

Het beschrijven van latente dimensies en complexe relaties

J.A. KNOTTNERUS
A. VOLOVICS

Deze aflevering is gewijd aan drie methoden die gebruikt worden om complexe relaties tussen waargenomen verschijnselen in kaart te brengen. Factoranalyse is gericht op het beschrijven van de relatie tussen waargenomen kenmerken en mogelijke, latente dimensies; principale-componentenanalyse en clusteranalyse worden toegepast om een grote hoeveelheid informatie over variabelen of personen samen te vatten in componenten respectievelijk clusters. Principes van deze methoden worden besproken, alsmede een aantal valkuilen en methodologische beperkingen.

Knottnerus JA, Volovics A. Het beschrijven van latente dimensies en complexe relaties [Huisarts, wetenschap en statistiek]. Huisarts Wet 1989; 32(10): 383-8.

Rijksuniversiteit Limburg, Postbus 616, 6200 MD Maastricht.

Prof. dr. J.A. Knottnerus, hoogleraar huisartsgeneeskunde, vakgroep Huisartsgeneeskunde; A. Volovics, vakgroep Medische Informatica en Statistiek.

Correspondentie: Prof. dr. J.A. Knottnerus.

Inleiding

In de huisartsgeneeskunde wordt gewerkt met zowel eenvoudig meetbare, als complexe, niet direct waarneembare grootheden. De onderwerpen die in het huisartsgeneeskundig onderzoek bestudeerd worden, zijn hiervan een afspiegeling. Het komt regelmatig voor dat men, uitgaande van een serie waargenomen kenmerken, al dan niet tevoren vermoede verbanden of factoren in kaart tracht te brengen. Vaak is men er dan op uit om in het onderzoeksmateriaal besloten liggende ordeningen of structuren te herkennen: latente dimensies, die een logische, natuurlijke indruk maken en dus goed interpreteerbaar zijn.¹

De voor dit doel beschikbare statistische methodieken gaan uit van andere uitgangspunten dan de eerder in deze serie besproken multiële regressie-analyse.² Met behulp van regressie-analyse onderzoekt men in de regel in hoeverre een variabele (bijvoorbeeld de bloeddruk) beïnvloed wordt door een aantal andere – verklarende – variabelen (zoals geslacht, leeftijd, gewicht, mate van lichaamsbeweging en gebruik van tabak, koffie en alcohol). Alle variabelen zijn scherp gedefinieerd en in principe direct en eenduidig te meten.

Bij *factoranalyse* ligt dit anders. De bedoeling is juist om uit het waargenomen materiaal achterliggende, niet rechtstreeks te meten dimensies te destilleren. Voorbeelden van het type vragen dat hierbij aan de orde is, zijn:

- Is het mogelijk om, aan de hand van het in een bepaalde populatie waargenomen gebruik van koffie, alcohol, sigaretten en andere genotmiddelen, een concept als 'verslavings-geneigdheid' te omschrijven ten opzichte van 'geen verslavings-geneigdheid'?
- Kunnen we, op basis van een registratie van op voorschrift van huisartsen door apothekers afgeleverde geneesmiddelen, bepaalde voorschrijfpatronen (bijvoorbeeld gericht en ongericht voorschrijfgedrag) ontdekken?^{1 3}
- Zijn er, uitgaande van een registratie van aandoeningen en handelingen in de huisartspraktijk, bepaalde algemene kenmerken van de huisartsgeneeskunde te herkennen?

In feite probeert men hier twee problemen het hoofd te bieden. Ten eerste is er het probleem van de definiëring en herkenning van de achterliggende factoren of dimensies. In de tweede plaats is er de kwestie van de meetbaarheid: het gebruik van de afzonderlijke genotmiddelen bijvoorbeeld is rechtstreeks en op een eenduidige schaal te meten; het concept verslavingsgeneigdheid kan alleen indirect worden gemeten, als afgeleide van het gebruik van genotmiddelen, zelfs al zou het concept zelf helder gedefinieerd zijn.

Daarnaast wordt in deze aflevering aandacht besteed aan twee andere, in dit tijdschrift weleens gebruikte technieken om een grote hoeveelheid informatie over variabelen en personen samen te vatten in componenten (*principale-componentenanalyse*) of clusters (*clusteranalyse*). Daarbij is op zichzelf geen sprake van het zoeken naar latente, niet direct meetbare dimensies.

