STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

LEIDING: PROF. DR D. VAN DANTZIG
ADVISEUR VOOR STATISTISCHE CONSULTATIE: PROF. DR J. HEMELRIJK

Rapport S 235 (V 21)

Statistische proefopzetten: bewijs en detectie

door

Prof. Dr J. Hemelrijk

maart 1958

Statistische proefopzetten: bewijs en detectie

De experimenten, waarmee statistici in aanraking komen, zijn gewoonlijk modellen van een klein stukje werkelijkheid. Zij berusten op vereenvoudigingen of analogieën en doen daarom de werkelijkheid min of meer geweld aan, omdat deze - naar men zou kunnen zeggen - niet altijd bereid is haar geheimen goedschiks prijs te geven en daartoe dus met zachte drang en kunstgrepen gedwongen moet worden.

De vereenvoudigingen, die veelal bestaan uit het kunstmatig scheppen van in alle opzichten constante en bekende omstandigheden, dienen om de te onderzoeken effecten en wetmatigheden duidelijker naar voren te doen komen. Onder de in werkelijkheid optredende omstandigheden (" in de practijk") kunnen vele bijkomstige factoren hun invloed doen gelden en het te onderzoeken wetmatigheidspatroon verdoezelen of zo zeer compliceren, dat het niet ineens te ontrafelen is. Deze vereenvoudigingen zijn dus essentieel bij een onderzoek en één der eerste stappen is dan ook vast te stellen welke vereenvoudigingen zullen worden aangebracht. Het werken met analogieën (b.v. met proefdieren in plaats van mensen) berust gewoonlijk op practische noodzaak. Zij brengen een extra onzekerheid in het experiment, nl. "of de analogie wel opgaat", en men zou ze liever vermijden. Voor experimenten, die op analogieën berusten geldt sterker nog dan voor andere, dat zij slechts door bevestiging in de practijk tot definitieve conclusies kunnen leiden.

Bij de vereenvoudiging van werkelijkheid tot experiment moet een compromis gevonden worden tussen twee tegenstrijdige doelstellingen. Vereenvoudigt men te sterk dan kunnen bepaalde effecten geheel verdwijnen, zodat het in feite onderzochte wetmatigheidspatroon te weinig overeenkomst meer vertoont met het werkelijke om tot voldoende generaliseerbare conclusies te kunnen leiden. Vereenvoudigt men niet genoeg, dan kan het patroon, dat dan het experiment beheerst, zo ingewikkeld blijven, dat men er niet uit komt en geen conclusies getrokken kunnen worden. Het is dan een schrale troost dat deze conclusies, als men ze wel verkregen had, een grotere algemene geldigheid zouden hebben gehad.

De kunst is dus te navigeren tussen deze beide klippen van duidelijke zichtbaarheid en eenvoud van de onderzochte effecten enerzijds en van voldoende generaliseerbaarheid anderzijds. Bij deze navigatie kan de statisticus (of althans de statistiek) als loods de koers uitzetten.

Ter illustratie beschouwen wij een eenvoudige situatie, die vele kenmerken van het typisch statistische onderzoek inhoudt. Om niet in een te specifieke gedachtegang te geraken houden wij de terminologie algemeen. Laten A en B twee middelen (of methoden) zijn, die, toegepast in een bepaald stadium van een procédé een kwantitatief meetbaar effect hebben. Men kan daarbij denken aan een fabricage-procédé of, zo men wil, aan een willekeurig ander proces. De kwantitatieve uitwerkingen van A en B geven wij aan met $\underline{x}$ resp. $\underline{y}$; dit zullen in de regel geen constanten, maar stochastische grootheden zijn. In de practijk zullen (of zouden), naar wij verder veronderstellen, de beide middelen A en B onder allerlei verschillende uitwendige omstandigheden worden gebruikt. Voor de eenvoud nemen wij aan, dat er slechts eindig veel van deze omstandigheden zijn, aan te geven met $U_1,\ldots,U_k$. Zij hebben invloed op de uitwerkingen $\underline{x}$ en $\underline{y}$, zodat wij deze moeten voorstellen door $\underline{x}_1,\ldots,\underline{x}_k$ resp. $\underline{y}_1,\ldots,\underline{y}_k$. Het gaat nu om het verschil, $\underline{x}-\underline{y}$, tussen de uitwerking van A en van B; preciezer gezegd om de verschillen $\underline{x}_i-\underline{y}_i$ $(i=1,\ldots,k)$ onder de verschillende omstandigheden.

Nu kunnen zich, in werkelijkheid, 3 verschillende situaties voordoen.

I De omstandigheden U_i hebben dezelfde invloed op $\underline{x}$ als op $\underline{y}$ en dus geen invloed op het verschil $\underline{x}-\underline{y}$; de indices i kunnen dus, voor zoverre het de verschillen $\underline{x}_i-\underline{y}_i$ betreft, wel weggelaten worden.

II De omstandigheden kunnen wel invloed hebben op $\underline{x}-\underline{y}$, maar alleen op de grootte van het verschil, niet op de richting. De kansverdelingen der $\underline{x}_i-\underline{y}_i$ zijn dus wel verschillend, maar hun verwachtingen (b.v.) zijn ofwel alle =0 ofwel hebben alle hetzelfde teken.

III De omstandigheden U_i kunnen niet alleen invloed hebben op de grootte van het verschil, maar ook op de richting. De ver-

wachtingen der verschillen $\underline{x}_i - \underline{y}_i$ kunnen verschillende tekens hebben.

Weet men van tevoren, welke van deze drie mogelijkheden vervuld is, dan kan men het experiment dienovereenkomstig opzetten.

In het eerste geval, inhoudende dat de U_i wel invloed hebben op de $\underline{x}_i$ en de $\underline{y}_i$ afzonderlijk, maar niet op hun verschil, kan men dan gemakkelijk aangeven welke vereenvoudiging aangebracht kan worden om de zichtbaarheid van het verschil te vergroten zonder de geldigheid der conclusie te schaden. Het is ieder statisticus bekend dat men in dat geval voor toetsing het grootste onderscheidingsvermogen en voor schatting de grootste nauwkeurigheid verkrijgt, indien men alle proeven neemt onder éénzelfde omstandigheid U en ze gelijkelijk verdeelt over de middelen A en B.

