen, maar berustten op belangrijke kwantitatieve verschillen.

In Nederland is de 9e editie van de ICD in gebruik. Op dit classificatiesysteem is kritiek geleverd (Slooff, 1981). In dit systeem worden fenomenologische, etiologische en beloopskenmerken ongesystematiseerd en ongedefinieerd gebruikt als indelingscriteria bij de classificatie van psychiatrische ziektebeelden. Betrouwbaarheidsonderzoek naar ICD-classificaties in Nederland is niet beschikbaar. Het is dus volstrekt onduidelijk welke waarde dit instrument heeft bij klinisch gebruik en bij onderzoek.
De PSE en het PSE-CATEGO-computerprogramma voorziet ons van een psychiatrisch beoordelingsinstrumentarium dat verschillende voordelen biedt. Met dit systeem is ten eerste een redelijk betrouwbare beschrijving van symptomatologische kenmerken van patiëntenpopulaties mogelijk. Ten tweede kan de relatie tussen ICD-categorieën en CATEGO-klassen van ziektebeelden bij patiënten worden onderzocht.
In 1977 werd door medewerkers van de psychiatrische kliniek van de RU te Groningen de 9de editie van de PSE vertaald ten behoeve van biologisch psychiatrisch onderzoek bij depressieve patiënten en sociaal psychiatrisch onderzoek bij patiënten, die voor de eerste keer werden opgenomen met een functioneel niet affectieve psychose (Giel e.a., 1980). Aan deze onderzoeken ging het onderhavige betrouwbaarheidsonderzoek van de PSE vooraf (deel I). In een tweede studie onderzoeken wij de overeenkomsten van enkele CATEGO-klassen met categorieën van de ICD (deel II).

Vraagstelling deel I

1. Hoe hoog is de overeenkomst tussen twee onafhankelijke beoordelaars van de PSE-symptomen, de PSE-syndromen en CATEGO-klassen op eenzelfde tijdstip (interviewer-observator-overeenkomst) en op twee opeenvolgende tijdstippen (test-retest-overeenkomst)?
2. Wat is de invloed van kenmerken van de beoordelaars en van patiënten op de interviewer-observator-overeenkomst?
3. Wat is de invloed van patiëntenkenmerken op de test-retest-overeenkomst?

Methode

(a) *PSE*. De PSE is een checklist van 140 klassieke psychiatrische symptomen waarvan 100 symptomen op hun aan- of afwezigheid en op hun ernst in de maand voorafgaande aan het interview worden nagegaan met behulp van een voorgestructureerd interview met de patiënt. Deze symptomen zijn verdeeld over secties betreffende neurotische en psychotische symptomatologie en vormen tezamen de vraagitems. Ze betreffen verbaal medegedeelde ervaringen van de patiënt (zie fig. 1). Daarnaast vindt beoordeling plaats van geobserveerd gedrag, affect en taal tijdens het interview aan de hand van een lijst van 33 items. De beoordelaar beschikt over de volgende score-alternatieven: 0 = afwezig, 8 = onzeker, 9 = niet

