Geneesmiddelenbulletin

Redactie-adres: Postbus 5406, 2280 HK Rijswijk (ZH), telefoon 070-407007. Abonnementen: telefoon 070-406477

Jaargang 21, nrs 14 en 15 **19 december 1987**

LEZEN TUSSEN DE REGELS VAN GERAPPORTEERD GENEESMIDDELENONDERZOEK*

INLEIDING

De medicus practicus wordt in toenemende mate geconfronteerd met resultaten van geneesmiddelenonderzoek; soms wordt hem ook gevraagd daaraan mee te doen. Bij het nemen van therapeutische beslissingen mag dan ook van hem worden verwacht dat feitelijke kennis over de werkzaamheid van geneesmiddelen een belangrijke rol speelt. Het is daarom van belang dat hij therapeutische experimenten ('clinical trials') goed begrijpt. Dit betreft vooral:
- de vraagstelling en de relevantie ervan voor de praktijk;
- de opzet van het onderzoek;
- de interpretatie van de gevonden resultaten.

Hij kan er niet mee volstaan conclusies uit tijdschriften, folders of voordrachten klakkeloos over te nemen. Steeds is het weer nodig de vraagstelling, de opzet van het onderzoek en de resultaten aan een kritische beschouwing te onderwerpen. Tevens dient hij zich een oordeel te vormen over de klinische relevantie zowel van de vraagstelling als van de gevonden effecten. Dat dit alles nodig is blijkt uit de vele publikaties (ook in vooraanstaande tijdschriften) van onderzoekingen die fouten vertonen in de opzet en analyse waardoor de conclusies ervan onbetrouwbaar en soms zelfs misleidend kunnen zijn.
Dit artikel is bedoeld als leidraad voor het beoordelen van publikaties over geneesmiddelenonderzoek. Tevens geeft het een overzicht van een aantal voorwaarden waaraan een goed geneesmiddelenonderzoek moet voldoen. Reeds eerder verscheen in het Geneesmiddelenbulletin een artikel over dit onderwerp getiteld 'Geneesmiddelenonderzoek in de huisartsenpraktijk' (Gebu 1980; 14: nr 4).
Dit artikel bestaat uit twee delen. In het eerste gedeelte wordt de onderzoeksmethodiek aan een beschouwing onderworpen. In het tweede gedeelte valt het accent op de interpretatie van de bevindingen van een onderzoek.

DEEL I

§ 1. DE COMPONENTEN VAN EEN BEHANDELINGSEFFECT

Wanneer een huisarts een patiënt die over hoofdpijn klaagt een geneesmiddel geeft en de pijn gaat over, dan is het verleidelijk te zeggen dat de pijn bij deze patiënt, dankzij een farmacologisch effect van het middel, is overgegaan. Met andere woorden, de pijn zou zijn overgegaan als gevolg van een specifiek *therapiegebonden effect*. Zo eenvoudig ligt het echter niet. Immers, het is mogelijk dat bij deze patiënt de pijn vanzelf zou zijn overgegaan, ook als het middel niet was gegeven, dat wil zeggen dat het te danken is geweest aan het *natuurlijke beloop* dat de pijn is verdwenen.
Het is ook mogelijk dat de patiënt alleen al door het ceremonieel van het nemen van een geneesmiddel zichzelf zodanig heeft beïnvloed dat de hoofdpijn juist daardoor is overgegaan (autosuggestie door het medisch ritueel en het daarin gestelde vertrouwen). Daarnaast is het mogelijk dat de arts de patiënt behalve het geneesmiddel ook rust heeft voorgeschreven waardoor de hoofdpijn is verdwenen. Verder kan men zich voorstellen dat een goed contact met een begripsvolle arts en het vertrouwen dat deze inboezemt de hoofdpijn doet verdwijnen ('the doctor as a drug'). In deze gevallen is er sprake van een *externe factor* (autosuggestie, rust, vertrouwen) die ervoor zorgt dat de hoofdpijn verdwijnt. Tenslotte is het mogelijk dat de patiënt (uit ontzag voor zijn dokter) alleen maar zegt dat de hoofdpijn over is of dat de arts (uit enthousiasme voor het gegeven middel) niet wil horen dat de hoofdpijn niet echt is verdwenen. In beide gevallen is er sprake van *foute informatie* over het ziektebeloop, dat wil zeggen over het verdwijnen van de hoofdpijn.
Wanneer de patiënt is genezen, kan men dat slechts achteraf constateren. Men kan echter niet vaststellen welke van de vier genoemde oor-

*Dr J.G.P. Tijssen, prof. dr E. van der Does en prof. dr J. Lubsen, onder medeverantwoordelijkheid van de Redactiecommissie

zaken daarvoor verantwoordelijk is geweest. In het waarneembare ziektebeloop zijn deze factoren namelijk niet van elkaar te onderscheiden. De individuele patiënt zal dat ook een zorg zijn, maar de arts behoort daar anders tegenover te staan: hij zal wel degelijk willen weten hoe groot het werkelijke therapiegebonden effect is bij zijn patiënten met de desbetreffende ziekte. Het bepalen daarvan is het doel van een geneesmiddelenonderzoek.

§ 2. NIET-VERGELIJKEND ONDERZOEK

Wanneer men een groep patiënten met hoofdpijn behandelt met een geneesmiddel, kan men waarnemen dat de pijn bij een deel van de patiënten overgaat. Men zou dit het waargenomen beloop kunnen noemen.
Zoals bij één patiënt niet kan worden uitgemaakt welke oorzaak voor zijn genezing verantwoordelijk was, kan ook bij een groep genezen patiënten die op dezelfde manier zijn behandeld niet zonder meer worden vastgesteld hoe groot elk van de vier genoemde samenstellende delen is van het waargenomen beloop.

In een zogenaamd *niet-vergelijkend onderzoek* wordt slechts één groep patiënten met een te onderzoeken geneesmiddel behandeld. Wanneer na een bepaalde tijd het ziektebeloop wordt beoordeeld, is het verleidelijk het waargenomen effect voor te stellen als therapiegebonden effect. Meestal is het resultaat van het *natuurlijke beloop* aanzienlijk groter dan het therapiegebonden effect; bovengenoemde voorstelling geeft dus een verkeerd beeld. Niettemin worden dergelijke onderzoekingen vaak uitgevoerd; deze zijn dan bedoeld om de deelnemende huisartsen met een middel vertrouwd te maken en/of de 'marketing'-afdeling van de farmaceutische industrie van materiaal voor reclamefolders te voorzien. Bij voorkeur worden ze verricht bij veel voorkomende ziekten met een goed of wisselend (niet-fataal) natuurlijk beloop. De folders suggereren dat het verdwijnen van de ziekte een gevolg is van het medisch handelen. Ten onrechte wordt op deze wijze het natuurlijke beloop voorgesteld als een therapiegebonden effect.
Soms zijn inzicht in de pathogenese en enkele getallen reeds voldoende om de gewekte indruk te ontmaskeren.

> Wanneer bijvoorbeeld in een niet-vergelijkend onderzoek op grond van een klacht over een zere keel na keelinspectie een antibioticum wordt gegeven, zal de keelpijn in 90% van de gevallen binnen enkele dagen verdwijnen. Het is bekend dat 70% van de 'zere kelen' berust op een virale infectie, die op klinische gronden vrijwel niet is te onderscheiden van een bacteriële infectie, maar waarbij het toedienen van antibiotica geen nut heeft. Temeer omdat zowel een virale als een bacteriële 'zere keel' na 5-7 dagen spontaan verdwijnt kan men stellen dat vrijwel alle genezingen eenvoudigweg tot stand kwamen via het natuurlijke beloop en niet het gevolg kunnen zijn van antibiotische therapie.

Een ander fenomeen waarop dit soort onderzoekingen in dit verband vaak inspeelt is dat van de '*regressie naar het gemiddelde*'. Dit fenomeen berust op het feit dat bepaalde grootheden een hoge mate van variabiliteit vertonen. De bloeddruk is daarvan een goed voorbeeld.