In de beide andere gevallen is de situatie gecompliceerder en valt niet zonder meer een eenvoudige stelregel voor de verdeling der proeven over de verschillende omstandigheden aan te geven. Wel heeft de statisticus een groot arsenaal van proefschema's tot zijn beschikking, waarmee hij kan trachten in deze gevallen "hoofdeffect" en "interactie" van elkaar te scheiden. Deze kunnen, bij een juiste proefopzet, afzonderlijk geschat en getoetst worden onder eliminatie van het effect der U_i op de $\underline{x}$ en $\underline{y}$ afzonderlijk. De keuze der proefopzet zal nu echter niet alleen beïnvloed worden door voorkennis omtrent de invloed der omstandigheden maar ook door het doel van het onderzoek. Kent men de verhoudingen, waarin de U_i zich in de praktijk voordoen, en is men hoofdzakelijk geïnteresseerd in het gemiddelde verschil onder practijk-omstandigheden, dan zal men wellicht de verdeling der proeven over de U_i in dezelfde verhouding kiezen. Heeft men echter niet alleen bij het experiment, maar ook in de practijk de omstandigheden in de hand, dan zal men zich wellicht speciaal interesseren voor de keuze van gunstige omstandigheden, waartoe men hoofdeffect en interactie afzonderlijk, en beide zo nauwkeurig mogelijk zal willen onderzoeken.

Het is niet noodzakelijk verder op dit voorbeeld in te gaan om tot de conclusie te komen, dat niet alleen het doel van het onderzoek en de voorkennis, voor zoverre aanwezig, van belang is bij de keuze der proefopzet, maar ook het arsenaal van statistische

methoden, dat de onderzoeker of de door hun geconsulteerde statisticus tot zijn beschikking heeft.

Want het heeft geen zin een koers uit te zetten, die alleen geschikt is voor een motorschip en deze dan te gaan varen met een zeilschip. Toch gebeurt dit, in figuurlijke zin dan, maar al te vaak. En het gevolg is dan - gelukkig niet altijd, maar toch veelvuldiger dan gewenst is - dat de statisticus achteraf gevraagd wordt te redden wat er te redden valt. De loods wordt dan aan boord gehaald als het schip al op de klippen zit en eigenlijk alleen nog geschikt is voor de sleepboot en de sloop.

Vrijwel iedere statisticus komt van tijd tot tijd in aanraking met een onderzoeker op ander gebied, die zijn hulp komt inroepen bij de analyse van reeds verzameld waarnemingsmateriaal. In de loop van het gesprek komt dit materiaal ter tafel en op de vraag, wat er nu eigenlijk mee moet gebeuren, met welk doel de gegevens zijn verzameld, wordt lang niet altijd een duidelijk en ondubbelzinnig antwoord verkregen. Veelal verwacht men van de statisticus dat deze "eruit zal halen wat erin zit" en soms blijkt zelfs, dat men hem, alsof hij een goochelaar is, die een konijn uit een lege hoed kan halen, verzoekt eruit te halen wat er niet in zit. Dit laatste dient hij beleefd, maar beslist te weigeren. Het eerste kan hij proberen, al is het gevaar aanwezig dat hij per ongeluk toch in het laatste vervalt. Statistische formules, die door leken vaak voor een soort toverformules aangezien worden, kunnen nl. de statisticus nolens, en misschien soms ook wel eens volens, maar al te gemakkelijk tot een goochelaar maken.

De redenen hiervan zijn velerlei. In de eerste plaats werkt de statisticus niet zonder meer met het materiële model van de werkelijkheid, dat de onderzoeker in de vorm van een experiment geschapen heeft, maar hij bootst dat model weer na met een wiskundig model, dat verdere vereenvoudigingen inhoudt: normaliteit, onafhankelijkheid, gelijke spreidingen, additiviteit van effecten, en andere veronderstellingen ten behoeve van de wiskundige analyse. Vooral bij niet zorgvuldig op bepaalde statistische verwerkingsmethoden ingestelde proefopzetten schuilt in deze verdere vereenvoudigingen het gevaar van onjuiste conclusies. Zo kan een toets tot ver-

werping van een getoetste hypothese leiden, terwijl eigenlijk de extra veronderstellingen verworpen zouden moeten worden. In de tweede plaats echter, en hier schuilt een veel groter gevaar, worden zeer vaak naast of in plaats van de vragen, waarvoor het experiment is opgezet, andere vragen onderzocht, die door het waarnemingsmateriaal gesuggereerd worden. Dit geschiedt vooral vaak bij uitgebreid materiaal, waarin bijna altijd wel eigenaardigheden voorkomen, die men niet zonder meer voorbij wil gaan. Het onderzoek van die eigenaardigheden leidt dan b.v. tot de berekening van overschrijdingskansen en als deze klein zijn vermeldt men de bijbehorende conclusies, meestal zonder ze - wat hun betrouwbaarheid betreft - te onderscheiden van de conclusies omtrent de oorspronkelijk gestelde vragen.

Deze procedure is, zoals eigenlijk iedereen wel weet, een uiterst hachelijke. Want het is bijna onmogelijk cijfermateriaal van enige omvang te verkrijgen, waarin geen eigenaardigheden optreden. Vooral bij zeer uitgebreid materiaal is dit het geval, daar ook zeldzame gebeurtenissen wel moeten optreden, als men maar genoeg waarnemingen doet. Anders zouden zij niet zeldzaam, maar onmogelijk zijn.

Een duidelijk voorbeeld hiervan is te vinden in de zorgvuldige pogingen, die door verschillende eminente statistici op grote schaal zijn verricht, om cijfermateriaal te vervaardigen, waar - althans in bepaalde opzichten - niets in zit: de zo uiterst nuttige aselecte getallen. Het klinkt wellicht paradoxaal, dat het zeer nuttig zou zijn cijfermateriaal te vervaardigen, waar niets in zit, maar de essentiële onmisbaarheid daarvan bij de opzet van experimenten is alle statistici bekend en behoeft geen commentaar. Niet paradoxaal, maar weinig aannemelijk, lijkt het op het eerste gezicht, dat het zó moeilijk zou zijn een dergelijke rij van aselecte getallen te vervaardigen. Toch is dit verre van eenvoudig, zoals uit het volgende mogen blijken.