van toepassing of niet gevraagd, 1 = duidelijk aanwezig in lichte mate, 2 = aanwezig in ernstige mate. 114 van de 140 symptomen en observaties worden met behulp van het CATEGO-computerprogramma gegroepeerd tot 36 CATEGO-syndromen. Deze syndromen hebben op zichzelf nog geen klinische betekenis. De syndromen zijn te verdelen in 4 categorieën, namelijk Delusional and Hallucinatory Syndromes (DAH), Behavior, Speech and Other Syndromes (BSO), Specific Neurotic Syndromes (SNR) en Nonspecific Neurotic Syndromes (NSN) (zie fig. 1).
Het CATEGO-programma deelt het ziektebeeld van de patiënt op grond van de syndroomscores bij één CATEGO-klasse in. Er zijn 12 van deze klassen en elk van hen heeft overeenkomst met bepaalde ICD-categorieën. Zo heeft klasse S overeenkomst met de classificatie Schizofrene Psychosen (ICD 295), klasse P met de ICD-klasse Paranoïde en Psychogene Psychosen (ICD 297, 298.3-9), M met het maniakaal toestandsbeeld (ICD 298.1, 296.0), D met psychotische depressie (ICD 296.1, 298.0), R met geremde depressie, N met neurotische depressie (ICD 300.4) en A met angsttoestand (ICD 300.0).
De verschillende CATEGO-klassen kunnen tevens beschreven worden met behulp van de syndroomscores die tot die klassen hebben geleid. Van een patiënt kan zo een "syndroomprofiel" worden gegeven.
(b) *Design*. De twee meest gebruikte designs bij betrouwbaarheidsonderzoek zijn het interviewer-observator design en het test-retest design. Bij het interviewer-observator design wordt de psychopathologie van een patiënt onafhankelijk gescored door twee beoordelaars op grond van een interview, dat één van deze beoordelaars (de interviewer) met de patiënt heeft. De andere beoordelaar (de observator) woont het interview hetzij direct hetzij via een monitor/video-opname bij. Bij het test-retest design wordt door twee onafhankelijke beoordelaars in twee onafhankelijke interviews bij dezelfde patiënt de psychopathologie van die patiënt gescored. Deze twee interviews vinden meestal zo kort mogelijk na elkaar plaats. Bij beide typen designs wordt de overeenstemming tussen beide beoordelaars bepaald door 1) de eigenschappen van het instrument, 2) de eigenschappen van de beoordelaars (algemene ervaring, mate van specifieke training met het instrument) en 3) de eigenschappen van de te beoordelen patiënt. Bij het interviewer-observator design wordt de overeenstemming verhoogd door het feit, dat beide beoordelaars de psychopathologie van de patiënt scoren op basis van hetzelfde interview en door het feit, dat de observator niet alleen de patiënt, maar ook de interviewer observeert. Derhalve zijn de scores van beide beoordelaars in het interviewer-observator design niet volledig onafhankelijk van elkaar. Bij het test-retest design bestaat de mogelijkheid dat de overeenstemming tussen beide beoordelaars wordt geschaad door instabiliteit van het toestandsbeeld van de patiënt. Bovendien kunnen alle bovengenoemde foutenbronnen elkaar wederzijds versterken of verzwakken met betrekking tot hun invloed op de uiteindelijk

gevonden overeenstemming.
Om een zo volledig mogelijk beeld te verkrijgen over de betrouwbaarheid van de PSE en de invloed van mogelijke foutenbronnen is in dit onderzoek van beide typen designs gebruik gemaakt. In de interviewer-observator-situatie observeert de observator via een video-monitor. Dit maakt de onafhankelijkheid van de observator groter dan wanneer deze er direct bij zou zitten. Het betrouwbaarheidsonderzoek werd uitgevoerd bij 100 patiënten in een interviewer-observator-situatie en 24 patiënten hiervan werden in de laatste fase van het onderzoek voor een tweede keer geïnterviewd.
Voor het onderzoek naar de overeenkomst tussen de ICD en CATEGO-categorieën werd aan elke patiënt direct na elk interview een ICD-categorie (9e editie) toegekend door het interviewer-observator-paar en bij 24 patiënten eveneens door de tweede interviewer. Daarnaast werden de ICD-categorieën van de patiënten na ontslag verzameld. De overeenkomst van de op deze wijze verkregen categorieën geven een indruk van de overeenkomst, waarmee ICD-categorieën worden toegekend. Niet meer dan een indruk overigens, want de beschikbare informatie vlak na opname (het toestandsbeeld) verschilt aanzienlijk van die bij ontslag (toestandsbeeld, etiologie, beloop en behandelingsresultaat). Het toewijzen van de ICD-categorieën bij ontslag geschiedde grotendeels onafhankelijk van die door de PSE-beoordelaar na opname. In een klein aantal gevallen was het onvermijdelijk dat de behandelaars als beoordelaar betrokken waren geweest bij het PSE-interview bij opname.
(c) *Patiënten*. In totaal werden 111 patiënten gevraagd medewerking te verlenen. 5 Patiënten weigerden hun medewerking en bij 6 patiënten werd het interview voortijdig afgebroken vanwege wijdlopigheid en vaagheid van de patiënt. Bruikbare gegevens werden uiteindelijk verkregen van 100 opgenomen patiënten (53 vrouwen en 47 mannen met een leeftijd variërende van 17-78 jaar) in een interviewer-observator-situatie en bij 24 patiënten werd een tweede interview afgenomen. De selectie-criteria waren: 1) in staat en bereid zijn aan het interview deel te nemen en 2) aanwezigheid van symptomatologie. In diagnostisch opzicht is sprake van een heterogene patiëntenpopulatie.
(d) *Beoordelaars*. Er namen 7 mannelijke beoordelaars deel aan de studie; zes psychiaters, van wie 3 ervaren en 3 minder ervaren, en een klinisch psycholoog. De beoordelaars werden in paren toegewezen aan de patiënten in elke mogelijke combinatie en rolverdeling in een zodanige volgorde, dat het leereffect voor allen gelijk was. Er werd zoveel mogelijk voor gewaakt, dat de informatie over de patiënten voor beide beoordelaars gelijk was. De tweede interviewer was niet op de hoogte van de resultaten van het eerste interview.
Voor het begin van de studie verrichtte elke beoordelaar minstens 5 PSE-interviews en van elke beoordelaar werd een interview uitvoerig bediscussieerd aan de hand van een video-opname.
Bovendien volgden twee psychiaters de PSE-trainingscursus in

