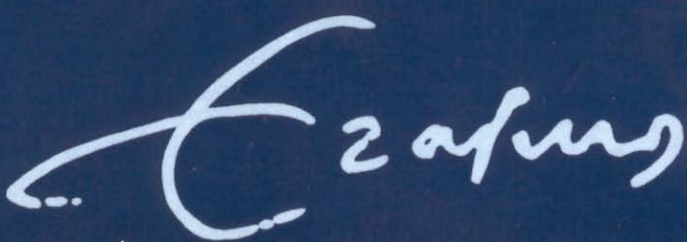
LOF DER BIOSTATISTIEK

PROF. R. VAN STRIK



ERASMUS UNIVERSITEIT ROTTERDAM

LOF DER BIOSTATISTIEK

AFSCHEIDSCOLLEGE VAN
PROF. R. VAN STRIK,
HOOGLERAAR BIOSTATISTIEK
ERASMUS UNIVERSITEIT ROTTERDAM

UITGESPROKEN OP 21 NOVEMBER 1996

Mijnheer de Rector Magnificus,
Mijnheer de Voorzitter van het College van Bestuur,
Mijnheer de Decaan van de Faculteit der Geneeskunde en Gezondheidswetenschappen,
en voorts gij allen, die door uw aanwezigheid blijk geeft van Uw belangstelling.
Zeer gewaardeerde toehoorders.

INLEIDING

Bij het voorbereiden van dit afscheidscollege kwam vrijwel meteen de vraag bij mij op: Hoe zou je een afscheidscollege moeten inrichten voor een zo gemêleerd gezelschap als hier aanwezig is? Gaat u maar na: hoogleraren, bestuurders, stafleden, dokters, docenten, statistici, epidemiologen, familie, vrienden en kennissen.
De vraag stellen is haar beantwoorden, zegt men. Een redelijk uitgangspunt lijkt: eenvoudig van opzet, rond een centraal thema en liefst begrijpelijk. Een normaal college dus waarin het gaat om overdracht van kennis, inzicht en ervaring. Al laat de titel van dit college wel vermoeden dat het nogal subjectief gekleurd zal zijn.
Wat die titel betreft, die heeft u allicht doen denken aan de grote Europese geleerde uit Rotterdam, naar wie onze universiteit genoemd is. Vooral naar een van zijn meer bekende geschriften, dat hij, naar men zegt, in 1508 in zeven dagen heeft voltooid.

Ongetwijfeld zijn veel uitspraken van de godin der Zotheid ook van toepassing op diverse terreinen in onze huidige maatschappij.
En natuurlijk weet iedereen dat er ook op het gebied van de statistiek talloze zotternijen zijn te vinden. Diverse spreekwoorden in vrijwel alle talen getuigen daarvan al eeuwen, zoals bv. deze uit de

vorige eeuw van de Engelse staatsman Benjamin Disraeli (1804 - 1881): "There are three kinds of lies: lies, damned lies and statistics." Die zotternijen nemen in aantal zelfs eerder toe dan af, dankzij de ruime beschikbaarheid van statistische programma's voor de Personal Computer. Mede om commerciële redenen, zijn die programma's zogenaamd gebruikersvriendelijk gemaakt voor leken.

Een fraai voorbeeld was enige tijd geleden te zien in het gezaghebbende tijdschrift Science. In een studie over relaties tussen bijenpopulaties werd de grafiek in fig. 1 gepresenteerd. De auteur merkt daarbij op dat de weergegeven parabool de beste aanpassing geeft bij deze punten, die de dichtheid van twee bijensoorten in 57 regio's representeren. Enige tijd later reageerde een collega hierop met de mededeling dat hij van deze curve niet echt onder de indruk was. Hij voegde daaraan toe dat, als je toch eenmaal zó bezig bent, de relatie door een véél betere curve kan worden beschreven, namelijk zoals in fig. 2.

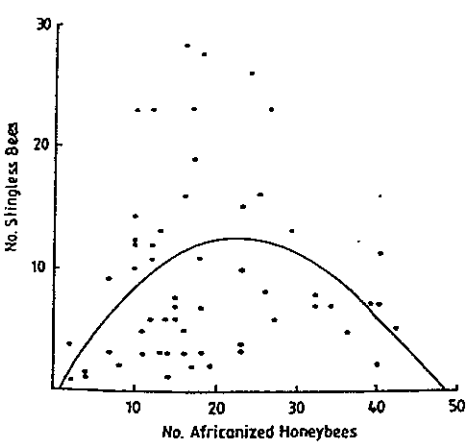


Fig. 1 (Science 201:1030)

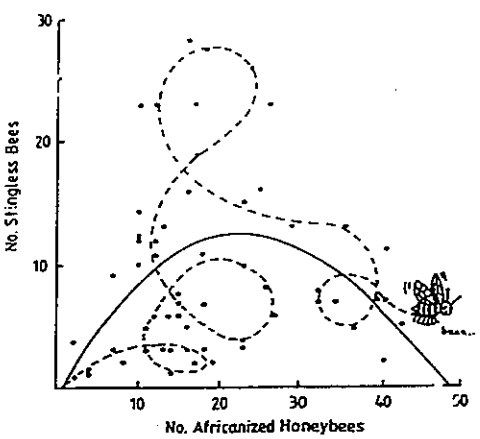


Fig. 2 (Science 202:823)

Het is niet moeilijk de nog resterende tijd voor dit college met soortgelijke sterke staaltjes te vullen. Maar verkneukelt u zich nog niet, ik zal daar vandaag niet verder op ingaan. Integendeel, ik wil U deelgenoot maken van enige voor de geneeskunde belangrijke ontwikkelingen in de biometrie, gedurende de afgelopen - laten we zeggen - 30 jaar. De volgende vraag was dus: wélk thema zal ik in dit college aansnijden? Er zijn legio statistische onderwerpen die juist in het medisch onderzoek een specifieke "impact" hebben en zich lenen voor een lofzang op de Biostatistiek. Het symposium, dat hier vandaag ter gelegenheid van mijn afscheid is gehouden over de Rol van Biostatistiek in Medisch Onderzoek, geeft een duidelijke indicatie van de verscheidenheid aan belangrijke "topics". Ik kom daar straks nog op terug.

VERGELIJKENDE KLINISCHE STUDIES

Het thema dat ik in dit college centraal wil stellen, betreft de prospectief vergelijkende klinische studie, of in het Engels, de "controlled clinical trial". En wel om diverse redenen; ik noem er twee. Ten eerste omdat de biostatistiek een belangrijke bijdrage heeft geleverd aan de ontwikkeling van de methodologie van klinische studies in de laatste 30 jaar. Ten tweede omdat strikte toepassing van die methodologie in het medisch wetenschappelijk onderzoek van direct belang is voor de volksgezondheid. Dat laatste is ook erkend door het College ter Beoordeling van Geneesmiddelen. Bij de besluitvorming over registratie van nieuwe geneesmiddelen zijn goede vergelijkende klinische studies voor het College een "conditio sine qua non". Na tien jaar lidmaatschap van dit College heb ik daarover geen enkele twijfel. Ook de farmaceutische industrie is daarvan overtuigd, zoals met een enkel voorbeeld is te illustreren.