Naar aanleiding van parapsychologische proeven in Engeland en Amerika, die vooral in Engeland sterk de aandacht hebben getrokken en waarbij op grote schaal van aselecte getallen gebruik

werd gemaakt, zijn door twee Engelse onderzoekers, G. SPENCER BROWN [1] en A.T. ORAM (de eerste een "ongelovige" en de tweede - vermoedelijk - een "gelovige" met betrekking tot parapsychologische verschijnselen), de gangbare tabellen van aselecte getallen, vervaardigd door TIPPETT en door KENDALL en BABINGTON SMITH, onderzocht alsof zijzelf de resultaten van een parapsychologisch experiment weergaven. Daarbij werd de bij parapsychologische onderzoekingen gebruikelijke methode gecopieerd: behalve de oorspronkelijk bedoelde tellingen van "treffers" in paren cijfers (waarbij alleen naar even en oneven werd onderscheiden) werden ook treffers bij verschuivingen e.d. onderzocht. In totaal was echter het aantal getoetste hypothesen niet zeer groot, vermoedelijk niet groter dan 10 of 20. Niettemin werden, tot kennelijke voldoening van de eerste der genoemde onderzoekers en tot verbazing van de tweede, naast grote ook kleine overschrijdingskansen gevonden; de kleinste bedroeg 0,00014.

De moraal hiervan is, dat men vooral bij onderzoekingen, die zo uitgebreid zijn, dat reeds geringe effecten tot kleine overschrijdingskansen leiden, zeer voorzichtig moet zijn in het trekken van conclusies. Herhaalde en onafhankelijke bevestiging van reeds gevonden resultaten - en niet onder weglating van mislukte experimenten - is daarvoor onmisbaar.

Aan welke eisen moet nu, statistisch gezien, een experiment voldoen om tot een definitieve uitspraak, een statistisch "bewijs", van een bepaalde conclusie te geraken? De term "bewijs" is eigenlijk a priori al te sterk, daar iedere statistische conclusie met een zekere onbetrouwbaarheid behept is. Wij verstaan daarom hier onder een statistisch bewijs een conclusie, waarvan de onbetrouwbaarheid inderdaad gelijk is aan de opgegeven waarde en niet, om één der bovengenoemde of om een nog andere reden, onbekend en wellicht vele malen groter dan de opgegeven waarde.

Zonder naar volledigheid te streven noemen wij hieronder enkele regels van algemeen karakter, die zeker vervuld dienen

---

1) G. SPENCER BROWN, Probability and scientific inference, Longmans, London - New York - Toronto 1957.

te zijn.

A. Het doel van het onderzoek dient van tevoren duidelijk te worden geformuleerd en tijdens het onderzoek niet te worden veranderd.

B. De te verrichten experimenten dienen van tevoren zorgvuldig te worden beschreven en hiervan dient niet te worden afgeweken.

C. De statistische analysemethoden dienen van tevoren te worden vastgesteld en hiervan dient niet te worden afgeweken.

D. Uit een reeks proeven, uitgevoerd volgens plan, dienen alleen conclusies te worden getrokken omtrent de in de doelstelling geformuleerde vragen. Met hun aantal en hun onderlinge relaties dient bij de keuze van de onbetrouwbaarheid, die voor iedere conclusie afzonderlijk wordt toegelaten, rekening te worden gehouden.

E. In die stadia der proeven, waarin onbekende of onberekenbare factoren van invloed kunnen zijn, dient statistische aselectering (loting) te worden toegepast om het optreden van "schijneffecten" (en "verdoezelingen") te vermijden.

De onvolledigheid van deze lijst blijkt o.a. reeds uit het ontbreken van eisen omtrent de boven gesignaleerde extra veronderstellingen, die aan de gekozen statistische analysemethoden ten grondslag liggen. Ook overigens zouden ongetwijfeld nog meer eisen geformuleerd kunnen worden. Aan de bovengenoemde is echter reeds zo zelden voldaan, dat wij er voor het ogenblik meer dan genoeg aan hebben.

Is wel aan al deze eisen voldaan en is ook overigens op het experiment niets aan te merken, dan blijft nog het probleem van extrapolatie in de tijd. Een conclusie, die alleen geldt voor het tijdstip van het experiment en die daarna zijn geldigheid ieder moment verloren kan blijken te hebben, is van weinig waarde. Hierdoor wordt nog een nieuwe eis toegevoegd:bij herhaald onderzoek van dezelfde vraag moet de conclusie, ook door andere onderzoekers, bevestigd kunnen worden. Dit is één van de redenen, waarom vaak het herhalen van reeds door anderen verrichte onderzoekingen verre van zinloos is. Uiteindelijk zal slechts bevestiging door de

practijk, onder niet vereenvoudigde omstandigheden, het onderzoek volledig kunnen afsluiten [2].

Is niet aan alle bovenstaande eisen voldaan - en dit is eer regel dan uitzondering - dan verliest een onderzoek daarmee gelukkig niet alle waarde, maar het verliest wel de kracht van bewijs. Wat overblijft zouden zij willen kwalificeren als detectie. Indien een statisticus waarnemingsmateriaal onderzocht op eigenaardigheden, die hem op het spoor van oorspronkelijk onvermoede effecten zouden kunnen brengen, kunnen zijn statistische analysemethoden hem helpen te onderscheiden tussen eigenaardigheden, die heel goed toevallig kunnen zijn ontstaan - en die dus weinig belovend zijn - en andere, die slechts zelden toevalligerwijze zouden ontstaan. Deze laatste worden aangewezen door het vinden van kleine overschrijdingskansen en na aanvulling door schattingen van deze eventuele effecten verkrijgt men dan veelal vruchtbare aanwijzingen voor verder, bevestigend (of ontkennend) onderzoek. Dit verdere onderzoek zal vaak geheel anders van opzet zijn dan het oorspronkelijke, omdat de doelstelling nu een andere is. Soms zal het niet eens meer een statistisch onderzoek zijn, maar b.v. een zuiver technisch of zelfs, in bepaalde situaties, een juridisch onderzoek. Het is echter steeds onmisbaar, wil men van het niveau van detectie terug naar dat van het bewijs.

Beide aspecten, detectie en bewijs, laten zich vaak zeer goed in één experiment verenigen in de vorm van een vooronderzoek van beperkte omvang - voor detectie - op grond waarvan een aan de boven gegeven strenge regels onderworpen definitief onderzoek kan worden opgezet. De doeltreffendheid van een dergelijke proefopzet is vaak aanzienlijk groter dan van een even groot of groter onderzoek zonder vooronderzoek. Bij publicatie dient dan niet alleen het uiteindelijke experiment, maar ook het vooronderzoek en de analyse daarvan beschreven te worden. Immers de wijze van ontstaan der onderzochte vragen - vóór het begin van het definitieve onderzoek - draagt aanzienlijk bij tot de bewijskracht van dit laatste.

---

2) Een fraai voorbeeld van een eenvoudig, maar zeer belangrijk en goed opgezet, experiment werd door Dr Chr.L. RÜMKE besproken in zijn Openbaar College: "De taak van de medische statistiek", Statistica Neerlandica 12 (1958) 1-16, bl. 4 en 5.