> Het is bekend dat niet alleen tussen personen maar ook bij één persoon de bloeddruk sterk varieert. Wanneer de huisarts bij een patiënt bijvoorbeeld voor het eerst een te hoge diastolische bloeddruk vaststelt, is dat mogelijkerwijs het gevolg van het feit dat de patiënt zich die ochtend erg heeft opgewonden. De diastolische bloeddruk ligt dan 'toevallig' wat hoger dan gebruikelijk. Het is waarschijnlijk dat deze bij hermeting een week later lager is. Na een relatief hoge waarde heeft de uitslag van de volgende meting dus de neiging terug te keren naar zijn gemiddelde, welk fenomeen wordt aangeduid als 'regressie naar het gemiddelde'. In een niet-vergelijkend onderzoek kan hierop worden ingespeeld door na die ene hoge uitslag een geneesmiddel te geven en op grond van de tweede uitslag, die in negen van de tien gevallen lager uitvalt, een bloeddrukverlagende werking aan het geneesmiddel toe te schrijven. Bij bloeddrukmetingen is dit fenomeen wel enigszins bekend; op grond van een eenmaal gevonden verhoogde waarde zal een huisarts dan ook geen bloeddrukverlagend middel voorschrijven.
> Ook in andere situaties speelt regressie naar het gemiddelde in het natuurlijke beloop van een ziekte een belangrijke rol. Een patiënt die zich bij zijn huisarts meldt in verband met klachten van depressiviteit, doet dit in de regel op een moment waarop de depressie relatief ernstig is. Door de min of meer extreme graad van depressie heeft deze eerder de neiging af dan toe te nemen. Wanneer onder deze omstandigheden een antidepressivum wordt gegeven, wordt het waargenomen effect in belangrijke mate bepaald door het natuurlijke beloop. Het is dan ook misleidend deze 'regressie naar het gemiddelde' toe te schrijven aan de werkzaamheid van het geneesmiddel.

Kortom, uit een niet-vergelijkend onderzoek kunnen in het algemeen geen conclusies worden getrokken over het therapiegebonden effect. Er zijn uiteraard uitzonderingen: bijvoorbeeld is het effect van penicilline bij bacteriële endocarditis destijds niet gevonden via vergelijkend onderzoek, maar zijn bacteriologische kennis en klinische ervaring hiervoor voldoende geweest.

§ 3. VERGELIJKEND ONDERZOEK

Door in een onderzoek een tweede groep patiënten te betrekken kan wèl het therapiegebonden effect worden bepaald. Een dergelijk onderzoek heet een vergelijkend onderzoek. Een vergelijkend onderzoek dient echter aan bepaalde voorwaarden te voldoen.
Daartoe dienen in eerste instantie de volgende vier vragen te worden gesteld:
1. Wat werd onderzocht?
2. Waarom werd onderzocht?
3. Hoe werd onderzocht?
4. Wat werd gevonden?

De eerste vraag 'wat werd onderzocht' betreft de vraagstelling van het onderzoek. Om deze goed tot zich te laten doordringen moet men de publikatie vaak zeer grondig lezen. In § 4 zal worden geprobeerd een structuur aan te brengen die bij het ontrafelen van de vraagstelling behulpzaam kan zijn. In het verlengde hiervan ligt de tweede vraag, namelijk die naar de motivatie voor het onderzoek. Immers, inzicht in de motivatie van de

onderzoekers leidt tot een beter begrip van de vraagstelling.
De derde vraag 'hoe werd onderzocht' betreft de gevolgde onderzoeksmethode. De aanwezigheid van een vergelijkingsgroep brengt niet automatisch met zich dat dan ook het therapiegebonden effect kan worden bepaald: alleen als de gevolgde methodiek correct is, kan dit zonder vertekening worden geschat; het onderzoek heet dan *intern valide*. Bij een foute onderzoeksmethodiek wordt van het therapiegebonden effect een vertekend beeld verkregen, waardoor niet-werkzame middelen werkzaam kunnen lijken en omgekeerd. De criteria waaraan een onderzoek moet voldoen om intern valide te zijn, worden behandeld in § 6.
De vierde vraag 'wat werd gevonden' betreft de resultaten van het onderzoek. Om deze vraag te kunnen beantwoorden is het nodig dat de uitkomsten op adequate wijze zijn weergegeven. Vaak is de hoeveelheid relevante informatie echter omgekeerd evenredig aan de hoeveelheid cijfermateriaal. In § 7 wordt aangegeven wat hiervan essentieel is. In het verlengde ligt de vraag: wat hebben de onderzoekers uit de bevindingen geconcludeerd? Vaak wordt geen onderscheid gemaakt tussen de resultaten en de conclusies van de auteurs.
Resultaten zijn evenwel objectief en de conclusies subjectief. *Conclusies zijn gebaseerd zowel op de feitelijke resultaten als op de (subjectieve) opvattingen van de onderzoekers voordat het onderzoek werd verricht.* Deze voordien aanwezige opvattingen kunnen dan zijn gebaseerd op eerder verricht onderzoek, resultaten van dierproeven, farmacologisch inzicht en klinische ervaring.

Tenslotte is het van belang dat de lezer zich een oordeel vormt over de klinische relevantie zowel van de vraagstelling als van de grootte van het therapiegebonden effect. De mate waarin de bevindingen uit een intern valide vergelijkend onderzoek van toepassing zijn voor de patiënten uit de eigen praktijk, wordt aangeduid met de termen *externe validiteit* of *generaliseerbaarheid*. Het spreekt voor zich zelf dat een onderzoek waarvan de onderzoeksmethodiek incorrect is, geen implicaties mee zal brengen voor de behandeling van de eigen patiënten. Met andere woorden: *interne validiteit is een absolute voorwaarde voor externe validiteit. Echter: interne validiteit impliceert nog geen externe validiteit.* Immers, een intern valide onderzoek met patiënten voor wie de criteria van de ziekte vaag of in het geheel niet zijn gedefinieerd, heeft geen externe validiteit. Externe validiteit wordt verder besproken in § 9.

§ 4. DE VRAAGSTELLING VAN EEN GENEESMIDDELENONDERZOEK

De vraagstelling van een geneesmiddelenonderzoek heeft drie kernelementen, te weten de ziekte en de betrokken patiëntengroep, de behandeling en het ziektebeloop. Deze dienen allereerst door de lezer te worden geïdentificeerd.

§ 4.1. Ziekte en patiëntengroep
Bij de definitie van de ziekte en de patiëntengroep is het in de eerste plaats van belang welke diagnostische criteria zijn gebruikt. In de tweede plaats is het belangrijk te weten op welke wijze patiënten voor het onderzoek werden geselecteerd.
Wanneer bijvoorbeeld in een onderzoek betreffende de behandeling van hypertensie een diastolische druk groter dan 95 mmHg als criterium is aangenomen kunnen de patiënten op verschillende manieren zijn geselecteerd. Er kunnen patiënten in het onderzoek zijn betrokken die niet bekend waren met hypertensie en bij wie tijdens een bezoek aan de huisarts de bloeddruk werd gemeten. Het is ook mogelijk dat patiënten voor het onderzoek zijn gekozen die reeds werden behandeld voor hypertensie. Als derde mogelijkheid zouden reeds behandelde patiënten kunnen zijn toegelaten nadat hun medicatie eerst enige tijd is gestaakt. Het is duidelijk dat elk van de drie manieren van selecteren een andere patiëntengroep oplevert. De toepasbaarheid en de interpretatie van de bevindingen hangen daarvan af.
Hiermee is de definitie van de patiëntengroep nog niet volledig. Een groep patiënten met gemiddeld een diastolische bloeddruk van 100 mmHg aan het begin van het onderzoek bijvoorbeeld representeert een minder ernstige vorm van hypertensie dan een groep met gemiddeld 130 mmHg.
Een profiel van de patiënten die feitelijk tot het onderzoek zijn toegelaten, vormt de afronding van de beschrijving van de ziekte. Hierop wordt teruggekomen in § 7

§ 4.2. Behandeling
De behandelingen die in een onderzoek met elkaar worden vergeleken, vervullen geen gelijkwaardige rol. In vrijwel ieder vergelijkend onderzoek gaat de expliciete belangstelling uit naar één van de twee behandelingen terwijl de andere geldt als referentiepunt. De behandeling met het te onderzoeken middel heet de *indexbehandeling*, de behandeling waarmee wordt vergeleken heet de *referentiebehandeling*. De betrokken patiëntengroepen heten de *indexgroep* en de *referentiegroep*. Deze laatste wordt ook de controlegroep genoemd. Bij de keuze van de referentiebehandeling zijn er vele mogelijkheden, bijvoorbeeld behandeling met placebo, behandeling met een veel gebruikt middel dan wel het achterwege laten van enige vorm van behandeling.
De keuze van de referentiebehandeling hangt af van de vraagstelling die de onderzoekers voor ogen hebben. Indien men is geïnteresseerd in de specifieke werkzaamheid van de chemische substantie is placebo de te kiezen referentie. Indien de belangstelling uitgaat naar het totale effect van de behandeling (incl. placebo-effecten) worden de patiënten uit de referentiegroep onbehandeld gelaten. Indien men is geïnteresseerd in het verschil tussen een nieuw middel en een veel

gebruikt standaardmiddel wordt dit laatste gekozen als referentie.
Ook de indexbehandeling brengt vele keuzemogelijkheden met zich, bijvoorbeeld ten aanzien van de toedieningsvorm en de (frequentie van) dosering.

