Application and efficiency of sequential tests in matched case-control studies

Ingeborg van der Tweel

Lay-out: Monique den Hartog
Omslag: Femke Bulten, afd. Vormgeving, faculteit Biologie
Druk: LABOR Grafimedia BV, Utrecht

CIP-gegevens Koninklijke Bibliotheek, Den Haag

Van der Tweel, Ingeborg
Application and efficiency of sequential tests in matched case-control studies
Utrecht: Universiteit Utrecht, Faculteit Geneeskunde
Thesis Utrecht University. - With ref. - With summary in Dutch.
ISBN 90-393-3582-6
Keywords: continuous sequential analysis, group sequential analysis, matched case-control data, efficiency

Application and efficiency of sequential tests in matched case-control studies

Toepassing en efficiëntie van sequentiële toetsen in gematchte case-controle studies

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de Rector Magnificus, prof. dr W.H. Gispen,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen
op dinsdag 2 maart 2004 des middags te 14.30 uur

door

Ingeborg van der Tweel

geboren op 12 januari 1955 te Bussum

Promotores:

Prof.dr. D.E. Grobbee

Julius Centrum voor Gezondheidswetenschappen en Eerstelijns Geneeskunde,
Universitair Medisch Centrum Utrecht

Prof.dr. Th. Stijnen

Instituut voor Epidemiologie en Biostatistiek
Erasmus Universitair Medisch Centrum Rotterdam

CONTENTS

Introduction 1

CHAPTER 1 Application of a sequential *t*-test in a cohort nested case-control study with multiple controls per case 7

CHAPTER 2 Efficient use of biological banks for biochemical epidemiology: exploratory hypothesis testing by means of a sequential *t*-test 23

CHAPTER 3 Comparison of one-sample two-sided sequential *t*-tests for application in epidemiological studies 45

CHAPTER 4 Sequential analysis of matched dichotomous data from prospective case-control studies 67

CHAPTER 5 Early stopping in clinical trials and epidemiological studies for futility: conditional power versus sequential analysis 91

CHAPTER 6 Sequential tests for gene-environment interactions in matched case-control studies 109

CHAPTER 7 Repeated looks at accumulating data: to correct or not to correct? 135

Summary 151

Nederlandse inleiding en samenvatting 159

Dankwoord 173

Curriculum vitae 177

Introduction

For large epidemiological observational studies banks of biological samples are sometimes created. To study specific research hypotheses blood, urine or tissue specimens can be obtained from participants to the study and stored for later analysis in a biological bank[1-5]. This was done, for example, in the DOM project. The DOM project was a breast cancer screening programme in the city of Utrecht (the Netherlands) and the region around it. The Dom project consisted of several birthcohorts of women. Women who volunteered to come to the screening were asked to bring their overnight urine or, in another cohort, their toenail clippings. Combination of the characteristics of the participants with information from the regional cancer registration led to the identification of cases, participants developing the disease of interest, after a shorter or longer follow-up. These cases can be contrasted to controls, participants who did not develop disease during the same follow-up period of the study.

Constantly new biochemical, molecular or genetic laboratory techniques are developed. These allow a large number of (new) etiologic hypotheses to be tested on the stored biological material of cases and controls to investigate interesting associations between an exposure and a disease. The amount of stored biological material however is, in general, limited, in particular for the cases, when the disease is not so common. Furthermore, with most laboratory techniques biological material is destroyed and cannot be used for another test. To combine the large number of interesting hypotheses with the limited number and amount of biological samples statistical methods are needed that can distinguish between more promising and less promising hypotheses at the expense of as little biological material as possible.

Sequential statistical methods offer a researcher the possibility to terminate an investigation as soon as sufficient evidence has accumulated to accept the null hypothesis (‘no association between exposure and disease’) or to reject it in favour of the alternative hypothesis (‘an association exists between exposure and disease’). After each new observation or group of observations the accumulated data are tested. Based on the cumulative test result the study is stopped or more information is obtained. A sequential analysis requires, on average, fewer observations to come to a decision (‘accept the null hypothesis or reject it’) than the corresponding fixed sample size analysis. Sequential methods are thus an efficient way to handle the available data.[6-10]

Cases and controls can be matched to control for possible confounding factors. These factors are related to both exposure and disease and may distort the size of the exposure-disease relation. Examples of possible confounding factors in epidemiological studies are age, ethnicity and menopausal status. In a case-control study one or more controls can be matched to a case based on the value of the confounding factor. When a disease is rather common, many cases are observed during follow-up. If enough cases are available, one case matched to one control gives the statistical test optimal power. However, when a disease is rare, few cases become apparent, but, in general, a lot of controls are available.

To achieve sufficient power in those situations, more than one control can be matched to a case. A matched design requires in general less cases and controls than an unmatched design. Matching, applied to control for confounding factors, enhances the efficiency of the analysis.[11]

In the design phase of a randomised clinical trial it is a matter of *Good Statistical Practice* to estimate the necessary number of patients before the data are obtained. This number of patients is determined by the size of the effect measure that is relevant to detect if it is present, the type I error α and the power 1-β.[12-14] In most epidemiological, observational studies the size of the study is determined by practical aspects like time, costs, availability of subjects, etc. and less by the effect size to detect or the power to detect this effect size. When an epidemiological study will be analysed sequentially, one has to specify beforehand α, 1-β and the effect size one would like to detect. The number of observations needed to come to a decision using a sequential analysis is not fixed, however, but is a stochastic variable. This implies that an average or median study size can be estimated beforehand, but that one also has to consider, for example, the 90th percentile of the expected study size.

When the exposure variable is continuous, such as the selenium content of toenail clippings, all values can be used. When, on the contrary, the exposure variable is dichotomous, such as the occurrence of a genetic mutation, it is possible that a matched case-control set does not contain any usable information. This happens when the case and its matched control(s) are all exposed or all unexposed, the so-called concordant sets. Only so-called discordant sets contain information for analysis. The total study size necessary thus depends on the probability of a discordant set. When this probability is small, a large number of matched case-control sets will have to be collected and analysed to obtain enough information for a decision. Matching more controls to a case, if possible, can increase the probability of a discordant set. This can be another way of handling the available data efficiently.[11]

The chapters of this thesis describe how sequential analysis on matched case-control sets can handle valuable biological samples efficiently, so as to be able to test a large number of interesting hypotheses. The developed sequential tests are illustrated by examples using data from the DOM cohorts. When the exposure variable in a matched case-control study is continuous and the difference between the value for the case and the (mean) value for the control(s) can be assumed normally distributed, this difference can be analysed using a one-sample t-test on paired observations. In chapter 1 two versions of a sequential one-sample t-test are compared. Subsequently one of the sequential one-sample t-tests from chapter 1 is compared to another version in chapter 2. In chapter 3 a standardized difference between the mean exposures of cases and controls is related to a minimum expected value for the odds ratio for the highest quintile versus the lowest quintile of the exposure distribution. An exposure variable can, however, not always be measured on a continuous scale. Often one is or is not exposed to some environmental factor, or one has or does not have a certain gene mutation. Chapter 4 describes the development of a sequential test for a dichotomous exposure variable with a fixed or a variable number of controls matched to each case. In chapter 5 the disadvantages of calculation of the conditional power as a decision tool for early stopping of a study are discussed. Use of a (group) sequential test is proposed as an alternative. Chapter 6 describes sequential designs to test for gene-environment interactions. In chapter 7 three approaches to statistical testing theory and their effects on sequential testing are compared.

References

1. De Waard F, Collette HJA, Rombach JJ, Baanders-van Halewijn EA, Honig C. The DOM project for the early detection of breast cancer, Utrecht, the Netherlands. J Chronic Dis 1984: 37: 1-44.
2. Wald NJ. Use of biological sample banks in epidemiological studies. Maturitas 1985; 7: 59-67.
3. Van Noord PAH. Selenium and human cancer risk (nail keratin as a tool in metabolic epidemiology). [thesis] Utrecht: University of Utrecht, the Netherlands, 1992.
4. Riboli E. Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC). Ann Oncol 1992; 3: 783-91.
5. Shuster J, et al. Minimax two-stage-designs with applications to tissue banking case-control studies. Stat Med 2002; 21: 2479-93.
6. Armitage P. Sequential medical trials. 2nd ed. Oxford: Blackwell Scientific Publications, 1975.
7. Wetherill GB, Glazebrook KD. Sequential methods in statistics. 3rd ed. London: Chapman and Hall, 1986.
8. Whitehead J. The design and analysis of sequential clinical trials, rev. 2nd ed. Chichester: John Wiley & Sons Ltd, 1997.
9. O'Neill RT, Anello C. Case-control studies: a sequential approach. Am J Epidemiol 1978; 108: 415-24.
10. Aplenc R, Zhao H, Rebbeck TR, Propert KJ. Group sequential methods and sample size savings in biomarker-disease association studies. Genetics 2003; 163: 1215-9.
11. Ury HK. Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. Biometrics 1975; 31: 643-9.
12. Pocock SJ. Clinical trials. A practical approach. Chichester: Wiley, 1983.
13. Piantadosi S. Clinical trials. A methodological perspective. New York: Wiley, 1997.
14. Friedman LM, Furberg CD, DeMets DL. Fundamentals of clinical trials. third ed. New York: Springer, 1998.

CHAPTER 1

Application of a sequential *t*-test in a cohort nested case-control study with multiple controls per case

I. van der Tweel*, P.A.H. van Noord[#], R. Kaaks[§]

**Centre for Biostatistics, Utrecht University, Utrecht, the Netherlands*
[#]Department of Epidemiology, Utrecht University, Utrecht, the Netherlands
[§]International Agency for Research on Cancer, Lyon, France

J Clin Epidemiol 1993; 46: 253-259

Abstract

Application of sequential analysis may avoid unnecessary experimentation and achieve economical use of available biomaterial stored in biological banks. When, as often in cohort case-control studies, cases are scarce, it may be possible to use multiple control observations per case to increase the power of a test for detecting differences between cases and controls.

Samples from a biological data bank were analysed. We compared results of a non-sequential analysis with results of sequential *t*-tests for 1 to 5 controls matched per case in a cohort nested case-control study. Simulations are performed to get an idea of the unreliability and the power of the sequential test.

In general the sequential *t*-tests are too conservative with respect to the achieved power. Average sample numbers are lower for the sequential tests and decrease with multiple controls. More than 3 or 4 controls per case does not give a meaningful increase in efficiency.

Keywords: sequential *t*-test, multiple controls, simulations, efficiency, biobanking, cohort nested studies

1.1 Introduction

Sequential analysis of quantitative data has never found wide application in clinical trial practice, even though considering its use might be worthwhile. For ethical reasons alone one may wish to minimize the expected number of exposed patients. From an experimental point of view, one may wish to avoid unnecessary experimentation. In cohort nested case-control studies exposures may be assessed in biological samples stored in a biological bank. In this situation, economy with material from the biological bank may be a reason to choose a sequential type of analysis. In a prospective study, cases are often detected sequentially during follow-up. A sequential analysis could then limit the total duration of the study.

In a sequential case-control analysis, the response of a case is compared with the response of a single control. O'Neill describes in a detailed way a sequential analysis of a matched pair case-control study with a dichotomous response.[1]

In a cohort study, usually there is only a limited amount of biological material per subject, and there are far more controls in the biobank for which such material can be analyzed than cases. Therefore it may be desirable to compensate for the loss of statistical power by comparing each case with more than one control.[2]

Ury[3] showed that, for non-sequential case-control studies with continuously distributed data, the efficiency of multiple ($k > 1$) controls relative to matched pairs ($k = 1$) is equal to $2k / (k+1)$ for equal case and control variability.

Gail *et al*[4] show that in (non-sequential) situations with a limited number of cases, more than four controls per case (or vice versa) gives no more meaningful power increase.

We are unaware of literature about the efficiency of multiple controls per case in sequential analyses. Therefore, we compared the effect of more controls per case in a sequential design with the results of a non-sequential analysis.

1.2 Materials and patients

We performed retrospective analyses on data from a cohort nested case-referent (control) study on breast cancer and the selenium content in ppm of toenails (Van Noord[5]). The aim of the study was to determine whether selenium, as available in the body, is already decreased before tumour occurrence.

Nail clippings had been collected since 1982 in a cohort of 8760 premenopausal (i.e. without menopausal signs) women (42-52 years of age), who attended to a breast cancer screening program. A total number of 64 premenopausal breast cancer cases were detected in this cohort. Controls were matched to cases for age. For 57 cases 5 controls per case were available; to 7 cases 3 or 4 controls could be matched per case.

Selenium content in the nails did not depend on age, probably due to the relatively small age-range in our data. No seasonal or other time trends were found in nail selenium contents during three years of investigation (unpublished results).
The data were analysed in the order the cases became available over time.

1.3 Statistical analysis

1.3.1 Non-sequential analysis

For matched case-control observations the minimal sample size n_1 (i.e. the number of case-control pairs necessary) for detecting a true difference between case and control observations of at least μ with a (two-sided) type I probability (or unreliability) α and a type II probability β (i.e. power 1-β) is[6]

$$n_1 = (t_\alpha + t_\beta)^2 * \sigma_1^2 / \mu^2$$

where

σ_1^2 is the variance of the difference between a case and a control observation,

t_α and t_β are values from the t-table with n_1-1 *df* corresponding to probabilities of $\alpha/2$ and β respectively.

The type I probability α is the risk one wants to accept that the null hypothesis of no difference between case and control observations is falsely rejected; the type II probability β is the risk of falsely not rejecting the null hypothesis when a true difference of at least μ exists between case and control observations.

In case of multiple (say k) control observations per case, assuming equal variances for cases and controls and, for the sake of argument, a negligible correlation between case and control observations, the variance of the difference between a case observation and the mean of the k control observations becomes

$$\sigma_k^2 = \{(k+1)/k\} * \sigma^2 = \{(k+1)/2k\} * \sigma_1^2,$$

($\sigma_1^2 = 2\sigma^2$, where σ^2 is the variance of a single case or control observation).

The minimal number of case-control sets for detecting the same difference μ then becomes

$$n_k = (t_\alpha + t_\beta)^2 * \sigma_k^2 / \mu^2 = n_1 * \{(k+1)/2k\}$$

N.B. We assumed (near) independence of case and control observations. In case of a positive correlation between case and control observations, the result will be a smaller σ_1^2 and σ_k^2 and a smaller sample size needed to detect the same difference μ.

1.3.2 Sequential analysis

Wald[7] developed the theory for the 'sequential probability ratio test' (SPRT). Rushton[8] further developed this theory to the one-sample, two-sided sequential *t*-test. This test is based on the probability ratio

$$l_n = \frac{\text{probability of observed results given } H_1 \text{ true}}{\text{probability of observed results given } H_0 \text{ true}}$$

for n observations processed so far. For our situation with case-control sets, we pose as null hypothesis H_0 :

$$\delta = \mu / \sigma_k = 0$$

and as alternative hypothesis H_1 :

$$|\delta| > 0$$

where μ is the minimal mean difference to be detected and σ_k is the theoretical standard deviation of the differences between the case and control observations. Because in most practical situations σ_k will be unknown and needs to be estimated from the data, the parameter $\delta = \mu / \sigma_k$ is used in the test. The test operates as follows :

- continue sampling as long as	$B < l_n < A$
- stop sampling and decide for H_0 as soon as	$l_n < B$
- stop sampling and decide for H_1 as soon as	$l_n > A$

To obtain approximately the *a priori* specified error probabilities α (two-sided type I error) and β (type II error), Wald stated the theorem that $A \sim (1-\beta)/\alpha$ and $B \sim \beta/(1-\alpha)$. The logarithm of the probability or likelihood ratio l_n can be calculated exactly using the series expansion of Kummer's function.[9]

Rushton[8] obtained a practical approximation to the logarithm of the likelihood ratio. See Appendix I for more details on Kummer's function, Rushton's approximation and our adaptation of the test statistic for k control observations per case.