Onlangs werd een aanvraag ingediend tot registratie van een nieuw bloeddrukverlagend middel. Om de werkzaamheid aannemelijk te maken werden in het dossier onder meer gegevens overgelegd van 15 gerandomiseerde en dubbelblind uitgevoerde klinische studies, bij in totaal meer dan 5000 patiënten, met licht tot matig verhoogde bloeddruk.

Zelfs in de Oudheid werd het belang van prospectief vergelijkend onderzoek blijkbaar al ingezien. Vermoedelijk de eerste studie volgens dit principe in het kader van de gezondheidszorg, dateert namelijk van ruim twee en een half duizend jaar geleden. U kunt deze beschreven vinden in het Oude Testament, om precies te zijn in het boek Daniel. Het betreft een vergelijkende studie van twee diëten, door Daniel opgezet omstreeks 580 voor Chr., tijdens zijn gedwongen verblijf aan het hof van Nebukadnezar. Alle jongelingen aan het hof waren verplicht het door de koning vastgestelde dagelijkse rantsoen bestaande uit vlees en wijn te consumeren. Daniel stelde daarop aan de kamerheer voor, hem en zijn drie vrienden uit Juda een heel sober dieet te verstrekken, namelijk groenten en water. Maar de kamerheer was bang voor ontdekking door de koning, als deze hun uiterlijk "minder welvarend zou vinden dan dat van de overige knapen". Toen zei Daniel: neem "tien dagen de proef; men geve ons groenten te eten en water te drinken; laat dan ons uiterlijk met dat van de knapen die de koninklijke spijze eten, door U vergeleken worden, ...". Hij gaf hem hierin gehoor ... "en na verloop van tien dagen bleek hun uiterlijk schoner en zagen zij er welvarender uit dan al de knapen die van de koninklijke spijze gegeten hadden". Nu, ruim 2500 jaar later, is het natuurlijk niet moeilijk om kritiek te leveren op de methodologie van deze studie.

Het is bijvoorbeeld nog onduidelijk of de twee groepen aan het begin statistisch gezien vergelijkbaar waren. Vermoedelijk zult U zelf nog meer punten hebben ontdekt, waarop wat af te dingen is.

Overigens, toen ik voor het eerst als statisticus te maken kreeg met een prospectief vergelijkende klinische studie van twee geneesmiddelen, helaas pas nadat alle resultaten al waren verzameld, kwamen in feite soortgelijke tekortkomingen in opzet en uitvoering ook aan het licht. Dát was in 1956, dus pas 40 jaar geleden. Sindsdien is er in allerlei opzichten natuurlijk veel veranderd.

KLINISCH GENEESMIDDELENONDERZOEK

De verandering geldt ook voor de kwaliteit van geneesmiddelenonderzoek. Die is mede gestimuleerd door een tragische gebeurtenis welke in het begin van de zestiger jaren wereldwijd bekend werd als het Softenon drama. Veel moeders die in een vroege fase van hun zwangerschap het bewuste slaapmiddel hadden gebruikt, kregen kinderen met ernstige congenitale misvormingen. In de hoop een dergelijke tragedie in de toekomst te voorkomen, werden door de overheden in vrijwel alle landen strengere eisen ingevoerd voor de kwaliteit, de veiligheid en de werkzaamheid van geneesmiddelen.
In Nederland is het College ter Beoordeling van Geneesmiddelen in 1963 ingesteld en belast met het op wetenschappelijke gronden beoordelen van alle aanvragen tot toelating en registratie van geneesmiddelen. Daartoe zijn objectieve criteria en richtlijnen ontwikkeld waaraan onderzoek met geneesmiddelen moet voldoen.

"Good Clinical Practice Guideline"

In het kader van de Europese Unie is al bijna vanaf het begin

gestreefd naar harmonisatie van deze criteria en richtlijnen in de betrokken landen. Dat heeft in 1991 geleid tot invoering van een gemeenschappelijke "Good Clinical Practice"- richtlijn voor geneesmiddelenstudies, afgekort GCP, die nu een wettelijke status heeft. De inbreng van de biostatistiek komt op diverse plaatsen in deze GCP-richtlijn naar voren. De samenvatting in het voorwoord vermeldt onder meer: "Pre-established ... written procedures for the organization, conduct, data collection, ... of clinical trials are necessary ... to establish the credibility of data and to improve the ... quality of trials. These procedures also include good statistical design as an essential prerequisite ..., it is unethical to enlist the co-operation of human subjects in trials which are not adequately designed".
Verder wordt o.m. een globale definitie van een studieprotocol gegeven: "a document which states the rationale, objectives and statistical design and methodology of the trial, ...".

Dat heeft schertsenderwijs al aanleiding gegeven tot een nieuwe definitie van de biostatisticus. Namelijk, als iemand die er niet van overtuigd is dat het Columbus was die Amerika heeft ontdekt, omdat in het protocol bij zijn subsidie-aanvraag aan koningin Isabella, geen sprake was van Amerika, maar alleen van Indië.

Overigens, wat deze strenge GCP-regels betreft, die gelden voornamelijk voor zogenoemde fase III-studies, die bedoeld zijn om de effectiviteit en de balans werkzaamheid / schadelijkheid van een geneesmiddel, overtuigend vast te stellen. In fase II-studies is er ruimte voor een meer exploratieve aanpak, zowel in de opzet als bij de analyse van de uitkomsten.

EEN CASUS

Na deze aanloop wil ik nu eerst een relatief eenvoudig en wat gestyleerd voorbeeld aan U voorleggen, van een klinische studie die beschouwd kan worden als op de grens van fase II en fase III. Het betreft een vergelijkende studie met twee varianten van eenzelfde bloeddrukverlagend middel, dat werd toegepast tijdens bepaalde operaties om de bloeddruk op het gewenste lage peil in te stellen. In dat opzicht werken beide varianten, kortweg aangeduid met A en B, even goed, zoals uit onderzoek was gebleken. Dat is verklaarbaar, omdat alleen de farmaceutische formulering iets verschilt en niet het werkingsmechanisme. De aanleiding tot dit onderzoek was de vraag of na het stoppen van de medicatie de nawerkingsduur van de nieuwe variant B, mogelijk korter zou zijn dan die van de bestaande formulering A. Die vraag is hier vertaald in: "Bij welke van de twee is de hersteltijd van de bloeddruk na afloop van de operatie het kortst en derhalve het gunstigst?". De hersteltijd werd in dit geval arbitrair gedefinieerd als de tijd na het stoppen van de medicatie tot de systolische bloeddruk weer het niveau van 100 mm kwikdruk heeft bereikt.

Opzet van de studie

De studie werd uitgevoerd bij 40 patiënten, waarvan de helft middel A kreeg en de andere helft middel B.