Factoranalyse

Via factoranalyse probeert men de correlatie tussen een groot aantal variabelen te verklaren in termen van een relatief klein aantal onderliggende factoren die niet (direct) waarneembaar zijn. De variabelen hangen in deze benadering af van (een lineaire combinatie van) de gevonden factoren, maar zijn ook onderhevig aan toevallige afwijkingen.

Factoranalyse is in het begin van deze eeuw bedacht en ontwikkeld door psychologen. Het doel was om, op basis van de scores van een individu op psychologische tests, de fundamentele componenten van cognitie, persoonlijkheid en gedrag op te sporen. Al gauw bleek dat scores op tests die geconstrueerd waren om 'intelligentie' te meten, redelijk correleerden met scores op tests die bedoeld waren voor meting van 'geheugen', terwijl beide scores weer correleerden met tests voor 'associatieve vermogens'. Een mogelijke verklaring was dat de tests gedeeltelijk hetzelfde maten (een gemeenschappelijke cognitieve 'factor') en gedeeltelijk iets anders (specifieke, testgebonden 'factoren').

Het werk van *Spearman* op dit gebied ligt aan de basis van zowel 'mental testing' als

van de factoranalyse.⁴ Zijn model, dat de nadruk legde op het vinden van één 'general factor', bleek echter te eenvoudig, en werd vervangen door een 'Algemeen Factoren Model', dat identificatie van een groter aantal algemene factoren toelaat. Deze factoren worden telkens afgeleid van min of meer sterke correlaties tussen afzonderlijke variabelen.

Indien men bijvoorbeeld, uitgaande van de 'datamatrix' in *tabel 1*, een sterke positieve correlatie vindt tussen het gebruik van koffie, alcohol en sigaretten, zou men het bestaan van een gemeenschappelijke factor aannemelijk kunnen achten. Als aanduiding hiervoor ligt 'geneigdheid tot verslaving aan genotmiddelen' voor de hand. Deze factor zal – per definitie – tot uiting komen (een hoge waarde hebben) bij personen die een bepaald minimum aan (gecombineerd) gebruik van de genoemde middelen overschrijden. Er zal ook een groep zijn die laag scoort op deze factor, met als uiterste de groep die geen van de drie middelen gebruikt.

Men kan zich nu afvragen of andere waargenomen kenmerken (variabelen) in de onderzochte populatie samenhangen met de geïdentificeerde factor, bijvoorbeeld de variabelen leeftijd, consultfrequentie bij de huisarts, aantal recepten of aantal verwijskaarten. Daarnaast kan men op analoge wijze onderzoeken of er tussen de variabelen consultfrequentie, geneesmiddelengebruik en aantal verkregen verwijskaarten een duidelijke positieve correlatie bestaat, terwijl deze variabelen minder sterk of nauwelijks gecorreleerd zijn met het gebruik van de diverse genotmiddelen. Indien dit het geval is, kan een tweede factor worden gepostuleerd, die men bijvoorbeeld 'medische consumptie' kan noemen. Ook de samenhang van deze factor met elk der waargenomen variabelen kan worden bestudeerd, zelfs tegelijkertijd met de invloed van de eerste factor hierop. Het Algemeen Factoren Model geeft nu voor elke oorspronkelijk gemeten variabele apart de samenhang weer met elk der gevonden factoren.

Schematisch is dit als volgt weer te geven. Stel dat men een serie waarnemingen doet met betrekking tot n variabelen $X_1, X_2, X_3, \dots, X_n$, bij alle leden van een

onderzoekspopulatie, en op grond van correlaties tussen deze variabelen k niet-rechtstreeks meetbare gemeenschappelijke factoren $F_1, F_2, \dots, F_k$ benoemt (waarbij in het algemeen k aanzienlijk kleiner zal zijn dan n). Het Algemeen Factoren Model ziet er dan als volgt uit:

$$X_1 = \mu_1 + b_{11}F_1 + b_{12}F_2 + \dots + b_{1k}F_k + S_1$$

$$X_2 = \mu_2 + b_{21}F_1 + b_{22}F_2 + \dots + b_{2k}F_k + S_2$$

$$\begin{array}{ccccccc} | & | & \cdot & | & & | & | \\ | & | & & | & & | & | \end{array}$$

$$X_n = \mu_n + b_{n1}F_1 + b_{n2}F_2 + \dots + b_{nk}F_k + S_n$$

In woorden: nagegaan wordt in welke mate elke waargenomen variabele X_i – steeds uitgaande van haar gemiddelde μ_i – samenhangt met elk van de factoren. De waarden van elke F en S variëren per individu, en worden aangegeven op een 'gestandaardiseerde' (van eenheden ontdane) schaal met gemiddelde 0 en standaardafwijking 1. De waarden van de coëfficiënten b_{ij} zijn constanten, *factorladingen* genaamd, die staan voor de correlatie tussen X_i en F_j . Zij geven de relatieve belangrijkheid van de bijdrage van de gemeenschappelijke factoren aan de waarnemingen $X_1, X_2, \dots, X_n$ weer. De factorladingen kunnen waarden aannemen van -1 (maximale negatieve correlatie tussen X_i en F_j) tot $+1$ (maximale positieve correlatie tussen X_i en F_j). Een factorlading van (ongeveer) 0 wijst erop dat die factor geen samenhang vertoont met de betreffende variabele. S_i vertegenwoordigt de voor een bepaalde variabele X_i specifieke 'restinvloed', dat wil zeggen de invloed die niet voor rekening komt van de gemeenschappelijke factoren F .

De bedoeling is nu een zodanig model te vinden dat de daarin geïdentificeerde (combinatie van) factoren gezamenlijk een zo groot mogelijk deel van de verschillen in waarnemingen tussen individuen 'verklaren' (gesproken wordt van verklaarde variantie). In dat geval is men immers het meeste te weten gekomen over relevante achterliggende dimensies, en heeft men de meeste garantie weinig hiervan gemist te hebben. Overigens roept de term 'verklaar-

de variantie' weleens misverstanden op. Het gaat hier primair om een gevonden associatie, maar dit behoeft geen causale relatie te zijn. Hiervan mag men pas spreken indien aan bepaalde voorwaarden is voldaan, bijvoorbeeld met betrekking tot de volgorde van de gebeurtenissen in de tijd en de correctie voor versturende invloeden.

Om de gezamenlijke invloed van de factoren F (in termen van hierdoor verklaarde variantie) op variabele X_i uit te drukken, gebruikt men het begrip *communaliteit*. De communaliteit van de variabele X_i is de variantie die X_i met de andere variabelen deelt via de gemeenschappelijke factoren. Daartegenover staat de *uniciteit*: de variantie die niet door andere variabelen wordt gedeeld.

Uit het voorgaande is duidelijk, dat voor een goed factorenmodel de communaliteit groot (zo dicht mogelijk bij 1) en de uniciteit klein (zo dicht mogelijk bij 0) dient te zijn.

Een recent voorbeeld van factoranalyse vinden we in een nadere analyse van de gegevens van het Monitoringproject.⁵ *Schadé en Lamberts* construeerden vanuit een waargenomen basisset van 25 voor de huisartsgeneeskundige relevante variabelen zeven factoren: chronische ziekte, probleemgedrag, veel voorkomende diagnoses, kinderziekten, onzekere diagnoses, verwijzing en interdokter-variatie. De relatie van elk van deze zeven factoren met de 25 oorspronkelijke variabelen werd conform het Algemeen Factoren Model weergegeven, onder vermelding van de communaliteit per variabele (*tabel 2*).

Factoranalyse is een statistisch onderbouwde techniek. Het is mogelijk om hypothesen te toetsen wat betreft het aantal relevante factoren en de factorladingen. Betrouwbaarheidsintervallen kunnen worden berekend, en populaties kunnen worden vergeleken.

Principale-componenten-analyse

In tegenstelling tot factoranalyse postuleert de principale-componentenanalyse vooraf

Tabel 1 Voorbeeld van een datamatrix met een serie waarnemingen in een onderzoekspopulatie van 100 personen, als uitgangspunt voor factoranalyse.