'The Maudsley Institute' in Londen.
(e) *Procedure*. Alle interviews vonden plaats in dezelfde ruimte aan het begin van de middag. De observator volgde het interview tegelijkertijd via een videoscreen. 98% van de interviews vond plaats in één sessie, de resterende interviews in 2 sessies. De duur van de interviews varieerde van 30 minuten tot 150 minuten, in de meeste gevallen werd het interview binnen een uur afgerond. In 62% van de gevallen vond het eerste interview plaats binnen de termijn van 3 dagen na opname en in alle gevallen binnen 12 dagen. Het tijdsverloop tussen het eerste en tweede PSE-interview varieerde van 1 tot 9 dagen, waarvan 67% binnen 3 dagen. De totale duur van het onderzoek bedroeg 11 maanden. De opnameduur van de patiënten varieerde van een week tot 1 jaar.
(f) *Statistische analyse*. De mate van overeenstemming is op drie verschillende niveaus berekend: over de symptomen, de syndromen en de klassen (vraagstelling 1). Bovendien is de invloed van enkele factoren op de mate van overeenstemming geëvalueerd (vraagstellingen 2 en 3). In het navolgende zal eerst ingegaan worden op enkele aspecten van het meten van overeenstemming. Daarna zal beschreven worden hoe de overeenstemming is bepaald met betrekking tot vraagstelling 1. Tenslotte zal worden aangegeven hoe de overeenstemming is bepaald met betrekking tot de vraagstellingen 2 en 3.
Als maat voor overeenstemming is gekozen voor kappa (Cohen, 1968). Kappa meet de mate van overeenstemming nadat overeenstemming op grond van kans is afgetrokken. Kappa wordt veel gebruikt als maat voor overeenstemming bij betrouwbaarheidsonderzoek van diagnostische instrumenten en maakt derhalve een vergelijking van de verschillende onderzoeksresultaten mogelijk. Bezwaren, die tegen kappa aangevoerd worden, zijn: 1) aan de ingebouwde statistische assumpties (de eenheden/subjecten zijn onderling onafhankelijk, de nominale schaalcategorieën zijn onafhankelijk, wederzijds uitsluitend en gezamenlijk uitputtend en de beoordelaars zijn onafhankelijk en even competent in het maken van hun beoordelingen) wordt lang niet altijd voldaan en 2) de berekening van overeenstemming op grond van kans is aanvechtbaar (Maxwell, 1977).
Het ontbreken van een redelijk alternatief voor kappa heeft in veel onderzoek echter geleid tot allerlei ad hoc oplossingen, hetgeen de onderlinge vergelijkbaarheid onmogelijk maakt. In het licht van bovenstaande bezwaren dient kappa beschouwd te worden als een zuiver descriptieve maat van overeenstemming.
Bij de berekening van kappa voor de *symptomen en syndromen* werden verschillende gewichten toegekend aan de verschillende vormen van niet-overeenstemming. Aan een verschil van opvatting over aan- of afwezigheid (0 vs. 1 of 2) werd een zwaarder gewicht toegekend dan aan een verschil van opvatting over de mate waarin het kenmerk aanwezig was (1 vs. 2). De gewichten werden toegekend overeenkomstig Kendell (1968) en Pen (1977).
De weging bij de berekening van de overeenstemming met betrek-

king tot de *CATEGO-klassen* is gebaseerd op de sequentie van de klassen in afname van ernst: S, O, P, M, D, R, N, A (zie fig. 1), waarbij aan de intervallen tussen deze klassen een gelijk gewicht werd toegekend, dus S - O = 1, S - P = 2, ... , S - A = 7; O - P = 1, O - M = 2, etc. De overeenstemming van de afzonderlijke klassen is berekend in een 2 x 2 tabel, waar de opsplitsing werd gemaakt in *behorende* tot de klasse en *niet behorende* tot de klasse met betrekking tot de scores van de interviewer resp. met betrekking tot de scores van de observator.