Om te zorgen dat beide groepen statistisch gezien in álle opzichten equivalent van samenstelling zouden zijn, werd randomisatie toegepast. Dat betekent dat elke patiënt die bereid was aan de studie mee te doen, aselect hetzij A óf B toegediend kreeg. In de praktijk is randomisatie op diverse manieren te realiseren, afhankelijk van locale faciliteiten. Daarbij wordt steeds meer gebruik gemaakt van

moderne hulpmiddelen. Maar de principiële achtergrond is nog dezelfde als 70 jaar geleden, toen dit concept werd geintroduceerd door de Engelse statisticus, Sir Ronald Fisher. Randomisatie is een universeel toepasbare methode om iedere denkbare vorm van systematische vertekening, of kortweg "bias", te vermijden in de samenstelling van de groepen. De methode werd ongeveer 20 jaar later ingevoerd in het klinisch geneesmiddelenonderzoek door een andere Engelse statisticus, Sir Austin Bradford Hill. Deze was betrokken bij de eerste zogenoemde "placebogecontroleerde en gerandomiseerde dubbelblinde studies" met Streptomycine bij patiënten met longtuberculose. De term dubbelblind houdt hier in, dat noch de patiënt noch de behandelend arts weten welk middel, P of Q, is toegediend. Dit uit voorzorg om elke vorm van "bias" bij de interpretatie van neveneffecten, resp. vermindering of verergering van klachten en symptomen, zoveel mogelijk te vermijden. Het dubbelblind principe werd ook toegepast in de genoemde studie met de twee varianten A en B, zowel om "bias" te vermijden bij de keuze van de dosering, als bij het eventueel geven van comedicatie, en bij het meten van de hersteltijd. Maatregelen als randomisatie en blinderen hebben vooral ten doel de interne validiteit van de studie te waarborgen. Tot zover deze globale beschouwing over de opzet van de studie. Andere op zich belangrijke aspecten, zowel in de opzetfase als bij de uitvoering, laat ik hier om tijdswil onbesproken.

Schets van de analyse

Laten we ons nu gaan richten op de analyse van de gevonden resultaten. Onder het motto "eerst tekenen en dan rekenen", begin ik met een grafische presentatie van de primaire uitkomst-variabele, zoals in figuur 3. Dat betreft dus de hersteltijden, bij de twee

groepen van 20 patiënten.

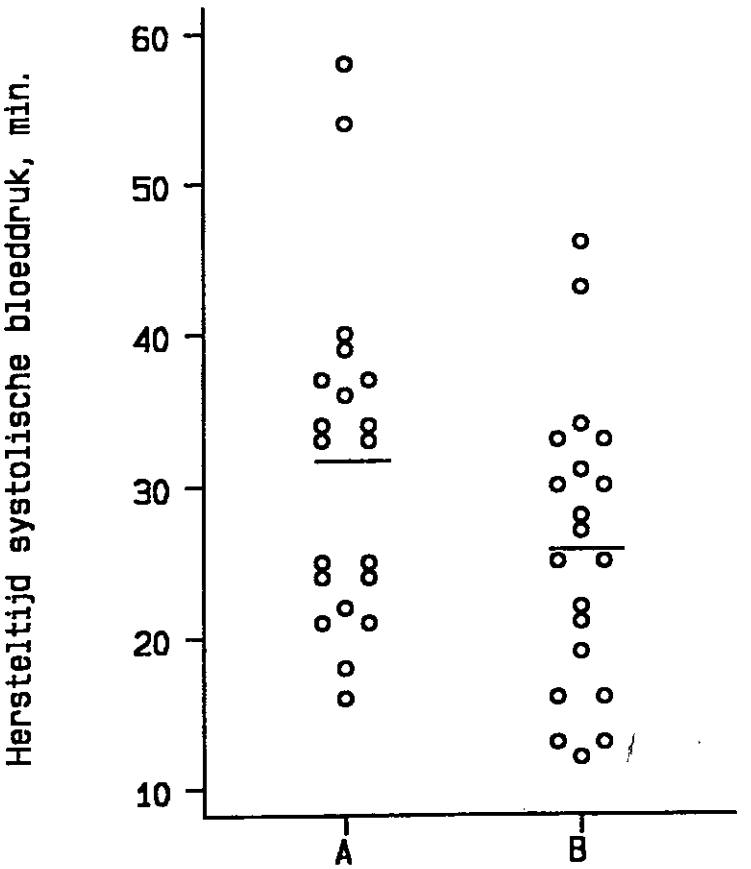


Fig. 3 Individuele uitkomsten in groepen A en B; de streepjes geven de gemiddelden weer.

Opvallend is de grote spreiding van ruim tien minuten tot bijna een uur. Daardoor heeft het waargenomen verschil tussen de groepsgemiddelden van ongeveer 6 min., mede wegens de kleine aantallen patiënten, weinig precisie. Als maat voor die precisie wordt in medische publicaties doorgaans de Engelse term "standard error" gebruikt, afgekort SE. Deze bedraagt voor het hier gevonden gemiddelde verschil, SE = 3,3 min.

Statistische toetsing van het gevonden verschil

Om na te gaan of uit dit verschil de conclusie kan worden getrokken, dat middel B, meer in het algemeen gesproken, gemiddeld gunstiger is dan A, zou men de nulhypothese kunnen toetsen, dat het verwachte echte, maar onbekende, verschil, nul is.

Bij toepassing van de toets van Student vinden we dan een tweezijdige overschrijdingskans, de zg. P-waarde, welke hier 9 % bedraagt. In tabel 1 is dit resultaat op de gebruikelijke manier samengevat. Gemiddeld verschilden A en B dus 5,7 minuten in deze studie.

Tabel 1

Gemiddelden (± SE) per groep voor de hersteltijd

Medicament	A	B	A-B	P*)
Hersteltijd (min.)	31,5(±2,5)	25,8(±2,1)	5,7(±3,3)	0,09
n	20	20		

*) Volgens t-toets van Student

De conclusie uit zo'n tabel wordt veelal geformuleerd in de trant van: "De beide middelen A en B verschillen ten aanzien van de gemiddelde hersteltijd in deze studie niet significant op het 5% niveau (P=0,09)", aangezien de P-waarde groter is dan de gebruikelijke significantie-drempel van 5 %. Soms echter wordt deze uitspraak ingekort tot: "Er is geen verschil in hersteltijd tussen de middelen A en B." Dát is niet alleen slordig, het is voorbarig en vaak onjuist! Dit heeft te maken met het feit dat ook in een gerandomiseerde studie, de interpretatie van "niet significant", niet eenduidig is. Dat blijkt zodra we de aandacht richten op een verwante, maar meer directe aanpak.

Onzekerheidsmarge: het betrouwbaarheidsinterval

Veel informatiever dan het resultaat van zo'n toets, is namelijk de mate van onzekerheid van het gevonden gemiddelde verschil. De onzekerheidsmarge kan uit de gegevens van een gerandomiseerd onderzoek direct worden afgeleid in de vorm van het betrouwbaarheidsinterval dat geldt voor het "echte" verschil, hier aangegeven met μ(A-B). In dit voorbeeld blijkt het 95 % betrouwbaarheidsinterval voor μ(A-B) tussen de grenzen - 1 min. tot + 12 min. te liggen. U ziet het al, het kan nog vriezen of het kan dooien.