Persoon (nummer)	Leeftijd	Variabelen					
		Sigaretten (per dag)	Koffie (koppen per dag)	Alcohol (glazen per dag)	Consult- frequentie per jaar	Prescripties per jaar	Verwijs- kaarten per 5 jaar
1	18	10	2	2	5	4	1
2	72	0	5	0	7	8	3
3	48	18	8	5	2	1	4
4	53	5	6	3	1	2	2
5	26	0	0	4	3	1	0
6	69	0	4	0	4	4	12
7	34	15	7	3	2	3	4
8	20	25	5	8	0	0	0
9	40	0	3	1	4	2	2
10	60	8	4	0	5	7	5
100	42	14	12	6	0	1	2

Tabel 2 Factoranalyse: model dat de samenhang aangeeft tussen de 25 gemeten variabelen en de benoemde 7 factoren.

Variables	1	2	3	4	5	6	7	Communalities
1. Encounters (abs)	.06	-.08	.93	-.06	-.03	-.19	-.12	0.93
2. 0- 4 year (%)	-.17	-.26	.06	.76	-.07	.03	.16	0.71
3. 5-14 year (%)	-.31	-.47	-.02	.69	.03	-.04	.06	0.80
4. 15-24 year (%)	-.54	-.52	-.10	-.10	.16	.27	-.09	0.69
5. 25-44 year (%)	-.64	.01	-.05	-.39	-.02	-.05	.34	0.68
6. 45-64 year (%)	.15	.57	.13	-.27	-.05	.04	-.35	0.56
7. 65+ years (%)	.83	.19	.01	-.01	-.03	.16	-.15	0.77
8. Sex (% male)	.27	-.11	-.01	.46	-.12	.21	-.21	0.48
9. Nation. insured (%)	.17	.14	-.01	-.10	-.13	-.02	-.09	0.08
10. Housecalls (%)	.53	.20	-.15	.48	.36	-.14	.03	0.72
11. Out of hours (%)	.03	.01	-.17	.67	.34	-.24	-.08	0.66
12. Defensive (%)	-.04	-.01	.08	.08	.70	.06	.02	0.51
13. No therapy (%)	-.06	.12	-.09	-.12	.33	.76	.20	0.73
14. Referral primary care (%)	.04	-.06	-.06	-.31	.59	.11	.32	0.57
15. Referral specialist (% of prevalence)	.05	-.19	-.18	-.01	.02	.80	-.01	0.71
16. Uncertain diagnosis (%)	.04	-.09	-.21	.13	.72	.13	-.24	0.66
17. New problem (%)	-.78	.02	.02	.30	.02	.14	-.21	0.76
18. Encounters per episode per year (abs)	.79	-.04	.19	-.21	.07	-.16	.25	0.80
19. Interdoctor variation per 1000 patients (abs)	.04	.19	.93	-.01	-.07	-.06	.07	0.92
20. Interdoctor variation (%)	.12	.38	-.04	.09	-.09	.22	.69	0.70
21. Concurrence psychologic problems (ratio)	.09	.82	.20	-.22	.01	-.08	.02	0.78
22. Concurrence social problems (ratio)	.06	.72	-.03	-.16	-.22	-.18	.12	0.64
23. Concurrence chronic diseases (ratio)	.82	.15	-.06	.02	-.05	.06	-.02	0.71
24. Utilization-encounters (t-value)	.60	.54	.09	-.12	.09	-.04	.22	0.73
25. Utilization-episodes (t-value)	.07	.80	-.12	.01	.12	.06	.22	0.73

Toelichting Factor 1 chronische ziekte; factor 2 probleemgedrag; factor 3 veel voorkomende diagnose; factor 4 kinderziekten; factor 5 onzekere diagnose; factor 6 verwijzen naar specialist; factor 7 interdokter-variatie.

geen onderliggende latente factoren die de correlaties tussen variabelen verklaren. Principale-componentenanalyse is in principe slechts bedoeld om de waargenomen n variabelen te transformeren tot een kleiner aantal nieuwe variabelen, zodanig dat deze niet van elkaar afhankelijk zijn en dat zo weinig mogelijk informatie verloren gaat. Met andere woorden: via principale-componentenanalyse zoekt men een aantal $k < n$ (lineaire) combinaties die gebruikt kunnen worden om data samen te vatten zonder informatie te verliezen (datareductie). Men kan zich hier iets bij voorstellen door te denken aan de Quetelet-index: een (overigens niet lineair) combinatie van lengte en gewicht, die voor allerlei doelen een zinvolle samenvatting is, zonder dat men denkt aan een onderliggende, latente dimensie.