1.3.3 Simulations

To examine the effect of multiple controls per case in a sequential t-test on its overall type I and type II error, simulation studies were performed. A simulation program was written in Turbo Pascal Version 5.0 (Borland). Random case and control observations were generated following a normal distribution with expectation μ_0 or μ_1 and theoretical standard deviation σ. The values chosen for μ_0, μ_1, σ and δ under H_1 are based on population values and a desirable shift in ppm of the selenium content (see Van Noord[10]). Both for case and control observations σ was chosen equal to 0.15. Under H_0: $\delta = 0$, μ_0 was chosen equal to 0.8. Under H_1: $|\delta| = \delta$, μ_1 was equal to $0.8 + \delta * \sigma * \sqrt{2}$.

Both under H_0: $\delta = 0$ and under H_1: $|\delta| = \delta$ ($\delta = 0.3$, 0.4 and 0.5 respectively), and with 1 to 5 controls per case, we ran a 1000 simulation runs ($\alpha = 0.05$, $1-\beta = 0.80$).

Per run the resulting decision ('accept H_0' or 'reject H_0 in favour of H_1') and the number of case-control sets necessary to come to that decision were recorded.
Simulations were performed using both Rushton's approximation to the logarithm of the likelihood ratio and the series expansion of Kummer's function.

1.4 Results

1.4.1 Non-sequential analysis

The results of a randomized blocks analysis of variance on the 'selenium and breast cancer' data for $n = 57$ cases and 5 control observations per case are shown in Table 1.
The mean difference between a case and the mean of the corresponding 5 control observations was 0.018 ppm with a SE = 0.029 ppm (NS).

Table 1 ‘Selenium and breast cancer’ study;
descriptive statistics and ANOVA table for 57 cases with 5 controls per case

	mean (ppm)	SD (ppm)	n
Cases	0.790	0.156	57
Controls	0.772	0.207	285

ANOVA table

Source	Sum of squares	Degrees of freedom	Mean squares	F	p
Between matched sets	2.20	56	0.04		
Within matched sets	0.16	5	0.03	<1	NS
Case-controls*	0.02	1	0.02		
Between controls	0.14	4	0.03		
Residual	11.24	280	0.04		

* Due to the difference between cases and the mean of the matched control observations.

Means, standard deviations and a randomized-blocks analysis of variance (ANOVA) table for n=57 cases with 5 controls per case. Data are the selenium content in ppm in toenails from the ‘selenium and breast cancer’ study.

Within matched sets the sum of squares, degrees of freedom and mean square are subdivided into two components: one that measures the variation because of a difference between cases and the mean of the matched control observations, and one that measures variation between controls. If we assume no differences between control observations, this last component can be combined with the residual sum of squares to give a (slightly) improved estimate of the residual mean square or error variance.

1.4.2 Sequential analysis

Sequential t-tests were performed on the ‘selenium and breast cancer’ data, using the available cases and a random sample of k (k = 1,...,5) control observations available in the matched set. (For each sequential test performed, control observations were replaced.) Both Kummer’s function and Rushton’s approximation were applied.
The number of cases (n) at which the decision ‘H_0 cannot be rejected’ was reached, is tabulated in Table 2 for several alternative hypotheses ($|\delta|$= 0.3, 0.4, 0.5).

N.B. None of the tests led to rejection of H_0; in case of H_1:$|\delta|$= 0.3, for some tests no conclusion could be reached with the available number of case-control sets.

Table 2 ‘Selenium and breast cancer’ study;
results of sequential *t*-tests for *k* controls per case

	H_1					
	$\lvert\delta\rvert = 0.3$		$\lvert\delta\rvert = 0.4$		$\lvert\delta\rvert = 0.5$	
k	R	K	R	K	R	K
1	21	--	12	23	9	13
1	25	--	30	30	11	15
1	27	62	21	21	10	13
1	22	48	23	24	10	14
1	26	50	13	25	8	18
2	25	50	17	21	9	14
3	22	--	18	21	12	16
4	22	--	12	21	8	13
5	22	--	13	21	9	13

Results of the sequential *t*-tests, given 57-64 cases and random samples of *k* controls per case, on the ‘selenium and breast cancer’ study ($\alpha = 0.05$ and $1\text{-}\beta = 0.80$);
R, Rushton’s approximation; K, Kummer’s function.

1.4.3 Simulations

The relative efficiency of more (*k*) controls per case is depicted graphically in Figures 1 and 2 for $\delta = 0.4$. (For $\delta = 0.3$ and $\delta = 0.5$ the course of the relative efficiency is similar). There the relative sample size n_k/n_1 is plotted against *k* for the median, mean and 95th-percentile number of cases required to reject H_0 in favour of H_1. The theoretical expected efficiency $(k+1)/2k$ is plotted as a comparison.

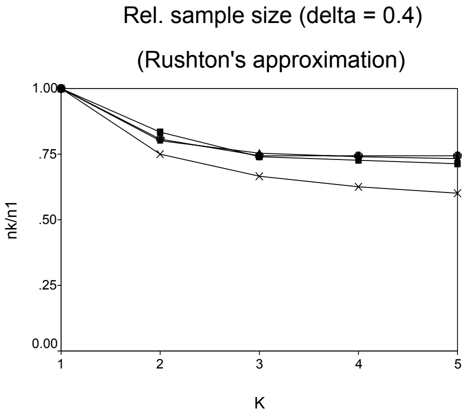


Figure 1 Relative sample size (n_k/n_1) for mean (▲), median (●) and 95-th percentile (■) number of cases necessary to reject H_0 in favour of H_1: $\lvert\delta\rvert = 0.4$ compared to the theoretical expected value (k+1)/2k (**x**), using Rushton’s approximation.

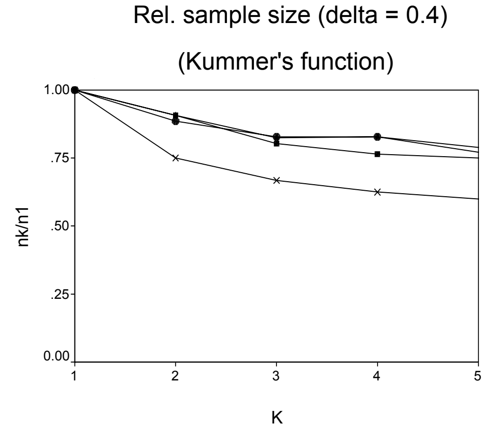


Figure 2 Relative sample size (n_k/n_1) for mean (▲), median (●) and 95-th percentile (■) number of cases necessary to reject H_0 in favour of H_1: $|\delta|$ = 0.4 compared to the theoretical expected value (k+1)/2k (**x**), using Kummer's function.

Appendix II shows data and calculations of one of the simulations as an example.

1.5 Discussion

Biological data banks contain valuable material that can be analysed to explore new hypotheses with possible important public health consequences. But, with most chemical analyses, these unique biological samples are destroyed and thus economical tests are preferable.[11]

While, in case-control studies, cases are mostly scarce, but control samples abundant, statistical efficiency of non-sequential tests can be increased by including multiple controls per case. If the power using equal allocation (k = 1) is greater than 0.9, this is of no practical importance. If the equal allocation power is less than 0.9, meaningful power increases may be obtained, but more than 4 controls per case are seldom worthwile.[4]

Retrospective analyses as well as prospective studies justify the use of sequential investigation to avoid unnecessary destruction of the biological material and to limit the total duration of the study. In prospective clinical trials ethical aspects may play a role. For example when chemotherapy is one of the trial arms in a trial comparing two cancer therapies, one wishes to expose as few patients as necessary in coming to a decision.

From an economical point of view we performed sequential *t*-tests with multiple control observations per case and compared the results with those of a non-sequential analysis and of simulation studies.

Table 3 Comparison of expected and observed sample size for one control matched per case ($k = 1$)

	H		
	$\lvert\delta\rvert = 0.3$	$\lvert\delta\rvert = 0.4$	$\lvert\delta\rvert = 0.5$
Fixed			
Paired *t*-test	88	50	32
Sequential			
Expected:			
Cox' approximation	57/34	34/20	22/14
Observed:			
Simulation results			
Rushton			
Mean	57/44	36/27	25/18
Median	50/33	31/20	21/13
Kummer			
Mean	64/54	39/31	26/21
Median	57/43	35/25	23/17

Sequential sample sizes are expressed as 'number of case-control pairs necessary to reject H_0/number of case-control pairs necessary to accept H_0'.

Expected sample size for a non-sequential paired *t*-test and expected and observed sample sizes for sequential *t*-tests with matched pairs (i.e. 1 control per case).

The expected average sample numbers (ASN) for a sequential *t*-test with one control observation per case are already smaller than the minimal sample size required for a corresponding non-sequential (=fixed sample size) paired *t*-test (Table 3). (See Appendix III for the calculation of the ASN according to Cox' approximation.[12]) Notable in Table 3 is the fact that both the mean number of case-control pairs required to reject H_0 using Rushton's approximation and the median number using Kummer's function almost equal Cox' approximated ASN. Only the median number of cases necessary to accept H_0 using Rushton's approximation resembles the corresponding ASN according to Cox. Our simulations indicate that Cox' approximation probably underestimates the average sample size, especially the expected ASN needed to accept H_0.

Most sequential *t*-tests of our 'selenium and breast cancer' data (Table 2) resulted in acceptance of H_0 at a considerable smaller number of case-control sets than necessary for a non-sequential analysis.

The simulations confirm these results even better. The largest gain in efficiency as compared to matched pairs is reached with 2 controls per case, when H_0 is rejected. When H_0 cannot be rejected, the gain in efficiency is smaller. The simulated power values are closer to each other for different values of δ using the exact Kummer function than they are using Rushton's approximation.

Rushton’s approximation, on the other hand, is less conservative with respect to the simulated power and thus more economical in its use of case-control sets. Only with the matched-pairs simulations Rushton’s approximation yields a simulated power significantly less than the theoretical power of 0.80. In general, the simulated unreliability using Rushton’s approximation is larger than that using Kummer’s function and more often even larger than the theoretical unreliability of 0.05.

Skovlund and Walløe[13] already drew attention to the conservatism of the sequential *t*-test when applied as a two-sample sequential test. Their smallest value for δ studied was 0.5, however. Neither did they simulate with more than 1 control matched per case.

In theory it is possible that a sequential test continues infinitely. To warrant that a decision is reached, albeit ‘no decision can be made’, it is recommended to set a restriction (e.g. once or twice the fixed sample size) to the total number of cases available for the test.

Our simulations illustrate that there is hardly any effect on the simulated power and unreliability when the sequential test procedure is truncated at twice the fixed sample size.

Truncating the procedure at the fixed sample size results in a simulated power that is still too large, except for the matched-pairs situation using Rushton’s approximation where it is too small. The unreliability resulting from the simulations using Rushton’s approximation with more than one control per case is often (significantly) too large.

When a sequential test is terminated at a small number of observations, point and interval estimates of the case-control difference are rather imprecise. We hold the view that these objections play a less important role when, as in our experimental set-up, a rather ‘qualitative’ answer (‘H_0 can/cannot be rejected’) suffices to distinguish promising new hypotheses from unfruitful ones (see for an example Van Noord[10]).

Group sequential procedures (for matched case-control sets)[15-18] also have the advantage of a reduction in the average sample size as compared to fixed-sample-size plans. There are some differences between group sequential procedures and a one-at-a-time SPRT, however. A one-at-a-time sequential approach can be stopped after every new case-control set, while a group sequential procedure can only be stopped after the next planned inspection. Furthermore, a group sequential procedure cannot come to the decision to accept the null hypothesis until after the last planned inspection. A SPRT can be stopped the very moment that evidence exists that the null hypothesis cannot be rejected anymore.

Therefore, the authors prefer a one-at-a-time SPRT over the group sequential procedure when ethical and/or economical motives play a role. Promising hypotheses as well as unfruitful ones are to be distinguished with as little as possible biological material destroyed or, for that matter, time and/or money spent.

Following Skovlund and Walløe[14], we hold the view that a sequential design might be considered more often in prospective clinical trials as well as in (cohort-nested) case-control studies.

Furthermore, we are of opinion that a sequential *t*-test with 2-4 controls per case is appropriate in case-control studies and other experimental designs where the case material must be used economically and the response is available (almost) immediately. In general the investigation can then be stopped at a lower average sample size as compared to one control per case or a non-sequential test.

The use of exact calculations (the series expansion of Kummer's function) is recommended, although less conservative procedures are to be developed.

Tables and figures summarizing the results from the computer simulations are available from the authors by written request.

1.6 Conclusions

1) A sequential *t*-test with 2-4 controls matched per case in general leads to lower average sample sizes than a matched-pairs sequential *t*-test or a non-sequential analysis. The largest gain in efficiency as compared to matched pairs is reached with 2 controls per case.
2) Rushton's approximation to the logarithm of the likelihood ratio is rather inaccurate and leads to a power that is significantly too small in case of a matched-pairs analysis.
3) The use of Kummer's function (the exact calculation) results in power values which are too conservative.
4) Cox' approximation to the expected average sample number probably underestimates the expected sample size needed to accept H_0.

1.7 Acknowledgement

The study was supported by the Praeventiefonds (the Netherlands), grant No 28-1560.

1.8 References

1. O'Neill RT, Anello C. Case-control studies: a sequential approach. Am J Epidemiol 1978; 108: 415-24.
2. Lachin J. Introduction to sample size determination and power analysis for clinical trials. Contr Clin Trials 1981; 2: 93-113.
3. Ury HK. Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. Biometrics 1975; 31: 643-9.
4. Gail M, Williams R, Byar DP, Brown C. How many controls? J Chron Dis 1976; 29: 723-31.
5. Van Noord PAH, Collette HJA, Maas MJ, Waard F de. Selenium levels in nails of premenopausal breast cancer patients assessed prediagnostically in a cohort-nested case referent study among women screened in the DOM project. Int J Epidemiol 1987; 16: 318-22.
6. Sokal RR, Rohlf FJ. Biometry. 2nd ed. New York: W.H. Freeman and Company, 1981.
7. Wald A. Sequential analysis, New York: John Wiley, 1947.
8. Rushton S. On a two-sided sequential *t*-test. Biometrika 1952; 39: 302-8.
9. Abramowitz M, Stegun IA. Handbook of mathematical functions. New York: Dover Publications Inc., 1968.
10. Van Noord PAH. Selenium and human cancer risk (nail keratin as a tool in metabolic epidemiology). [thesis] Utrecht: University of Utrecht, the Netherlands, 1992.
11. Wald NJ. Use of biological sample banks in epidemiological studies. Maturitas 1985; 7: 59-67.
12. Wetherill GB, Glazebrook KD. Sequential methods in statistics. 3rd ed. London: Chapman and Hall, 1986.
13. Skovlund E, Walløe L. A simulation study of a sequential *t*-test developed by Armitage. Scand J Stat 1987; 14: 347-52.
14. Skovlund E, Walløe L. Sequential or fixed sample trial design? A case study by stochastic simulation. J Clin Epidemiol 1991; 44: 265-72.
15. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977; 64:191-9.
16. Pasternack BS, Shore RE. Group sequential methods for cohort and case-control studies. J Chron Dis 1980; 33: 365-73.
17. Pasternack BS, Shore RE. Sample sizes for group sequential cohort and case-control study designs. Am J Epidemiol 1981; 113: 182-91.
18. Pasternack BS, Shore RE. Sample sizes for individually matched case-control studies: a group sequential approach. Am J Epidemiol 1982; 115: 778-84.