(g) *Factoren, die de mate van overeenstemming kunnen beïnvloeden*. Elke beoordeling bevat een interactie tussen een patiënt, een beoordelaar en een item van psychopathologie. De overeenstemming wordt beïnvloed door elk van deze drie factoren. Om een indruk te krijgen over de mate van overeenstemming per *item*, zijn de gewogen kappa's berekend voor de symptomen, syndromen en klassen voor de interviewer-observator-situatie over 100 patiënten (K_{IO}) en voor de interviewer-herinterviewer-situatie over 24 patiënten (K_{IH}). De overeenstemmingsresultaten in de interviewer-herinterviewer-situatie worden vergeleken met de overeenstemmingsresultaten over dezelfde 24 patiënten in de interviewer-observator-situatie. Om een indruk te krijgen of er een verschil bestaat in mate van overeenstemming voor verschillende *kenmerken van de beoordelaars* (mate van training in de PSE, ervaren psychiater vs. onervaren psychiater, interviewer vs. observator en interviewer vs. herinterviewer) is voor elke patiënt afzonderlijk de overeenstemming tussen de syndroomscores van beide beoordelaars berekend. Dezelfde procedure is gevolgd om de invloed van *diagnostische kenmerken* van de patiënten op de mate van overeenstemming tussen beide beoordelaars te evalueren.