Nu werd volgens het studieprotocol bij de zogenoemde "power"-analyse voor de bepaling van het vereiste aantal patiënten, er van uitgegaan dat een werkelijk verschil μ van 5 minuten of meer, klinisch relevant is. De onzekerheidsmarge is dus te breed om over de relevantie van het echte verschil μ(A-B) al een conclusie te kunnen trekken op basis van deze studie. En hiermee is dus tevens een nadere interpretatie van "niet significant" aan de orde.

Nu zult U natuurlijk denken, dat dit al met al een tamelijk onbevredigend slot is, na alle geleverde inspanningen van het onderzoeksteam. Maar ik kan U op voorhand al enigszins gerust stellen. De biostatisticus in het team was bij de al genoemde "power"-analyse uitgegaan van een kleinere spreiding in hersteltijden tussen de patiënten per groep. Om precies te zijn een factor 2 kleiner, in termen van de standaardafwijking. Dat was inderdaad terecht, zoals we zullen zien. Want de tot nu toe gepresenteerde analyse is weliswaar niet onjuist, maar ook niet optimaal.

Alternatieve analyse

Er is onder meer geen rekening gehouden met diverse variabelen die ook de hersteltijd kunnen beïnvloeden, behalve middel A of B. Dat kunnen patiënt-kenmerken zijn, zoals leeftijd, gewicht, begin-bloeddruk, en gaat U zo maar door. Deze zogenaamde covariabelen dienen in principe vóór de randomisatie gemeten te worden. Voor zover ze invloed hebben op de hersteltijd, betekent dit in een gerandomiseerde studie in de eerste plaats, dat ze de spreiding van de hersteltijden zullen vergroten.

Een wijd verbreid misverstand onder niet-statistici is, dat het alleen zin heeft bij de analyse met die variabelen rekening te houden, als de gemiddelden duidelijk verschillend zijn uitgevallen. In een gerandomiseerd onderzoek zal er per definitie slechts een toevallige onbalans zijn in de samenstelling van de groepen. Vandaar dat ook tussen de groepsgemiddelden van gemeten covariabelen alleen toevallige verschillen optreden. Het toetsen van de significantie van dergelijke verschillen, wat soms gedaan wordt, is derhalve een uiting van onvoldoende begrip van de betekenis van randomisatie.

Slechts als zo'n verschil onverwacht groot én klinisch relevant is voor een belangrijke covariabele, is het nuttig na te gaan of de conclusie anders zou uitvallen wanneer dit verschil "post hoc" in rekening gebracht wordt. Als dat zo blijkt te zijn, is er voor een fase III studie wel een probleem. Want volgens de GCP-richtlijn blijft de nadruk liggen op de in het protocol voorziene analyse.

"Post hoc" analyse is nu eenmaal niet conform deze richtlijn, waarin gesteld wordt dat ook de statistische methodologie van de studie in

het protocol gedocumenteerd moet worden. Dit alles om iedere schijn te vermijden dat bij tegenvallende conclusies een "fishing expedition"of "data dredging" heeft plaatsgevonden, door achteraf de uitkomsten zo lang en zo vaak via een serie computerprogramma's te "martelen", totdat ze hebben bekend.

Keren we nu weer terug naar het voorbeeld.
Behalve covariabelen zijn er ook variabelen die na randomisatie gemeten worden, zoals hier de duur van de operatie, de totale dosis van het bloeddrukverlagend middel en het daarmee bereikte gemiddelde bloeddrukpeil tijdens de operatie. Van de laatstgenoemde twee valt te verwachten dat ze óók de hersteltijd zullen beïnvloeden en als zodanig voorspellende waarde hebben ten aanzien van die hersteltijd. Echter, het betrekken van dergelijke variabelen in de analyse alsof het covariabelen betreft, is in zijn algemeenheid niet verantwoord. Als voorwaarde geldt namelijk dat covariabelen niet beïnvloed mogen zijn door de te vergelijken behandelingen zelf. Om dat te kunnen garanderen moeten die variabelen daarom vóór de randomisatie zijn gemeten.
In dit speciale geval evenwel, was het "a priori" - zoals gezegd, op een aantal gronden - aannemelijk dat beide prognostische variabelen tijdens de operatie op dezelfde wijze en in dezelfde mate zullen zijn beïnvloed door de twee formuleringen A en B. Dit uitgangspunt blijkt door de resultaten ook niet te worden weersproken, zoals we nu zullen zien.
Om dit te verifiëren zijn de uitkomsten voor beide variabelen eerst grafisch weergegeven in fig. 4 en tevens samengevat in tabel 2.

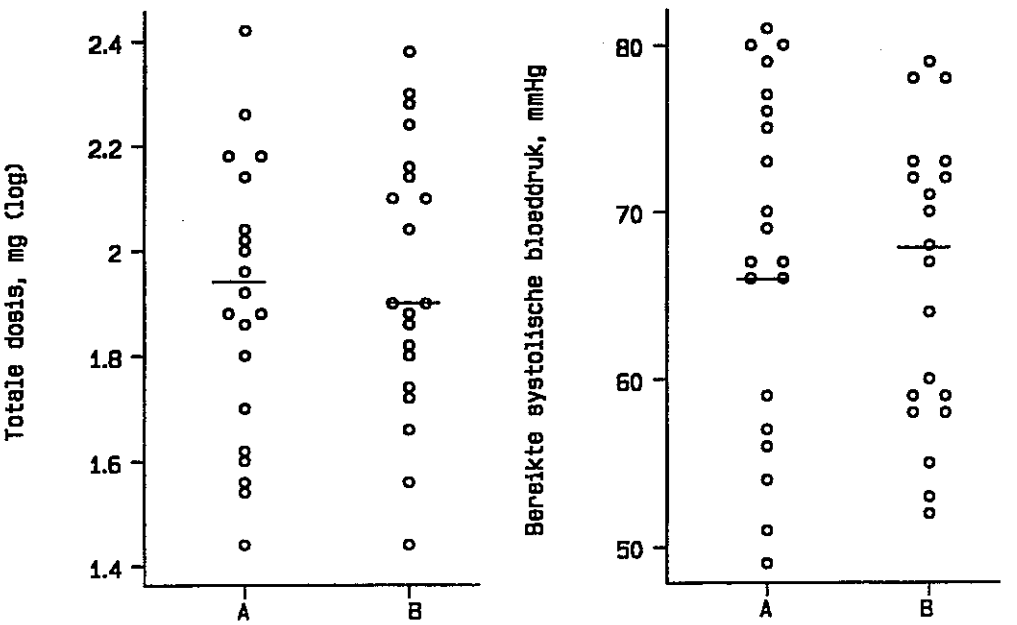


Fig. 4 Individuele uitkomsten in groepen A en B; de streepjes geven de gemiddelden weer.

Tabel 2.

