

EEN PRAKTISCHE STOPREGEL VOOR KLINISCHE PROEVEN

Hubert J.A. Schouten

Samenvatting

Voor groep-sekwente klinische proeven, ter vergelijking van medische behandelingswijzen, wordt een algemeen toepasbare stopregel met een groeiende significantiegrens voorgesteld. Het maximale aantal patiënten bij deze stopregel is slechts 8% (9%) groter dan het vaste aantal patiënten bij een conventionele toets met vereist onderscheidend vermogen 0,90 (0,80), wat van doorslaggevende betekenis kan zijn voor de uitvoerbaarheid van een experiment. Omdat de significantiegrens nooit minder is dan 0,005 (0,01 bij ten hoogste vier interim analyses), is er steeds een realistische mogelijkheid met het experiment te stoppen. Met betrekking tot het in de stopregel gebruikte evaluatiecriterium is een geringe correctie nodig voor de overschrijdingskans en het 95% betrouwbaarheidsinterval.

Vakgroep Medische Informatica en Statistiek
Rijksuniversiteit Limburg
Postbus 616, 6200 MD Maastricht

1. Inleiding

In zijn boek geeft Pocock (1983, Chapter 10) een heldere en uitgebreide inleiding betreffende de statistische en organisatorische aspecten van sekwente klinische proeven; de lezer wordt ook verwezen naar het meer recente overzicht van Geller en Pocock (1987). In deze paragraaf volsta ik met het noemen van de belangrijkste aspecten.

Wanneer twee geneeswijzen in een therapeutisch experiment worden vergeleken, moet vertekening zoveel mogelijk worden tegengegaan door onder meer aselectering, blinding, analyse volgens het "intention to treat" principe, en door rekening te houden met de ernstige gevolgen van veelvuldig statistisch toetsen. Om ethische en economische redenen wil een onderzoeker met zijn experiment stoppen zodra een statistisch significant verschil in behandelingsresultaat wordt bereikt. Echter, bij een conventionele statistische toets wordt een vast aantal patienten verondersteld, en herhaaldelijk conventioneel toetsen heeft tot gevolg dat de onbetrouwbaarheid van de toetsingsprocedure ontoelaatbaar groot wordt. Bij een significantiegrens van 0,05, en een (asymptotisch) normaal verdeelde toetsingsgrootheid onder de nulhypothese, heeft de toetsingsprocedure een totale onbetrouwbaarheid $\alpha = 0,08$ bij twee keer toetsen, $\alpha = 0,14$ bij vijf keer toetsen en $\alpha = 0,19$ bij tien keer toetsen. De onbetrouwbaarheid van de toetsingsprocedure kan op $\alpha = 0,05$ worden gehouden door bij tussentijdse analyses een sekwente stopregel toe te passen. Om logistieke redenen verdient een groep-sekwente toets de voorkeur boven een continu schema. De stopregel moet worden toegepast op niet meer dan één evaluatie-criterium; Pocock, Geller en Tsiatis (1987) beschrijven echter een allesomvattende toets waarmee meerdere evaluatie-criteria simultaan kunnen worden vergeleken (op een verstandiger manier dan bijvoorbeeld met Hotelling's T^2 het geval is).

In de volgende paragraaf wordt een stopregel met een groeiende significantie-grens beschreven, behorend bij een sekwente toets met totale onbetrouwbaarheid $\alpha = 0,05$. Deze stopregel is eenvoudig toe te passen, hetgeen van belang is voor de vele experimenten die zonder noemenswaardige hulp van een statisticus plaats vinden: het maximale aantal patienten volgt eenvoudig uit het vereiste aantal patienten bij een conventionele toets, uit een tabel kan de gecorrigeerde P-waarde

worden gelezen, en het 95% betrouwbaarheidsinterval kan gemakkelijk worden gecorrigeerd.