Resultaten

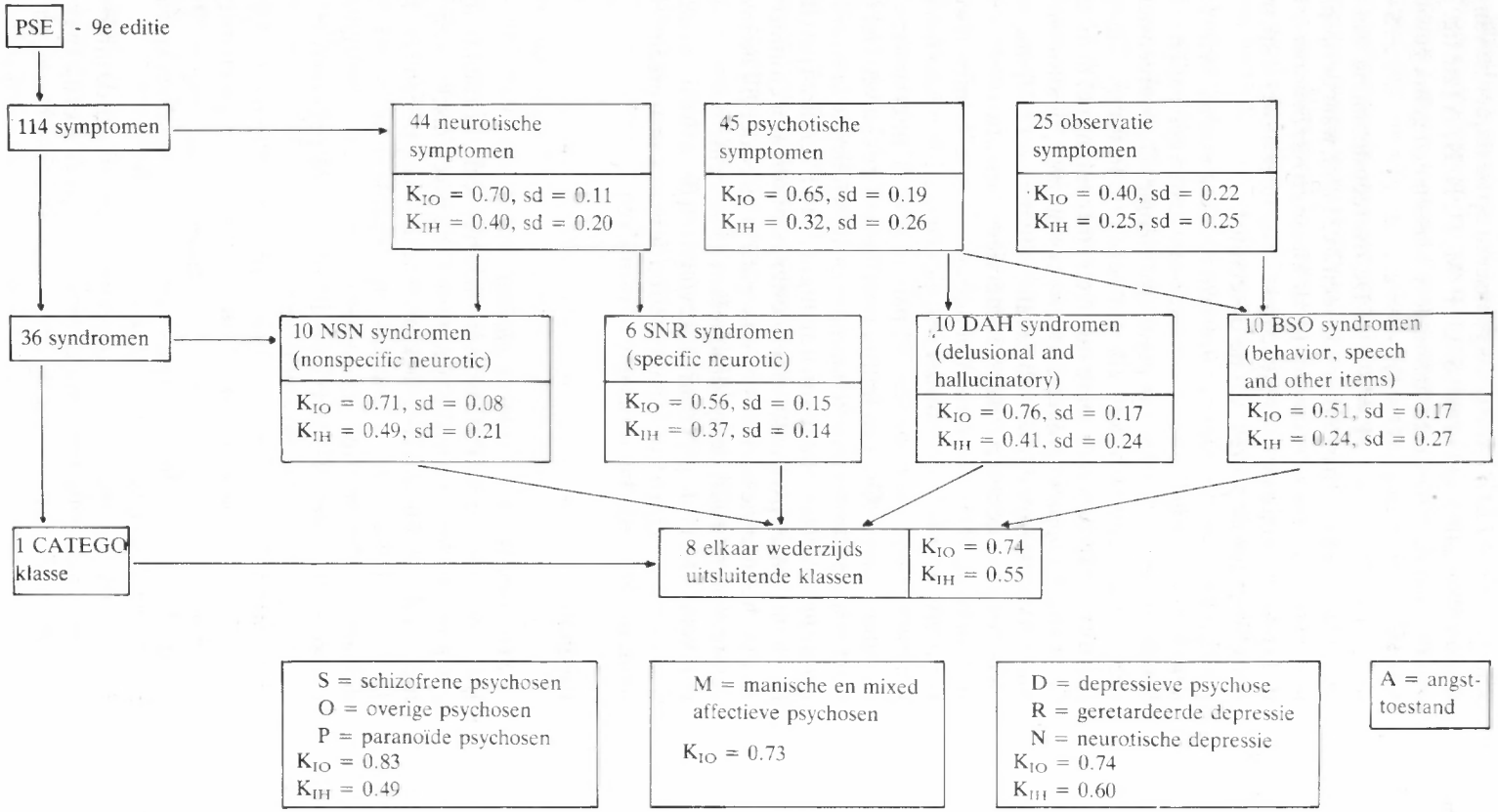
1. *Symptomen, syndromen en klassen*

De symptomen in de PSE kunnen onderverdeeld worden in drie categorieën: de neurotische symptomen, de psychotische symptomen en de gedrag-affect-spraak(observatie)-symptomen (zie fig. 1). Per categorie is de gemiddelde kappa-waarde en de bijbehorende standaarddeviatie weergegeven voor respectievelijk 100 patiënten in de interviewer-observator-situatie, en 24 patiënten in de interviewer-herinterviewer-situatie.

In de interviewer-observator-situatie (100 patiënten) blijkt, dat de gemiddelde kappa-waarden van de neurotische en psychotische symptomen niet sterk van elkaar verschillen (de K_{IO} resp. 0.70 en 0.65). De gemiddelde kappa-waarde van de gedrag-affect-spraak-symptomen ligt echter aanzienlijk lager (K_{IO} 0.40).

De gemiddelde kappa-waarden over de 24 patiënten in de interviewer-observator-situatie stemmen redelijk overeen met die over de 100 patiënten. Wat betreft de mate van overeenstemming in de interviewer-observator-situatie is deze groep dus representatief voor de totale populatie. Echter, de gemiddelde kappawaarden

Figuur 1: Schematische weergave van de structuur van de PSE en de voornaamste resultaten



K_{IO} = gewogen kappa in de interviewer-observator-situatie
K_{IH} = gewogen kappa in de interviewer-herinterviewer-situatie
K = gemiddelde kappa; sd = standaard-deviatie

over deze 24 patiënten in de interviewer-herinterviewer-situatie liggen aanmerkelijk lager (neurotische symptomen: $K_{IH} = 0.40$; psychotische symptomen: $K_{IH} = 0{,}32$ en observatie-symptomen: $K_{IH} = 0.25$). Dit verschil is ongeveer even groot voor de psychotische als voor de neurotische symptomen. Bij de observatie-symptomen is het verschil veel kleiner.

Evenals Wing (1977) hebben wij de syndromen onderverdeeld in vier categorieën: 10 NSN-syndromen (non specific neurotic), 6 SNR-syndromen (specific neurotic), 10 DAH-syndromen (delusional and hallucinatory) en 10 BSO-syndromen (behavior, speech and other items). De lage overeenstemming van de observatie-symptomen in de interviewer-observator-situatie wordt gereflecteerd in de BSO-syndromen ($K_{IO} = 0.51$). De gemiddelde overeenstemming van de SNR-syndromen is echter ook vrij laag ($K_{IO} = 0.56$); voornamelijk ten gevolge van de lage overeenstemming op de syndromen betreffende dwangverschijnselen en hysterie. De gemiddelde overeenstemming voor de DAH- en NSN-syndromen is aanzienlijk hoger ($K_{IO} = 0.76$ resp. $K_{IO} = 0.71$). In de interviewer-herinterviewer-situatie is de mate van overeenstemming wederom aanzienlijk lager. Dit verschil is het grootst bij de DAH-syndromen ($K_{IH} = 0.41$) en het kleinst bij de SNR- en NSN-syndromen ($K_{IH} = 0.37$ resp. $K_{IH} = 0.49$). Het verschil bij de BSO-syndromen ligt hier tussenin. Uit deze resultaten kan geconcludeerd worden, dat het verschil in overeenstemming in de interviewer-herinterviewer-situatie het grootst is bij de psychotische syndromen.