Uitgangspunt is weer een datamatrix met gegevens over een aantal variabelen of scores ($X_1, X_2, \dots, X_n$) gemeten in een onderzoekspopulatie. In tegenstelling tot de werkwijze bij het Algemene Factoren Model, waarbij men de waargenomen X_i -waarden uitdrukt als resultaat van de geïdentificeerde factoren, gaat men nu in feite in omgekeerde richting te werk: getracht wordt om een beperkt aantal componenten te vinden als lineaire combinaties van de gemeten variabelen $X_1, X_2, \dots, X_n$. De k nieuw gedefinieerde factoren noemt men principale componenten ($Z_1, Z_2, \dots, Z_k$). Een en ander is weer te geven in formule-vorm:

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \dots + a_{1n}X_n$$

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \dots + a_{1n}X_n$$

$$\begin{vmatrix} | & | & | & | \\ | & | & | & | \\ | & | & | & | \\ | & | & | & | \end{vmatrix}$$

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \dots + a_{1n}X_n$$

Hierbij is a_{ij} de coëfficiënt die de invloed van variabele x_j op de principale component Z_i weergeeft.

De techniek is zodanig ontworpen dat die component die de grootste variantie heeft van alle mogelijke lineaire combinaties van de X -variabelen, het eerste aan bod komt. Dat houdt in dat deze component,

Z_1 , de meeste informatie bevat. Z_2 is de component die de grootste variantie heeft van alle niet met Z_1 gecorreleerde lineaire combinaties; Z_3 is de component die de grootste variantie heeft van alle niet aan Z_1 noch aan Z_2 gecorreleerde lineaire combinaties, enz. Men kan nu doorgaan tot de volgende component geen noemenswaardige variantie meer heeft, dat wil zeggen: geen informatie meer bevat.

Sommige mensen spreken ook bij principale-componentenanalyse van ladingen, namelijk van de componenten ten aanzien van de oorspronkelijke variabelen. Hiervoor worden meestal de correlatie-coëfficiënten tussen de geconstrueerde componenten en de oorspronkelijke variabelen gebruikt. Voor een relevante samenhang kan men een minimaal te behalen 'lading' vaststellen. Bij het gebruik van het begrip lading kan echter verwarring met factoranalyse optreden; daarom kan men beter de bedoelde correlatie omschrijven.

Ter illustratie van principale-componentenanalyse kan het onderzoek van Kleynen *e.a.* dienen.⁶ Uitgangspunt was een registratie van een groot aantal door huisartsen al dan niet verrichte handelingen met betrekking tot lichamelijk onderzoek. Het ging hierbij om 25 gemeten variabelen. De onderzoekers trachtten op basis hiervan, door middel van principale-componentenanalyse, een beperkt aantal handelingspatronen van huisartsen te identificeren. Voor de samenhang van componenten en variabelen werd een lading van .50 als minimum-eis gehanteerd.

Er konden vijf principale componenten worden onderscheiden; deze werden benoemd op grond van de aard van die samenstellende variabelen die aan de minimum-eis voldeden:

- 1 inspectie, palpatie en bewegingsonderzoek;
- 2 percussie en auscultatie;
- 3 testen en bewegingsonderzoek;
- 4 scopie;
- 5 kwantitatieve meting.

Evenals factoranalyse is principale-componentenanalyse een statistische techniek, die zich – gegeven de bijbehorende aannames – leent voor toetsing van hypothesen

en berekening van betrouwbaarheidsintervallen voor diverse grootheden. Principale-componentenanalyse wordt ook wel gebruikt als tussenstap bij het gebruik van andere technieken, bijvoorbeeld als er in het kader van regressieanalyse behoefte is aan het reduceren van een groot aantal variabelen tot een beperkt aantal van elkaar onafhankelijke variabelen.