2. Een praktische stopregel

De hier beschreven stopregel heeft de volgende structuur. Van te voren wordt het aantal evaluatie-momenten in het onderzoeks-protocol vastgelegd; dit aantal duiden we aan met k en we beperken ons tot de situatie dat $2 \leq k \leq 10$. Op elk evaluatie-moment wordt een conventionele toets toegepast op een enkele uitkomst-variabele. Op het i -de evaluatie-moment ($i = 1, 2, \dots, k-1$) wordt besloten het experiment te stoppen als de waargenomen overschrijdings-kans P kleiner is dan de in tabel 1 vermelde significantie-grens. Bij een grotere P -waarde wordt het experiment gecontinueerd. Tussen twee evaluatie-momenten zitten steeds een zelfde aantal patiënten, of een zelfde aantal doden bij overlevings-analyse. Ter verduidelijking: bij $k=4$ evaluaties worden achtereenvolgens de grenzen 0,01 0,01 0,02 en 0,031 gebruikt.

Tabel 1

Significantie-grenzen voor de conventionele P -waarde; deze sekwentetoeets heeft een onbetrouwbaarheid $\alpha = 0,05$

evaluatie- moment	k = (maximaal) aantal evaluatie-momenten						
	2	3	4	5	6	8	10
$i \leq k-4$				,005	,005	,005	,005
$i = k-3$			,01	,01	,01	,01	,01
$k-2$		,01	,01	,01	,01	,01	,01
$k-1$	,02	,02	,02	,02	,02	,02	,02
k	,038	,035	,031	,032	,033	,032	,031

Voor $k = 5$ interim analyses liet Pocock (1983, p. 154) zien dat het onderscheidend vermogen $1-\beta$ in sterke mate bepaalt welke significantiegrenzen het gemiddeld te verwachten aantal patienten zo laag mogelijk houden. Bij $1-\beta = 0,90$ of $0,95$ zijn constante grenzen bijna optimaal, voor $1-\beta = 0,50$ zijn de grenzen van O'Brien en Fleming (1979) bijna optimaal en voor $1-\beta = 0,70$ of $0,80$ zijn de grenzen in tabel 1 bevredigend. Andere criteria zijn echter ook van belang:

Wanneer de stopregel van O'Brien en Fleming wordt gebruikt, is stoppen bij de eerste interim analyse bijna onmogelijk vanwege de extreem kleine significantiegrens.

Voor de haalbaarheid van een klinische proef is van belang hoe groot het aantal patienten wordt als bij de laatste (k -de) analyse wordt gestopt, en in dit opzicht zijn constante grenzen ongunstig. Een ander nadeel van constante grenzen is dat de uiteindelijk verkregen conventionele P -waarde substantieel gecorrigeerd dient te worden, wat de interpretatie van de uiteindelijke analyse kan bemoeilijken; de uiteindelijke analyse betreft bovendien variabelen die niet in de stopregel werden gebruikt en houdt rekening met covariabelen. Voor de nieuwe stopregel worden deze aspecten in de volgende paragrafen behandeld.

De totale onbetrouwbaarheid bij de groeiende significantiegrens ($0,005 \dots 0,005 \quad 0,01 \quad 0,01 \quad 0,02 \quad 0,03+$) in tabel 1 is $\alpha = 0,05$. Hierbij wordt verondersteld dat de toetsingsgrootte bij benadering normaal verdeeld is. In voldoende grote steekproeven geldt dit voor de Student t -toets, de Wilcoxon-Mann-Whitney toets, de logrank toets, de ongecorrigeerde chi kwadraat toets voor het vergelijken van twee onafhankelijke percentages (de Yates correctie is vooral bij sekwent toetsen conservatief met betrekking tot de nulhypothese; bovendien is het moeilijk om in een sekwente toets het aantal successen op ieder evaluatie-moment als vast te beschouwen, zoals gedaan wordt bij de toets van Fisher waarvoor de chi kwadraat toets met Yates correctie een benadering is), Mantel-Haenszel en nog meer toetsen.