Aan het onderscheid tussen detectie en bewijs, wordt bij statistische ond rzoekingen en publicaties te weinig aandacht besteed. Het is voor de beoordeling van de overtuigingskracht van een statistische conclusie van groot belang te weten of de vraag van te voren was gesteld of tijdens de analyse en naar aanleiding van het waarnemingsmateriaal naar voren is gekomen. Het verdient daarom aanbeveling dit te vermelden en ook een chronologisch overzicht van het onderzoek te schetsen. Analoge beschouwingen gelden voor de aangebrachte vereenvoudigingen en de gebruikte veronderstellingen. Zijn deze ad hoc aangebracht en opgesteld, dan geven zij meer reden tot voorzichtigheid bij de interpretatie dan wanneer zij van tevoren zijn overwogen en het experiment zo is opgezet, dat hun toelaatbaarheid op grond van het onderzoek bevestigd of althans onderzocht kan worden.

(De ongebruikelijke nummering der figuren en bladzijden - bl. 10 en fig. 1 ontbreken - is een gevolg van stencil-technische omstandigheden: bl. 11-19 zijn ontleend aan een ander rapport, S 228 (V 18), van de Statistische Afdeling van het Mathematisch Centrum).

Ter illustratie bespreken wij [3] een passage uit het proefschrift van K.H. BRANDT: "Over de plaats van vorming der urobilinogenen in het menselijk organisme", Utrecht 1957.

De passage betreft het verschil tussen de z.g. B-gal, die wordt uitgescheiden na de galblaas te zijn gepasseerd, en de C-gal welke rechtstreeks uit de lever komt.

Doordat de gal in de galblaas door resorptie van water wordt ingedikt is bij eenzelfde persoon de bilirubineconcentratie in de B-gal aanmerkelijk hoger dan in de C-gal. Bij 21 patiënten [4] van BRANDT varieerde de verhouding van beide concentraties van ruim 1,5 tot 50 [5]. Ook voor de urobilinogenen vond BRANDT bij al deze personen, op één uitzondering na, de hoogste concentratie in de B-gal.

Van de waarnemingen, bij deze 21 personen verricht (l.c., p. 116-117 [6]), zijn enkele gegevens in tabel 2 overgenomen. Daarbij zijn concentraties, zoals gebruikelijk, door vierkante haken aangegeven.

---

3) De volgende beschouwingen zijn ontwikkeld in samenwerking met Prof.Dr C.G.G. VAN HERK, medewerker van de Statistische Afdeling van het Mathematisch Centrum, Amsterdam.

4) Bijna allen hadden een onbelaste anamnese wat betreft lever- en galwegen; één patiënt had een steen, één had een hepatitis infectiosa doorgemaakt, maar tijdens het onderzoek een goede leverfunctie.

5) Misschien zijn deze zeer uiteenlopende getallen ten dele toe te schrijven aan een meer of minder sterke vermenging van gal met vocht in het duodenum. Zolang dit vocht zelf geen galkleurstoffen bevat is een dergelijke verdunning voor de verdere discussie van geen belang..

6) Om ons onbekende redenen zijn de gegevens over patiënt 21 hier niet vermeld. Deze zijn ons door de schrijver verstrekt en ook in deze beschouwingen opgenomen. Weglating zou echter geen essentiële wijziging tot gevolg hebben. De tabel bevat enkele onbetekende rekenfouten.

Tabel 2

Concentraties van urobilinogenen (in mg/100 cc) en bilirubine (in E/100 cc) in B- en C-gal

nr. patiënt	$[ur]_B$	$[ur]_C$	$[bil]_B$	$[bil]_C$
1	0,37	0,06	96,96	15,70
2	0,108	0,005	99,26	2,82
9	0,312	0,208	98,7	8,7
12	1,38	0,57	46	9,74
19	0,01	0,005	20,6	13,1

Van belang is verder dat geen bilirubine in de galblaas wordt gevormd of afgescheiden. Wel kan deze stof daar verdwijnen, althans onder pathologische omstandigheden, maar voor ons betoog speelt dit (gelukkig) geen rol.

Voor iedere patiënt beschouwt BRANDT nu de beide quotiënten:

(9) $$q_{bil} = [bil]_B / [bil]_C \ , \ q_{ur} = [ur]_B / [ur]_C \ .$$

Indien in de galblaas alleen indikking plaats vindt is er geen reden waarom deze quotiënten onderling zouden verschillen, behalve dan door onvermijdelijke onnauwkeurigheden bij de chemische analyse [7]. Zou echter blijken dat q_{ur} systematisch kleiner is dan q_{bil}, dan zou dit wijzen op resorptie of afbraak van urobilinogenen in de

---

7) Met een mogelijke storende invloed van het leverrhythme op de samenstelling van de gal is geen rekening gehouden. Hierover hebben wij nl. geen gegevens kunnen verkrijgen. JORES (Tabulae biologicae 14 ('37) p. 95-98) geeft een dagkromme voor het bilirubinegehalte van het bloed (maximum om ca. 0^h, minimum ca. 4^h) en voor de urobilinogeenuitscheiding in de urine (maximum ca. 12^h, minimum ca. 4^h), maar spreidingen worden daarbij niet vermeld. Bij het gebruikelijke onderzoek worden de B- en C-gal op verschillende delen van een etmaal gesecerneerd. Voor de statistische aspecten van deze discussie is dit overigens van geen belang.

galblaas. Omgekeerd zou door uitscheiding of aanmaak van urobilinogenen (al dan niet bilirubine) in de galblaas, q_{ur} systematisch groter worden dan q_{bil} [8].

Statistisch bezien komt het er dus op neer of van de beide quotiënten het ene bij een voldoend aantal patiënten voldoende kleiner is dan het andere. Dit houdt nl. een aanwijzing in voor een systematisch verschil, en daarmee voor ten minste één der verschijnselen: resorptie, aanmaak etc. van hetzij urobilinogenen, hetzij bilirubine in de galblaas.

In fig. 2 zijn de beide concentratiequotiënten voor deze 21 patiënten grafisch tegen elkaar uitgezet (zie diss. p. 131).

Ligt een punt in deze figuur boven de bissectrice door de oorsprong, dan is $q_{ur} > q_{bil}$, terwijl voor punten eronder het omgekeerde geldt. Geen der punten ligt precies op deze lijn, wat trouwens ook niet kon worden verwacht.