Op grond van zijn scores op de 36 syndromen wordt elke patiënt door het CATEGO programma toegewezen aan één CATEGO-klasse. Er worden bij onze populatie acht elkaar uitsluitende hoofdklassen onderscheiden (zie fig. 1). De overeenstemming tussen de interviewer en de observator met betrekking tot alle klassen gezamenlijk is zeer bevredigend ($K_{IO} = 0.74$). De overeenstemming tussen interviewer en herinterviewer is wederom lager ($K_{IH} = 0.55$). Voor de berekening van de overeenstemming van de afzonderlijke klassen zijn de niet affectief psychotische klassen bij elkaar genomen, ter vergelijking met de wel depressieve klassen omdat niet alle klassen voldoende gevuld waren om een zinvolle uitspraak te kunnen doen over de overeenstemming voor elke klasse apart. De S-, O- en P-klassen bij elkaar genomen, bereikten een K in de interviewer-observator situatie van 0.83 en in de interviewer-herinterviewer-situatie een marginale waarde van 0.49. Voor de D-, R- en N-klassen is K in de interviewer-observator situatie 0.74 en in de interviewer-herinterviewer-situatie 0.60. Het verschil in de interviewer-herinterviewer-situatie is dus aanzienlijk groter voor de S-, O- en P-klasse dan voor de D-, R- en N-klasse. Voor de M-klasse is K in de interviewer-observator-situatie 0.73. Voor de M-klasse is in de interviewer-herinterviewer-situatie geen K berekend wegens het geringe aantal patiënten in de M-klasse in deze situatie. Om dezelfde reden is voor de A-klasse apart geen K berekend in de interviewer-observator-situatie en evenmin in de interviewer-herinterviewer-situatie.

2. *Interviewer-variabelen*
Ter evaluatie van mogelijke trainingseffecten werden de 100 interviews in de interviewer-observator-situatie opgedeeld in vier groepen van elk 25 interviews overeenkomstig de chronologische volgorde (dus in groep 1 de 25 interviews, in groep 2 de daaropvolgende 25 interviews, etc.). Over elke groep van 25 interviews werd een gemiddelde K berekend. De verschillen in gemiddelde K deze vier groepen waren niet groot (voor de vier achtereenvolgende groepen werden de volgende K-waarden gevonden: 0.72, 0.76, 0.75 en 0.69).
De beoordelaars vervulden even vaak de interviewerrol als de observatorrol. Bij de vergelijking van de totaal-scores voor elk interview bleken de interviewers systematisch meer symptomatologie te scoren dan de observatoren; het gemiddelde verschil (2.34; sd = 6.05) was echter gering. Onder de zeven beoordelaars bevonden zich zes psychiaters; drie ervaren en drie onervaren. De onervaren psychiaters vertonden vertoonden de interviewerrol de neiging meer symptomatologie te scoren in vergelijking met de observatoren van hun interviews (gemiddeld verschil = 3.32; sd = 6.90) dan de ervaren psychiaters deden als interviewers in vergelijking met de observatoren van hun interviews (gem. verschil = 1.73; sd = 5.47). Men zou kunnen verwachten dat de herinterviewer systematisch minder zou scoren. Behandelingen hebben immers reeds een aanvang genomen. Er bleek echter geen systematische bias verbonden te zijn aan de herinterviewerrol.

3. *Patiënt-variabelen*
Om een indruk te krijgen over de invloed van diagnostische kenmerken van de patiëntenpopulatie op de mate van overeenstemming werden de patiënten ingedeeld naar de CATEGO-klassen, verkregen met de scores van de interviewer. Patiënten uit de klassen S, O en P te zamen ($n = 38$, $K_{IO} = 0.72$, sd = 0.12) zijn minder betrouwbaar te scoren dan de patiënten uit de klassen D, R en N te zamen ($n = 50$, $K_{IO} = 0.78$, sd = 0.12). Het verschil tussen beide groepen is echter klein.
In de interviewer-herinterviewer-situatie is dit verschil echter aanmerkelijk groter. De overeenstemming bij patiënten uit de S-, O- en P-klasse ($K_{IH} = 0.46$, sd = 0.15) is aanzienlijk lager dan de overeenstemming bij patiënten uit de D-, R- en N-klasse ($K_{IH} = 0.71$, sd = 0.10, $n = 7$). De duur van het tijdsinterval tussen het eerste en tweede interview bleek niet van invloed op deze resultaten.