Gemiddelden (± SE) per groep (n=20), voor de tijdens de operatie
(a) totaal toegediende dosis in mg per patiënt (op log schaal) en
(b) gemiddeld bereikte systolische bloeddruk per patiënt (in mm Hg)

Medicament	A	B	A - B
(a) Dosis (log)	1,90 (±0,06)	1,95 (±0,06)	-0,05 (±0,08)
(b) Bloeddruk	67,6 (±2,3)	66,0 (± 1,9)	1,6 (± 3,0)

Zowel de grafieken als de tabel geven aan dat de twee varianten A en B in gelijke mate de gemiddelde bloeddruk tijdens de operatie beïnvloeden. Voorts is er ook weinig of geen verschil in de gemiddelde dosering.

De biostatisticus had dit in de protocolfase al gepeild binnen het onderzoeksteam en voor beide variabelen de invloed op de hersteltijd gekwantificeerd op basis van eerder verkregen gegevens. In het protocol was, conform de GCP-richtlijn, vastgelegd hoe de beide variabelen in de analyse zouden worden betrokken. Laten we nu eens nagaan of die analyse in dit geval ook rendement oplevert. We starten weer met een grafische presentatie in figuur 5. Voor elke groep afzonderlijk is de relatie tussen hersteltijd en totale dosis in kaart gebracht.

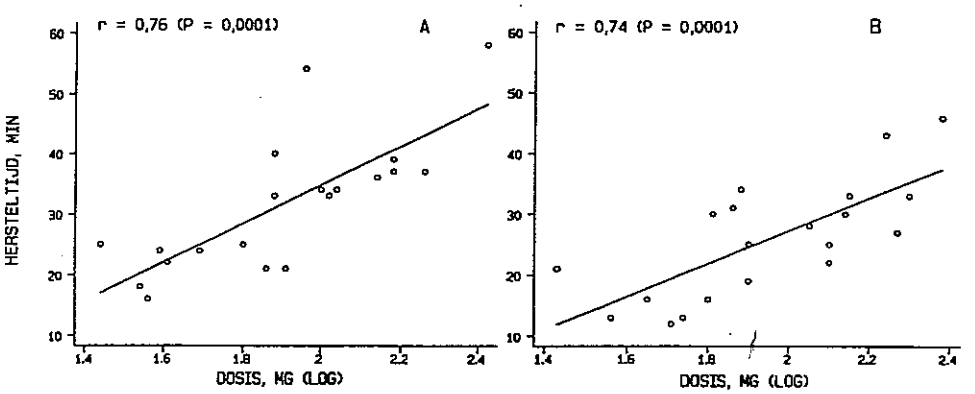


Fig. 5 Relatie tussen hersteltijd en totale dosis per groep (n=20); de best passende lineaire regressielijn en de Pearson correlatie-coëfficient met bijbehorende P-waarde zijn aangegeven in de grafiek.

Uit een statistische analyse van deze gegevens blijkt dat er een duidelijk significant verband bestaat tussen hersteltijd en dosis. Voorts kon worden vastgesteld dat tussen A en B de hellingen van de beide regressie-lijnen niet significant verschillen. Dat vormt nog een additioneel argument voor equivalente effectiviteit van A en B tijdens de operatie. Een en ander impliceert dat de variantie tussen hersteltijden per groep voor een deel te wijten is aan de variantie

tussen de doses. Hoe groot dat deel is, wordt bepaald door het kwadraat van Pearson's correlatie-coëfficiënt, die in de grafieken is aangeduid met de kleine letter r.

Voor de hersteltijden in groep A bijvoorbeeld, betekent dit dat de fractie van de variantie tussen hersteltijden die verklaard wordt door de spreiding tussen doses gelijk is aan r^2 is $(0,76)^2$, dit is 0,58. Voor de gemeten bloeddrukwaarden tonen soortgelijke grafieken eveneens een duidelijk significant verband met de hersteltijden, zoals fig. 6 laat zien.

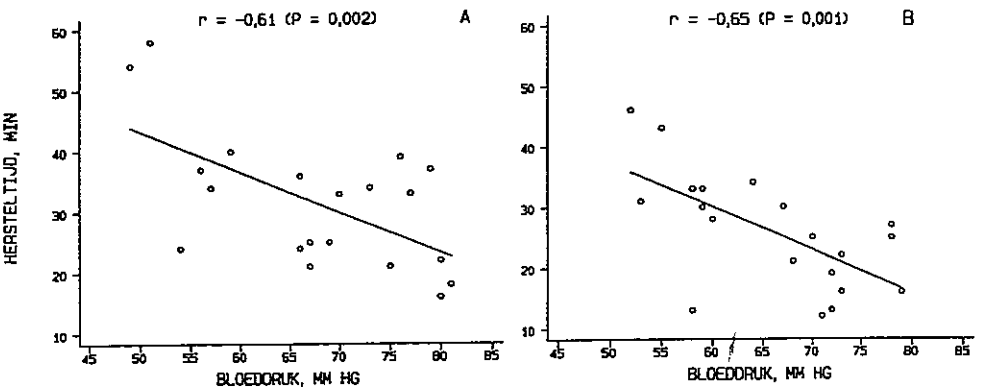


Fig. 6 Relatie tussen hersteltijd en het bereikte systolische bloeddrukpeil per groep (n=20); zie verder bij fig. 5.

De spreiding tussen de bloeddrukwaarden in groep A bijvoorbeeld verklaart een fractie r^2 is $(-0,61)^2$, dit is 0,37, van de variantie in hersteltijden. In hoeverre beide variabelen als ze simultaan in rekening worden gebracht, een grotere fractie zullen verklaren, hangt onder meer af van de correlatie tussen deze twee variabelen onderling. De grafieken in fig. 7 tonen dat deze samenhang gering is en dat de correlatie-coëfficiënt laag en niet significant is, zodat de beide variabelen als bijna onafhankelijk zijn te beschouwen.

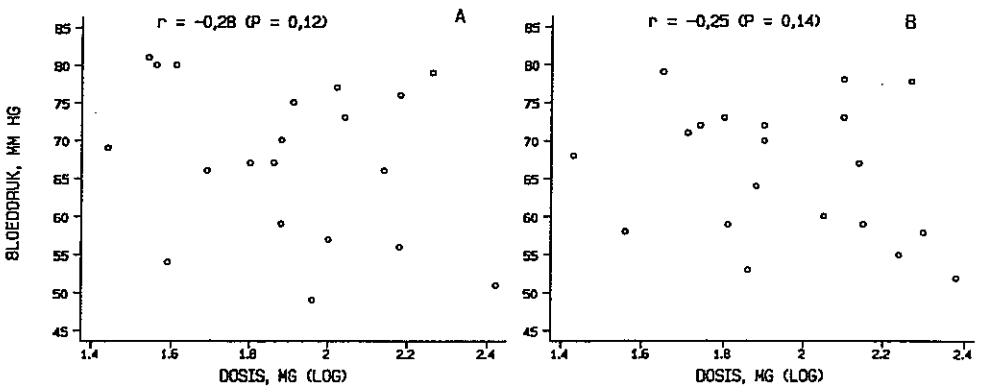


Fig.7 Relatie tussen het bereikte systolische bloeddrukpeil en de toegediende totale dosis per groep (n=20); zie verder bij fig. 5.