§ 4.3. Ziektebeloop
Het ziektebeloop van de patiënt wordt zowel door somatische factoren als door psychische en sociale factoren bepaald. Vooralsnog is het gebruikelijk het ziektebeloop vrijwel uitsluitend in somatische termen te definiëren. Ter vereenvoudiging concentreert men zich in een onderzoek op één of meer parameters die eenvoudig zijn vast te stellen. Deze parameters worden ook wel *uitkomstvariabelen* genoemd. We onderscheiden twee categorieën:
1. waarbij voor elke patiënt de uitkomst wordt omschreven door het al dan niet optreden van een bepaalde gebeurtenis, bijvoorbeeld het verdwijnen van de hoofdpijn, het optreden van een bloeding of het overlijden van de patiënt;
2. waarbij de uitkomsten na een bepaalde periode kwantitatief worden vastgelegd, bijvoorbeeld de diastolische bloeddruk, depressiviteitsscore volgens een bepaalde schaal, of levensverwachting.

§ 5. DE INTERPRETATIE VAN DE FEITELIJK ONDERZOCHTE VRAAGSTELLING

Het is van zeer groot belang zich af te vragen of de feitelijk onderzochte vraagstelling dezelfde is als die welke de onderzoekers voor ogen hebben gehad. Dat deze soms moeilijk is te achterhalen moge blijken uit de volgende voorbeelden.

Een eerste voorbeeld betreft een onderzoek van de (Engelse) Medical Research Council (MRC) waaraan 3000 huisartsen hebben meegewerkt.[1] [2] Aan 1249 patiënten bij wie door de huisarts een vers hartinfarct werd vermoed, werd thuis eenmaal 300 mg acetylsalicylzuur (ASA) gegeven en aan 1281 patiënten een placebo. De patiënten werden vervolgens op een hartbewakingsafdeling opgenomen; nagegaan werd of ze na 28 dagen nog in leven waren.
De met ASA behandelde patiënten vormden dus de indexgroep, de patiënten die placebo ontvingen de referentiegroep. Het onderzoek was gericht op de specifieke werking van de chemische substantie ASA. De uitkomstvariabele was 'dood binnen 28 dagen'. Het percentage patiënten dat binnen 28 dagen overleed kan worden geïnterpreteerd als het sterfterisico bij een enkele behandeling met ASA respectievelijk met placebo.
Het onderzoek betreft de vraag in welke mate een eenmaal verstrekte gift van ASA het sterfterisico (binnen 28 dagen) verlaagt bij patiënten *bij wie de huisarts een vers hartinfarct heeft vermoed.* Het onderzoek 'meet' dus het relatieve sterfterisico bij een behandeling met 300 mg ASA ten opzichte van placebo bij *deze* groep patiënten.
De vraagstelling van het onderzoek is dus niet, zoals men wellicht is geneigd te denken, 'het nut van ASA bij patiënten met een bewezen hartinfarct'. De onderzoekers hebben deze vraagstelling zelf ook fout geïnterpreteerd. In hun publikatie zijn namelijk slechts de uitkomsten gerapporteerd van patiënten die achteraf een hartinfarct bleken te hebben. Met andere woorden: ook zij zijn ervan uitgegaan dat het onderzoek patiënten betreft met een bewezen infarct. De ziekte waar het hier om gaat is een *vermoed vers hartinfarct* zoals dat op grond van anamnese en lichamelijk onderzoek door de huisarts is vastgesteld, zonder dat deze de beschikking had over ECG's of enzymwaarden. De huisarts die met een vergelijkbare patiënt wordt geconfronteerd, is ook per definitie geïnteresseerd in het effect van ASA bij de patiënt die hij op dat moment wegens een *vermoed* vers hartinfart behandelt, ongeacht of deze nu achteraf wel of geen vers infarct blijkt te hebben. De onderzoekers hadden dan ook de uitkomsten van *alle* patiënten moeten rapporteren.

Een tweede voorbeeld betreft het '60-plus'-onderzoek dat in de jaren zeventig in Nederland werd uitgevoerd bij 878 oudere patiënten die wegens een doorgemaakt hartinfarct reeds jarenlang onder behandeling waren met orale anticoagulantia.[3] Van deze patiënten werd een deel doorbehandeld met anticoagulantia en een ander deel ontving placebo. In het verloop van twee jaar werd nagegaan hoeveel patiënten in elk van beide groepen waren overleden, dan wel een reïnfarct of een cerebrovasculair accident hadden gekregen.
Bij oppervlakkige lezing zou men misschien menen dat de vraagstelling van het onderzoek was na te gaan of het langdurig geven van antistolling leidt tot een betere tweejaarsoverleving bij patiënten die een hartinfarct hadden doorgemaakt. Nauwkeurige lezing zowel van de selectieprocedure op zich zelf als van het resultaat ervan (d.w.z. het profiel v.d. toegelaten patiënten) laat zien dat àlle in het onderzoek opgenomen patiënten reeds lange tijd (gem. 6 jaar) bij een trombosedienst onder behandeling met anticoagulantia waren. De vraagstelling betreft daarom het staken van langdurige antistollingsbehandeling en niet de instelling ervan. Men kan zelfs stellen dat vervanging van de standaardbehandeling door placebo de indextherapie en voortgezette antistolling de referentietherapie is. De juiste perceptie van de feitelijke vraagstelling heeft belangrijke consequenties voor de extrapolatie van de onderzoeksbevindingen naar de dagelijkse praktijk. *Immers, de bevinding dat staken van een langdurig gegeven therapie schadelijk is, impliceert niet zonder meer dat instelling van die therapie nuttig is.* Het is namelijk zeer wel mogelijk dat na zes jaar therapie alleen die patiënten die voordeel bij de behandeling hadden, over zijn. Het ligt dan voor de hand dat het staken van de therapie voor deze patiënten nadelige gevolgen heeft. Het onderzoek zoals uitgevoerd zegt niets over het nuttige effect van langdurige antistollingsbehandeling direct na het hartinfarct bij ontslag uit het ziekenhuis.

§ 6. INTERNE VALIDITEIT

Een vergelijkend geneesmiddelenonderzoek heet, zoals vermeld, intern valide indien de methodiek dusdanig is dat uit de onderzoeksresultaten een goede schatting van het therapiegebonden effect kan worden verkregen. Alleen in aanwezigheid van een referentiegroep is het mogelijk het therapiegebonden effect van de indexbehandeling te schatten, mits aan bepaalde voorwaarden is voldaan.
Het waargenomen beloop in de indexgroep (WB_1) is samengesteld uit het therapiegebonden effect (TGE), het natuurlijke beloop (NB_1), externe factoren (EF_1) en waarnemingsfouten (WF_1).
Het waargenomen beloop in de referentiegroep (WB_0) bestaat uit het natuurlijke beloop (NB_0), externe factoren (EF_0) en de waarnemingsfouten (WF_0); het therapiegebonden effect is daarbij uiteraard afwezig. De vergelijking tussen de indexgroep en de referentiegroep kan dan worden weergegeven als:

$$WB_1 = TGE + NB_1 + EF_1 + WF_1 \text{ (indexgroep)}$$
$$WB_0 = NB_0 + EF_0 + WF_0 \text{ (referentiegroep)}$$
$$WB_1\text{-}WB_0 = TGE + (NB_1\text{-}NB_0) + (EF_1\text{-}EF_0) + (WF_1\text{-}WF_0)$$

Hieruit blijkt dat de vergelijking tussen de index- en referentiegroep (WB_1-WB_0) mag worden geacht het therapiegebonden effect (TGE) te representeren indien de andere componenten van het waargenomen beloop in beide groepen van gelijke invloed zijn. Met het *randomiseren* wordt beoogd de invloed van het natuurlijke beloop in beide groepen gelijk te schakelen. Door een *placebo* te geven wordt datzelfde nagestreefd voor de externe factoren. Door het *blinderen* van de waarnemer worden de waarnemingsfouten gelijkgemaakt.
Samenvattend: voorwaarden voor interne validiteit zijn dat het natuurlijke beloop, de invloed van externe factoren en de wijze van informatieverzamelen in beide groepen gelijk zijn.

§ 6.1. Vergelijkbaarheid van het natuurlijke beloop
Om het natuurlijke beloop in de te behandelen groepen zo goed mogelijk vergelijkbaar te maken past men randomisatie ('at random' = volgens het toeval) toe. Dit betekent dat door middel van een loting wordt beslist in welke van de te vergelijken behandelingsgroep de patiënten worden ingedeeld. Het kan echter gebeuren dat ondanks randomisatie toch verschillen in het natuurlijke beloop ontstaan. Bekende risicofactoren voor de uitkomstvariabele dienen over de groepen gelijk te zijn verdeeld. In een onderzoek van bijvoorbeeld postinfarctpatiënten met als uitkomstvariabele 'dood binnen twee jaar', is het van belang dat het aantal patiënten dat reeds eerder een infarct heeft doorgemaakt, gelijk over de twee groepen is verdeeld. Deze patiënten hebben namelijk een relatief hoog risico.
Indien ondanks randomisatie ten aanzien van een risicofactor een belangrijke onvergelijkbaarheid is opgetreden, kan hiervoor met een statistische methode (stratificatie of multivariate-analyse) worden gecorrigeerd. Onvergelijkbaarheid doet zich vooral voor bij relatief kleine onderzoekingen. Bij grote tot zeer grote onderzoekingen is de kans op een ernstige onevenwichtigheid in de verdeling van risicofactoren zeer gering.