Clusteranalyse

Door middel van clusteranalyse probeert men de onderzochte elementen (personen of variabelen) op een natuurlijke wijze in te delen in een beperkt aantal groepen of 'clusters', op een zodanige wijze dat de elementen in één groep op de een of andere manier op elkaar lijken, en wel meer dan op de elementen in de andere groepen. Er zijn dan ook twee criteria voor een goede clustering: een goede interne samenhang binnen een cluster, en een goede afgrenzing (isolatie) ten opzichte van andere clusters. De eisen die men wat dit betreft stelt, kunnen vooraf worden gespecificeerd.

Stel dat men van een grote groep huisartsen gegevens heeft over voorschrijfgedrag (al dan niet up-to-date voorschrijven), verwijsgedrag (verwijsfrequentie, verwijzing naar wie), de kwaliteit van het medisch handelen (onder meer eventuele bijzondere verrichtingen), consultvoering (de mate van gespreksvaardigheid), het gebruik van de sociale kaart, de praktijkinrichting en het eventueel bestaan van samenwerkingsrelaties. Via clusteranalyse kan nu bestudeerd worden of er, uitgaande van deze gegevens, subgroepen (clusters) van huisartsen kunnen worden onderscheiden, bijvoorbeeld meer medisch-technisch georiënteerden, psychosociaal georiënteerden en 'neutrale' huisartsen. Hierbij kan worden nagegaan in hoeverre er sprake is van 'hiërarchische' clustering, met subclusters. Zo zouden medisch-technisch georiënteerden (gesteld dat ze bestaan) nog onderscheiden kunnen worden in huisartsen die snel en veel naar de medisch specialist verwijzen, en huisartsen die zelf veel medisch-technische problemen afhandelen.

Aldus kan een zich steeds verder vertakende boom zichtbaar worden, waarbij de

eind-takjes in principe gevormd worden door de afzonderlijke individuen in de onderzochte groep. Uit een oogpunt van overzichtelijkheid, en vanwege de noodzaak van voldoende grote subgroepen voor een zinvolle clustering, blijft het aantal 'vertakkingen' meestal beperkt. In veel gevallen is het verstandig dit aantal van tevoren te bepalen.

De hiërarchische clustermethode is veel toegepast in de biologie. Men denke aan de plant- en diertaxonomie, waarbij elk object deel uitmaakt van een soort, een soort deel uitmaakt van een genus, en een genus weer onderdeel is van een familie.

Er bestaan diverse methoden voor clusteranalyse, waarop wij in dit verband niet verder zullen ingaan.⁷ Van belang is dat men zich realiseert dat verschillende methoden en verschillende criteria tot uiteenlopende wijzen van clustering kunnen leiden.

Clusteranalyse is, in tegenstelling tot de eerder besproken methoden, eigenlijk geen echte statistische techniek. Men is er namelijk nog niet in geslaagd om op een bevredigende manier systematische en toevallige effecten van elkaar te onderscheiden door middel van statistische toetsing of berekening van betrouwbaarheidsintervallen. Voordat men vertrouwen stelt in een bepaalde structuur, zal deze een- of meermalen geverifieerd moeten worden op onafhankelijke, vergelijkbare datasets.

Beschouwing

Duidelijk is geworden dat technieken die gebruikt worden om latente dimensies in kaart te brengen, niet alleen complexe relaties beschrijven, maar ook zelf zeer complex zijn. Men zal dan ook terdege rekening moeten houden met een aantal overwegingen.

In de eerste plaats dient onderscheid gemaakt te worden tussen drie, impliciet of expliciet, veel gebruikte toepassingen:

- Bevestiging van het bestaan van – op grond van deskundigheid – a priori verwachte doch niet rechtstreeks meetbare dimensies of samenhangen, en het eventueel nader beschrijven of kwantificeren ervan: hypothesen-toetsende, respectievelijk kwantificerende analyse (confirmatieve

factoranalyse). In dit eerste geval wordt gericht gezocht en wordt een beperkt aantal vooraf geformuleerde opties nagegaan. Dit is een zinvolle toepassing, waardoor theorie-ontwikkeling kan worden ondersteund.