Discussie

De PSE blijkt toepasbaar bij een heterogeen samengestelde patiëntenpopulatie waarvoor opname noodzakelijk was. Slechts bij 6 van de 111 gevraagde patiënten bleek een interview niet uitvoerbaar.
Na een korte trainingsperiode bereikten de 7 beoordelaars een

redelijk betrouwbaarheidsniveau (althans v.w.b. hun beoordelingen op eenzelfde tijdstip), dat gedurende het verloop van de studie constant bleef. Interviewers scoorden significant méér symptomatologie dan observatoren. Minder ervaren beoordelaars stelden meer symptomatologie vast dan ervaren beoordelaars. De verschillen zijn echter klein. Ook in dit opzicht kan de PSE een bruikbaar instrument genoemd worden. Daar voor voldoende betrouwbaarheid geen criterium voor handen is, kunnen wij voor de beoordeling van de resultaten ons slechts richten op vergelijkbare studies. In het algemeen liggen de resultaten van de betrouwbaarheidsmetingen op het niveau van de in de literatuur vermelde. Ook bij ons bereiken de PSE-observaties betreffende gedrag, affect en spraak een onvoldoende betrouwbaarheidsniveau. Dit geldt zowel voor metingen op hetzelfde tijdstip als voor metingen op twee opeenvolgende tijdstippen. Deze onbetrouwbaarheid kan niet worden toegeschreven aan de lage frequentie, waarin deze symptomen voorkomen in dit onderzoek, daar de frequentie niet verschilt van die van psychotische symptomen, waarbij een veel hoger resultaat wordt bereikt. Daar aan de aanwezigheid van sommige observatiegegevens belangrijke waarde wordt toegekend bij de beslissingen in het CATEGO-programma, zal in de toekomst aandacht moeten worden besteed aan de verhoging van de betrouwbaarheid waarmee afwijkend gedrag, affect en taal wordt vastgesteld. Dit kan mogelijk worden bewerkstelligd door beter uitgewerkte definities en scoreregels. Een andere mogelijkheid is CATEGO-classificatie op grond van uitsluitend symptomen ontleend aan mededelingen van de patiënt. De test-retest resultaten zijn veel minder bevredigend dan de interviewer-observator-resultaten. Dit is hoofdzakelijk te wijten aan de lagere overeenkomsten, die bereikt werden bij acuut psychotische patiënten uit de CATEGO S-, O- en P-klassen. Deze verschillen zijn niet uitsluitend toe te schrijven aan de invloed van interview-stijlen, het tijdsverloop tussen het eerste en tweede interview en intussen ingestelde behandelingen. Bij depressieve patiënten is dit verschil immers veel minder groot. Er werd evenmin een systematisch verschil gevonden in het aantal aanwezige symptomen tussen het eerste en tweede interview bij patiënten uit de S-, O- en P-klasse. Deze bevindingen leiden tot de conclusie, dat acuut psychotische patiënten op twee opeenvolgende tijdstippen een verschillend verslag geven van hun symptomatologie in de maand voorafgaande aan de interviews. De waarde van de CATEGO-classificatie bij deze categorie patiënten is dus beperkt. Voor depressieve patiënten geldt dat in mindere mate.

Literatuur

Beck, A. T. (1962). Reliability of psychiatric diagnoses: 1. A critique of systematic studies. *Am. J. Psychiatry* 119, p. 210-16.

Brockington, I. F., Kendell, R. E. and Leff, J. P. (1978). Definitions of schizophrenia, concordance and prediction of outcome. *Psychol. Med.* 8, p. 387-98.

Cohen, J. (1968). Nominal scale agreement with provision for scaled disagreement

or partial credit. *Psychol. Bull.* 70, p. 213-20.

Cooper, J. E., Kendell, R. E., Gurland, B. J., Sharpe, L., Copeland, J. R. M. and Simon, R. (1972). *Psychiatric diagnosis in New York and London*. Maudsley Monographs, Oxford University Press.

Cooper, J. E., Copeland, J. R. M., Brown, G. W., Harris, T. and Gourlay, A. J. (1977). Further studies on interviewer training and inter-rater reliability of the Present State Examination (PSE). *Psychol. Med.* 7, p. 517-23.