Van de 21 punten liggen er nu 6 boven en 15 onder de bissectrice. Evenals bij een eerder genoemd voorbeeld verkeren wij daarom in een situatie waarin geen voldoende aanwijzingen voorhanden zijn voor een systematisch verschil, hoewel fig. 2 toch wel een vermoeden in die richting kan doen ontstaan.

Nu wijken een aantal punten, zowel boven als onder de bissectrice, ver van deze lijn af. Dit doet vermoeden dat de nauwkeurigheid der quotiënten gering is. Immers als in de galblaas alleen indikking plaats vindt, zijn alle afwijkingen van de bissectrice aan onnauwkeurigheden van de analyse te wijten. Maar ook als dit niet het geval is en er b.v. ook afbraak van urobilinogenen plaats vindt, zijn deze punten met een redelijk uniforme galblaasfunctie van de onderzochte personen onverenigbaar. De verhouding q_{ur}/q_{bil} varieert van 0,132 tot 5,74, zodat de grootste waarde hiervan ruim 43 maal de kleinste bedraagt. Voor personen met ogenschijnlijk normale lever en galwegen, - zij het ook met mogelijke andere afwijkingen - , lopen deze getallen toch wel erg ver uiteen om zonder meer aan functieverschillen te worden toegeschreven. Er is trouwens een veel meer voor de hand liggende verklaring.

Over de nauwkeurigheid der bepalingen waren wij niet ingelicht,

---

8) Dit zou ook het geval zijn bij afbraak of selectieve resorptie van bilirubine aldaar.

vermoedelijk was deze onbekend. Deze onbekendheid, - een lacune in de opzet van het onderzoek -, had tot gevolg dat de statistiek niet als bewijsmiddel kon worden gehanteerd, zodat het waarnemingsmateriaal geen duidelijke conclusie toeliet.

Maar het is niet onaannemelijk, - zoals nader zal blijken -, dat dit anders zou zijn geweest indien ook de nauwkeurigheid van de analyse behoorlijk was onderzocht, o.a. door proeven te nemen met afgewogen hoeveelheden urobilinogeen en door het verrichten van een voldoend aantal duplowaarnemingen. Bij ontbreken hiervan kan de statistiek altijd nog fungeren als detectiemiddel.

Redelijke speculaties over de analysenauwkeurigheid kunnen nl. het effect, dat in fig. 2 in zwakke mate aanwezig schijnt te zijn, duidelijker naar voren brengen. Als regel is immers de relatieve nauwkeurigheid van een quantitatieve bepaling (bij eenzelfde methodiek) geringer, wanneer het kleinere hoeveelheden van de onderzochte stof betreft. De relatieve verliezen zijn dan i.h.a. groter, de meeste aflezingen naar verhouding onnauwkeuriger.

Hieruit volgt, dat ook de verhouding van twee hoeveelheden van een zelfde stof des te onnauwkeuriger wordt bepaald, naarmate deze hoeveelheden kleiner zijn. Omgekeerd zal van twee ongeveer gelijke, aldus verkregen quotiënten datgene met de grootste teller en noemer het betrouwbaarst zijn.

Bovendien zijn de door verliezen veroorzaakte fouten éénzijdig. Bij de zo juist beschouwde quotiënten zullen daarom, in het algemeen gesproken, zowel de teller als de noemer te klein uitvallen. Is één van deze grootheden aanmerkelijk kleiner dan de andere, dan zal de relatieve invloed van de verliezen op het kleinste getal het sterkst zijn. Is dit getal de noemer, dan zal voor het quotiënt een te grote, en soms zelfs een exorbitant grote waarde worden gevonden.

Wij vermoeden dat een en ander zich bij de boven beschouwde grootheden voordoet, speciaal bij de quotiënten q_{ur}. Immers de bilirubineconcentratie is in de gal groter dan die van het urobilinogeen, en de voor q_{bil} verrichte aflezingen waren blijkbaar steeds in ten minste 2 cijfers mogelijk; vermoedelijk is daarom q_{bil} nauwkeuriger bepaald dan q_{ur}.

Daarentegen is in het bijzonder de noemer $[ur]_C$ van q_{ur} soms zeer klein (d.w.z. it niet meer dan één cijfer aangegeven), en vermoedelijk dus zeer onnauwkeurig. Dat de drie kleinste waarden van

$[ur]_C$ in de (volledige) tabel alle gelijk zijn aan 0,005 ondersteunt dit vermoeden in niet geringe mate. Wellicht liet de chemische analyse een verdergaande specificatie van het begrip "zeer klein" niet toe.

Het ligt voor de hand dat ecarteren van de minst goede waarnemingen een onderzoek als het nu besprokene ten goede kan komen. De onnauwkeurige waarnemingen kunnen een eventueel aanwezig effect gemakkelijk verdoezelen.

De meest voor de hand liggende wijze van ecarteren bestaat in het weglaten van die gevallen waarbij $[ur]_C$ klein en q_{ur} dus onbetrouwbaar is. Wel is het gewenst dat daarbij genoeg punten uit fig. 2 overblijven om over de structuur van de resterende puntenwolk een behoorlijke indruk te kunnen krijgen. Nu is bij 13 van de 21 beschouwde patiënten $[ur]_C \leq 0,104$; bij de 8 overigen varieerde deze grootheid van 0,208 tot 1,31 [9]. De met deze laatsten corresponderende betrouwbaardere punten zijn in fig. 3 weergegeven.

Deze liggen nu alle duidelijk beneden de bissectrice, hetgeen een sterke aanwijzing voor een systematisch verschil tussen q_{bil} en q_{ur} zou inhouden, als bovenstaande procedure verantwoord kon worden geacht. Was een uitkomst van de aard van fig. 3 verkregen na ecartering op grond van nauwkeurigheidsbepalingen, dan was een ondubbelzinnig resultaat bereikt.

Bij de gevolgde methode verdwijnen evenwel juist die punten, waarvoor $[ur]_C$ klein, dus q_{ur} i.h.a. groot is, d.w.z. overwegend hooggelegen punten in fig. 2. En inderdaad zijn ook de hooggelegen punten onder de bissectrice verdwenen. Toch was het niet vanzelfsprekend dat de overgebleven punten alle onder de bissectrice zouden liggen; q_{ur} kan nl. óók groot zijn door een grote teller (waar over later). Maar het trekken van een conclusie uit het verkregen resultaat alleen is zeker niet verantwoord.