APPENDIX I

The logarithm of the likelihood ratio l_n is a function of δ, n and u^2 and equal to

$$L = \ln(l_n) = \ln M\left(n/2\,;\,\tfrac{1}{2}\,;\,\tfrac{1}{2}\cdot\delta^2\cdot u^2\right) - \tfrac{1}{2}\cdot n\cdot\delta^2 \tag{1}$$

For the nth case-control pair ($n = 1,2,3,...$ successively and one control observation per case) u^2 is equal to

$$u^2 = \left(\sum d_i\right)^2 \Big/ \sum d_i^2 = n\cdot t^2/\left(n-1+t^2\right),\ i = 1, \ldots, n$$

where

$$t^2 = n\cdot\text{mean}(d)^2/\text{var}(d),$$

d_i is the difference between the observation for the case and the control observation, and mean(d) and var(d) stand for the mean and variance of these differences. For every n L is compared to $\ln(\beta/(1\text{-}\alpha))$ and $\ln((1\text{-}\beta)/\alpha$. $M(a;b;x)$ is the confluent hypergeometric function, which can be calculated using Kummer's function[9], a series expansion:

$$M(a;b;x) = 1 + ax/b + a(a+1)x^2/\{b(b+1)2!\} + a(a+1)(a+2)x^3/\{b(b+1)(b+2)3!\} + \ldots$$

We involved 30 terms of this expansion. Rushton's approximation[8] to L is equal to

$$l_1 = \tfrac{1}{2}\cdot\delta\cdot u^3\big/\sqrt{n} + \sqrt{\left(n\cdot\delta^2\cdot u^2\right)} - \left(\tfrac{1}{2}\cdot n\cdot\delta^2 + \ln(2)\right). \tag{2}$$

For k control observations per case the variance of the difference between the case observation and the mean of the k control observations is estimated using the cumulating case-control variance-covariance matrix. This estimate is then substituted as s^2 in the equations mentioned below. (The variance-covariance matrix takes the correlations among cases and controls into account. If we assume negligible correlations among control observations, equal variance for the control observations and equal correlations between the case and each of the controls, s^2 can be approximated by the variance of the differences between the case and the mean of the control observations.) Then Rushton's approximation to L can be calculated by

$$l_1 = \tfrac{1}{2}\cdot\delta\cdot u_k^3\big/\sqrt{n} + \sqrt{\left(n\cdot\delta^2\cdot u_k^2\right)} - \left(\tfrac{1}{2}\cdot n\cdot\delta^2 + \ln(2)\right) \tag{3}$$

with

$$u_k^2 = n\cdot t_k^2\big/\left(n-1+t_k^2\right)$$

and

$$t_k^2 = n\cdot mean(d)^2\big/s^2$$

N.B. For matched case-control observations ($k = 1$) equation (3) is equal to equation (2).

APPENDIX II

Data and calculations of one of the simulations with $\alpha = 0.05$, $1\text{-}\beta = 0.80$, $\delta = 0.5$, $\mu_0 = \mu_1 = 0.8$, $\sigma = 0.15$ and 2 controls per case (see Appendix I for the notation used):

n	Case	Control	Control	s^2	t_2^2	u_2^2	M	L
1	0.911	0.912	0.891					
2	0.919	1.044	0.518	0.008	1.312	1.135	1.312	0.022
3	0.628	0.867	1.029	0.056	0.180	0.248	1.095	-0.284
4	0.781	0.759	0.861	0.037	0.273	0.334	1.174	-0.340
5	0.947	0.740	0.600	0.049	0.022	0.028	1.018	-0.608
6	0.527	0.728	0.791	0.050	0.084	0.099	1.075	-0.677
7	0.814	1.053	0.771	0.042	0.223	0.250	1.230	-0.668
8	0.784	0.730	0.877	0.036	0.263	0.290	1.308	-0.731
9	0.908	0.860	0.826	0.033	0.151	0.167	1.195	-0.947
10	0.745	0.637	0.580	0.032	0.018	0.020	1.025	-1.226
11	0.846	0.659	0.672	0.032	0.032	0.035	1.048	-1.328
12	0.650	0.762	0.919	0.032	0.019	0.020	1.031	-1.470
13	0.898	0.896	0.778	0.030	0.001	0.002	1.002	-1.623

After 13 case-control sets are evaluated, M equals 1.002 and therefore $L = -1.623$ becomes smaller than the lower boundary, $\ln(\beta/(1\text{-}\alpha)) = -1.558$, and thus H_0 cannot be rejected.

When Rushton's approximation to L is applied, the sequential analysis can be stopped after the 10th case-control set, where $l_1 = -1.719$.

APPENDIX III

For matched case-control observations, the average sample number (ASN) for a sequential *t*-test with unknown variance is approximately $(1+\delta^2/2)$ times the ASN for a test with known variance (Cox' approximation, Wetherill and Glazebrook[12]).
Under H_0 this ASN (unknown variance) is about

$$-2/\delta^2 * \{\alpha' \cdot \ln((1-\beta)/\alpha') + (1-\alpha') \cdot \ln((\beta)/(1-\alpha'))\}$$

and under H_1 this ASN is about

$$(1+2/\delta^2) * \{\beta \cdot \ln((\beta)/(1-\alpha')) + (1-\beta) \cdot \ln((1-\beta)/(\alpha'))\}$$

(with $\alpha' = \alpha/2$).

We recognize that Cox' approximation is an asymptotic result and that it is currently unknown how accurate it is.

CHAPTER 2

Efficient use of biological banks for biochemical epidemiology: exploratory hypothesis testing by means of a sequential *t*-test

R. Kaaks[§], I. van der Tweel*, P.A.H. van Noord[#], E. Riboli[§]

[§]International Agency for Research on Cancer, Lyon, France
**Centre for Biostatistics, Utrecht University, Utrecht, the Netherlands*
[#]Department of Epidemiology, Utrecht University, Utrecht, the Netherlands

Epidemiology 1994;5:429-438

Summary

In view of recent advances in molecular and biochemical epidemiology, there is growing interest in the creation of biological banks of blood, urine, tissue, or other biological specimens collected from participants in prospective cohort studies. The existence of biological banks may make it possible to study a multitude of etiologic hypotheses, by comparing biochemical parameters measured in the biological specimens of subjects who will eventually develop the disease of interest ('cases') and of control subjects, using a nested case-control or a case-cohort design. In practice, however, the amount of biological material available per subject (in particular, that of cases) will limit the number of hypotheses that can be tested. The present paper discusses the use of a sequential *t*-test which, compared with an analogous fixed sample procedure, will on average require fewer biological specimens before a given study hypothesis can he accepted or rejected. The sequential test should thus facilitate an early decision on whether a new hypothesis is worth further investigation, while avoiding wasting too much biological material on testing hypotheses that may eventually prove unfruitful. If the test reveals an exposure difference of interest, the study may be extended so that relevant epidemiologic effect measures can be estimated more accurately.

Keywords: biological banks, nested case-control studies, sequential methods, epidemiologic methods.

2.1 Introduction

Following recent developments in 'biochemical' and 'molecular' epidemiology, there is growing interest in the creation of banks of biological samples of material, such as blood or urine specimens, collected from participants in prospective cohort studies.[1,2] After detection of a sufficient number of cases of a given disease (during a given follow-up period), parameters measured in their biological specimens can be compared with those of controls to study specific etiologic hypotheses. Since new laboratory techniques are constantly being developed for the assessment of specific biochemical or molecular parameters, the number of new hypotheses that can be tested is also increasing rapidly. In practice, however, the amount of biological material stored (in particular, that of cases) will limit the number of possible studies.[3] It would therefore be useful to have a statistical method which, at the expense of as little biological material as possible, will distinguish between promising hypotheses, which may be worth further investigation, and less promising ones. Such a method may be particularly useful in exploratory investigations, when there is only limited prior evidence to justify a study based on a large number of biological specimens.

Using *sequential* statistical designs,[4,5] it is theoretically possible to terminate an investigation on a specific hypothesis as soon as sufficient evidence has accumulated for it to be accepted or rejected. On average, sequential analysis will arrive at a decision after substantially fewer observations than equally reliable test procedures based on a fixed sample size. The first sequential procedures were developed during the Second World War,[6] when Wald described the theoretical basis for a sequential probability ratio test (SPRT), and it almost immediately became an important tool for efficient quality control in wartime factories. Nowadays, sequential methods have also been adopted for use in medical research, in particular for the design and analysis of clinical trials.[7,8] So far, however, sequential methods have not been used to a large degree in epidemiologic studies, outside of clinical trials.

The present paper discusses the use of a sequential *t*-test for exploratory hypothesis testing, in cohort-nested case-control studies where the exposure assessment is based on a biochemical marker, obtained by laboratory analysis of stored biological specimens. (To simplify, we shall refer to the biological marker as a measurement of an internal or external 'exposure', although it is clear that markers can also be a measure of individual susceptibility or of intermediate endpoints.)[9,10] The application of the sequential *t*-test will be illustrated using data from a study conducted to examine whether selenium is a potentially protective agent against breast cancer.[11]

2.2 The sequential *t*-test

We shall assume that the biomarker measurements, M, can be considered as values drawn from two normal distributions, for cases and for controls, respectively. We also assume that both distributions have an equal variance, σ^2, but that their means may be different; that is:

$$M \mid \text{case} \sim N(\mu_1, \sigma^2)$$

and

$$M \mid \text{control} \sim N(\mu_0, \sigma^2)$$

The null hypothesis to be tested is that the mean exposures of cases and controls are equal; that is:

$$\text{H}_0\text{: } \mu_1 = \mu_0$$

or

$$\mu_1 - \mu_0 = 0$$

If σ is not known *a priori*, but must be estimated, the magnitude of the mean difference $\mu_1 - \mu_0$ that can be detected with a given power is unknown. The null hypothesis, however, can be re-defined in terms of a *standardized* difference, $\theta = (\mu_1 - \mu_0)/\sigma$ between the mean exposures of cases and of controls:

$$H_0\text{: } \theta = \frac{\mu_1 - \mu_0}{\sigma} = 0$$

If the standard deviation σ is high, then, for a given number of observations, only very large differences will be detectable with sufficient statistical power. Inversely, the power will be higher if σ is small.

A *t*-test can be used to evaluate the null hypothesis against an alternative. In the case of a well defined biological hypothesis, a one-sided alternative may be reasonable; that is:

$$H_1\text{: } \theta \geq \theta_R$$

Here, θ_R is the minimum standardized difference, $(\mu_1 - \mu_0)/\sigma$, that one would find relevant enough to be detected, with a power of at least 1-β and a significance level α. (If the exposure is expected to be higher for controls than for cases, the standardized difference can also be defined as $\theta = (\mu_0 - \mu_1)/\sigma$.) A two-sided alternative can be specified as:

$$H_1\text{: } |\theta| \geq \theta_R$$

Most epidemiologists are familiar with the traditional, fixed sample *t*-test, based on the comparison of the mean exposures of predetermined numbers of cases and controls. The procedure described here, however, uses a sequential sampling of cases and controls within the cohort. This sequential sampling may follow the detection of cases over time. Alternatively, if a large number of cases has already accrued, the sequential sampling can also be performed retrospectively. In the latter situation, the order in which cases are selected does not need to follow the chronological order in which they were detected but can also be based on a random selection process. For each case selected, a random subset of *k* controls is drawn from the disease-free subjects in the cohort. If there are many cases, and if the major concern is to limit the expenditure on laboratory analyses, 1:1 matching (*k* = 1) will give optimal statistical power at a given total cost. When disease incidence rates are low (for example, for a given type of cancer), however, cohort studies must be very large to observe a sufficient number of cases. Additional costs for laboratory assessments, even though sometimes considerable, may then still be low in comparison with the initial investments in the study, and priority may be given to the possibility of studying as many hypotheses as possible with the biological material available. In this case, a higher matching ratio will be more efficient (*k* > 1), as this design will increase the power of the test while keeping constant the number of specimens from cases. A matching ratio greater than 5 will seldom be worthwhile, however.[12] After every new set of one case plus corresponding controls is sampled, the biochemical measurements are compared for all cases and controls processed up to that point to determine whether there is sufficient evidence either to reject or to accept the null hypothesis H_0.

The earliest theory for sequential test procedures (that of the sequential probability ratio test) was initially developed by Wald.[6] According to this theory, a sequential test was based on the logarithm of the following likelihood ratio, L_n, which can be computed after every new case-control set has been sampled:

$$L_n = \frac{\text{the probability of observing the case and control measurements if } H_1 \text{ is true } (\textit{ie,} \text{ if } \theta \geq \theta_R)}{\text{the probability of observing the case and control measurements if } H_0 \text{ is true } (\textit{ie,} \text{ if } \theta = 0)}$$

where n is the number of case-control sets processed so far. A high value of the logarithm of the likelihood ratio, l_n, indicates that, given the measurements observed, the alternative hypothesis H_1 is more likely to be true than the null hypothesis H_0, whereas a low value of l_n indicates that the null hypothesis is more likely to be true. The testing process will continue until one of the following arises:

1. The log-likelihood ratio l_n becomes smaller than a critical minimum value A. In this case, the conclusion is that the standardized difference θ is unlikely to be as large as θ_R, and the null hypothesis H_0 will not be rejected.

2. The log-likelihood ratio l_n becomes larger than a critical maximum value B. In this case, it will be concluded that there is a standardized difference between the average exposures of cases and controls as large as or larger than θ_R, and the null hypothesis H_0 is rejected in favour of the alternative hypothesis.

Whitehead[8] developed a more general approach to sequential test procedures, which includes procedures that are equivalent to Wald's sequential probability ratio tests, and which is based on a log-likelihood function (with unknown parameter θ) rather than on a log-likelihood ratio. The log-likelihood function can be expressed in terms of the parameter θ (for our comparison of two mean exposures still defined as $\theta = (\mu_1 - \mu_0)/\sigma$ as well as of two test statistics, Z and V, which are both computed at each stage of the sequential test procedure. Formulas for the computation of Z and V are given in Appendix I. Z is the so-called 'efficient score for θ' and, for the comparison between quantitative exposures of cases and controls discussed here, is computed as the cumulative difference in exposure divided by an estimate of the unknown standard deviation σ. V is a measure of the amount of information about θ contained in Z, also referred to as 'Fisher's information', and increases as the sequential test procedure progresses. Whitehead has shown that, when θ is small and samples are large, then, at any stage in the sampling process, Z follows approximately a normal distribution with mean θV and variance V.[8p60]

In practice, the sequential testing process can be conveniently presented in the form of a graph, plotting Z against V. The testing process then continues until:

1. Z becomes smaller than the critical value A* = $-a + bV$, in which case H_0 cannot be rejected, or
2. Z becomes larger than the critical value B* = $a + bV$, in which case H_0 will be rejected.