Duckworth, G. S. and Kedward, H. B. (1978). Man or machine in psychiatric diagnosis. *Am. J. Psychiatry* 135, p. 64-8.

Frank, G. H. (1975). *Psychiatric diagnosis. A review of research*. Pergamon Press, Oxford.

Giel, R., Sauer, H. C., Slooff, C. J. and Wiersma, D. (1980) Over epidemiologie van functionele psychosen en invaliditeit. *Tijdschrift voor Psychiatrie* 22, p. 710-722.

Helzer, J. E., Robins, L. N., Taibleson, M., Woodruff, R. A., Reich, T. and Wish. E. D. (1977). Reliability of psychiatric diagnosis, I. A methodological review. *Arch. Gen. Psychiatry* 34, p. 129-33.

Helzer, J. E., Clayton, P. J., Pambakian, R., Reich, T., Woodruff, R. A. and Reveley, M. A. (1977). Reliability of psychiatric diagnosis. II. The test/retest reliability of diagnostic classification. *Arch. Gen. Psychiatry* 34, p. 136-41.

Kendell, R. E., Everett, B., Cooper, J. E., Sartorius, N. and David, M. E. (1968). The reliability of the 'Present State Examination'. *Soc. Psychiatry* 3, p. 123-29.

Kendell, R. E. (1975). *The role of diagnosis in psychiatry*. Oxford Blackwell, Scientific Publications.

Kendell, R. E., Brockington, J. F. and Leff, J. P. (1979). Prognostic implications of six alternative definitions of schizophrenia. *Arch. Gen. Psychiatry* 36, p. 25-31.

Kreitman, N. (1961). The reliability of psychiatric diagnosis. *J. Ment. Sci.* 107, p. 876-887.

Kreitman, N., Sainsbury, P., Morrissey, J., Towers, J. and Scrivener, J. (1961). The reliability of psychiatric assessment: an analysis. *J. Ment. Sci.* 107, p. 887-910.

Luria, R. E. and McHugh, P. R. (1974). Reliability and clinical utility of the 'Wing' Present State Examination. *Arch. Gen. Psychiatry* 30, 866-71.

Maxwell, A. E. (1977). Coefficients of Agreement between observers and their interpretation. *Brit. J. Psychiatry* 130, 79-83.

Pen, A., Arrindell, W. A., Lawant, W. and Pols, J. (1977). De betrouwbaarheid van de 'Present State Examination', een replicatiestudie. *Tijdschrift voor Psychiatrie* 19, p. 375-87.

Slooff, C. J., Giel, R., De Jong, A. en Wiersma, D. (1981) Klassificatie van functionele niet-affectieve psychosen, ICD of DSM. *COBO-bulletin* 14, 3, p. 105-110.

Spitzer, R. L. and Fleisch, J. L. (1974). A re-analysis of the reliability of psychiatric diagnosis. *Br. J. Psychiatry* 125, 341-47.

Spitzer, R. L., Fleisch, J. L. and Endicott, J. (1978). Problems of classification: reliability and validity. In *Psychopharmacology: a generation of progress* (eds. Lipton, DiMascio and Killam), Raven Press, New York.

Verhey, F. H. M. (1981). *Langzame hersenpotentialen en gemeten gedrag*. Leiter Nijpels BV, Maastricht (proefschrift).

WHO Geneva (1973). *The international pilot study of schizophrenia*.

WHO (1979). *Schizophrenia, an international follow-up study*. John Wiley and Sons Ldr, Lat., etc.

Ward, C. H., Beck, A. T., Mendelson, M, Mock, J. E. and Erbaugh, J. K. (1962). The psychiatric nomenclature. Reasons for diagnostic disagreement. *Arch. Gen. Psychiatry* 7, p. 198-205.

Wing, J. K., Birley, L. T., Cooper, J. E., Graham, P. and Isaacs, A. D. (1967). Reliability of a procedure for measuring and classifying 'Present Psychiatric

State'. *Br. J. Psychiatry* 113, p. 499-515.

Wing, J. K., Cooper, J. E. and Sartorius, N. (1974). *Description and classification of psychiatric symptoms*. Cambridge University Press, Cambridge.

Wing, J. K., Nixon, J. M., Mann, S. A. and Leff, J. P. (1977). Reliability of the PSE (ninth edition) used in a population study. *Psychol. Med.* 7, p. 505-16.