Het is van het allergrootste belang dat de randomisatie procedureel correct is uitgevoerd. Essentieel hierbij is dat de arts die de patiënt tot het onderzoek toelaat, niet op de hoogte is van de eerstvolgende behandelingstoewijzing. Is hij dat wel, dan doet zich de mogelijkheid van selectieve toelating voor.

> In een niet-gepubliceerd onderzoek bijvoorbeeld werd voor postinfarctpatiënten via randomisatie beslist of ze aan een programma van fysieke rehabilitatie zouden deelnemen. Nadat het onderzoek was voltooid, bleek dat de patiënten die aan het trainingsprogramma hadden deelgenomen, gemiddeld tien jaar jonger waren dan de patiënten van de referentiegroep. Er was weliswaar 'gerandomiseerd', maar de procedure was incorrect. De randomisatie gebeurde namelijk via een doos met enveloppen die bij de arts op het bureau stond. Verondersteld werd dat de arts *nadat* hij had besloten dat een patiënt voor het onderzoek in aanmerking kwam, de eerstvolgende enveloppe zou openen en de patiënt dienovereenkomstig zou indelen. In de praktijk echter bleek eerst de enveloppe te zijn geopend, waarna een geschikte patiënt werd uitgezocht. Het is zelfs niet uitgesloten dat meer enveloppen tegelijk werden geopend, waarna de jongere patiënten aan het trainingsprogramma werden toegewezen en de oudere aan de controlegroep. De procedure zou correct zijn geweest indien iedere patiënt eerst centraal was geregistreerd en de gerandomiseerde behandelingstoewijzing vervolgens was meegedeeld aan de arts.

Vergelijkend onderzoek tussen groepen patiënten is ook mogelijk zonder dat specifiek in de behandeling van patiënten wordt ingegrepen.

> Bijvoorbeeld werd in het kader van een onderzoeksproject van het Interuniversitair Cardiologisch Instituut Nederland een grote groep patiënten die een 'bypass'-operatie hadden ondergaan, na één jaar opnieuw gecatheteriseerd om te zien of de aangelegde 'grafts' nog open waren.[4] Sommige patiënten waren behandeld met cumarinederivaten, anderen met bloedplaatjesaggregatieremmers, terwijl weer anderen in het geheel geen profylaxe hadden gekregen. Vergelijking van deze drie groepen patiënten zou een indruk kunnen geven van de (relatieve) effectiviteit van deze behandelingen. Voor onvergelijkbaarheid van sommige risicofactoren zou via bepaalde statistische methoden kunnen worden gecorrigeerd. De vraag blijft echter of de drie groepen patiënten wat het natuurlijke beloop betreft vergelijkbaar zijn. Er kan immers een met behulp van de statistische methode niet te achterhalen reden zijn geweest waarom de behandelende arts de ene patiënt cumarine voorschreef, de ander een bloedplaatjesaggregatieremmer en de derde juist niets. Uit dit onderzoek kunnen dan ook slechts voorlopige uitspraken worden geformuleerd.

Bij de introductie van een nieuwe therapie worden de op dit ogenblik behandelde patiënten soms vergeleken met die uit een eerdere periode. Men spreekt dan over een onderzoek met *historische* controles. Dergelijke onderzoekingen geven, met name door verschillen in de gehanteerde criteria, onvoldoende waarborg dat de te vergelijken groepen inderdaad vergelijkbaar zijn. Ook een behandelingstoewijzing op grond van geboortedatum of datum van ziekenhuisopname (bv. even dag leidt tot indextherapie, oneven tot referentietherapie) geeft onvoldoende garantie voor de vergelijkbaarheid van beide groepen en is dus ondeugdelijk.

§ 6.2. Vergelijkbaarheid van externe factoren
Om de invloed van externe factoren op de index- en referentiegroep zoveel mogelijk gelijk te maken, zorgt men er voor dat de beide behandelingen hetzelfde aanzien hebben. Daarbij zijn bovendien noch de patiënt noch de behandelende arts van de feitelijke therapietoewijzing op de hoogte. Dit heeft tot gevolg dat noch het handelen van de arts noch dat van de patiënt hierdoor wordt beïnvloed. In dat geval spreekt men van een dubbelblind onderzoek.
Een dubbelblinde opzet is echter geen absolute voorwaarde voor interne validiteit van een geneesmiddelenonderzoek. In sommige gevallen is een dubbelblinde opzet ook niet goed uitvoerbaar, bijvoorbeeld als het effect van een operatie wordt vergeleken met een medicamenteuze therapie;

'placebo-operaties' worden in het algemeen als onethisch ervaren. In deze situatie is men meer geïnteresseerd in het effect van de behandeling in zijn geheel, dat wil zeggen met inbegrip van alle daarmee samenhangende ingrepen.

Een voorbeeld hiervan is een in ons land verricht onderzoek naar het effect van intracoronair toegediend streptokinase bij patiënten met een vers hartinfarct.[5] Het onderzoek was open; alleen de met streptokinase behandelde patiënten werden in de acute fase gecatheteriseerd hetgeen als onderdeel werd gezien van de gehele therapeutische procedure. In een Amerikaans onderzoek werden daarentegen àlle patiënten in de acute fase gecatheteriseerd en werd aan de ene helft streptokinase en aan de andere helft het oplosmiddel toegediend. Het Nederlandse onderzoek richtte zich dus op de klinische toepasbaarheid van intracoronair toegediend streptokinase als procedure terwijl in het Amerikaanse alleen het effect van de chemische substantie werd bepaald.

§ 6.3. *Vergelijkbaarheid van informatieverzameling*

Om een juist beeld van het therapiegebonden effect te verkrijgen dient uiteraard het ziektebeloop met de grootst mogelijke nauwkeurigheid te zijn vastgesteld. Essentieel is dat de wijze van verzamelen van informatie tussen de groepen vergelijkbaar is.
Bijvoorbeeld dient, wanneer een geneesmiddel wordt gegeven om een recidiefinfarct te voorkómen, de controlefrequentie in beide groepen gelijk te zijn.

Stel dat in een dergelijk postinfarctonderzoek de patiënten van de indexgroep om de drie weken worden teruggezien door de huisarts, terwijl de patiënten van de controlegroep deze bezoeken niet behoeven af te leggen. Wanneer het optreden van een recidiefinfarct de uitkomstvariabele is, is de indexgroep in het nadeel. Immers, het optreden van lichte pijn op de borst vlak voor een controlebezoek zal aan de huisarts worden gemeld. Wanneer deze een ECG maakt en enzymwaarden bepaalt, kan een (klein) recidiefinfarct worden vastgesteld. Een vergelijkbare patiënt uit de de controlegroep die dezelfde pijn op de borst bemerkt, onderneemt wellicht geen verdere actie. Bij hem kan dat kleine recidief misschien niet meer worden vastgesteld, nog afgezien van de mogelijkheid dat hij zich de pijn later ook niet meer herinnert.

In een dubbelblind onderzoek is vergelijkbaarheid van informatie in het algemeen gegarandeerd; immers, het schema van waarnemingen en controles is gelijk en bovendien is de waarnemer niet op de hoogte van de behandelingstoewijzing. Een onbevooroordeelde waarneming is daarvan het gevolg.
Ook voor open onderzoekingen is vergelijkbaarheid van informatie te garanderen door het kiezen van een zogenaamd hard criterium. Dit zou in bovengenoemd voorbeeld het geval zijn geweest indien niet het optreden van een recidiefinfarct maar het al dan niet overlijden van de patiënt in de observatieperiode als uitkomstvariabele zou zijn gekozen.
Ten aanzien van de vergelijkbaarheid van informatie is ook de vaststelling van het begin van de observatieperiode van belang. Dit valt samen met het moment van de randomisatie.

Wanneer bijvoorbeeld het effect van een chirurgische behandeling wordt vergeleken met dat van een voortgezette medicamenteuze therapie, behoort de observatie te beginnen op het moment dat - via randomisatie - tot de operatie wordt besloten en niet op het moment dat de operatie wordt uitgevoerd. Door de randomisatie worden de patiënten die overlijden tussen het moment van de beslissing en de operatie (m.a.w. de categorie v.d. zeer zieke patiënten) gelijkelijk over de beide behandelingsgroepen verdeeld. Wanneer deze patiënten niet meer in de chirurgische, maar wel in de medicamenteuze groep worden meegeteld zijn de groepen niet meer vergelijkbaar, met als gevolg dat geen juist beeld van het behandelingseffect wordt verkregen. Sommige auteurs gaan zelfs zover dat ze deze patiënten meetellen bij de medicamenteus behandelde groep. Een groep patiënten met een zeer hoog risico wordt dan overgeheveld naar de andere behandelingsgroep. Uiteraard zijn de groepen met betrekking tot het natuurlijke beloop dan niet meer vergelijkbaar.