The critical values A* and B* are both linear functions of V. The slope (b) and intercepts ($\pm a$) of these linear functions depend on the values chosen for α, β, and θ_R (see Appendix I). An example of the graphic presentation of the sequential t-test is shown in Figure 1 (further discussed in the next section). The computations for this example, including those for determination of the critical values A* and B*, were performed using the computer program PEST, developed by Whitehead and Brunier.[13]

2.3 An example

Within a cohort of participants in the DOM project, a population-based breast cancer screening program in Utrecht, The Netherlands, toenail clippings were collected and stored in a biological bank. After an average follow-up of 25.7 months, a total of 61 cases of premenopausal breast cancer were detected.[11] Results were reanalyzed using a sequential t-test. The null hypothesis of an equal selenium content in toenails of cases and of controls (H_0: $(\mu_0 - \mu_1)/\sigma = 0$) was tested against the one-sided alternative of a higher selenium content in the control group (H_1: $(\mu_0 - \mu_1)/\sigma \geq \theta_R$). The θ_R-value was chosen equal to 0.25. The significance level and the statistical power were fixed at $\alpha = 0.05$

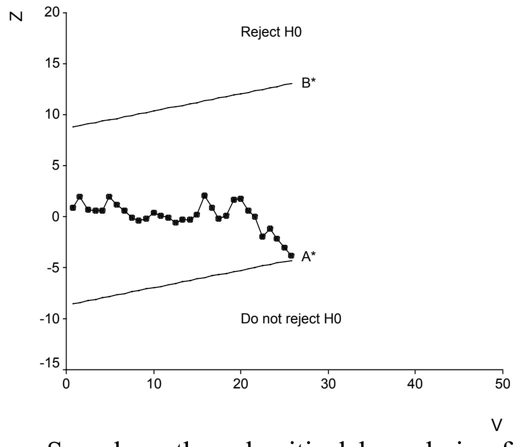


Figure 1 Sample path and critical boundaries for the Selenium and Breast Cancer data (one-sided sequential t-test without matching; $\alpha = 0.05$, $1-\beta = 0.8$, and $\theta_R = 0.25$). A* and B* are the critical boundaries of the test; Z is the so-called 'efficient score' for θ, computed as the cumulative standardized difference between the exposures of cases and controls; V is a measure of the amount of information about θ contained in Z, also referred to as 'Fisher's information' statistic.

(one-sided) and 1-β = 0.80, respectively. Case-control subsets consisted of one case and five controls each and were analyzed in the chronological order in which the cases had been diagnosed.

The results of the sequential testing procedure are shown in Figure 1. After a total number of 31 case-control sets (that is, 31 cases and 155 controls), the sample path of the efficient score Z plotted against V crossed the critical boundary corresponding to no rejection of the null hypothesis.

2.4 Gain in efficiency: the expected sample size

The advantage of sequential procedures is that the expected number of observations (average sample size) needed to reject a given study hypothesis, or not, is smaller than when the test is based on a fixed sampling procedure (that is, with predetermined sample size). Indeed, it has been shown that, when either the null hypothesis H_0 or the alternative hypothesis H_1 is true, the sequential probability ratio test is a more efficient test.[4]

Table 1 shows, for different values of θ_R, the expected sample size for the sequential t-test used in the previous example, as compared with that for a test with a fixed sample size (these expected sample size values can be computed by the PEST program). It can be seen from Table 1 that, for a sequential t-test with the given specifications (one-sided α = 0.05, and 1-β = 0.80), the expected sample size under H_0 is approximately 0.46 times the fixed sample size at all values of θ_R. The expected sample size under H_1 is approximately 0.66 times the fixed sample size. For the open sequential test procedure described here, the expected sample size of the sequential t-test reaches its maximum in situations where the true θ-value is approximately equal to 0.75θ_R, but even then, it remains below the sample size for a classical, fixed sample test of equal reliability.

Table 1 Expected number of case-control sets N in a sequential test for a standardized exposure difference θ, when in reality $\theta = 0$, $\theta = \theta_R$, or $\theta = 0.75\theta_R$:
Test without matching; α = 0.05, 1-β = 0.80

Number of controls per case		Sequential test			Fixed sample test
	θ_R	$\theta = 0$	$\theta = \theta_R$	$\theta = 0.75\theta_R$	
1 (k = 1)	0.15	255	361	414	550
	0.25	92	130	149	198
	0.35	47	67	76	101
5 (k = 5)	0.15	153	216	249	330
	0.25	55	78	90	119
	0.35	28	40	46	61

2.5 Choice of the alternative hypothesis

In sequential test procedures, an explicit definition of the alternative hypothesis H_1 is required, specifying the minimum standardized exposure difference θ_R high enough to be detected with a given statistical power. Specification of the alternative hypothesis, in addition to the null hypothesis H_0, results in a rule that defines at which stage there is sufficient evidence for not rejecting H_0 or for rejecting H_0 in favour of H_1. If there were no such rule for stopping a sequential test procedure without rejection of H_0, the sampling of cases and controls could continue infinitely in those situations where no difference in exposure between cases and controls exists, without ever reaching a conclusion.

The probability that, at a given stage in the sequential testing process, sufficient evidence will have accumulated on whether or not to reject the null hypothesis, H_0, depends on the specific alternative hypothesis against which H_0 is tested. For example, imagine a situation in which, at a given number of observations, there appears to be little difference between the mean exposures of cases and of controls. In such situations, the log-likelihood ratio l_n would tend to be small if the alternative hypothesis were defined by a relatively extreme θ_R -value, and H_1 would appear less likely to be true than H_0 given the case and control observations. At a small value of θ_R specified, however, the same set of case and control observations would have led to a higher log-likelihood ratio. The probability of concluding the test procedure with no rejection of the null hypothesis would therefore be higher in the first case (high value of θ_R) than in the second (small value of θ_R). Of course, this phenomenon is not specific for sequential tests in particular but occurs also in statistical procedures based on a fixed sample size. The example does underscore, however, that the choice of the alternative hypothesis (that is, the value for θ_R) should be well motivated, in terms of potential public health impact or strength of the biological relation to disease.

For the sequential *t*-test discussed here, θ_R is specified as a standardized difference between the mean exposures of cases and of controls. For epidemiologists, who are more familiar with the definition of study hypotheses in terms of measuring disease risk, this specification of the alternative hypothesis may be difficult to interpret. If the disease incidence is low over the entire range of exposures (that is, the 'rare disease' assumption), however, and the alternative hypothesis is true, it is possible to compute a minimum expected odds ratio value, OR_R, for different quantile levels of the distribution of exposure measurements within the cohort (from which cases and controls were drawn). For instance, the expected odds ratio for the highest *vs* the lowest quintile of the exposure distribution equals:

$$OR_R(Q_5 - Q_1) = e^{2.80\,\theta_R}$$

(See Appendix IIA.)

Thus, for an alternative hypothesis defined as $\theta_R \geq 0.25$, the minimum expected odds ratio of disease for the highest *vs* the lowest quintile of exposure measurements approximately equals 2.0. An extended list of expected odds ratio estimates, for different values of θ_R, is given in Table 2.

Table 2 Expected Odds Ratios, $\mathrm{OR_R}(Q_5 - Q_1)$, for the highest vs the lowest quintile of the exposure distribution in the cohort, under the alternative hypothesis $\theta = \theta_R$: study without matching

θ_R	$\mathrm{OR_R}(Q_5\text{-}Q_1)$
0.15	1.5
0.20	1.8
0.25	2.0
0.30	2.3
0.35	2.7
0.40	3.1

2.6 Analysis of matched studies: the pairwise sequential *t*-test

The sequential test procedure described thus far did not take account of any potential confounding factors. In many situations, however, it may be necessary to adjust for potential confounding factors such as age, length of follow-up, or additional risk factors such as body weight and menopausal status. Using the sequential procedure described here, adjustments for confounding can be made by matching cases and controls for such additional risk factors. In case-control studies nested within a cohort, this may not be too complicated, since there will be a pool of disease-free subjects in which to find matched controls (unless there are many matching criteria). Whenever a matched study design is used, however, the matching should be reflected in the analysis to obtain unbiased results.

A matched sequential *t*-test can be based on the pairwise differences between the exposure measurement of a case, and the mean exposure measurement of *k* controls belonging to the same matched subset.

We shall assume that these differences, D_i (where $i = 1,\ldots,n$ is the number of case-control sets evaluated), will be normally distributed:

$$D_i \sim N\left(\delta, \tau_k^2\right)$$

where δ is the mean, and τ_k^2 is the variance of the differences D_i. As in the unmatched situation, the hypotheses H_0 and H_1 can then be defined in terms of a standardized difference θ:

$$\mathrm{H}_0\text{: } \theta = \frac{\delta}{\tau_k} = 0$$

and, for the one-sided alternative of a higher exposure for cases than for controls,

$$\mathrm{H}_1\text{: } \theta = \frac{\delta}{\tau_k} \geq \theta_R$$

(Note: If the exposure is expected to be higher for controls, the alternative hypothesis may be defined as H_1: $\theta = \delta \,/\, \tau_k \leq -\theta_R$. A two-sided alternative may be specified as H_1: $|\theta| \geq \theta_R$.)

The computation of the statistics *Z* and *V* is slightly different from that for the unmatched situation (see Appendix IB). Nevertheless, the critical boundaries of the test, A* and B*, remain the same (since these depend only on the values chosen for α, β, and θ_R). Also, with respect to a fixed sample test, efficiency gains will be made similar to those in the unmatched situation, in terms of a decrease in the expected sample size.

Again, with some additional assumptions, it is possible to compute expected odds ratio values under the alternative hypothesis, $\theta = \theta_R$, for quantile levels of the within-stratum exposure distribution (strata being defined by the matching variables). As before, it will be assumed that for cases and controls the exposure measurements have an equal variance, σ^2, and that the overall incidence of disease is low. In addition, it will be assumed that, after matching, exposure measurements are equally correlated between controls or between cases and controls. The variance of the exposure differences D_i, between a case and *k* matched controls, can then be written as:

$$\tau_k^2 = \frac{k+1}{k}\left(\sigma^2 - \gamma\right) = \frac{k+1}{k}\sigma'^2,$$

where γ is the covariance between the exposure measurements of cases and controls (due to the matching), and σ'^2 is the average variance of exposure among controls (and thus, approximately, in the full cohort) within strata defined by the matching variables. The expected odds ratio for the within-stratum difference between the upper and the lower quintiles of the exposure distribution will be approximately equal to:

$$\mathrm{OR}_\mathrm{R}\left(\mathrm{Q}_5 - \mathrm{Q}_1\right) = e^{2.80\sqrt{[(k+1)/k]}\;\theta_R},$$

where *k* is the control-to-case matching ratio, and with $\theta_R = \delta/\tau_k$ (see Appendix IIB). In Table 3, some expected odds ratio values are given for different values of θ_R and *k*.

Table 3 Expected Odds Ratios, $OR_R(Q_5–Q_1)$, for the highest vs the lowest quintile of the exposure distribution of the cohort within strata of the matching variables, under the alternative hypothesis that $\theta = \theta_R$ (1:*k* matching)

	$OR_R(Q_5–Q_1)$				
θ_R	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
0.15	1.8	1.7	1.6	1.6	1.6
0.20	2.2	2.0	1.9	1.9	1.8
0.25	2.7	2.4	2.2	2.2	2.2
0.30	3.3	2.8	2.6	2.6	2.5
0.35	4.0	3.3	3.1	3.0	2.9
0.40	4.9	3.9	3.6	3.5	3.4

2.7 Discussion

We have shown how a sequential *t*-test can be applied in case-control studies where the exposure measurement is a continuous variable. The use of the sequential probability ratio method in epidemiologic studies has been suggested previously by O'Neill and Anello,[14] who described a sequential test for analyzing (matched pair) case-control studies, with a dichotomous exposure variable. Thus far, however, this has not been put into practice widely. An explanation may be that the advantage of a smaller expected sample size does not outweigh certain drawbacks in the use of a sequential probability ratio procedure, particularly in studies where (dichotomous) exposure assessments are based on information derived from questionnaires. One such drawback may be the fact that epidemiologists are not familiar enough with sequential statistical methods, and, until recently, no simple computer software for sequential analysis was widely available. Another drawback may be that the sequential probability ratio procedure does not allow flexible, multivariate data modelling for the control of varying sets of confounding factors. In spite of these various drawbacks, a strong argument in favour of the use of sequential methods is the desire to make optimal use of material from biological banks, reducing the number of biological samples needed to test a given hypothesis.

In a sequential design, the number of case-control sets that will be sampled before a conclusion is reached is a random variable. On average, the sample size needed will be smaller than that of an equivalent fixed sample test, but occasionally, larger numbers of observations may be needed. This feature may introduce some uncertainty into the process of setting a budget for grant requests. Nevertheless, budgets can be reasonably planned on the basis of the 90th percentiles of the sample size distribution. The PEST program contains a subroutine for the computation of these percentiles, at the planning stage of a study. Further details about these computations are given in Whitehead's book on sequential clinical trials.[8]

The sequential *t*-tests described in this paper can be useful, especially in exploratory studies, to decide, at the expense of as little biological material as possible, whether a new hypothesis seems worth further investigation, or whether it is more likely that it will eventually prove unfruitful. It is generally agreed, however, that the use of hypothesis testing is an unsatisfactory way of assessing and presenting epidemiologic findings, and that results should rather be presented as estimates of relevant measures of exposure-disease association and their confidence intervals.[15,16] Therefore, after terminating the sequential test, and irrespective of whether the null hypothesis is rejected or not, final results should always include such point and interval estimates, describing the association between the marker values and disease risk (for instance, in terms of relative risks for different quantiles of the marker assessments). Since, on average, a sequential test will terminate at a smaller sample size than an equivalent fixed sample procedure, estimates of epidemiologic effect measures may be relatively imprecise. Once a given hypothesis has proved of interest, however, (that is, if the null hypothesis of ‘no difference’ in exposure is rejected), the investigator may decide to extend the number of laboratory assessments, so as to increase the precision of the study. The number of additional assessments needed to reach sufficient precision can then be determined from the standard error of effect estimates at the end of the sequential test, as in a double sampling design.[17]

The combination of sequential testing and subsequent estimation of epidemiologic effect measures -with or without further extension of the study- can be seen as a two-step procedure, which will tend to result in effect estimates with the desired degree of precision if there is a dear difference in exposure, or in less precise estimates if no exposure difference of interest exists. In the latter case, on average more biological samples will be saved for the investigation of other hypotheses.

O'Neill and Anello[14] have described how, for a dichotomous exposure variable and for matched case-control pairs, the critical values of a sequential test can be interpreted in terms of odds ratio values. We have shown that, under the rare disease assumption, and for a matched or an unmatched case-control design, similar interpretations can be given to the critical θ_R-value of a sequential *t*-test for comparison of cases and controls by a continuous exposure variable. Due care must be taken to avoid misinterpretation, however. The sequential procedures described in this paper essentially provide a test for a difference between the mean exposures of cases and of controls and do not give the same results as a test of statistical significance for odds ratios at different quantile levels of exposure. It is possible to compute *expected* odds ratio values for different quantile categories of exposure, such as quartiles or quintiles, under the assumption that the alternative hypothesis, $\theta = \theta_R$ is true (that is, that a certain standardized difference in mean exposure actually exists). The relation between a θ_R -value chosen and expected odds ratio values for different quantile levels of exposure is of interest only insofar as it may help define a reasonable θ_R -value for the alternative hypothesis. Within this context, the choice of quintile levels of exposure was, of course, quite arbitrary; computation of expected odds ratio values for tertiles or quartiles could be equally informative.