- Opsporing van vooraf niet bekende, noch vooronderstelde dimensies: exploratieve analyse. In dit tweede geval worden allerlei – denkbare en ondenkbare – samenhangen afgetast. Men ziet dan vanzelf wel wat er uitkomt. De kans is dan groot dat men in feite toevallige, irrelevante c.q. 'fout-positieve' samenhangen vindt, waaraan men achteraf spitsvondige betekenissen meent te kunnen verbinden. De zo gewillige computer is in dit verband een gevaarlijk instrument. Een parallel dringt zich op met het toepassen van allerlei screenende diagnostiek bij een lage a priori kans op ziekte, met een groot risico op fout-positieve uitkomsten en dus overschatting van de waarde van screening.⁸ Een 'blinde', dat wil zeggen, vooraf niet inhoudelijk onderbouwde en geplande exploratieve analyse moet dan ook worden afgeraden.

- Het samenvatten van informatie van een groot aantal afzonderlijke variabelen in een beperkt aantal overzichtelijke, samengepaste dimensies: datareductie. Ook bij deze toepassing geldt dat men toevallige, irrelevante ordeningen moet vermijden, zeker indien informatieverlies optreedt. Anders doet men niet alleen aan datareductie, maar ook aan 'inhoudsreductie'.

Op de achtergrond van deze overweging speelt het probleem dat er bij deze vormen van analyse in het algemeen geen meetbare 'standaard' beschikbaar is, waartegen men gevonden resultaten (bijvoorbeeld het gevonden factorenmodel) kan afzetten. Bij onderzoek naar de determinanten van hoge bloeddruk kan men de consequenties van het gevonden regressiemodel steeds vergelijken met de aangetroffen reële waarden. Voor een concept als 'verslavingsgeneigdheid' bestaat zo'n standaard niet, en is men uiteindelijk aangewezen op de geloofwaardigheid van het gevondene. Bij de beoordeling hiervan zijn subjectieve elementen niet te vermijden, en is er sprake van zogenaamde 'face validity'.

Een extra probleem is, dat de besproken

technieken in vergelijkbare situaties tot verschillende oplossingen kunnen leiden, mede afhankelijk van de gekozen criteria (bijvoorbeeld aantal toegelaten factoren, componenten of clusters, minimaal vereiste correlatie, etc.). Relevant in deze context is het probleem van de 'identificeerbaarheid' van een model: zo er een oplossing is van het model, is er dan ook maar één oplossing of zijn er twee of meer niet (essentieel) van elkaar verschillende oplossingen? Met name indien er diverse, essentieel verschillende, oplossingen zijn, is niet duidelijk hoe de resultaten van een analyse geïnterpreteerd moeten worden. Maar ook als er één oplossing gevonden is, kan men er nooit geheel zeker van zijn dat die ook 'de juiste' is.

Een manier om toch enigszins aan de genoemde problemen tegemoet te komen, is het herhalen van het onderzoek en de analyse in verschillende situaties, teneinde na te gaan hoe stabiel een oplossing is. Dit kan meer inzicht verschaffen in het bestudeerde 'mechanisme' en de consistentie van een model.

In de regel zal een model aanzienlijk minder factoren, componenten of clusters (moeten) omvatten dan het aantal oorspronkelijk gemeten variabelen of het aantal te clusteren personen. Als dit niet het geval zou zijn, zou men immers weinig zijn opgeschoten en rijst de vraag of de situatie misschien niet juist ingewikkelder is gemaakt.

Omdat een complex van relaties wordt bestudeerd zal een zinvolle analyse in het algemeen een vrij grote onderzoekspopulatie vergen. Hoe groot precies, is afhankelijk van de specifieke eisen per situatie. Als de onderzoeksgroep te klein is, zal de analyse een te geringe 'power' hebben, waardoor relevante verbanden niet aangetoond kunnen worden. Als de populatie extreem groot is, moet men weer meer bedacht zijn op het vinden van weinig praktijkrelevante samenhangen.⁹