Verder mag de vaststelling van het einde van de observatieperiode niet afhangen van de toestand van de patiënt. Patiënten die uitvallen bij een onderzoek zijn vaak degenen bij wie het goed gaat, of bij wie het juist slecht gaat. Wanneer deze patiënten niet in de analyse worden betrokken, kan een onjuist beeld ontstaan van het therapiegebonden effect. Alleen als het al dan niet overlijden van de patiënt de uitkomst is van het onderzoek, is dit voor alle patiënten te achterhalen, ongeacht of ze op dat moment nog met de onderzoeksmedicatie werden behandeld.
Wanneer op deze wijze de sterfte per behandelingsgroep wordt vastgesteld, spreekt men van een analyse volgens het '*intention-to-treat*'-principe. Dit betekent dus dat de patiënt, wat er ook met hem gebeurt (bv. overlijden, beëindigen v.d. therapie) blijft meetellen bij de (index- of controle-) groep waarbij hij (of zij) bij het begin van het onderzoek was ingedeeld.
Ook wanneer de uitkomst minder 'hard' is, bijvoorbeeld het optreden van reïnfarcering, kan een analyse volgens het 'intention-to-treat'-principe plaatsvinden.
Het spreekt voor zich zelf dat degenen die het onderzoek hebben ontworpen ervan uit gaan dat de deelnemende (huis)artsen zich volledig houden aan het protocol en de gevraagde informatie naar behoren geven. De interne validiteit hangt in hoge mate hiervan af.

DEEL II

In het vorige deel zijn de vraagstelling en de methodiek van een onderzoek onderwerp van bespreking geweest. Bijzondere aandacht is besteed aan de drie vergelijkbaarheidscriteria: het natuurlijke beloop, de externe factoren (d.w.z. de bijkomende factoren die op het behandelingseffect van invloed kunnen zijn) en de wijze waarop in de observatieperiode de informatie is verzameld. In dit deel valt het accent op de weergave en de interpretatie van de resultaten.

§ 7. BESCHRIJVING VAN DE RESULTATEN

Onder de kop 'Resultaten' treft men in een artikel de beschrijving van de onderzoeksbevindingen aan.

In deze context onderscheiden we drie onderdelen, te weten een profiel van de feitelijk onderzochte patiënten, een beschrijving van het ziektebeloop in de behandelingsgroepen en de zogenaamde statistische grootheden.

De in § 4.1. aangegeven definitie van de onderzochte groep patiënten wordt afgerond met het schetsen van een profiel van de patiënten die aan het onderzoek hebben deelgenomen. Meestal worden geslacht, leeftijd en een aantal klinische kenmerken die de medische status van de patiënten op het moment van randomisatie beschrijven per behandelingsgroep gegeven. Een tabel is vooral als een nadere karakterisering van de onderzochte groep patiënten van belang.

Het belangrijkste onderdeel is een overzichtelijke weergave van het ziektebeloop per behandelingsgroep. Indien de uitkomstvariabele het al dan niet optreden van een gebeurtenis (bv. een reïnfarct) tijdens de observatieperiode is, wordt het beloop in een groep patiënten gekarakteriseerd door het percentage dat een reïnfarct heeft gekregen. Voor de individuele patiënt is dit percentage de uitdrukking van het risico om gedurende de observatieperiode een reïnfarct te krijgen. Verder dient aantekening te zijn gemaakt van eventuele veranderingen in de gekozen dosering of van mogelijke aanvullende medicaties en is het van belang (bv. door telling van resterende tabletten) na te gaan in hoeverre de patiënten de voorschriften hebben opgevolgd ('patient compliance'). Een tabel kan het beloop op zeer eenvoudige wijze weergeven.

Het effect van de index- ten opzichte van de referentiebehandeling wordt uitgedrukt in een zogenaamde *effectmaat*.
Eén van de mogelijkheden hiertoe is het risico in de controlegroep af te trekken van dat in de indexgroep: het *risicoverschil*.
Een andere mogelijkheid is de risico's op elkaar te delen: aldus ontstaat de *risicoratio*, ook wel het relatieve risico genoemd.
Wanneer sterfte de uitkomstvariabele is wordt vaak de (relatieve) *mortaliteitsreductie* gebruikt; deze is gelijk aan 100 x (1-risicoratio).

> Stel, als voorbeeld, dat in de indexgroep 30 van de 200 patiënten en in de controlegroep 40 van de 200 patiënten binnen één jaar overlijden. Het risico van overlijden binnen één jaar in de indexgroep is 30/200, dus 15%; in de controlegroep is dit 40/200, dus 20%. Het risicoverschil is 15-20%, dus -5%. De interpretatie is dat door de indextherapie van iedere 100 behandelde patiënten er vijf minder binnen één jaar overlijden.
> De risicoratio is 15%: 20%, dus 0,75. De interpretatie is dat van iedere vier patiënten die onder de referentiebehandeling zouden zijn overleden, er in de indexgroep slechts drie overlijden.
> De (relatieve) mortaliteitsreductie is 100 x (1-0,75) dus 25%. De interpretatie is, dat van de patiënten die onder de referentiebehandeling zouden zijn overleden, 25% dankzij de indextherapie in leven blijft.

Wanneer de uitkomstvariabele kwantitatief is (bv. de diastolische bloeddruk) wordt het beloop in een groep patiënten gekarakteriseerd door de *gemiddelde* of door *de mediane* (= middelste v.d. naar grootte gerangschikte) waarden. De effectmaat wordt verkregen door de gemiddelden (of medianen) van elkaar af te trekken, of op elkaar te delen.

Het derde deel van de beschrijving van de onderzoeksbevindingen bestaat uit statistische grootheden, die de uitdrukking zijn van toevalsvariatie in de uitkomsten. Hiervoor worden het betrouwbaarheidsinterval en/of de p-waarde en statistische significantie gebruikt. De interpretatie van deze grootheden wordt besproken in § 9.

Het is van belang zich rekenschap te geven van het feit dat de statistische grootheden slechts hulpmiddelen zijn bij de interpretatie van de uitkomsten en niet de uitkomsten zelf van het onderzoek zijn. Alleen de vermelding dat het nieuwe middel een statistisch significant beter effect had dan het standaardmiddel beschrijft de waarnemingen niet, maar geeft een interpretatie die de lezer geen indruk verschaft van de grootte van het effect. Ook de bevindingen zelf, in de vorm van percentages of gemiddelden (of medianen) behoren te zijn beschreven.

§ 8. EXTERNE VALIDITEIT

Wanneer de lezer voldoende inzicht heeft verworven in de vraagstelling, de methodiek en de uitkomsten van het onderzoek, blijft de vraag in hoeverre de bevindingen van dit onderzoek ook van toepassing zijn voor de eigen patiënten. In deze context spreekt men, zoals vermeld, over de externe validiteit van het onderzoek. Soms wordt ook de term generaliseerbaarheid gebruikt. Een klinisch geneesmiddelenonderzoek is extern valide indien de feitelijk onderzochte groep patiënten in biologische zin representatief is voor een klinisch herkenbare indicatie. Het is een wijd verbreid misverstand dat, naarmate de insluitingscriteria vager zijn gedefinieerd, de draagwijdte van de onderzoeksbevindingen groter is. Het tegendeel is eerder waar. Immers, een onderzoek met strikte criteria voor het infarct geeft bijvoorbeeld meer informatie over het effect van ß-blokkade na een hartinfarct dan een onderzoek met vage definities.
De aanwezigheid van een geloofwaardig werkingsmechanisme bevordert in het algemeen de externe validiteit, maar absoluut noodzakelijk is dit niet. Het voorschrijven van bepaalde ß-blokkerende stoffen na een hartinfarct wordt bijvoorbeeld door een veelheid van onderzoeksresultaten ondersteund, hoewel het inzicht in de precieze wijze waarop deze middelen als secundair preventicum werkzaam zijn op dit moment ontbreekt.
Externe validiteit geldt uiteraard alleen voor de onderzochte vraagstelling. In § 5 werden reeds enige voorbeelden gegeven waarin goed lezen nodig is om de bestudeerde vraagstelling te achterhalen. In het algemeen kan niet worden volstaan met de naam van een ziekte alleen. Wanneer bijvoorbeeld in een onderzoek wordt ge-

rapporteerd over reumatoïde aandoeningen terwijl vage spierpijnen of andere pijnen van het bewegingsapparaat worden bedoeld is er sprake van een zekere mate van misleiding, aangezien 'reumatoid' volgens de criteria van de American Rheumatism Association zeer strikt is gedefinieerd.