The exact value that should be chosen as a reference odds ratio value OR_R (as defined, for instance, for quintiles) may depend on the specific hypothesis to be tested, as well as on the potential relevance of the exposure in terms of attributable risk (that is, also taking into account the prevalence of exposure within a population). O'Neill and Anello[14] recommend specifying that the alternative hypothesis should correspond to an odds ratio not greater than about 2.0 for exposed *vs* nonexposed subjects (the exposure in their paper being defined as a dichotomous variable). We agree that the value of θ_R should always correspond to relatively small expected odds ratio values, so that a failure to reject the null hypothesis can be interpreted as the absence of any relevant association between exposure and disease risk. Of course, it should also be kept in mind that, due to intra-individual variation over time, many biochemical markers will provide only an approximate estimate of the true risk factor of interest, and that the observed association with disease risk (also in terms of a standardized difference between mean exposures) may therefore be attenuated.

In this paper, it was assumed that the sequential testing process proceeds in steps corresponding to case-control sets consisting of only one case and its k controls. It will often be more practical, however, to run laboratory analyses in batches of more than only one case-control set at a time. It is possible to perform the sequential probability ratio test on case-control sets each comprising multiple cases. The only disadvantage of such larger inspection intervals is that there can be some ‘overrunning’ of the critical boundary, by the sample path of Z plotted against V. The number of observations may thus exceed the number that was actually required to reach a conclusion, and part of the advantage of sequential methods, in terms of a reduction in expected sample size, will be lost. This loss of efficiency resulting from overrunning can be limited by including only a relatively small number of cases in each group of observations. We have discussed only so-called ‘open’ or ‘nontruncated’ procedures, in which no upper limit has been set to the number of observations needed before a conclusion is reached. Therefore, although sequential procedures will on average require fewer case-control comparisons than equivalent tests based on a fixed sample size, there may be occasions on which the sequential procedure terminates after a much larger number of observations than would have been required for a classical, fixed sample test. In ‘closed’ or ‘truncated’ sequential procedures, an upper limit is fixed for the actual number of observations that may be needed to reach a conclusion. For instance, it may be decided that the null hypothesis will not be rejected if the number of case-control comparisons becomes larger than twice the normal sample size for a fixed sample test without reaching the critical boundaries, A* or B*. Such an additional stopping rule will then affect α and β to some extent. If the maximum number of observations chosen is sufficiently large, however, the effects on these error probabilities will be relatively small. Whitehead and Brunier's PEST computer program[13] provides an option for the analysis of sequential studies with a truncated design. More extensive discussions of truncated sequential procedures are given in Whitehead's textbook on sequential clinical trials,[8] as well as by Wetherill and Glazebrook[5] and by Armitage.[7]

Aliquots of biological specimens such as blood serum cannot be thawed and refrozen too frequently without potentially causing changes in the biochemical parameters of interest. The volume of aliquots, however, may often be sufficiently large to allow more than one type of biochemical analysis within the same laboratory. It would thus be possible to study several etiologic hypotheses in parallel, based on different biochemical markers measured in the same aliquot. The simple sequential tests described in this paper are based on the concept of studying only one type of exposure measurement in relation to a single type of disease. Further development of sequential statistical methods is needed, so that such multiple, parallel hypotheses can be evaluated simultaneously with minimal loss of biological material.

2.8 Acknowledgements

The authors thank Annie Sasco and Paul Demers for critically reading earlier drafts of this manuscript.

2.9 References

1. De Waard F, Collette HJA, Rombach JJ, Baanders-van Halewijn EA, Honig C. The DOM project for the early detection of breast cancer, Utrecht, the Netherlands. J Chronic Dis 1984: 37: 1-44.
2. Riboli E. Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC). Ann Oncol 1992; 3: 783-91.
3. Van Noord PAH. Selenium and human cancer risk (nail keratin as a tool in metabolic epidemiology). [thesis] Utrecht: University of Utrecht, the Netherlands, 1992.
4. Kendall MG, Stuart A. Sequential methods. In: Kendall MG, Stuart A. The advanced theory of statistics. vol. 2. London: Griffin, 1967; 624-58.
5. Wetherill GB, Glazebrook KD. Sequential methods in statistics. 3rd ed. London: Chapman and Hall, 1986.
6. Wald A. Sequential analysis, New York: John Wiley, 1947.
7. Armitage P. Sequential medical trials. 2nd ed. Oxford: Blackwell Scientific Publications, 1975.
8. Whitehead J. The design and analysis of sequential clinical trials, 2nd ed. Chicester: Ellis Horwood Limited, 1992.
9. Hulka BS, Wilkosky TC, Griffith JD. Biological markers in epidemiology. New York: Oxford University Press, 1990.
10. Riboli E, Rönnholm H, Saracci R. Biological markers of diet. Cancer Surv 1987; 6: 685-718.
11. Van Noord PAH, Collette HJA, Maas MJ, Waard F de. Selenium levels in nails of premenopausal breast cancer patients assessed prediagnostically in a cohort-nested case referent study among women screened in the DOM project. Int J Epidemiol 1987; 16: 318-22.
12. Ury HK. Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. Biometrics 1975; 31: 643-9.
13. Whitehead J, Brunier H, Department of Applied Statistics. PEST Programme. Version 2.2. Reading: University of Reading, U.K., 1992.
14. O'Neill RT, Anello C. Case-control studies: a sequential approach. Am J Epidemiol 1978; 108: 415-24.
15. Rothman KJ. A show of confidence. N Engl J Med 1978; 299: 1362-3.
16. Gardner MJ, Altman D. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J 1986; 292: 746-50.
17. Govindarajulu Z. Sequential estimation. ch. 4. In: Sequential statistical procedures. London: Academic Press, 1975.
18. Rosner B, Hennekens CH. Analytical methods in matched pair epidemiological studies. Int J Epidemiol 1978; 7: 367-72.

APPENDIX I

Computation of the statistics *Z* and *V*, and formulas for critical boundaries A* and B*

A. Analysis without Matching

After the *n*th case-control subset, the following sample statistics will be available:

	Cases	Controls	All
Number of observations	n	nk	$n(k + 1)$
Sum of observed exposures	S_1	S_0	S
Sum of squares	Q_1	Q_0	Q

The statistics *Z* and *V* are computed from the cumulative sums, S_1 and S_0, and cumulative sums of squares, Q_1 and Q_0, of the exposure measurements of cases and controls, respectively. (See Whitehead.[8,pp57-62]) The efficient score statistic *Z* is computed as:

$$Z = \frac{nkS_1 - nS_0}{n(k+1)C},$$

where

$$C^2 = \frac{Q}{n(k+1)} - \left(\frac{S}{n(k+1)}\right)^2.$$

It is to be noted that C^2 is a maximum likelihood estimate of σ^2, under the null hypothesis $\theta = 0$. Thus, *Z* is equivalent to the cumulative difference between the exposure measurements of cases and of controls, divided by a maximum likelihood estimate of the standard deviation σ.

Fisher's information statistic *V* is computed as:

$$V = \frac{n \cdot nk}{n(k+1)} - \frac{Z^2}{2n(k+1)}.$$

Coefficients for critical boundaries A* and B* are computed as:

$$a = \frac{\ln\left(\frac{1-\beta}{\beta}\right) + \ln\left(\frac{1-\alpha}{\alpha}\right)}{2\theta_R}$$

and

$$b = \frac{\theta_R}{\ln\left(\frac{1-\beta}{\beta}\right) + \ln\left(\frac{1-\alpha}{\alpha}\right)} \ln\left(\frac{1-\alpha}{\alpha}\right)$$

For $\alpha = 0.05$, $\beta = 0.2$ and $\theta_R = 0.25$, this leads to $a = 8.661$ and $b = 0.17$.

B. Matched

The following sample statistics can be computed after each new case-control subset:

Number of case-control sets	n
Sum of exposure differences D_i	S
Sum of squared differences D_i^2	Q

The efficient score statistic Z is computed as:

$$Z = \frac{S}{C}$$

where

$$C^2 = \frac{Q}{n}$$

Again, C^2 corresponds with a maximum likelihood estimate of the variance of the exposure differences D_i, under the null hypothesis ($\theta = 0$). Fisher's information statistic is computed as:

$$V = n - \frac{Z^2}{2n}$$

(See Whitehead[8,pp67-68])

APPENDIX II

Relation between θ_R and the expected odds ratio for the upper *vs* the lower quintile of exposure

A. Analysis without matching

Suppose that, both among cases and among controls, exposure measurements M have normal distributions with different means but with an equal variance:

$$M \mid \text{case} \sim N\left(\mu_1, \sigma^2\right),$$

and

$$M \mid \text{control} \sim N\left(\mu_0, \sigma^2\right)$$

If the probability density functions of both exposure distributions are given by φ_1(M) = Pr(M | case) and φ_0(M) = Pr(M | control), respectively, then, for a given difference in exposure, $\Delta = m_1 - m_0$, the odds ratio of disease can be written as:

$$\text{OR}(\Delta) = \frac{\phi_1(m_1)/\phi_0(m_1)}{\phi_1(m_0)/\phi_0(m_0)} = e^{(\mu_1-\mu_0)(m_1-m_0)/\sigma^2} = e^{\theta\Delta/\sigma}$$

Here, the standard deviation σ is unknown.

If the disease incidence in the cohort is low, however, the distribution of exposure measurements of the controls will be approximately identical to the exposure distribution in the entire cohort. Then, for subjects belonging to different *quantile* categories of this distribution, the expected difference in exposure Δ can be expressed as a number of unknown standard deviations σ. The expected exposure measurement above a given cutpoint value L can be computed as the mean of a truncated normal distribution:

$$E(M|M > L) = \mu_0 + \sigma\left(\frac{\phi[(L-\mu_0)/\sigma]}{1-\Phi[(L-\mu_0)/\sigma]}\right),$$

where $\phi(u)$ is the probability density function, and $\Phi(u)$ is the cumulative distribution function of the standard normal distribution at the point u. If L is chosen to be the cutpoint for the highest *quintile* of exposure, we find from the normal distribution table that $(L\text{-}\mu_0)/\sigma = 0.84$. The average exposure in the highest quintile is thus expected to be equal to:

$$E\left(M|M > L\right) = \mu_0 + \sigma\left(\frac{\phi(0.84)}{1-0.80}\right) = \mu_0 + 1.40\sigma$$

Likewise, the average exposure in the lowest quintile is expected to he equal to $E(M \mid M < -L) = \mu_0 - 1.40\sigma$. Thus, the difference between the average exposures in the highest and lowest quintiles will be equal to $\varDelta = 2.80\sigma$.

The expected odds ratio for the highest *vs* the lowest quintile of exposure can now be written as a function of θ:

$$\mathrm{OR}_{\mathrm{R}}\left(\mathrm{Q}_5 - \mathrm{Q}_1\right) = e^{\theta\varDelta/\sigma} = e^{2.80\theta}$$

Inversely, this function can be used to compute the value for θ_R that corresponds with a minimal expected odds ratio, $\mathrm{OR}_{\mathrm{R}}(\mathrm{Q}_5 - \mathrm{Q}_1)$ for the upper *vs* the lower quintile of exposure. For instance, an expected odds ratio of 2.00 corresponds with a standardized difference between the mean exposures of cases and controls equal to $\theta_R = \ln(2.00)/2.80 \approx 0.25$.

B. Matched analysis

Suppose that, in a matched pairs study, $\mathrm{D_i}$ is the difference between the exposure measurement for the *i*th case and its matched control, and let the distribution of such differences be given by $\mathrm{D}_i \sim N\left(\delta, \tau_1^2\right)$. Then, as was previously derived by Rosner and Hennekens,[18] the odds ratio for a difference $D_i = \varDelta$ can be computed as:

$$\mathrm{OR}\left(\varDelta\right) = \frac{\Pr\left(D_i = \varDelta\right)}{\Pr\left(D_i = -\varDelta\right)} = \frac{e^{-\frac{1}{2}(\varDelta-\delta)^2/\tau_1^2}}{e^{-\frac{1}{2}(-\varDelta-\delta)^2/\tau_1^2}} = e^{2\theta_1\varDelta/\tau_1},$$

with

$$\theta_1 = \frac{\delta}{\tau_1}$$

Assume, moreover, that, for unmatched cases and controls, the exposure distributions have an equal variance and that, after matching, the exposure measurements are equally correlated between controls (if more than one control is matched per case) or between cases and controls. The variance of the exposure differences, D_i, between a case and a single matched control will then be equal to:

$$\tau_1^2 = 2\left(\sigma^2 - \gamma\right) = 2\sigma'^2$$

where γ is the covariance between the exposures of cases and of controls (due to the matching). In this case, $\sigma'^2 = \sigma^2 - \gamma$ can be interpreted as the average variance of exposure measurements among controls (and thus, approximately, in the full cohort) within strata defined by the matching variables. The expected difference between the top and bottom quintiles of the within-stratum exposure distribution can then be written as:

$$\varDelta = 2.80\sigma' = 2.80\frac{\tau_1}{\sqrt{2}}.$$

The odds ratio corresponding with this difference in exposure equals:

$$\mathrm{OR}_{\mathrm{R}}\left(\mathrm{Q}_5 - \mathrm{Q}_1\right) = e^{2\theta_1 \Delta / \tau_1} = e^{2.80\sqrt{2}\theta_1}$$

with

$$\theta_1 = \frac{\delta}{\tau_1}$$

If $k > 1$ controls are matched per case, the variance of the exposure differences D_i becomes smaller:

$$\tau_k^2 = \frac{k+1}{k}\sigma'^2,$$

and we can write:

$$\theta_1 = \delta / \tau_1 = \sqrt{\left[(k+1)/2k\right]}\,\delta / \tau_k = \sqrt{\left[(k+1)/2k\right]}\,\theta_k$$

Thus, with k controls per case,

$$\mathrm{OR}_{\mathrm{R}}\left(\mathrm{Q}_5 - \mathrm{Q}_1\right) = e^{2.80\sqrt{2}\theta_1} = e^{2.80\sqrt{\left[(k+1)/k\right]}\theta_k}$$

CHAPTER 3

Comparison of one-sample two-sided sequential *t*-tests for application in epidemiological studies

I. van der Tweel*, R. Kaaks§, P.A.H. van Noord#

**Centre for Biostatistics, Utrecht University, Utrecht, the Netherlands*
§International Agency for Research on Cancer, Lyon, France
#Department of Epidemiology, Utrecht University, Utrecht, the Netherlands

Statistics in Medicine 1996; 15: 2781-2795

Summary

In epidemiological prospective cohort studies exposure levels of cases with disease and disease-free control subjects can be measured by laboratory analysis of previously stored biological specimens. In such studies, a sequential *t*-test can be used for preliminary evaluations, at the expense of the smallest possible number of specimens, of whether a new aetiological hypothesis is worth further investigation or whether specimens should rather be spared to test other, more fruitful, hypotheses. For this purpose, we recently compared two sequential probability ratio tests (SPRTs), in which the log-likelihood ratio was either based on an approximation, or computed exactly, and which were adapted to account for various control-to-case matching ratios. The tests turned out to be relatively conservative, particularly in terms of the significance level achieved. In the present paper, we compare an SPRT for matched or paired data based on Rushton's approximation to the log-likelihood ratio with a profile log-likelihood method developed by Whitehead. The comparison is partly mathematical, and partly based on computerized simulations. Average sample size for a sequential test is already smaller than for the equivalent fixed sample test. Increasing the number of controls matched per case further reduces the average sample size necessary to come to a decision. We show that, irrespective of the number of controls per case, pre-specified levels of statistical power and significance are respected closely by Whitehead's method, but not by Rushton's SPRT. This last procedure can lead to a significant loss in power. Since, in addition, Whitehead's method has been implemented in a commercially available computer program ('PEST'), we conclude that this method can be preferred above the methods we described earlier. Moreover, compared with the method of Rushton, Whitehead's method has the advantage that it can also be applied to groupwise inspection of the data and that it can also be converted easily into a truncated procedure

3.1 Introduction

During the Second World War, Wald[1] developed his theory for sequential statistical test procedures as a tool for efficient quality control in wartime factories. Nowadays, sequential statistical methods have also found important applications in medical research, in particular, in randomized clinical trials evaluating the effect of alternative medical treatments, where, for ethical reasons, it is desirable to reach a conclusion with the smallest possible number of patients enrolled. When the effect of the treatment is measured as a continuous variable with an approximately normal distribution, a simple sequential *t*-test on paired or unpaired observations can be used to evaluate whether the mean effect differs between the treatments. In the first clinical application of sequential analysis a one-sample sequential *t*-test was performed on paired observations in the same subject[2]. Facey[3] applied a sequential *t*-test to design a phase II efficacy trial with a normally distributed outcome variable.