In zo'n geval ligt het voor de hand de twee variabelen tegelijk in rekening te brengen, via multipele covariantie-analyse.
De fractie van de variantie in hersteltijden die door deze variabelen samen wordt verklaard, blijkt dan 0,77 te bedragen, derhalve een aanzienlijke reductie van 77%. De resterende variantie van de hersteltijden is minder dan een kwart van de oorspronkelijke variantie. Dit betekent dat de "standard error" tot de helft is teruggebracht, dus tot het niveau zoals de biostatisticus dat tevoren had ingeschat.

Deze resultaten kunnen nu worden vertaald in het betrouwbaarheids-interval voor μ(A-B), zoals figuur 8 laat zien. De onzekerheids-marge wordt duidelijk smaller als met één van beide variabelen rekening wordt gehouden, en tot ongeveer de helft gereduceerd als beide tegelijk in rekening worden gebracht.

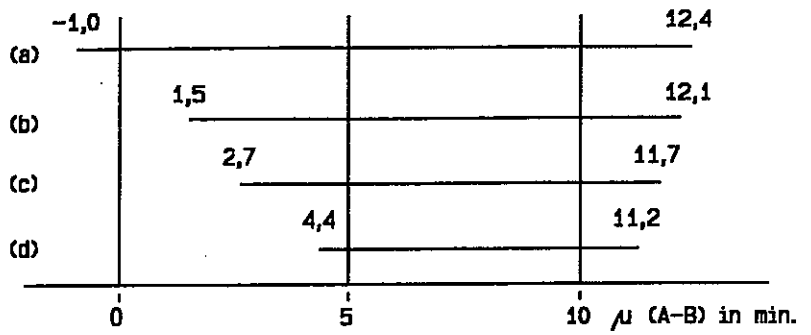


Fig. 8 Betrouwbaarheidsintervallen (95%) voor μ(A-B), afhankelijk van prognostische variabelen;
a) behandeling sec; b) + systolische bloeddruk; c) + totale dosis;
d) + beide.

Het verschil in gemiddelde hersteltijd is met de alternatieve analyse niet alleen statistisch significant, maar blijkt bovendien klinisch relevant, aangezien de ondergrens voor μ(A-B) nu vrijwel gelijk is aan 5 min. In tegenstelling tot de weinig bevredigende uitslag van de eerste analyse, kan na deze verfijning een duidelijke conclusie worden getrokken. Om dit te bereiken zonder de twee prognostische variabelen in rekening te brengen, zouden dus viermaal zoveel patiënten vereist zijn geweest, d.w.z. totaal 160 i.p.v. 40.
Tot zover deze uitweiding over een voorbeeld van een prospectief vergelijkende klinische studie, waarin de lof der biostatistiek kon worden verwoord.

BIOSTATISTISCHE RICHTLIJN

In aansluiting op de GCP-richtlijn werd in 1993 een meer specifieke biostatistische richtlijn ingevoerd in de Europese Unie. Daarin zijn diverse aspecten van de biostatistische methodologie nader uitge-

werkt, zowel de opzet van een klinische studie betreffend als de statistische analyse. Ikzelf heb destijds aan deze richtlijn ook een steentje mogen bijdragen. In die richtlijn zijn de ontwikkelingen in de biostatistiek gedurende de afgelopen 30 jaar voor een groot deel verwerkt. Om u daarvan een indruk te geven, heb ik de belangrijkste hoofdstukken, met een globale aanduiding van de inhoud, in een overzicht samengevat (tabel 3).

Tabel 3.
Overzicht van enige hoofdstukken uit de EU-Richtlijn:
BIOSTATISTICAL METHODOLOGY IN CLINICAL TRIALS

1. Overall considerations	5. Design techniques to avoid bias
Confirmatory studies	Blinding
Exploratory studies	Randomisation
2. Clinical trial designs	6. Prespecified data analysis
The parallel group design	Study populations
The cross-over design	1) The intention to treat population
	2) The per protocol population
3. Other design issues	Adjustment of P-values
Multicentre trials	
Trials to show equivalence	7. The conduct and monitoring phase
Group sequential designs	Interim analysis and early stopping
4. The sample size	8. Overal summary of several trials

Uiteraard kan ik binnen het bestek van dit college niet al deze punten stuk voor stuk bespreken. Alleen op twee ervan zal ik nu nader ingaan, namelijk van hoofdstuk 6 het concept "intention-to-

treat" populatie en van hoofdstuk 7 de interim analyse problematiek. Beide betreffen tamelijk recente ontwikkelingen in de biostatistiek, gericht op specifieke toepassing in klinische studies.

"Intention-to-treat" analyse

Ter illustratie van het begrip "intention-to-treat", eerst een voorbeeld. In een dubbelblinde gerandomiseerde "eind fase II studie", werd een nieuw anti-depressivum in twee doseringen vergeleken met een standaard middel, namelijk amitriptyline. Van de 50 gerandomiseerde patiënten, staakten 15 voortijdig de behandeling, allen wegens bijwerkingen van het gegeven middel. Voor de overige 35 patiënten werd het effect van de behandeling blind beoordeeld als ++ : zeer effectief, respectievelijk + : effectief en - : niet-effectief. De uitkomsten daarvan zijn in tabel 4 samengevat.

Tabel 4
Behandelingsresultaten van twee anti-depressiva bij 35 psychiatrische patiënten.

Medicament Dosis		Nieuw Laag	Nieuw Hoog	Standaard
Zeer effectief	++	2	8	6
Effectief	+	4	2	8
Niet effectief	-	3	2	0
Beoordeeld		9	12	14

In eerste instantie werd uit deze resultaten geconcludeerd dat de behandeling met de hoge dosis van het nieuwe middel, zéér effectief was bij 8 van de 12 patiënten, d.w.z. bij 67%. Voor de standaard was dit percentage lager, namelijk 6 van de 14, dat wil zeggen 43%.

Hoewel klinisch relevant, was het verschil - mogelijk wegens de kleine aantallen - statistisch niet significant.

De adder onder het gras, is hier uiteraard het weglaten van de uitvallers wegens bijwerkingen. Deze dienen meegeteld te worden in de analyse, namelijk bij de categorie niet-effectief. Immers ook bij hen heeft de behandeling gefaald. Aldus wordt een ander beeld van de resultaten verkregen voor alle gerandomiseerde patiënten (zie tabel 5).

Tabel 5

Behandelingsresultaten van twee antidepressiva bij alle 50 gerandomiseerde patiënten.