Bij de beoordeling of het nuttig is een behandeling ook bij de eigen patiënten toe te passen is in de eerste plaats de grootte van het behandelingseffect - en niet zozeer de statistische significantie ervan - relevant.

> Bijvoorbeeld werd in de Verenigde Staten enkele jaren geleden het BHAT-onderzoek (ß-Blocker Heart Attack Trial) uitgevoerd.[7] Van de 1916 postinfarctpatiënten die met propranolol werden behandeld, overleden er 136 in een periode van gemiddeld twee jaar (7,2%); van de 1921 patiënten die met placebo werden behandeld, overleden er 188 (9,8 %). De resultaten van het onderzoek waren hoogst significant. Het risicoverschil was -2,6% (= 7,2%-9,8 %). De risicoratio was 0,74 (= 7,2%:9,8%). De mortaliteitsreductie was 26% (= 100 x (1-0,74)).
> Het effect, een reductie van de mortaliteit met 26%, is desondanks niet erg groot. Indien 1000 patiënten gedurende twee jaar na hun infarct worden behandeld met propranolol, wordt daarmee bewerkstelligd dat niet 98 maar 72 patiënten in die periode overlijden. Om 26 patiënten in leven te houden moet men er 1000 gedurende twee jaar onderwerpen aan een behandeling met propranolol. Wat precies het lange termijneffect is in gewonnen levensjaren, kan niet zonder meer uit de onderzoeksresultaten worden afgeleid.

§ 9. STATISTISCHE ASPECTEN

Zoals vermeld is toevalsvariabiliteit in de uitkomsten niet geheel vermijdbaar.
Bijvoorbeeld wordt onder invloed van een ß-blokkerende stof ten opzichte van placebo een gemiddelde bloeddrukdaling van 20 mmHg gemeten. Door de toevalsvariabiliteit zal, indien het onderzoek een aantal malen zou worden herhaald, niet elke keer hetzelfde effect worden geregistreerd; de volgende keer zou het 15 mmHg of 25 mmHg kunnen zijn.
Het is mogelijk de onzekerheid over de werkelijke grootte van het behandelingseffect die hiervan het gevolg is, te kwantificeren. Hiervoor gebruikt men het *95%-betrouwbaarheidsinterval*, dat uit de gegevens kan worden berekend. De procedure garandeert dat het 95%-betrouwbaarheidsinterval in 95 van de 100 gevallen het werkelijke behandelingseffect omsluit.
Wanneer in het voorbeeld het 95%-betrouwbaarheidsinterval loopt van 17 tot 23 mmHg, kan men er vrij zeker van zijn dat het werkelijke behandelingseffect inderdaad tussen deze waarden ligt. In het eerder genoemde BHAT-onderzoek bijvoorbeeld was de mortaliteitsreductie 26%. Het 95%-betrouwbaarheidsinterval loopt van 10% tot 40%, hetgeen wil zeggen dat men op grond van de bevindingen van het onderzoek er voor 95% zeker van kan zijn dat de werkelijke mortaliteitsreductie tussen deze percentages ligt.
Het toevalselement neemt af naarmate de groepen groter zijn; de grenzen van het betrouwbaarheidsinterval versmallen naarmate de groepsgrootte toeneemt.

In het algemeen geven een effectmaat (bv. risicoratio) en het 95%-betrouwbaarheidsinterval een direct interpreteerbare beschrijving van de relevante informatie uit een onderzoek. Door eerstgenoemde wordt, zoals vermeld, het behandelingseffect gekwantificeerd, terwijl de laatste een beeld geeft van de onzekerheid. Onzes inziens is dit de enig juiste methode die bij elk geneesmiddelenonderzoek zou moeten worden gehanteerd. Helaas vindt deze methode geen algemene toepassing, maar wordt vaak gebruik gemaakt van statistische toetsen met bijbehorende begrippen zoals statistische significantie en p-waarde. De nu volgende uiteenzetting wordt gegeven om duidelijk te maken wat de essentie van deze methode is en met name wat de beperkingen ervan zijn.
Een statische toets kan het beste worden vergeleken met een diagnostische test. In een diagnostische test wordt gebruik gemaakt van een laboratoriumwaarde ten einde na te gaan of een patiënt een bepaalde ziekte heeft. In een statistische toets wordt de p-waarde (een getal tussen 0 en 1, berekend uit de gegevens v.h. onderzoek) gebruikt om na te gaan of de indexbehandeling (t.o.v. de referentiebehandeling) werkzaam is. De statistische toets heet positief (ofwel statistisch significant) als de p-waarde kleiner is dan 5%; hij is negatief indien de p-waarde groter is dan 5%. Statistische significantie van onderzoeksresultaten betekent dus niet meer maar ook niet minder dan dat de 'test op werkzaamheid' positief is. Een positieve diagnostische test 'bewijst' echter niet dat de patiënt de ziekte heeft, evenmin impliceert een negatieve test afwezigheid van ziekte. Dit geldt mutatis mutandis ook voor de statistische toets. Bij het al dan niet aannemen of een middel werkzaam is spelen ook andere factoren een rol. Wanneer door reeds beschikbare wetenschappelijke gegevens een zeer groot 'à priori'-geloof bestaat in de onwerkzaamheid van een middel dan zal een enkele significante p-waarde in een onderzoek niet bij machte zijn dit geloof te veranderen. Wanneer daarentegen een 'à priori' groot geloof bestaat in de werkzaamheid van een middel dan kan zelfs een niet-significante p-waarde een voldoende argument zijn om aan te nemen dat de gegevens in die richting wijzen. Evenals een laboratoriumtest is de p-waarde slechts een hulpmiddel die niet automatisch tot een beslissing leidt: het is altijd de arts die beslist. Verder spelen ook de testeigenschappen een rol; deze omvatten de sensitiviteit (het percentage positieven onder de ziekten) en de specificiteit (het percentage negatieven onder de niet-zieken). Een statistische test heeft (per definitie!) een specificiteit van 95%. De sensitiviteit van een statistische toets hangt af van de grootte van de onderzoeksgroepen: hoe groter de groepen, hoe groter de sensitiviteit. De sensitiviteit van een statistische toets wordt meestal aangeduid als het onderscheidingsvermogen (Engels: power) van de toets (of v.h. onderzoek). Met deze vergelijking voor ogen dient de lezer

begrippen als statistisch significant te waarderen.
Problemen ontstaan wanneer men probeert te begrijpen wat de p-waarde (de universele testvariabele van het 'statistisch laboratorium') voorstelt. De p-waarde wordt gedefinieerd als de kans op het vinden van een bepaald (of verdere afwijkend) onderzoeksresultaat *onder aanname* dat de index- en referentietherapie even effectief zijn. Deze definitie is evenwel moeilijk te doorgronden en wordt vaak op onjuiste wijze geïnterpreteerd, zelfs door statistici. De p-waarde is niet de kans dat de resultaten aan het toeval zijn te wijten, ook niet de kans dat de indextherapie niet werkzaam is. Helaas wordt de p-waarde meestal aangezien voor de laatste. Men kan zelfs stellen dat de 'populariteit' van de p-waarde het gevolg is van deze foute interpretatie.
Het heeft daarom de voorkeur bovengenoemde definitie te vergeten en de p-waarde te zien als een testvariabele die positief ($p < 0,05$) of negatief ($p > 0,05$) kan zijn, waarvan de specificiteit bekend en hoog is (95%), maar waarvan de sensitiviteit afhangt van onder meer de grootte van het onderzoek. In de diagnostiek geldt als regel dat een ongevoelige test (d.w.z. met lage sensitiviteit) nietszeggend is, ongeacht of de uitkomsten ervan positief dan wel negatief zijn. Iets dergelijks geldt met name voor bevindingen van een klein onderzoek, waarvan immers de sensitiviteit laag is.
De p-waarde is dus alleen een bruikbaar hulpmiddel bij het trekken van conclusies, wanneer de onderzoeksgroepen voldoende groot zijn. De criteria voor voldoende groot zijn niet gemakkelijk te geven. Daardoor is ook de betekenis van 'statistisch significant' en van 'statistisch niet-significant' vaak moeilijk te schatten.
Deze problemen zijn te vermijden door gebruik te maken van wel direct interpreteerbare effectschattingen aangevuld met een betrouwbaarheidsinterval. Bovendien hangt toekomstig gebruik van de onderzochte middelen in de praktijk af van de grootte van het verwachte behandelingseffect; dit kan op zijn beurt worden afgewogen tegen de kosten en de mogelijke, verwachte bijwerkingen. De voor de klinische praktijk relevante vraag is hoé effectief een middel is en niet óf het een positief effect heeft. Alleen effectmaat en betrouwbaarheidsinterval geven hierop een expliciet antwoord. Van de statistische toets kan dat niet worden gesteld.
De volgende twee voorbeelden, beide met te kleine groepen, laten nog eens zien dat de p-waarde de lezer op het verkeerde been kan zetten, terwijl effectmaat en betrouwbaarheidsinterval wel inzicht verschaffen.