Recently we have described how a sequential *t*-test can also be applied usefully in prospective epidemiological studies, where exposure levels of cases with disease and of matched controls are measured by laboratory analyses of blood or tissue specimens stored previously in a biological bank[4-6] (this type of measurement is often (log-)normally distributed). The continuous development of new laboratory techniques for measurement of biochemical or molecular parameters has allowed evaluation of a steadily increasing number of aetiological hypotheses. Nevertheless, the number of hypotheses that can be evaluated in practice will often be limited because only small amounts of biological material are available and because laboratory tests often lead to its destruction. It is thus useful to have a statistical method which, using the smallest possible number of biological specimens, can distinguish between promising and less promising aetiological hypotheses. Using a sequential *t*-test to evaluate whether or not there is a relevant mean exposure difference between cases and controls, up to 50 percent of the biological material of cases can be spared for the investigation of potentially more fruitful hypotheses if in reality the null hypothesis of ‘no difference’ is true[5]. On the other hand, if the null hypothesis is rejected, additional specimens may be analysed to allow more precise estimation of relative risks or other epidemiological measures of association[6].

Within this context, we[5] evaluated the operating characteristics of two types of sequential probability ratio test (SPRT) in which the log-likelihood ratio was either based on an approximation developed by Rushton or calculated exactly using the so-called Kummer function. In prospective cohort studies on relatively rare forms of disease there are many more potential control subjects than cases, and thus the biological specimens will be particularly precious for cases. Therefore, to reduce further the number of specimens to be analysed for cases, multiple controls were matched to each case, and the sequential *t*-tests were modified to account for control-to-case matching ratios greater than one. The two types of *t*-test employed were found to be relatively conservative, particularly in terms of the significance level achieved. On the contrary, Rushton's approximation led to a power which was significantly too small for matched pairs. For one-to-one matching Skovlund and Walløe[7] also concluded that exact calculation of the log-likelihood ratio resulted in an over-conservative test.

In the meantime, Whitehead[8] published his theory for the planning and evaluation of sequential trials based on a profile log-likelihood approach, implemented in practice using the computer program ‘PEST’[9]. In the present paper we derive the test statistics for studies with more than one control matched per case and compare a one-sample two-sided sequential *t*-test following Whitehead's approach with an SPRT using Rushton's approximation. The comparison is made mathematically and by simulations.

3.2 Two types of sequential *t*-test

Let X_i be the observation for the *i*th case, Y_{ij} the observation for the *j*th control ($j = 1,2,...k$) matched to the *i*th case, $Y_{i.}$ the mean over the *k* controls matched to the *i*th case (within strata of potential confounding factors), and $D_i = X_i – Y_{i.}$, for $i = 1,2,...$ (each case is matched to a fixed number, *k*, of controls). Further, let $X_i \sim N(\mu_x; \sigma^2)$ and $Y_{ij} \sim N(\mu_y; \sigma^2)$, then $D_i \sim N(\mu; \sigma_k^2)$ where $\mu = \mu_x - \mu_y$, $\sigma_k^2 = \sigma^2(1-\rho)\times(k+1)/k$ and ρ is the correlation between observations in a case-control set. The parameter θ_k, as a measure for the standardized difference, is defined as μ / σ_k, where σ_k is equal to $\sigma_1\sqrt{\{(k+1)/2k\}}$. The null hypothesis to be tested is formulated as H_0: $\theta_k = 0$ and the two-sided alternative hypothesis as H_1: $|\theta_k| \geq \theta_R$.Tests are performed with a two-sided type I error 2α and a type II error β.

We have shown elsewhere[6] how for various control-to-case matching ratios the value of θ_R can be related to an expected odds ratio, for example, between the top and bottom quintiles of measured exposure levels. The alternative hypothesis can thus be specified as a mean, standardized difference in exposure level θ_R corresponding to a minimum odds ratio considered of aetiological or public health relevance, and important enough to be detected with given significance level and power.

3.3 The SPRT using Rushton's approximation

The test statistic for Wald's sequential probability ratio test[1] is based on the likelihood ratio

$$L_n = \frac{\text{likelihood of observed results given } H_1 \text{ is true}}{\text{likelihood of observed results given } H_0 \text{ is true}}$$

for *n* data sets processed up to a given point.

For a one-sample two-sided sequential *t*-test the likelihood ratio can be written as

$$L_n = e^{-\frac{1}{2}n\theta_R^2} M\left(\frac{n}{2}; \frac{1}{2}; \frac{1}{2}\theta_R^2 U^2\right) \qquad (1)$$

where

$$U^2 = \frac{\left(\sum_{i=1}^{n} D_i\right)^2}{\left(\sum_{i=1}^{n} D_i^2\right)}$$

and $M(a;b;x)$ is the confluent hypergeometric function[10]. For pairwise matching (k = 1) Rushton[11] developed an approximation to the log-likelihood ratio, while for k controls per case, we extended Rushton's test statistic to

$$l_n = \ln(L_n) = \frac{\theta_R^2}{4} U^2 \frac{2k}{k+1} + \sqrt{\left(n\theta_R^2 U^2 \frac{2k}{k+1}\right)} - \frac{n}{2}\theta_R^2 \frac{2k}{k+1} - \ln(2) \tag{2}$$

The null hypothesis is tested against the composite alternative hypothesis H$_1$: $|\theta_k| \geq \theta_R$, making use of two critical boundaries:

$$l_R = \ln(\beta/(1-2\alpha)) \text{ and } u_R = \ln((1-\beta)/2\alpha)$$

The sampling process continues as long as l_n (eq.(2)) remains between these boundaries. When l_n becomes larger than u_R, the process is stopped and H$_0$ rejected. Likewise, the process is stopped, but with the acceptance of H$_0$, when l_n becomes smaller than l_R.

3.4 The sequential test following Whitehead's approach

Rather than based on a log-likelihood ratio, Whitehead's procedure is based on a profile log-likelihood function, considering σ_1 as a nuisance parameter. The test statistics are Z (the 'efficient score for θ_k') and V ('Fisher's information about θ_k contained in Z'). For k controls matched per case we extended the test statistics for a one-sample two-sided sequential t-test (see Appendix I), to

$$Z = \sum_{i=1}^{n} D_i \sqrt{\left(\frac{2k}{k+1}\frac{n}{\sum_{i=1}^{n} D_i^2}\right)}, \qquad V = n\frac{2k}{k+1} - \frac{Z^2}{2n} \tag{3}$$

For comparison we considered an open double sequential test, which is a combination of two one-sided tests leading to stopping boundaries of the form

$$\begin{aligned} Z &= a + bV(u_{W1}) \qquad & Z &= -a + bV(u_{W2}) \\ Z &= a - bV(l_{W2}) \qquad & Z &= -a - bV(l_{W1}) \end{aligned}$$

The slopes ±b and intercepts ±a of the boundaries are functions of the error probabilities 2α and β, as well as of the choice of a minimum standardized exposure difference, θ_R, that defines the alternative hypothesis H$_1$: $|\theta_k| \geq \theta_R$

$$b = \frac{\theta_R}{\ln\left(\frac{1-\beta}{\beta}\right) + \ln\left(\frac{1-\alpha}{\alpha}\right)} \ln\left(\frac{1-\alpha}{\alpha}\right) \tag{4}$$

$$a = \frac{\ln\left(\frac{1-\beta}{\beta}\right) + \ln\left(\frac{1-\alpha}{\alpha}\right)}{2\theta_R} \tag{5}$$

Throughout this paper, we shall assume that the boundaries of the test are placed symmetrically about the horizontal axis.

The sequential test following Whitehead's approach is conveniently presented in the form of a graph, plotting Z against V (see Figure 1). The testing process continues as long as the sample path formed by successive Z-values plotted against corresponding V-values remains between the boundaries u_{W1} and u_{W2} or between l_{W1} and l_{W2}. The sampling is stopped and H_0 rejected when the sample path crosses u_{W1} or l_{W1}. The test is stopped with acceptance of H_0 when the sample path crosses u_{W2} or l_{W2}. The test is also stopped with the acceptance of H_0 when the sample path has crossed the first parts of both u_{W2} and l_{W2}. To adjust for the discrete monitoring of a strictly continuous process, Whitehead[8] recommends replacing the intercept a by $a - 0.583\sqrt{(V_i - V_{i-1})}$ for $i = 1,2,...$ with $V_0 = 0$ (this adjustment is called the 'Christmas tree' correction).

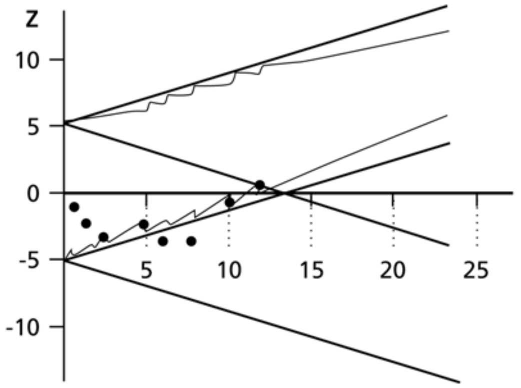


Figure 1 Sample path of successive ($Z;V$)-values for three controls matched per case in an open double sequential t-test with 2α =0.05, power $1-\beta$ = 0.80 and θ_R= 0.5. Values for the intercepts ±a are ±5.05 and values for the slopes ±b are ±0.3627. Data come from a cohort-nested case-control study relating selenium levels in toe-nails to the occurrence of colon cancer[4].

In Figure 1, an example of such presentation is given. Women from a cohort of participants in a population-based breast-cancer screening programme were also monitored for the occurrence of colon and rectal cancer. Three control women without colorectal tumours were matched by age to each case in the order in which cases were notified. The selenium content in toe-nails, stored previously in a biological bank, was used as a biomarker of selenium status. The purpose of the investigation was to detect a difference in selenium content between cases and controls and relating this difference to the occurrence of colorectal cancer. Selenium values can be considered to follow a normal distribution (details and other examples are given in references 4-6). Applying the sequential *t*-test, the null hypothesis could be accepted after the 8th case-control set was processed. For a fixed sample paired *t*-test at least 32 case-control sets would have been necessary to satisfy the same requirements, in terms of 2α, β and θ_R.

Although the program PEST does not contain standard features for the design and analysis of paired data, it can be used for a matched sequential *t*-test by choosing the 'DEFAULT-response' option, and using precalculated values of the *Z*- and *V*-statistics as data input.

3.5 Comparison of the two types of sequential *t*-test

We compared the test statistic l_n according to Rushton's approach to the test statistics *Z* and *V* following Whitehead's method. For a one-sided test a mathematical comparison shows that Rushton's approach (2) is identical to Whitehead's procedure (3), if the latter is conducted without continuity correction and if $\alpha = \beta$(see Appendix II). Likewise, for a two-sided test the rejection boundaries for the two methods can also be shown to be identical, irrespective of the value of *k* (the number of controls matched to each case), when Whitehead's procedure is conducted without continuity correction, $\alpha = \beta$ and in case of rejection of the null hypothesis. The two methods can be shown to differ with respect to their acceptance boundaries, both for $\alpha = \beta$ and for $a \neq \beta$ (see Appendix III).

We conducted computer simulations to evaluate the type I and type II error probabilities. Simulations were performed applying Rushton's approximation in a SPRT, and Whitehead's procedure for an open double sequential *t*-test either with or without continuity correction. Simulation programs were written in Turbo Pascal Version 5.0 (Borland). Uncorrelated, random 'case' and 'control' observations were generated. Both under H_0: $\theta_k = 0$ and under H_1 (with $\theta_R = 0.25$ and 0.5) and with 1, 2 and 3 controls per case, 2500 simulations were run with a type I error $2\alpha = 0.10$and a type II error $\beta = \alpha = 0.05$. In terms of the median or average sample size *n*, the results of these simulations can be summarized as follows, irrespective of *k*:

under H_0: $n_R < n_{W+} < n_{W-}$ under H_1: $n_{W+} < n_R \approx n_{W-}$

(where the subscripts R stand for Rushton, and W+ and W- for Whitehead's approach with and without continuity correction).

In Figures 2 and 3 the simulation results for $\alpha = \beta$ are depicted graphically in terms of the achieved type I and type II error probabilities.