De onbetrouwbaarheid van de sekwente toets werd berekend door te simuleren met een miljoen herhalingen. Standaardnormale waarden $X_1, X_2, \dots, X_k$ werden gegenereerd met behulp van de NAG subroutine

G05DDF. Voor het i -de evaluatie-moment ($i = 1, 2, \dots, k$) werd de standaardnormale toetsingsgrootheid $Z_i = (X_1 + X_2 + \dots + X_i)/\sqrt{i}$ berekend. De waarden Z_i werden gebruikt om de sekvente toets uit te voeren. Een significant resultaat werd verkregen als de met Z_i corresponderende P -waarde kleiner was dan de in tabel 1 vermelde significantiegrens. De fractie significante toetsings-resultaten, van een miljoen herhalingen, is een tamelijk nauwkeurige schatting van de onbetrouwbaarheid van de sekvente toets. De simulatie werd zowel voor tweezijdig als voor eenzijdig toetsen uitgevoerd.

In de praktijk kan het gebeuren dat een experiment om goede redenen wordt gestopt zonder dat de significantie-grens werd overschreden, bijvoorbeeld wegens onverwacht ernstige bijwerkingen, of wegens nieuwe wetenschappelijke kennis die buiten het betreffende experiment werd verkregen (in een recent artikel, tijdens of na een lezing, of bij een werkbezoek). De stopregel wordt daardoor een vuistregel waarvan de statistische eigenschappen niet precies bekend zijn.

3. Maximaal aantal patienten

Het aantal patienten is maximaal als niet voor het k -de evaluatie-moment wordt gestopt. Dit maximale aantal patienten $n_{\max}$ bepaalt in veel gevallen de haalbaarheid van een therapeutisch experiment. Daarom mag $n_{\max}$ niet veel groter zijn dan het vereiste aantal patienten n_{fix} bij een conventionele toets, waarbij het totale aantal patienten van tevoren wordt vastgelegd, met dezelfde onbetrouwbaarheid $\alpha = 0,05$ en hetzelfde onderscheidend vermogen $1-\beta$. Dit wordt bereikt door de onbetrouwbaarheid van de sekvente toets vooral aan het eind te besteden, en dus groeiende significantie-grenzen te hanteren. Hieronder wordt de verhouding $n_{\max}/n_{\text{fix}}$ onderzocht. Bij gelijke α en gelijke $1-\beta$ is het onvermijdelijk dat $n_{\max} > n_{\text{fix}}$. Dit is de tol die moet worden betaald voor de mogelijkheid vroeg met een experiment te kunnen stoppen.

Hieronder wordt met z_γ het γ -fractiel van de standaard-normale verdeling aangegeven. Bij een tweezijdige conventionele toets met onbetrouwbaarheid α en onderscheidend vermogen $1-\beta$, voor een zekere alternatieve hypothese, is het vereiste aantal patienten bij goede benadering evenredig met $(z_{\alpha/2} + z_\beta)^2$; daarbij wordt verondersteld dat de toetsings-grootheid normaal verdeeld is en onder de nulhypothese en

onder de alternatieve hypothese dezelfde standaardafwijking heeft. Uit de laatste regel in tabel 1 volgt dat het vereiste aantal patiënten bij een conventionele toets met onbetrouwbaarheid 0,03 een bovengrens is voor $n_{\max}$. Dus geldt

$$\frac{n_{\max}}{n_{\text{fix}}} < \frac{(z_{\beta} + z_{0,015})^2}{(z_{\beta} + z_{0,025})^2} = \frac{(z_{\beta} - 2,17)^2}{(z_{\beta} - 1,96)^2}$$