Het eerste voorbeeld is een in de jaren zestig door Meuwissen et al. in Nederland uitgevoerd gerandomiseerd onderzoek naar het effect van langdurige behandeling met anticoagulantia bij 138 postinfarct patiënten.[8] Van de 68 met antistolling behandelde patiënten overleed er één en van de 70 placebopatiënten zeven. De p-waarde werd berekend op 0,04 waarmee dus volgens de conventionele criteria zou zijn 'aangetoond' dat langdurige behandeling met anticoagulantia sterfte bij postinfarctpatiënten voorkómt. De effectmaat met het 95%-betrouwbaarheidsinterval geeft een ander beeld. De relatieve mortaliteitsreductie bedraagt 85% (immers, de placebomortaliteit is 10% (7/70) en wordt onder antistolling met 85% 'gereduceerd' tot 1,5% (1/68)). Deze mortaliteitsreductie representeert een behandelingseffect van enorme omvang. Geen enkel ander geneesmiddel dat wordt gebruikt in de secundaire preventie na het hartinfarct, heeft ooit een effect van een dergelijke omvang getoond in een voldoende groot onderzoek. Een effect van deze omvang is dus niet erg geloofwaardig. Het 95%-betrouwbaarheidsinterval loopt van 16% toename tot 98% reductie. Dit zeer brede interval illustreert de instabiliteit van de bepaling van het behandelingseffect en geeft aan hoe onzeker men over de werkelijke grootte daarvan nog is. Het lijkt dus waarschijnlijk dat de uitkomsten van dit onderzoek in zekere zin een toevalsbevinding vormen. Ze schetsen waarschijnlijk een te optimistisch beeld van het effect van antistolling. Op grond van deze overweging kan dit onderzoek niet worden gezien als 'bewijs' voor de werkzaamheid van antistolling bij de secundaire preventie van het hartinfarct.
Het tweede voorbeeld betreft de in 1971 door Hill et al. gepubliceerde resultaten van een onderzoek waarin thuisbehandeling (indexbehandeling) werd vergeleken met opname op een hartbewaking (referentiebehandeling) bij patiënten met een acuut infarct.[9] Van de 132 thuis behandelde patiënten overleden er 17 (13%) binnen zes weken, van de 132 opgenomen patiënten 14 (11%). De p-waarde wordt berekend op 0,57. Deze is (aanzienlijk) groter dan 0,05. Hieruit krijgt men de indruk dat dit onderzoek laat zien dat thuisbehandeling equivalent is aan opname. Thuisbehandeling heeft er echter toe geleid dat 17 in plaats van 14 van de 132 patiënten zijn overleden en dat er dus een mortaliteitstoename is van 3 op de 14 patiënten (= 21%).

Het uit deze gegevens berekende betrouwbaarheidsinterval laat zien dat zowel 37% mortaliteitsreductie als 136% mortaliteitstoename mogelijk is. De breedte van het betrouwbaarheidsinterval geeft dus aan dat we van het werkelijke effect op de mortaliteit van thuisbehandeling van het acute infarct eigenlijk vrijwel niets weten. Dit onderzoek is dan ook 'inconclusive' en niet negatief. Deze 'inconclusiveness' is het gevolg van het feit dat het onderzoek te klein was.

§ 10. KRUISPROEVEN

Bij de beschrijving van een geneesmiddelenonderzoek is steeds uitgegaan van de situatie dat de helft van de patiënten met de indextherapie is behandeld en de andere helft met de referentietherapie. Een dergelijk onderzoek heeft een opzet met *parallelle groepen* ('parallel group design').
Een andere, veel gebruikte, onderzoeksopzet is die van de *kruisproef* ('cross-over design'). In deze onderzoeksvorm wordt de helft van de patiënten eerst enige tijd met de indextherapie behandeld en vervolgens een bepaalde tijd met de referentietherapie. De overige patiënten ondergaan ook beide therapieën, maar dan in de andere volgorde. De keuze voor één van de behandelingsvolgorden gebeurt meestal op basis van het toeval. Het is gebruikelijk het onderzoek te beginnen met een *uitwasperiode* ('wash out'), waar-

in de bestaande behandeling geleidelijk wordt verminderd en tenslotte wordt gestaakt. Ook bij de wissel is er een uitwasperiode. Het voordeel van deze opzet is dat iedere patiënt in de analyse met zich zelf kan worden vergeleken.
Aan deze opzet kleeft echter een zeer groot nadeel. Het is voor de geldigheid van de kruisproef een absolute voorwaarde dat het behandelingseffect niet afhangt van de behandelingsvolgorde. Dit veronderstelt dat geen van beide therapieën irreversibele veranderingen bij de patiënt teweegbrengt.
Aan deze voorwaarden is onder meer voldaan indien de uitkomstvariabele na de therapie telkens terugkeert op een stabiel uitgangsniveau. De kruisproef is daarom alleen geschikt voor het bestuderen van palliatieve effecten bij chronische aandoeningen.
Kruisproeven worden onder meer gebruikt om de bloeddrukverlagende werking van geneesmiddelen te vergelijken, of om het anti-angineuze effect van middelen te vergelijken bij patiënten met *stabiele* angina pectoris. Bij de behandeling van *onstabiele* angina pectoris kan men geen gebruik maken van een kruisproef. Het is immers onwaarschijnlijk dat een patiënt chronisch onstabiel is: de onstabiele periode degenereert in een infarct of verbetert tot stabiliteit. Met betrekking tot het gebruik van de kruisproef bij stabiele angina pectoris zijn er in feite ook al problemen. Stabiele angina pectoris kan vrij plotseling degenereren tot onstabiele angina pectoris of zelfs tot een infarct. Het valt in het geheel niet te voorspellen wanneer dat zal gebeuren. Het is gebruikelijk een dergelijke patiënt als uitvaller te boeken. Men vertrouwt er dan op dat dat onder beide behandelingen even vaak optreedt. Aangezien niet kan worden uitgesloten dat de medicatie de stoot heeft gegeven tot de degeneratie, is deze procedure een bron van onvergelijkbaarheid. De kruisproef tolereert in het algemeen geen uitvallers. Het optreden van uitvallers kan namelijk leiden tot een ernstige vertekening zowel ten gunste als ten nadele van het onderzochte middel. Veronderstel dat een nieuw antidepressivum wordt vergeleken met een bestaand middel. Indien bij het nieuwe middel veel patiënten met name in de eerste periode uitvallen omdat de depressie onhanteerbaar wordt, zou een analyse waarvan de uitgevallen patiënten zijn uitgesloten tot een veel te optimistisch beeld van het nieuwe middel kunnen leiden. Bij het lezen van een artikel over kruisproeven lette men altijd op de uitvallers. Eén van de manieren om hiervoor te corrigeren is de uitvallers te voorzien van een arbitrair gekozen ongunstige score en te kijken of het gunstige effect aanwezig blijft. Een foute, maar helaas veel gehanteerde, methode is die waarbij uitvallers worden vervangen door nieuwe patiënten, die in de analyse de plaats innemen van de uitgevallen patiënten. Ook bij deze methode worden patiënten met wie het slecht ging niet in de analyse betrokken. Dit kan leiden tot onvergelijkbaarheid van de groepen en daardoor tot een vertekende weergave van het behandelingseffect.