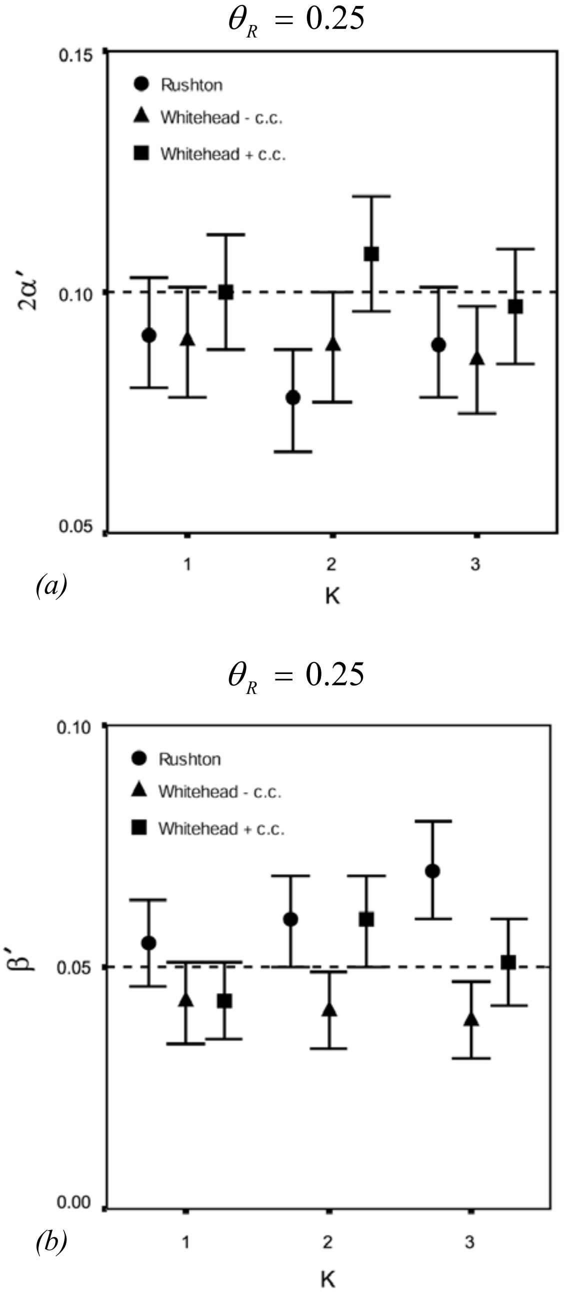
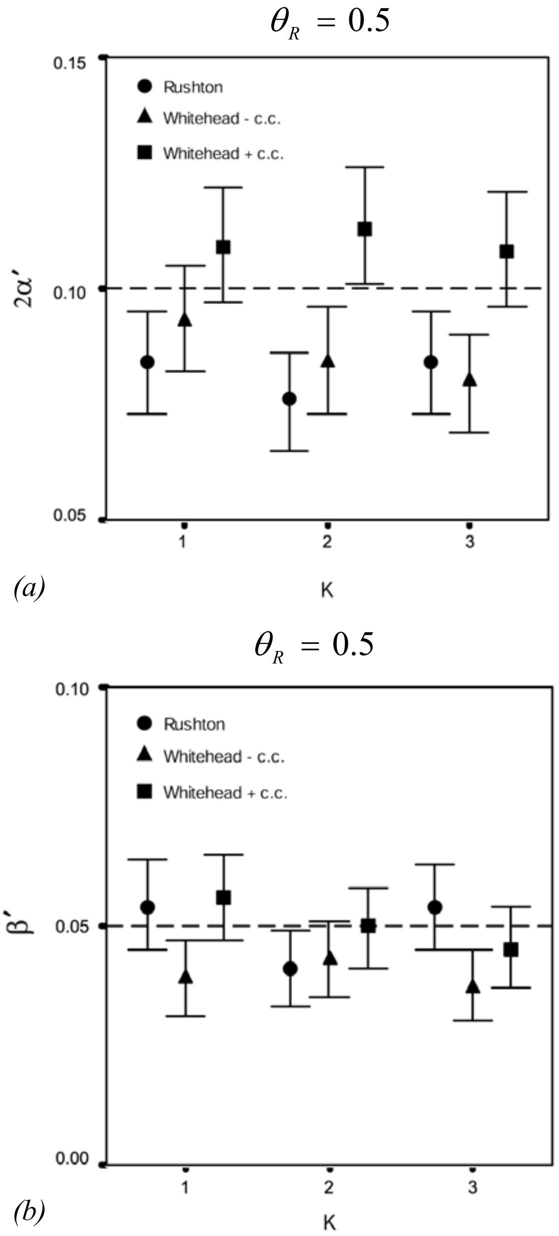


Figure 2 Achieved type I and type II error probabilities $2\alpha'$ (*a*) and β' (*b*) versus *k*, the number of controls matched per case for $2\alpha = 0.10$, $\beta = 0.05$, $\theta_R = 0.25$. The 0.95-confidence intervals are calculated using the normal approximation to the binomial distribution.

As shown by these figures, the achieved type I error probabilities for Rushton's approximation and for Whitehead's approach without continuity correction are smaller than the prespecified ones. Using Whitehead's approach with continuity correction, however, the achieved error probabilities are not significantly different from the prespecified, theoretical values, although the type I error is a little large for $\theta_R = 0.5$ (that is, for small *n*). For $\theta_R = 0.25$, Rushton's type II error is on the large side.


$\theta_R = 0.5$

(a)

$\theta_R = 0.5$

(b)

Figure 3 Achieved type I and type II error probabilities $2\alpha'$ (*a*) and β' (*b*) versus *k*, the number of controls matched per case for $2\alpha = 0.10$, $\beta = 0.05$, $\theta_R = 0.5$. The 0.95-confidence intervals are calculated using the normal approximation to the binomial distribution.

For $2\alpha = 0.05$, $\beta = 0.20$, $\theta_R = 0.25$ and 1, 2 or 3 controls per case, results of 2500 simulation runs for Rushton's procedure and Whitehead's approach with continuity correction are depicted in Figure 4.

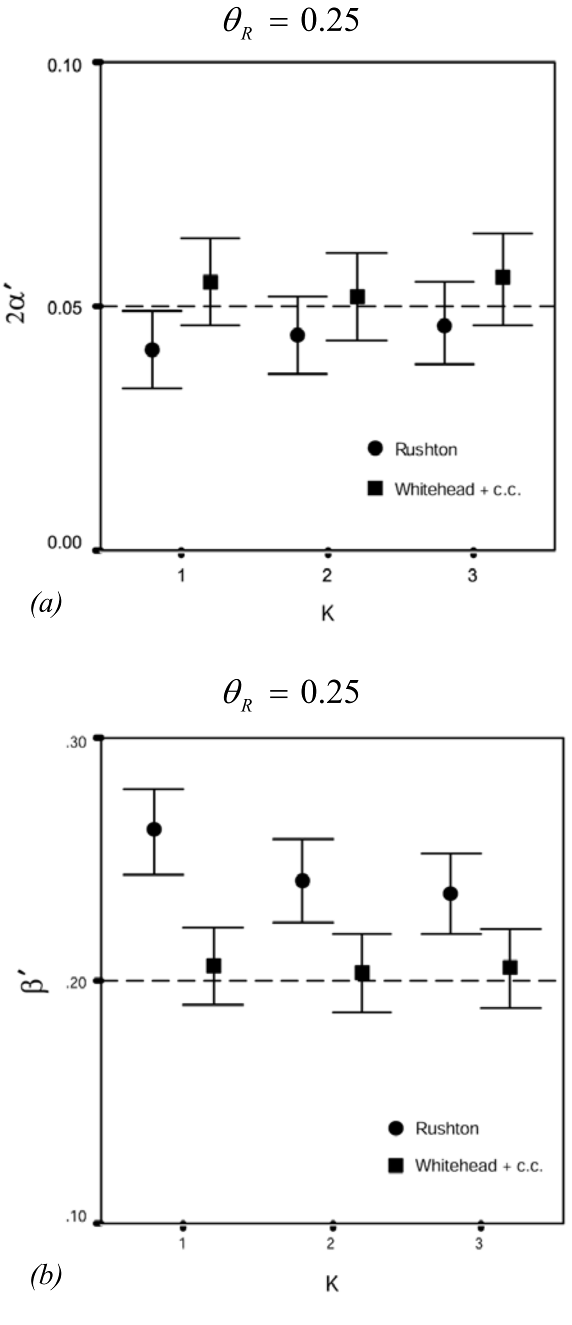


Figure 4 Achieved type I and type II error probabilities $2\alpha'$ (*a*) and β' (*b*) versus *k*, the number of controls matched per case for $2\alpha = 0.05$, $\beta = 0.20$, $\theta_R = 0.25$. The 0.95-confidence intervals are calculated using the normal approximation to the binomial distribution.

As for $\alpha = \beta$, Rushton is somewhat conservative in terms of the achieved type I error, while Whitehead's approach with continuity correction tends to result in somewhat higher

$2\alpha'$-values , especially for $\theta_R = 0.5$ (not shown). For $\theta_R = 0.25$ Rushton's type II error is now significantly too large. In terms of the number of case-control sets *n* processed, the results can be summarized as follows:

under H$_0$: $n_R < n_{W+}$ and under H$_1$: $n_{W+} \leq n_R$

3.6 Discussion

We have compared two approaches to perform a sequential *t*-test, with a view to applying such a test in prospective cohort studies for preliminary evaluations whether or not there is a relevant difference in mean exposure level between cases with disease and disease-free control subjects. The first approach is a classical sequential probability ratio test (SPRT) using an approximation to the log-likelihood ratio derived by Rushton. The other procedure, developed by Whitehead, is analogous to the SPRT but is based on a profile likelihood rather than a likelihood ratio. We extended both procedures for studies with more than one control matched per case.

We have compared Rushton's approach only to the symmetric, open version of the *t*-test given by Whitehead. Using Whitehead's method, a two-sided test with the boundaries placed symmetrically (at ±*a*) about the horizontal axis and with $\alpha = \beta$ has an expected maximum sample size when $\theta_k = \theta_R / 2$. The restriction $\alpha = \beta$ does not correspond to the usual choice of prespecified values of significance level and power in clinical or epidemiological studies: generally these values are chosen equal to $2\alpha = 0.05$ and $1-\beta = 0.90$ or $1-\beta = 0.80$, respectively. If $\alpha \neq \beta$, but the expected maximum sample size should occur at $\theta_k = \theta_R / 2$, test boundaries can be developed with slope *b* equal to $\theta_R / 2$, but with asymmetric intercepts.[12] Such an asymmetric test (without the 'Christmas tree' correction) equals Rushton's procedure for the one-sided version and for the two-sided version when H$_0$ is rejected. Because the program PEST provides only for symmetric versions of the SPRT-like procedure, we only evaluated and discussed these symmetric tests.

Whitehead[8] states that the open, one-sample sequential *t*-test using test statistics *Z* and *V* is equivalent to the test based on the approximation by Rushton.[13] Our mathematical comparison shows, however, that this is true only (for the symmetric versions of Whitehead's test) when:

1) α is equal to β, in case of a one-sided test; or, for a two-sided test, α is equal to β and H$_0$ is rejected;
2) no 'Christmas tree' corrections are made on the critical boundaries.

The difference between the two-sided 'open' versions of the sequential *t*-test following Rushton, or following Whitehead, can be explained partly by the fact that there are two possible approaches to design a two-sided test,[14-16] namely:

(i) to combine two one-sided tests, one to test H_0 versus H_+: $\theta_k \geq \theta_R$ and one to test H_0 versus H_-: $\theta_k \leq -\theta_R$, both with prespecified error probabilities α and β;

(ii) to compare H_0 with a composite alternative hypothesis, giving equal weights to H_+ and H_-, and using specified error probabilities 2α and β.

On the one hand, when H_0 is rejected in favour of, for example, H_+, then under (i) L_+/L_0 is greater than or equal to $(1-\beta)/\alpha$, and under (ii) $(L_++L_-)/2L_0$ is greater than or equal to $(1-\beta)/2\alpha$ (where L_0, L_+ and L_- denote the likelihood under H_0, H_+ and H_- respectively). Then the two approaches are practically equal because L_- is negligible in comparison to L_+. On the other hand, when H_0 is accepted, under (i) L_+/L_0 and L_-/L_0 are both smaller than $\beta/(1-\alpha)$ and under (ii) $(L_++L_-)/2L_0$ is smaller than $\beta/(1-2\alpha)$. In the latter case, these two approaches will differ for all values of α and β, the second approach (ii) being more conservative for $\alpha < 1/3$ (i.e. for all practical values of α). Rushton[11] mentions that his likelihood ratio (equation (1)) for a two-sided test is '... obtained as the simple average of two likelihood ratio's appropriate for one-sided *t*-tests ...'. Rushton's likelihood ratio is, however, expressed in terms of U^2 rather than of $\pm U$:

$$\frac{1}{2} e^{\frac{1}{4}\theta_R^2 U^2 - \frac{1}{2} n\theta_R^2} e^{\sqrt{(n\theta_R^2 U^2)}} \tag{6}$$

If the likelihood ratio were obtained as the average of two likelihood ratio's, it would have been

$$\frac{1}{2} e^{\frac{1}{4}\theta_R^2 U^2 - \frac{1}{2} n\theta_R^2} (e^{\theta_R U\sqrt{n}} + e^{-\theta_R U\sqrt{n}}) \tag{7}$$

and then would have been a test of type (ii). It can easily be seen that, for $n > 0$, (6) is always smaller than (7). Thus when H_0 is rejected, a test based on (6) is more conservative than a test based on (7), and when H_0 is accepted, a test based on (6) is less conservative than a test based on (7). The open double sequential tests as implemented by Whitehead in PEST are tests of type (i), which is generally considered the best approach to two-sided tests[14]. Our results for $\alpha = \beta$ are in agreement with these theoretical considerations, since we observed that: (1) $n_R \approx n_{W-}$, when H_0 is rejected; and (2) $n_R < n_{W-}$, when H_0 is accepted. Furthermore, from the mathematical comparison (see Appendix III, equations (20) and (21)) it follows that, when $\alpha = \beta$, Rushton's test will lead to the acceptance of H_0 more often than Whitehead's approach, that is for smaller n, and thus in general Rushton's type I error probabilities will be smaller and type II error probabilities will be larger than their theoretical values (see also Figures 2 and 3). For $\alpha \neq \beta$ (and especially $\beta > \alpha$) the difference between Rushton's and Whitehead's procedure increases, leading to more frequent acceptance of H_0 and thus to power values which are significantly too small, especially for larger sample sizes (see Figure 4).

Our simulations show that Whitehead's approach with continuity correction results in type I error probabilities which are on the large side for small sample sizes $(\theta_R = 0.5)$. This might be due to the fact that, because of these small sample sizes, the normality assumption for the test statistic Z was not met. This same phenomenon can be observed in the simulation results for a triangular test with larger θ_R-values[3]. For larger samples type I error probabilities are not significantly different from their prespecified, theoretical values. Irrespective of the choice of 2α, β and θ_R the achieved type II error probabilities resulting from Whitehead's procedure are close to the theoretical ones.

As expected, our simulations showed that the average sample size for a sequential test with pairwise matching (k = 1) is smaller than that for the equivalent fixed sample test procedure, while increasing the number of controls per case further reduced mean, median and 90th-percentile of the number of case-control sets necessary to reach a decision. In addition, using Whitehead's approach, the relative efficiency of the sequential t-test using multiple (k) controls per case instead of pairwise matching appears to be approximately equal to the theoretical value $2k/(k+1)$[17].

Theoretically it is possible for an open SPRT to process an infinite number of observations without ever coming to a decision whether to accept or reject the null hypothesis. This can be prevented by limiting the maximum allowed number of case-control sets sampled for analysis (that is usually referred to as ‘truncation’). The slopes and intercepts of the critical boundaries must then be adjusted to maintain the same significance level and power, which is accomplished automatically using the program PEST, following Whitehead’s approach. For instance, in the example shown in Figure 1 the coefficients of the critical boundaries a and b should be changed from 5.05 to 5.10 and from 0.3627 to 0.3601, respectively, if the test is truncated at twice the fixed sample size. The expected sample size under H_0 was 19.4 and then becomes 19.7; the expected sample size under H_1 was 21.6 and becomes 21.9. Using Rushton’s approach with truncation, similar adjustments of the critical boundaries have not been described so far, and the probabilities of type I and II errors will thus change. In previous simulations we have shown[5], however, that the error probabilities are hardly affected when the truncation point is greater than or equal to twice the required sample size for an equivalent fixed-sample test procedure, suggesting that at this level the adjustments of critical boundaries become as small as to be negligible in practice.

Besides the truncated SPRT-like test procedure, the PEST-program provides routines for other types of sequential tests originally designed as ‘closed’ procedures. These procedures are still based on the test statistics Z and V, but use different types of critical boundaries. An example is the so-called ‘triangular’ test. The triangular test is more efficient, in terms of the amount of information used, when the true value of θ_k is close to $\pm\theta_R/2$. On the other hand, not only open but also truncated, SPRT-like procedures tend to require less information than the equivalent triangular test when θ_k is anticipated to be equal to 0 (the null hypothesis) or when $|\theta_k| \geq \theta_R$ (the alternative hypothesis)[8]. In all instances, whether using a triangular test or a SPRT-like procedure, less information is needed than in an equivalent fixed sample size procedure.