voor de verhouding tussen het maximale aantal in de sekvente toets en het vaste aantal in de conventionele toets met onbetrouwbaarheid 0,05. Voor een onderscheidend vermogen $1-\beta = 0,90$ geldt $n_{\max}/n_{\text{fix}} < (3,82/3,61)^2 = 1,12$ en voor $1-\beta = 0,80$ geldt $n_{\max}/n_{\text{fix}} < (3,45/3,24)^2 = 1,13$. Om deze bovengrens nader te onderzoeken in een simulatie-experiment beschouwen we een toetsingsgrootheid Z die onder de nulhypothese standaardnormaal verdeeld is. We veronderstellen dat Z ook onder de alternatieve hypothese normaal verdeeld is met variantie 1, maar met een verwachting die evenredig is met de wortel uit de steekproefgrootte; dit geldt asymptotisch voor de twee-steekproeven t -toets (als beide steekproeven m keer zo groot worden, dan wordt de alternatieve verwachting van Student's t $\sqrt{m}$ keer zo groot) en bij goede benadering voor andere toetsen. Bij een onbetrouwbaarheid α en vereist onderscheidend vermogen $1-\beta$ wordt de vaste steekproefgrootte n_{fix} zo gekozen dat Z onder de alternatieve hypothese een verwachting $|z_{\alpha/2} + z_{\beta}|$ heeft. Voor de corresponderende sekvente toets moeten dan, analoog aan de simulatie in de vorige paragraaf, grootheden X_i ($i=1, 2, \dots, k$) worden gegenereerd met variantie 1 en verwachting $|z_{\alpha/2} + z_{\beta}| \sqrt{n_{\max}/(kn_{\text{fix}})}$. De toetsingsgrootheid $Z_i = (X_1 + X_2 + \dots + X_i)/\sqrt{i}$ ($i=1, 2, \dots, k$) heeft dan variantie 1 en verwachting $|z_{\alpha/2} + z_{\beta}| \sqrt{(in_{\max})/(kn_{\text{fix}})}$. Het uitproberen van verschillende verhoudingen $n_{\max}/n_{\text{fix}}$, in een simulatie-experiment met een miljoen herhalingen, rechtvaardigt de volgende aanbeveling.

AANBEVELING:

Bereken eerst het vereiste aantal patiënten n_{fix} voor een conventionele toets. Het maximale aantal n_{max} voor de sekweite toets is dan gelijk aan $n_{\text{max}} = 1,08n_{\text{fix}}$ bij een onderscheidend vermogen $1-\beta = 0,90$ en $n_{\text{max}} = 1,09n_{\text{fix}}$ bij $1-\beta = 0,80$. Er moet slechts een geringe tol worden betaald om het experiment vroeg te kunnen stoppen als er een onverwacht groot verschil in therapeutisch effect blijkt te zijn.

4. Overschrijdingskans

Het is zeker niet zo dat een enkele klinische proef het definitieve antwoord moet geven op de vraag welke van twee geneeswijzen het beste resultaat oplevert. Een dergelijke conclusie wordt gebaseerd op meerdere experimenten. Daarom is het niet zinvol om een strikte scheiding aan te brengen tussen wel en niet significante experimenten. Een overschrijdingskans bevat meer informatie dan het resultaat van een pure significantie-toets, zoals ook Cox en Hinkley (1974, section 3.2) betogen. Omdat de conventionele P-waarde slechts geldig is bij eenmalig toetsen, moet de stopregel verdisconteerd worden in een gecorrigeerde P-waarde.

Veronderstel dat een sekweite proef niet voor het i -de evaluatiemoment wordt gestopt en dat op het i -de evaluatiemoment een conventionele overschrijdingskans P wordt berekend; $1 \leq i \leq k$. De werkelijke overschrijdingskans, dat wil zeggen de voor de stopregel gecorrigeerde P-waarde, is dan de kans om op een eerder moment te stoppen of op moment i een kleinere overschrijdingskans te krijgen. Tabel 2 laat de werkelijke P-waarde zien, voor bepaalde conventionele P-waarden en stop-momenten, gebaseerd op simulaties met een miljoen herhalingen. Bij een conventionele P-waarde van 0,05 hoort een gecorrigeerde P-waarde van hoogstens 0,065.

Tabel 2 toont de werkelijke tweezijdige overschrijdingskans. Bij eenzijdig toetsen is het verschil tussen conventionele en gecorrigeerde overschrijdingskans nog iets kleiner.