Men kan echter trachten op grond van andere criteria te ecarteren, en zien of men aldus aanwijzingen verkrijgt die in dezelfde richting gaan. Dit blijkt inderdaad het geval. Daarbij valt te bedenken dat iedere methode van ecarteren willekeurig is (zoals de grens voor $[ur]_C$, die aan fig.3 ten grondslag lag, vrij arbitrair

---

9) De lacune tussen 0,104 en 0,208 in de waarden van $[ur]_C$ verschafte een ongedwongen criterium voor het ecarteren.

was). Er zijn immers geen voldoende gegevens over de nauwkeurigheid om tot niet arbitraire richtlijnen te komen.

Om hier een beter inzicht te krijgen kan men de tellers $[ur]_B$ en de noemers $[ur]_C$ tegen elkaar uitzetten, zoals in fig. 4 is gedaan (terwille van de duidelijkheid met verschillende schalen op de horizontale en verticale as).

De punten die in deze figuur onder de met I aangegeven stippellijn liggen beantwoorden dan aan de hierboven geëcarteerde waarnemingen. Daarbij werd alleen rekening gehouden met de grootheden $[ur]_C$. We zagen evenwel dat ook kleine waarden van $[ur]_B$ tot vermoedelijk onbetrouwbare uitkomsten leiden. In het bijzonder was er reden een quotiënt q_{ur} te wantrouwen indien zowel de teller als de noemer klein zijn, wat voor die punten van fig. 4 geldt die dichtbij de oorsprong liggen. Een hierop gebaseerde ecartering doet fig. 5 ontstaan.

Ditmaal beantwoorden de weggelaten waarnemingen aan die punten in fig. 4, die liggen onder de (alweer tamelijk willekeurige) stippellijn II.

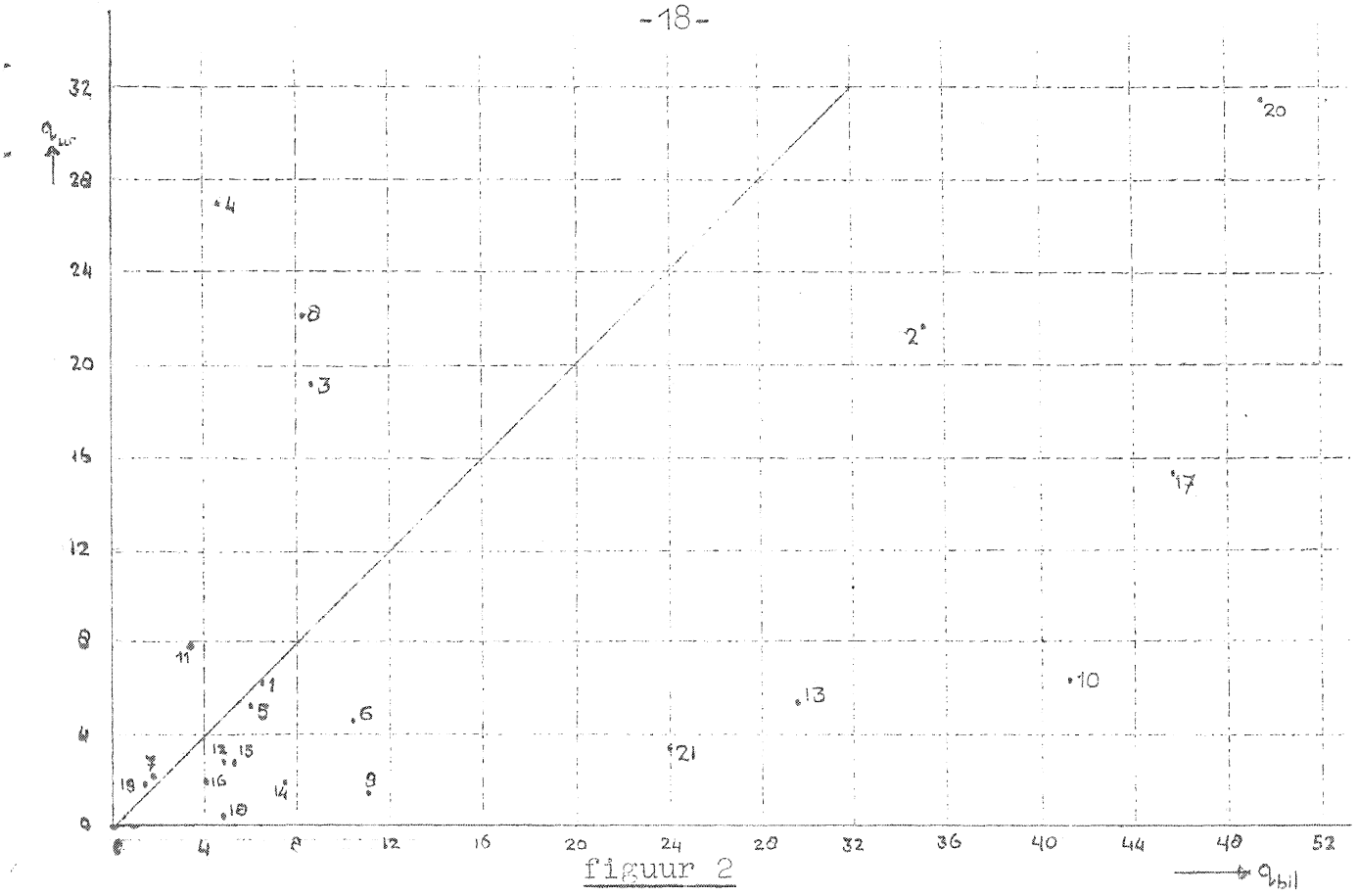
Voor het merendeel zijn de overgebleven punten in fig. 5 dezelfde als in fig. 3. Alle op één na liggen weer onder de bissectrice. Ook nu is de aanwijzing voor een systematisch verschil aanzienlijk duidelijker dan in fig. 2, zij het zwakker dan in fig. 3, vanwege het ene punt ver boven de bissectrice.

Een derde methode komt tegemoet aan het bezwaar dat in fig.3 vrij opzettelijk hooggelegen punten waren geëlimineerd. Men bevordert het tegendeel door juist die waarnemingen uit te zoeken met de grootste tellers $[ur]_B$. Evenals bij fig. 3 zullen ook ditmaal i.h.a. betrouwbaarder punten uit fig. 2 overblijven. Kiest men weer 8 punten, nu de 8 met de grootste tellers $[ur]_B$, dan zijn dit (min of meer toevallig) precies dezelfde als in fig. 5 aangegeven, zodat ook nu weer 7 van de 8 punten onder de bissectrice liggen. Met name liggen de punten, beantwoordend aan de 3 hoogste waarden van de teller, eronder. Onze poging bij voorkeur boven de bissectrice over te houden heeft dus een averechts effect!