§ 11. ETHISCHE OVERWEGINGEN BIJ MEDEWERKING AAN EEN ONDERZOEK

Aan een huisarts kan, zoals eerder gesteld, worden gevraagd deel te nemen aan een onderzoek, hetgeen inhoudt dat hem wordt verzocht patiënten te behandelen volgens een onderzoeksprotocol.
Bij de gebruikelijke patiënt/arts-relatie kan de arts een keuze maken uit verschillende therapeutische opties op grond van wat naar zijn mening het beste is, al dan niet in overleg met de patiënt. Bij deelname aan een geneesmiddelenonderzoek daarentegen wordt een keuze tussen index- en referentiebehandeling gemaakt volgens een op randomisatie berustende procedure, met het speciale oogmerk over het relatieve effect van beide behandelingen te leren.
Wanneer de methodiek fout of de omvang te klein is, leidt het onderzoek per definitie niet tot een betere behandeling van toekomstige patiënten. In dat geval vervalt iedere grond om een patiënt aan het onderzoek te laten deelnemen. Fout opgezet onderzoek is per definitie onethisch. Dit geldt dus voor vrijwel alle niet-vergelijkende klinische geneesmiddelenonderzoekingen. Wanneer de arts zelf niet weet welke van de twee te vergelijken behandelingen is te prefereren, ontstaat de meest geschikte situatie voor het verrichten van vergelijkend onderzoek. Deelname is dan zelfs in het belang van de patiënt.
Een arts die aan een onderzoek deelneemt waarin een pas ontwikkeld geneesmiddel wordt vergeleken met een standaardgeneesmiddel, onthoudt zijn patiënten niet een mogelijk effectievere therapie; hij stelt een gedeelte van zijn patiënten in staat voortijdig van dat middel te profiteren. Ook hier vindt een afweging plaats: het is immers ook mogelijk dat het middel helemaal niet zo effectief is of dat het ernstige bijwerkingen heeft.
Het is gebruikelijk patiënten vooraf in te lichten over de vraagstelling van het onderzoek, de voor- en nadelen van beide behandelingen en de te verrichten waarnemingen. Nadat de patiënt is ingelicht, wordt hem toestemming gevraagd. Deze procedure heet 'informed consent'. Weigert de patiënt, dan wordt hij doorgaans op de gebruikelijke wijze behandeld. De regels waaraan men zich te houden heeft zijn verwoord in de zogenaamde Declaratie van Helsinki van de Wereldgezondheidorganisatie.

§ 12. SAMENVATTING EN CONCLUSIE

Onder vrijwel alle omstandigheden is vergelijkend onderzoek nodig om de effectiviteit van een geneesmiddel te onderzoeken. Bij het lezen van een publikatie vormt men zich eerst een beeld van de feitelijk onderzochte vraagstelling. Vervolgens wordt de gevolgde onderzoeksmethodiek aan een kritische beschouwing onderwor-

pen. Men lette op de drie vergelijkbaarheidscriteria: natuurlijk beloop, externe factoren en informatieverzameling.
Vervolgens bestudeert men het waargenomen ziektebeloop in de index- en controlegroep. Het effect van de behandelingen op het ziektebeloop is in essentie kwantitatief. Het verdient derhalve aanbeveling het effect in een getal (maat) uit te drukken, zoals in het risicoverschil, de risicoratio of de (relatieve) mortaliteitsreductie. De met de randomisatie samenhangende onzekerheid wordt het best uitgedrukt in een 95%-betrouwbaarheidsinterval voor het behandelingseffect.

Literatuur

1. Elwood PC, Williams WO. A randomized controlled trial of aspirin in the prevention of early mortality in myocardial infarction. J R Coll Gen Pract 1979; 29: 413-416.
2. Elwood PC, Sweetnam PM. Aspirin and secondary mortality after myocardial infarction. Circulation 1980; 62 (suppl V): 53-58.
3. The Sixty Plus Reinfarction Study Research Group. A double-blind trial to assess long-term oral anticoagulant therapy in elderly patients after myocardial infarction. Lancet 1980; ii: 989-994.
4. Lubsen J. Effect of antithrombotic therapy on bypass graft patency: evidence from a non-randomized study. Eur Heart J 1984; 5 (suppl I): 325.
5. Simoons ML, Serruys PW, Brand M van der et al. Improved survival after early thrombolysis in acute myocardial infarction. Lancet 1985; ii: 578-582.
6. Khaja F, Walton JA, Brymer JF et al. Intracoronary fibrinolytic therapy in acute myocardial infarction. N Engl J Med 1983; 308: 1305-1311.
7. ß-Blocker Heart Attack Trial Research Group. A randomized trial of propranolol in patients with acute myocardial infarction. JAMA 1982; 247: 1707-1714.
8. Meuwissen OJAT, Vervoorn AC, Cohen O et al. Double-blind trial of long-term anticoagulant treatment after myocardial infarction. Acta Med Scand 1969; 186: 361-368.
9. Hill JD, Hampton JR, Mitchell JRA. A randomized trial of home-versus-hospital management for patients with suspected myocardial infarction. Lancet 1978; i: 837-841.
10. Klinisch geneesmiddelenonderzoek, een praktische leidraad. Red. Lang R de, Lubsen J. Wetenschappelijke Uitgeverij Bunge, Utrecht 1987.

Aanbevolen literatuur

Onlangs is onder redactie van R. de Lang en J. Lubsen een Nederlandstalig boek verschenen waarin door een tiental deskundigen verschillende aspecten van klinisch geneesmiddelenonderzoek uitvoerig worden belicht.[10]

Trefwoorden: geneesmiddelenonderzoek, onderzoek (geneesmiddelen), geneesmiddelen (onderzoek)

AAN DE LEZERS VAN HET GENEESMIDDELENBULLETIN

De beide laatste nummers van het Gebu zijn gewijd aan het geneesmiddelenonder zoek en als zodanig van een iets andere inhoud dan u gewend bent.
Niettemin is er nauwelijks een onderwerp in de geneeskunde denkbaar dat zo be langrijk is voor hen die geneesmiddelen voorschrijven en afleveren als dit. Aan dat voorschrijven behoort een rationele basis ten grondslag te liggen. Het afge ven van een recept zonder voldoende argumenten stelt de hulpvrager bloot aan bijwerkingen, interacties en mogelijke ziekten terwijl ook onnodige kosten wor den gemaakt: niet alleen door het voorschrift maar ook de mogelijke onderzoekin gen en zelfs opneming in het ziekenhuis die van al dan niet begrepen bijwerkin gen en interacties het gevolg kunnen zijn.
Het doel van dit artikel is het geven van informatie over rationele farmacothe rapie, om als leidraad te dienen bij de vraag of men als voorschrijver kan in gaan op het verzoek om deel te nemen aan een geneesmiddelenonderzoek. Tevens kan het bij het lezen van een artikel over geneesmiddelenonderzoek een hulpmiddel zijn om te toetsen of het beschreven onderzoek aan de eisen voldoet. Ook kunnen deze afleveringen van het Gebu dienstig zijn bij het onderwijs.

In de afgelopen tijd is er veel te doen geweest over het voortbestaan van het Gebu en in welke vorm dat dan zou zijn. Het verheugt ons u te kunnen meedelen dat wij, dat wil zeggen de Redactiecommissie en -staf en de Adviesraad, met de overheid en de verschillende beroepsorganisaties zijn begonnen gezamenlijk naar wegen te zoeken om het voortbestaan van het Gebu veilig te stellen, met zoveel mogelijk behoud van de doelstellingen, namelijk gratis verspreiding en onafhan kelijke informatie over geneesmiddelen onder auspiciën van de overheid aan al degenen die geneesmiddelen voorschrijven of verstrekken. Wat dit betreft zien wij 1988 met enig optimisme tegemoet.
Wij hopen dat u het komende jaar geheel en al met optimisme tegemoet kunt zien en wensen u, namens de Redactiecommissie en -staf, een goede dosis voorspoed.

CORRECTIE GEBU 1987; 21: NR 12 - ANIHISTAMINICA

Blz. 62, 2e kolom, 5 -6 regel boven het opschrift 'CARA' dient als volgt te worden gelezen:
Uit de schaarse beschikbare gegevens komen astemizol en terfenadine ook bij de behandeling van hooikoorts bij kinderen als bruikbaar naar voren.

Op blz. 65, 2 kolom, 8 regel van boven dient 'luchtweginfecties' te worden vervangen door 'luchtwegaandoeningen'.

GENEESMIDDELENBULLETIN

Adviesraad:

Prof. dr E. van der Does, (voorzitter), Rotterdam
P.C.M. van den Berg, Amsterdam
S. Flikweert, Nijkerk
Dr H.A. van Geuns, Rijswijk
Prof. dr F.W.J. Gribnau, Nijmegen
Prof. dr C.J. de Groot, Amsterdam
Dr F. Kalsbeek, 's-Gravenhage
Dr A.L.M. Kerremans, Helmond
Dr J.F.F. Lekkerkerker, Enschede
Dr. H. Mattie, Leiden
Prof. dr M.F. Michel, Rotterdam
Prof. dr A.S.J.P.A.M. van Miert, Utrecht
Mw dr B.C.P. Polak, Rotterdam
Prof. dr F. Schwarz, Bilthoven
R.W. Zaadnoordijk, 's-Gravenhage

Redactiecommissie: Prof. dr E. van der Does (voorzitter); Prof. dr M.N.G. Dukes (adv. lidj, Kopenhagen; Mw L.T.W. de Jong-van den Berg, Groningen; Prof. dr J. Lubsen (adv. lid), Rotterdam; Prof. dr J.P. Nater, Groningen; Mw M. Pannevis, Rotterdam, Dr C.A. Teijgeler (adv. lid), Rijswijk

Redactiestaf/-secretariaat: Mw H.H. Kortland-Brinkman / Mw M. Brouwer-Klopper, Mw J.J. Doorschodt-van der Steenhoven

ISSN: 0304-4629