Tabel 2

Voor de stopregel gecorrigeerde P-waarde op het moment dat het experiment wordt gestopt.

conven- tionele P-waarde	stop- moment i	k = maximaal aantal evaluaties						
		2	3	4	5	6	8	10
,005	k-7						,005	,012
,005	k-6						,009	,015
,005	k-5					,005	,012	,017
,005	k-4				,005	,009	,015	,019
,01	k-3			,010	,014	,016	,021	,024
,01	k-2		,010	,018	,020	,022	,025	,028
,01	k-1	,010	,018	,024	,025	,026	,028	,031
,02	k-1	,020	,027	,032	,032	,033	,035	,037
,02	k	,035	,038	,042	,041	,041	,042	,043
,03	k	,043	,046	,049	,048	,048	,049	,049
,04	k	,052	,054	,057	,056	,056	,056	,057
,05	k	,061	,063	,065	,064	,064	,064	,065
,10	k	,108	,109	,110	,109	,109	,110	,111

5. Betrouwbaarheids-interval

Tsiatis, Rosner en Mehta (1984) beschreven hoe een betrouwbaarheids-interval kan worden gecorrigeerd voor het toepassen van een sekwente stopregel met de t-toets. Hun tamelijk ingewikkelde procedure kan voor elke mogelijke stopregel worden toegepast en een computerprogramma kan bij hen worden aangevraagd. Hieronder geef ik een eenvoudige alternatieve procedure.

Naar mijn mening is het vooral van belang te zorgen dat het toetsings-resultaat en het 95% betrouwbaarheids-interval met elkaar in

overeenstemming zijn. Wanneer bijvoorbeeld het experiment gestopt wordt op een evaluatie-moment met tweezijdige significantiegrens 0,01 en het waargenomen verschil is positief, vervang dan de ondergrens van het 95% betrouwbaarheids-interval door de ondergrens van het 99% betrouwbaarheids-interval en laat de bovengrens ongewijzigd. Het zo verkregen 95% interval zal iets te ruim uitvallen, maar sluit volledig aan bij de sekvente toets.

6. Discussie

Als in een experiment een sekvente stop-regel wordt gebruikt, dan is een correctie nodig voor de overschrijdings-kans en het 95%-betrouwbaarheids-interval betreffende het in de stopregel gebruikte evaluatie-criterium. Omdat de correctie gering is bij de hier voorgestelde stopregel, lijkt geen correctie nodig met betrekking tot de variabelen die niet in de stopregel worden gebruikt, mits deze variabelen niet al te sterk samenhangen met het in de stopregel gebruikte evaluatie-criterium. Dit onderwerp werd behandeld door Whitehead (1986). De eenvoud van het toepassen van de hiervoor beschreven stopregel, inclusief het berekenen van de maximale steekproef-grootte, is in het bijzonder van belang voor de talrijke experimenten die zonder veel hulp van een statisticus worden uitgevoerd.

De vermelde eigenschappen van de voorgestelde stopregel, bijvoorbeeld de mate waarin de P-waarde moet worden gecorrigeerd, gelden alleen als deze stopregel strikt wordt gehanteerd. Toch kunnen ook andere overwegingen van belang zijn bij een zorgvuldige beslissing om met een experiment te stoppen of door te gaan: nieuwe resultaten in de literatuur of op een congres, onverwacht ernstige bijwerkingen of een zeer recent ontwikkeld geneesmiddel. Deze andere overwegingen maken de stopregel onvermijdelijk tot een vuistregel waarvan de eigenschappen niet precies bekend zijn.

Literatuur

- Cox, D.R and Hinkley, D.V. (1974). Theoretical Statistics.
London: Chapman and Hall.
- Geller, N.L. and Pocock, S.J. (1987). Interim analysis in
randomized clinical trials: ramifications and guidelines for
practioners. Biometrics 43, 213-223.
- O'Brien, P.C. and Fleming, T.R. (1979). A multiple testing
procedure for clinical trials. Biometrics 35, 549-556.
- Pocock, S.J. (1983). Clinical Trials: A Practical Approach. New
York: Wiley.
- Pocock, S.J., Geller, N.L. and Tsiatis, A.A. (1987). The analysis
of multiple endpoints. Biometrics 43, 487-498.
- Tsiatis, A.A., Rosner, G.L. and Mehta, C.R. (1984). Exact
confidence intervals following a group sequential test.
Biometrics 40, 797-803.
- Whitehead, J. (1986). Supplementary analysis at the conclusion of
a sequential clinical trial. Biometrics 42, 461-471.

Ontvangen: 21-06-1988
Geaccepteerd: 25-11-1988