Al deze argumenten wijzen dus in dezelfde richting: dat q_{ur} systematisch kleiner is dan q_{bil}, dus dat er afbraak of resorptie van urobilinogenen plaats vindt in de galblaas.

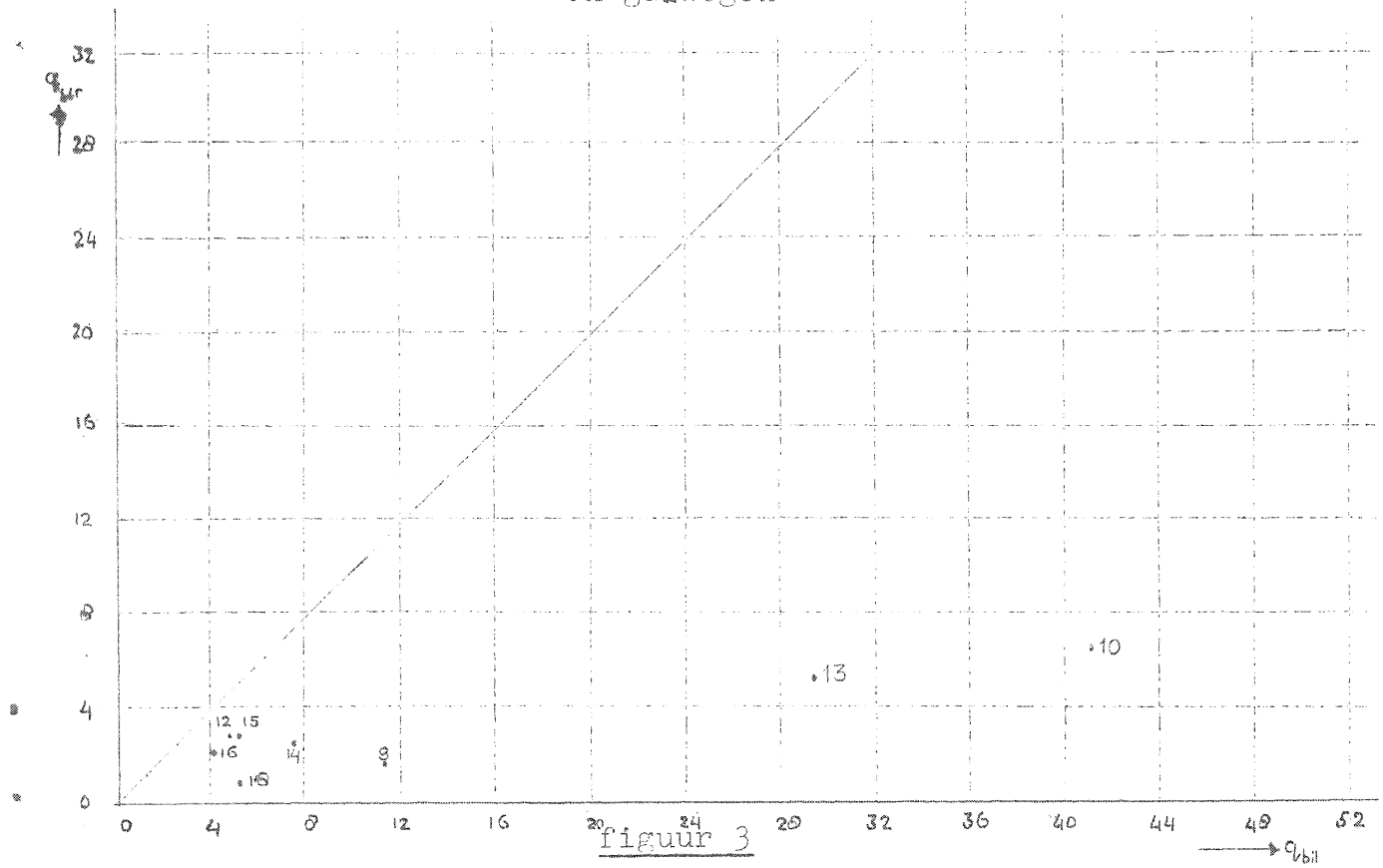
Het speculatieve karakter van deze beschouwingen maakt het

onmogelijk dit als een "bewijs" op te vatten. Men kan slechts de tweeledige conclusie trekken dat mogelijk wel een duidelijke aanwijzing zou zijn verkregen als meer aandacht aan de nauwkeurigheid was geschonken, en dat een verder onderzoek, - indien deze kwestie voldoende belangrijk wordt geacht -, aanbevelenswaardig is.