Medicament Dosis		Nieuw Laag	Nieuw Hoog	Standaard
Zeer effectief	++	2	8	6
Effectief	+	4	2	8
Niet-effectief	-	3	2	0
Uitvallers	(-)	6	8	1
Gerandomiseerd		15	20	15
% succes		40%	50%	93%

Aangezien het aantal uitvallers bij het nieuwe middel zoveel groter was dan bij de standaard, zijn de percentages met een zeer effectief resultaat nu zelfs gelijk voor hoge dosis en standaard, nl. 8 van de 20 en 6 van de 15, d.w.z. 40%. Bovendien, als we zeer effectief en effectief beide als een succes beschouwen, is het percentage successen voor de standaard 14 van de 15, d.i. 93%, tegen 10 van de 20, dus 50% voor de hoge dosis en 6 van de 15, dus 40% voor de lage dosis. Het succespercentage voor de standaard is statistisch

significant hoger dan elk van de beide percentages voor het nieuwe middel met $P < 0,01$.
De eerdere interpretatie van de uitkomsten blijkt na het meetellen van uitvallers, in dit geval radicaal bijgesteld te moeten worden. Deze laatste analyse met alle gerandomiseerde patiënten wordt aangeduid als de "intention-to-treat" analyse. De analyse, met weglating van o.a. uitvallers, staat bekend als de "per protocol" analyse en ook wel als de analyse van evalueerbare patiënten.

De algemene gedachtengang achter de "intention-to-treat" analyse kan als volgt geïllustreerd worden. Stel dat alle gerandomiseerde patiënten voldoen aan álle inclusie criteria, daarbij alle procedures in de studie tot het einde hebben gevolgd en geen enkele controle in de kliniek hebben gemist, zodat er geen ontbrekende gegevens zijn. Dan kan de analyse volgens plan worden uitgevoerd met alle gerandomiseerde patiënten. De ideale studie, waarnaar altijd wordt gestreefd ! Maar meestal steekt Murphy ergens wel een spaak in het wiel, geheel in overeenstemming met de beide door hem geformuleerde hoofdwetten.
De eerste luidt: "In any field of scientific endeavour, anything that can go wrong, will go wrong". En de tweede: "If everything seems to be going well, you have obviously overlooked something!"

Vandaar dat voor een fase III studie steeds een "intention-to-treat" analyse met alle gerandomiseerde patiënten wordt gevraagd onder het motto: "eenmaal gerandomiseerd, altijd geanalyseerd".
Deze pragmatische aanpak, die gericht is op het vermijden van "bias", wordt o.m. geprefereerd omdat het resultaat beter zal aansluiten bij de verwachte uitkomst in de gangbare medische

praktijk. Uiteraard wordt hiermee de "power" van de studie gereduceerd. Dat is de premie die verlangd wordt voor het feit dat het om patiënten gaat, en niet om proefpersonen of proefdieren. Vandaar dat vaak ook een per-protocol analyse wordt uitgevoerd, die gebaseerd is op alleen die patiënten welke voldoen aan de criteria van de "ideale" studie. De studie wint in het algemeen aan betrouwbaarheid naarmate het percentage patiënten dat voor de per-protocol analyse is weggelaten, kleiner is, en het verschil tussen de conclusies uit beide analyses geringer is.

In een exploratieve fase II studie wordt in het analyseplan de toe te passen methodologie alleen in grote lijnen beschreven. Maar voor een fase III studie is een gedetailleerde beschrijving nodig, inclusief de wijze waarop diverse problemen, zoals protocolafwijkingen, zullen worden gehanteerd, bijv. uitvallers, partieel ontbrekende gegevens en uitschieters. Dat laatste leidt doorgaans tot minstens twee analyseprocedures, volgens een pragmatische aanpak via "intention-to-treat" analyse en een verklaringsgerichte aanpak via "per protocol" analyse. Als deze analyses tot duidelijk verschillende resultaten leiden, wordt veelal een derde analyse toegevoegd die als een tussenvorm van deze twee kan worden beschouwd. In een fase III studie krijgt dan de "intention-to-treat" analyse de hoogste prioriteit.

Interim analyse

Laten we nu de overstap maken naar de interim analyse.

In sommige klinische studies beslaat de instroom van patiënten een lange periode. Als de observatieduur per patiënt relatief kort is, zal halverwege die periode dus de uitkomst al bekend zijn voor

ongeveer de helft van het geplande aantal patiënten. Het ligt voor de hand dat de onderzoeker, na alle inspanning bij voorbereiding en uitvoering van de studie, méér dan nieuwsgierig zal zijn naar het verschil in resultaat bij de tot dusver behandelde groepen patiënten. Om die nieuwsgierigheid te bevredigen zal hij de nu verkregen resultaten willen analyseren. Echter, ongecoördineerd tussentijdse analyses uitvoeren, om op grond van de uitslag de studie eventueel al of niet te continueren, brengt een aantal risico's met zich. Eén van die risico's houdt verband met de P-waarde.

Stel dat in een vergelijkende klinische studie het échte verschil $\mu(A-B)$ inderdaad nul is, m.a.w. de nulhypothese is correct. De vraag is dan, hoe groot is de kans dat bij herhaald tussentijds analyseren, toch onterecht een significant verschil zal worden gevonden. D.w.z. minstens éénmaal een P-waarde kleiner dan de significantie-drempel α van 5%. Bijvoorbeeld voor 4 tussentijdse en 1 eindanalyse blijkt die kans al te zijn opgelopen naar 14%, in plaats van de beoogde 5%. Tabel 6 geeft een indruk van de toename van die kans bij toenemend aantal tussen-analyses.

Tabel 6

Risico m.b.t. inflatie van de P-waarde bij herhaald tussentijds toetsen van een correcte nulhypothese: $\mu(A-B) = 0$.

Aantal toetsen, steeds bij $\alpha = 0,05$	Kans op verwerpen van correcte nul-hypothese (in %)
1	5
2	8
3	11
4	13
5	14
10	19
20	25

Uit de tabel blijkt dat die kans na bijvoorbeeld 20 analyses is gegroeid naar 25%.

Een dergelijk risico kan worden omzeild door bij elke toetsing als grens voor de P-waarde een lagere drempelwaarde aan te houden. Tabel 7 geeft u een indruk waar die grens ligt afhankelijk van het aantal analyses.

Tabel 7

Nominale significantie drempel, zodat de kans op verwerpen van een correcte nul-hypothese bij herhaald tussentijds toetsen 5% blijft.

Aantal analyses	Nominale significantie drempel
2	0,029
3	0,022
4	0,018
5	0,016

Deze aanpak houdt in dat tevoren in het protocol moet worden

vastgelegd hoeveel interim analyses maximaal voorzien worden, met de bijbehorende nominale drempel voor de P-waarde bij elke analyse. Recent biostatistisch onderzoek heeft er toe geleid dat in plaats van zo'n rigide planning, een veel flexibeler aanpak mogelijk is geworden door een zg. "α-spending function" toe te passen. Die functie dient uiteraard ook in het protocol gespecificeerd te worden. Op de vraag naar de beste vorm voor deze functie, kan ik in verband met de tijd niet nader ingaan.