Ook moet aandacht worden besteed – en dat gebeurt vaak niet – aan de zogenaamde interne en externe validiteit van de analyse-resultaten: is het gevonden model niet een resultaat van een vertekende waarneming of een niet-valide onderzoekopzet? Speel-

den *selectiebias* of verstorende variabelen een mogelijk relevante rol? Soms leiden dergelijke vragen ertoe, dat een – op basis van de verzamelde data – goed passend model toch niet aanvaard kan worden. Deze vragen zijn achteraf zelden bevredigend te beantwoorden als men vooraf, bij het plannen van onderzoek en dataverzameling, onvoldoende voorzorgsmaatregelen heeft getroffen. Dit geldt niet alleen indien causale relaties worden bestudeerd, maar ook als andere – bijvoorbeeld diagnostische – verbanden aan de orde zijn. Hetgeen hierover besproken is in eerdere afleveringen, gaat ook op dit gebied onverkort op.¹⁰

Een apart hoofdstuk vormen de aannames die ten grondslag liggen aan de analysemodellen. Bij het Algemene Factoren Model en de principale-componentenanalyse wordt er bijvoorbeeld van uitgegaan dat de bestudeerde variabelen normaal verdeeld zijn. Voorts zijn de modellen in de regel lineair, en worden niet-lineaire relaties niet (gemakkelijk) ontdekt. Vaak zullen deze aannames niet opgaan of niet getoetst kunnen worden.

Samenvattend kunnen we stellen dat Factoranalyse mogelijkheden biedt om op het eerste gezicht latente dimensies en ordeningen in kaart te brengen, terwijl principale-componentenanalyse en clusteranalyse

in de eerste plaats bedoeld zijn voor datareductie. Degenen die deze theoretisch en qua toepassing zeer complexe en soms verraderlijke technieken zelf willen gebruiken, wordt aangeraden zich terdege voor te bereiden en bij voorkeur een professionele statisticus met ervaring in het gebruik van deze methoden te consulteren.^{7 11 12} Men dient deze technieken in ieder geval toe te passen uitgaande van een vooraf inhoudelijk en data-analytisch goed onderbouwde aanpak. Blind screenen van data bestanden levert niet goed interpreteerbare uitkomsten op. Daarnaast is er een aantal methodologische en statistische vraagstukken waarvan men zich terdege bewust moet zijn, en waarop men zo goed mogelijk moet anticiperen. Men bedenke dat het niet alleen gaat om het toepassen van analytische technieken of het aanvragen van het juiste computerprogramma. Men is bezig met de laatste fase van een onderzoek, waarvan ook een verantwoorde probleemstelling, studie-opzet en gegevensverzameling deel uitmaken.

Literatuur

- ¹ Van Eijk JThM, Gubbels JW Wetenschappelijk onderzoek in de huisartsgeneeskunde. Lelystad: Meditekst, 1987.
- ² Knottnerus JA, Volovics A. Het analyseren van de invloed van meer dan één variabele. Huisarts Wet 1988; 31: 207-11.
- ³ Mekkink H. Ziekenfondscijfers als parameters voor het handelen van huisartsen [Dissertatie]. Nijmegen: Katholieke Universiteit Nijmegen, 1986.
- ⁴ Spearman C. General intelligence objectively determined and measured. Am J Psychol 1904; 15: 201-93.
- ⁵ Schadé E, Lamberts H. Het clusteren van diagnoses. 2. Een andere kijk op huisartsgeneeskundig handelen. Huisarts Wet 1988; 31: 196-201.
- ⁶ Kleijnen JGVM, Metsemakers JFM, Van der Horst FG. Het lichamelijk onderzoek door de huisarts. Huisarts Wet 1986; 29: 201-3.
- ⁷ Mardia KV, Kent JT, Bibby JM. Multivariate analysis. London: Academic Press, 1979.
- ⁸ Knottnerus JA, Volovics A. Het onderscheidend vermogen van diagnostische tests. Huisarts Wet 1989; 32: 338-46.
- ⁹ Knottnerus JA, Volovics A. Statistisch toetsen. Huisarts Wet 1988; 31: 100-5.
- ¹⁰ Knottnerus JA, Volovics A. Over validiteit, vergelijkbaarheid en verstorende variabelen. Huisarts Wet 1988; 31: 135-140.
- ¹¹ Armitage P, Berry G. Statistical methods in medical research. 2nd ed. London: Blackwell, 1987.
- ¹² Van Knippenberg A, Siero F. Multivariate analyse. Beknopte inleiding en toepassingen. Deventer: Van Loghum Slaterus, 1980. ■