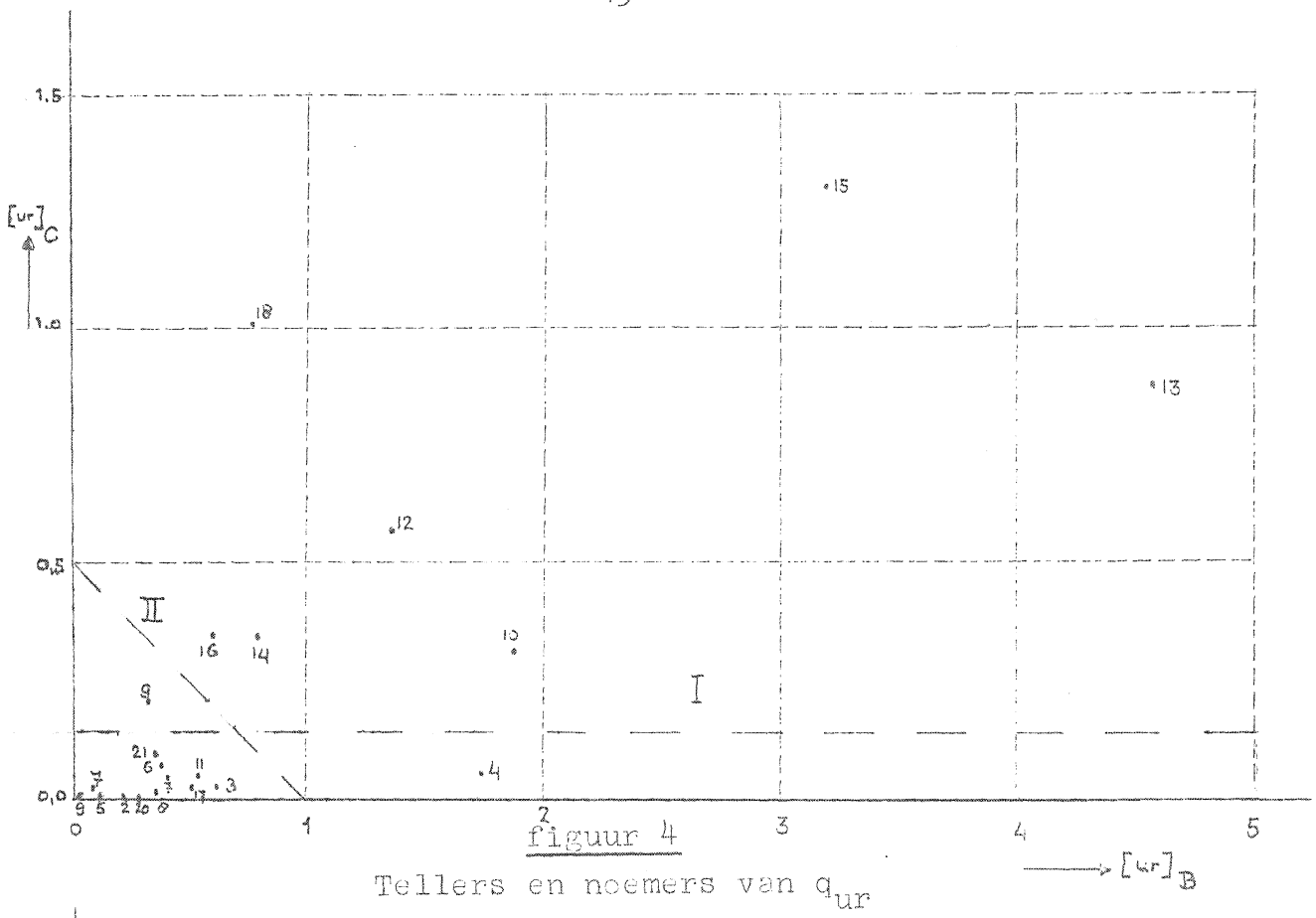
figuur 2

Concentratiequotiënten van 21 patiënten met vrijwel normale lever en galwegen



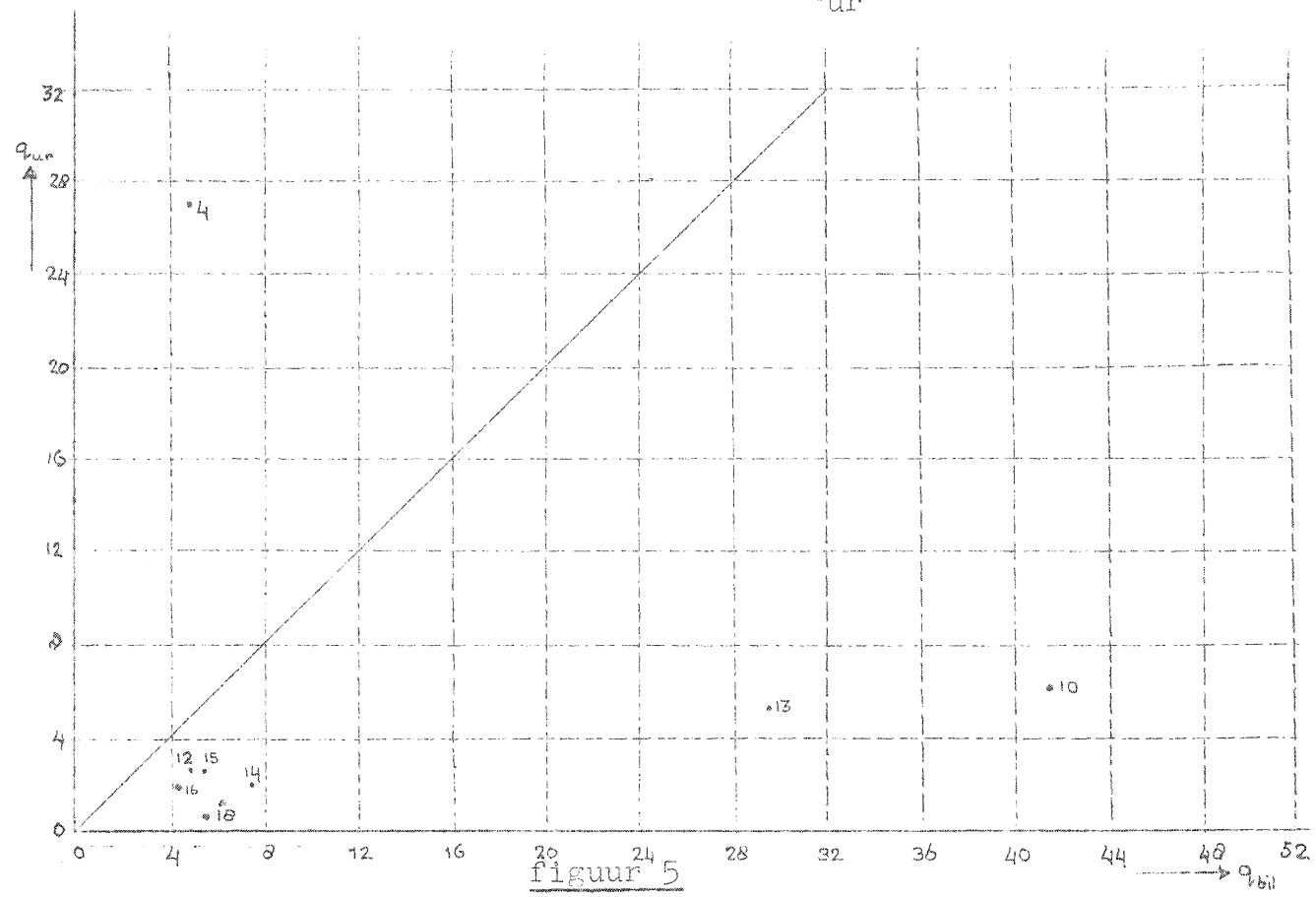
figuur 3

Concentratiequotiënten van 8 patiënten met de grootste waarden van $[ur]_C$



figuur 4

Tellers en noemers van q_{ur}



figuur 5

Concentratiequotiënten van 8 patiënten met de grootste waarden van $[ur]_B + 2[ur]_C$, resp. van $[ur]_B$

The Mathematical Centre at Amsterdam, founded the 11th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications, and is sponsored by the Netherlands Government through the Netherlands Organization for Pure Research (Z.W.O.) and the Central National Council for Applied Scientific Research in the Netherlands (T.N.O.), by the Municipality of Amsterdam and by several industries.