Het doel van geplande interim-analyses is op een zo vroeg mogelijk tijdstip verantwoord te kunnen stoppen met de studie wegens drie mogelijke redenen.
De eerste reden, als overtuigend is aangetoond dat de behandelingen A en B significant verschillen t.a.v. de primaire uitkomstvariabele. De tweede, als het aantonen van een klinisch relevant verschil in deze studie niet meer haalbaar lijkt te zijn. En de derde, eventueel wegens onverwachte bijwerkingen. Daarbij dient opgemerkt te worden dat, zeker in een dubbelblinde studie, elke interim analyse door een onafhankelijke biostatisticus moet worden uitgevoerd. De onderzoekers worden vervolgens alléén geïnformeerd of stoppen al of niet verantwoord is, volgens de vooraf geplande stopregel. Tot zover deze beschouwing over de biostatistische richtlijn in de Europese Unie.

EEN KORTE BLIK IN HET STATISTISCH "ARSENAAL"

Zoals ik aan het begin al aangaf, er zijn legio statistische onderwerpen die in het medisch wetenschappelijk onderzoek een specifieke impact hebben.
Ik kan in de tijd die nog rest voor dit college, nog hoogstens enige

ervan kort aanstippen. Ten eerste geldt dat de recente ontwikkelingen in de te volgen methodologie bij de selectie van prognostische variabelen, uit de veelheid van potentiële kandidaten. Het gaat daarbij vooral om het gericht toetsen van de validiteit en de betrouwbaarheid van het daarop gebaseerde statistische model.
Een ander thema betreft het analyseren van longitudinale onderzoeksresultaten, waar gedurende de follow-up bij elke patiënt de uitkomstvariabele herhaaldelijk is gemeten. Veel naïeve analyse-methoden, die in medische publicaties nog té vaak worden aangetroffen, houden onvoldoende rekening met het longitudinale karakter van de gegevens. Voor een adequate analyse komen diverse geavanceerde statistische modellen in aanmerking. De beste keuze daaruit hangt samen met een aantal factoren. Onder meer of het eventueel optreden van uitvallers gerelateerd is aan de behandeling. Deze meer complexe methoden zijn overigens alleen door professionele statistici optimaal toe te passen en slechts dank zij de beschikbaarheid van zeer geavanceerde statistische programmatuur.

Een ander belangrijk aspect betreft de reproduceerbaarheid van de meetmethode in een klinische studie. Naarmate die methode minder goed reproduceerbaar is, zijn er meer patiënten per groep nodig voor dezelfde "power" van de studie. Als te voorzien is dat de reproduceerbaarheid nogal matig zal zijn, is het raadzaam vooraf een daarop gerichte studie uit te voeren. Veronderstel dat uit de resultaten van zo'n onderzoek is gebleken dat de reproduceerbaarheidsvariantie bijvoorbeeld 22% bedraagt van de totale variantie tussen patiënten per groep op basis van de gemeten waarden.
Zou men nu besluiten om in de klinische studie elke meting onafhankelijk door een ander te laten herhalen en beide waarden te

middelen, dan zal dit percentage dalen van 22% naar 12%.
Daaruit valt af te leiden dat het vereiste aantal patiënten in de studie dan zou dalen met 11%.
Een bescheiden winst in vergelijking met die in het eerste voorbeeld, maar wellicht de extra inspanning waard.

SAMENVATTING

Laat ik proberen het voorgaande nu kort samen te vatten. In dit college heb ik vooral getracht U enig inzicht te geven in statistische aspecten van prospectief vergelijkende klinische studies en iets van de ontwikkelingen in de afgelopen 30 jaar. Ik hoop dat ik U enigszins duidelijk heb kunnen maken dat de biostatistiek - alleen al binnen dit beperkte kader - beschikt over een heel arsenaal van methoden welke, mits oordeelkundig toegepast, een efficiënte en effectieve aanpak van klinische vraagstellingen mogelijk maken, zowel in het ziekenhuis als in het laboratorium. Het streven daarbij is het aantal individuen dat in een studie zal worden betrokken zoveel mogelijk te beperken, met behoud van de "power" van de studie. Zodoende worden minder mensen, patiënten en proefpersonen, resp. proefdieren belast met interventies, waarvan achteraf soms zal blijken dat deze deels onvoldoende rendement hebben. Dat vergt echter vooraf expliciete overweging van o.m. alle kennis omtrent prognostische variabelen en storende bronnen van variabiliteit.

Vanwaar deze loftuiting?

Wellicht hebt u zich tevoren afgevraagd of ik in de voetsporen van Erasmus zou treden en de biostatistiek met zotheid zou vereenzelvigen. Ik neem aan dat het U inmiddels duidelijk is

geworden dat de lof echt letterlijk en inhoudelijk bedoeld is.
Mogelijk vraagt u zich dan nu af, of het niet méér in de lijn had gelegen, iemand anders te vragen om de loftrompet te hanteren. Dat is ook gebeurd, zoals uit het volgende citaat mag blijken. Het betreft de laatste zin in een uitzending op 6 juli jl. van de "Open University" voor de BBC over de "History of Clinical Trials". "Biostatistics has a major influence on everybody's health and well-being". Deze uitspraak was de voornaamste aanleiding tot de titelkeuze voor dit college.

Slotwoord

Tot slot zou ik nog graag een aantal personen willen bedanken.

In de eerste plaats gaat mijn dank uit naar mijn vroegere leermeester, wijlen Professor Huug Hamaker, die mijn interesse voor de statistiek heeft gewekt, voor de theoretische basis, maar vooral voor de toepassingen ervan op practische probleemstellingen. Ik betreur het dat hij er niet meer is.
Ten tweede bedank ik degenen die zich beijverd hebben om deze leerstoel onverkort gehandhaafd te houden. Ruim 23 jaar nadat Hare Majesteit de Koningin, nu Prinses Juliana, mij benoemde en ik als eerste deze plaats mocht bezetten, draag ik de leerstoel graag over aan mijn opvolger.
Voorts dank ik mijn vroegere en mijn huidige medewerksters en medewerkers, zowel die tijdens mijn loopbaan in het bedrijfsleven, als degenen daarna bij mijn functioneren aan deze faculteit. Zij allen hebben ertoe bijgedragen dat ik elke werkdag met steeds evenveel plezier naar mijn werkplek ben gegaan.
De groep biostatistici in de faculteit is in de loop der jaren

uitgebreid van drie naar zeven. Het aantal onderzoeksprojecten in faculteit en ziekenhuis waarin de groep jaarlijks participeert is gestegen tot boven de 200. Het aantal publicaties met een biostatisticus als co-auteur, in vooral klinische, internationale tijdschriften, is gegroeid tot ca. 60 per jaar. Ik beschouw het als een groot voorrecht dat ik hieraan mijn steentje heb mogen bijdragen.

Tenslotte, "last but not least", wil ik mijn vrouw bedanken voor de "support" en toewijding, die ik gedurende nu al meer dan 42 jaar van haar heb mogen ondervinden. Ik hoop dat we nog lang gezond zullen blijven en genieten van het leven, samen met onze kinderen en kleinkinderen.
Rest mij nog U allen te danken voor Uw aanwezigheid.
Ik heb gezegd.