For the analysis of specimen stored in a biobank it may be more practical to retrieve and analyze biological specimens of cases in batches (groups) of multiple cases (and their matched controls). Likewise, in large-scale, multi-centre clinical trials it may also be more feasible in practice to process and analyse data for groups of patients. In the past, various approaches for group sequential tests were proposed. In the seventies Pocock[18] developed tables for group sequential tests based on the prior specification of a small number of equally spaced inspections of the cumulating data, applying the same nominal significance value at each inspection. Pasternack and Shore[19] amongst others used Pocock's tables in the group sequential analysis of cohort data from a toxicological/epidemiological study. O'Brien and Fleming[20] proposed increasing nominal significance levels at equally spaced inspection intervals to reduce the chance of very early stopping. In practice, however, the approach of prespecified, equally spaced inspection intervals may be cumbersome. A more flexible approach is based on the so-called ‘alpha-spending function’ (see for example DeMets and Lan[21] and Whitehead[8]), that allows the cumulating data to be analysed at arbitrary inspection intervals. At each inspection of the data the nominal significance level corresponds to the amount of information ‘used’ (for example, the fraction of total trial duration expired). All group sequential tests offer the advantage of smaller expected sample size requirements compared with traditional, fixed-sample counterparts when the alternative hypothesis is true (although the efficiency gain is usually smaller than that for continuous sequential procedures as there may be some ‘overrunning’). The choice of appropriate nominal significance levels at intermediate inspections guarantees that the desired overall significance level is maintained.

A major disadvantage of the group sequential methods developed by Pocock,[18] O'Brien and Fleming[20] and DeMets and Lan[21] is that the null hypothesis H_0 cannot be accepted until the last planned inspection of the data. In Rushton's and in Whitehead's procedures the critical boundaries are chosen so that both the overall significance level 2α and the type II error β are guaranteed. Whitehead's sequential tests[8] can be easily adapted for groupwise inspection at fixed or arbitrary inspection intervals using the, already mentioned, ‘Christmas tree’ correction of the critical boundaries. Thus, Whitehead’s sequential procedure as discussed in this paper can be used for groupwise sequential testing. In addition, both the SPRT-like test and the triangular test following Whitehead's approach allow earlier stopping when enough information has been gathered to conclude that H_0 can be accepted. Although Rushton's procedure also allows early stopping with acceptance of H_0, an adjustment of its boundaries for groupwise inspection has not been described.

In conclusion, although PEST does not contain default features for the sequential analysis of matched or paired data, it can be used for such tests if precalculated *Z*- and *V*-values are used as data input. Our mathematical comparisons and simulations show that, when the continuity correction is applied (as is done in the PEST-program), type I and type II error probabilities are very close to the theoretical (prespecified) 2α- and β-values, in particular for small values of θ_R. On the contrary, Rushton's approximation can lead to a large loss in power. Storing blood or tissue specimens from members of a cohort in a biological bank enables researchers to test aetiological hypotheses about supposed associations between biological markers and diseases such as cancer. Using a sequential *t*-test to distinguish between promising and less promising hypotheses can save precious biological material. Matching more than one control to each case leads to an even more economical test. Practical considerations can necessitate grouped sampling and groupwise inspection of the specimens. To prevent the occurrence of a very large sample size in an open SPRT a truncated procedure can be considered. Contrary to Rushton's procedure, Whitehead's procedure can be easily adapted for truncation and/or groupwise inspection of the data without affecting the type I and the type II error of the sequential test. (More tables and figures summarizing the computer simulations are available from the first author on written request.)

3.7 Acknowledgements

The study was supported by the Praeventiefonds (the Netherlands), grant no. 28-1560. We thank Cas Kruitwagen for a critical evaluation of an earlier version of the manuscript and the two referees for their detailed and helpful comments.

3.8 References

1. Wald A. Sequential analysis, New York: John Wiley, 1947.
2. Kilpatrick GS, Oldham PD. Calcium chloride and adrenaline as bronchial dilators compared by sequential analysis. BMJ 1954; ii: 1388-91.
3. Facey KM. A sequential procedure for a phase II efficacy trial in hypercholesterolemia. Contr Clin Trials 1992; 13: 122-33.
4. Van Noord PAH. Selenium and human cancer risk (nail keratin as a tool in metabolic epidemiology). [thesis] Utrecht: University of Utrecht, the Netherlands, 1992.
5. Van der Tweel I, Noord PAH van, Kaaks R. Application of a sequential t-test in a cohort nested case-control study with multiple controls per case, J Clin Epidemiol 1993; 46: 253-9.
6. Kaaks R, Tweel I van der, Noord PAH van. Efficient use of biological banks for biochemical epidemiology: exploratory hypothesis testing by means of a sequential t-test. Epidemiology 1994; 5: 429-38.
7. Skovlund E, Walløe L. A simulation study of a sequential t-test developed by Armitage. Scand J Stat 1987; 14: 347-52.
8. Whitehead J. The design and analysis of sequential clinical trials, 2nd ed. Chicester: Ellis Horwood Limited, 1992.
9. Brunier H, Whitehead J. PEST3 Operating manual. Reading: University of Reading, U.K., 1993.
10. Abramowitz M, Stegun IA. Handbook of mathematical functions. New York: Dover Publications Inc., 1968.
11. Rushton S. On a two-sided sequential t-test. Biometrika 1952; 39: 302-8.
12. Whitehead J. The design and analysis of sequential clinical trials, 1st ed. Chicester: Ellis Horwood Limited, 1983.
13. Rushton S. On a sequential t-test. Biometrika 1950; 37: 326-3.
14. Armitage P. Sequential medical trials. 2nd ed. Oxford: Blackwell Scientific Publications, 1975.
15. Cox DR, Hinkley DV. Theoretical statistics. London: Chapman and Hall, 1974.
16. Wetherill GB, Glazebrook KD. Sequential methods in statistics, 3rd ed. London: Chapman and Hall, 1986.
17. Ury HK. Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. Biometrics 1975; 31: 643-9.
18. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977; 64:191-9.
19. Pasternack BS, Shore RE. Group sequential methods for cohort and case-control studies. J Chron Dis 1980; 33: 365-73.
20. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979; 35: 549-56.
21. DeMets DL, Lan KKG. Interim analysis: the alpha spending function approach. Stat Med 1994; 13: 1341-52.

APPENDIX I

For one control matched to each case (k = 1) the difference D_i between the observation for the ith case (X_i) and that for the ith control (Y_{i1}) is Normally distributed with variance $\sigma_1^2 = 2\sigma^2(1-\rho)$.

For k controls matched to each case, the variance of the difference between X_i and the mean over the k controls $Y_{i.}$ becomes

$$\sigma_k^2 = \frac{k+1}{k}\sigma^2(1-\rho) = \frac{k+1}{2k}\sigma_1^2 \tag{8}$$

After n case-control sets processed the likelihood function is

$$L_n = \frac{1}{(\sigma_k\sqrt{(2\pi)})^n}\, e^{-\frac{1}{2\sigma_k^2}\sum_{i=1}^{n}(D_i-\mu)^2}$$

Writing l_n for the logarithm of L_n and substituting (8) and $\mu = \theta_1\sigma_1$, the first-order derivatives of l_n with respect to θ_1 and σ_1 are

$$l_{\theta_1}(\theta_1,\sigma_1) = \left(\frac{2k}{k+1}\right)\frac{\sum D_i}{\sigma_1} - \left(\frac{2k}{k+1}\right)\theta_1 n \tag{9}$$

and

$$l_{\sigma_1}(\theta_1,\sigma_1) = -\frac{n}{\sigma_1} + \left(\frac{2k}{k+1}\right)\frac{\sum D_i^2}{\sigma_1^3} - \left(\frac{2k}{k+1}\right)\frac{\theta_1\sum D_i}{\sigma_1^2}$$

The ML-estimate of σ_1, σ_1^*, can be derived by solving $l_{\sigma_1}(\theta_1 = 0, \sigma_1) = 0$

$$\sigma_1^* = \sqrt{\left(\frac{\sum D_i^2}{n}\frac{2k}{k+1}\right)} \tag{10}$$

Substituting $\theta_1 = 0$ and (10) into (9) leads to the test statistic Z:

$$Z = \sum D_i \sqrt{\left(\frac{2k}{k+1}\frac{n}{\sum D_i^2}\right)}$$

The second-order derivatives of l_n with respect to θ_1 and σ_1 are

$$l_{\theta_1\theta_1}(\theta_1,\sigma_1) = -\left(\frac{2k}{k+1}\right)n$$

$$l_{\theta_1\sigma_1}(\theta_1,\sigma_1) = -\left(\frac{2k}{k+1}\right)\frac{\sum D_i}{\sigma_1^2}$$

and

$$l_{\sigma_1\sigma_1}(\theta_1,\sigma_1) = \frac{n}{\sigma_1^2} - \left(\frac{2k}{k+1}\right)\frac{3\sum D_i^2}{\sigma_1^4} + \left(\frac{2k}{k+1}\right)\frac{2\theta_1\sum D_i}{\sigma_1^3}$$

Now

$$\left\{l^{\theta_1\theta_1}(\theta_1,\sigma_1)\right\}^{-1} = l_{\theta_1\theta_1}(\theta_1,\sigma_1) - l_{\theta_1\sigma_1}(\theta_1,\sigma_1)\frac{1}{l_{\sigma_1\sigma_1}(\theta_1,\sigma_1)}l_{\theta_1\sigma_1}(\theta_1,\sigma_1) \tag{11}$$

Substituting $\theta_1 = 0$ and (10) into (11) leads to the test statistic *V*:

$$V = \left\{- l^{\theta_1\theta_1}\left(\theta_1 = 0, \sigma = \sigma_1^*\right)\right\}^{-1} \quad = n\left(\frac{2k}{k+1}\right) - \left(\frac{2k}{k+1}\right)\frac{\left(\sum D_i\right)^2}{2\sum D_i^2}$$

$$= n\left(\frac{2k}{k+1}\right) - \frac{Z^2}{2n}$$

APPENDIX II: The one-sample one-sided sequential *t*-test

For a one-sided test[13] Rushton's approximation to the log-likelihood ratio substituting $U = \sum D_i / \sqrt{\left(\sum D_i^2\right)}$ leads to

$$l_n = \theta_R \frac{\sum D_i}{\sqrt{\left(\sum D_i^2\right)}} \sqrt{n}\sqrt{f_k} - \frac{n}{2}\theta_R^2 f_k + \frac{\theta_R^2}{4}\frac{\left(\sum D_i\right)^2}{\sum D_i^2} f_k$$

with $f_k = 2k/(k+1)$.

The null hypothesis H$_0$: $\theta_k = 0$ is rejected in favour of the alternative hypothesis H$_1$: $\theta_k \geq \theta_R$ when

$$l_n \geq \ln((1-\beta)/\alpha) \qquad (12)$$

Using Z and V in a SPRT leads to rejection of H$_0$ when

$$Z \geq a + bV \qquad (13)$$

Substituting (3), (4) and (5) into (13) leads to

$$\theta_R \frac{\sum D_i}{\sqrt{\left(\sum D_i^2\right)}} \sqrt{n}\sqrt{f_k} - \frac{2\ln\left(\frac{1-\alpha}{\alpha}\right)}{\ln\left(\frac{1-\beta}{\beta}\right) + \ln\left(\frac{1-\alpha}{\alpha}\right)} \left[\frac{n}{2}\theta_R^2 f_k - \frac{\theta_R^2}{4}\frac{\left(\sum D_i\right)^2}{\sum D_i^2} f_k\right] \geq \frac{\ln\left(\frac{1-\beta}{\beta}\right) + \ln\left(\frac{1-\alpha}{\alpha}\right)}{2} \qquad (14)$$

Equations (12) and (14) are only equal when $\alpha = \beta$.

In Rushton's test H$_0$ is not rejected when

$$l_n \leq \ln(\beta/(1-\alpha)) \qquad (15)$$

Using Z and V in a SPRT leads to acceptance of H$_0$ when

$$Z \leq -a + bV \qquad (16)$$

Substituting (3), (4) and (5) into (16) leads to

$$\theta_R \frac{\sum D_i}{\sqrt{\left(\sum D_i^2\right)}} \sqrt{n}\sqrt{f_k} - \frac{2\ln\left(\frac{1-\alpha}{\alpha}\right)}{\ln\left(\frac{1-\beta}{\beta}\right) + \ln\left(\frac{1-\alpha}{\alpha}\right)} \left[\frac{n}{2}\theta_R^2 f_k - \frac{\theta_R^2}{4}\frac{\left(\sum D_i\right)^2}{\sum D_i^2} f_k\right] \leq \frac{\ln\left(\frac{\beta}{1-\beta}\right) + \ln\left(\frac{\alpha}{1-\alpha}\right)}{2} \qquad (17)$$

Equations (15) and (17) are only equal when $\alpha = \beta$. A similar derivation can be made for H$_0$ versus H$_1$: $\theta_k \leq -\theta_R$.

APPENDIX III: The one-sample two-sided sequential *t*-test

For a two-sided test Rushton's approximation is given by equation (2). Substituting U^2 from equation (1) into equation (2) leads to

$$l_n = \sqrt{\left\{ n\theta_R^2 \frac{\left(\sum D_i\right)^2}{\sum D_i^2} f_k \right\}} - \frac{n}{2}\theta_R^2 f_k + \frac{\theta_R^2}{4}\frac{\left(\sum D_i\right)^2}{\sum D_i^2} f_k - \ln(2)$$

with $f_k = 2k/(k{+}1)$.

The null hypothesis H_0: $\theta_k = 0$ is tested with significance level 2α and rejected in favour of H_1: $|\theta_k| \geq \theta_R$ when

$$l_n \geq \ln((1-\beta)/2\alpha) = \ln((1-\beta)/\alpha) - \ln(2)$$

Thus

$$\sqrt{\left\{ n\theta_R^2 \frac{\left(\sum D_i\right)^2}{\sum D_i^2} f_k \right\}} - \frac{n}{2}\theta_R^2 f_k + \frac{\theta_R^2}{4}\frac{\left(\sum D_i\right)^2}{\sum D_i^2} f_k \geq \ln\left(\frac{1-\beta}{\alpha}\right) \tag{18}$$

Using Z and V in a SPRT leads to rejection of H_0 when

$$|Z| \geq a + bV \tag{19}$$

Now (18) and (19) are only equal when $\alpha = \beta$.

In Rushton's test H_0 is accepted when $l_n \leq \ln(\beta/(1-2\alpha))$, that is

$$\sqrt{\left\{ n\theta_R^2 \frac{\left(\sum D_i\right)^2}{\sum D_i^2} f_k \right\}} - \frac{n}{2}\theta_R^2 f_k + \frac{\theta_R^2}{4}\frac{\left(\sum D_i\right)^2}{\sum D_i^2} f_k \leq \ln\left(\frac{2\beta}{1-2\alpha}\right) \tag{20}$$

Using Z and V in a SPRT leads to rejection of H_0 when

$$|Z| \leq -a + bV \tag{21}$$

Equation (20) can never be equal to equation (21).

CHAPTER 4

Sequential analysis of matched dichotomous data from prospective case-control studies

I. van der Tweel*, P.A.H. van Noord[#]

**Centre for Biostatistics, Utrecht University, Utrecht, the Netherlands*
[#]Julius Center for Patient-Oriented Research, Department of Epidemiology,
Utrecht University, Utrecht, the Netherlands

Statistics in Medicine 2000; 19: 3449-3464

Summary

Sequential analysis of randomized controlled clinical trials and epidemiological prospective (matched) case-control studies can have ethical or economical advantages above a fixed sample size approach. It offers the possibility to stop early when enough evidence for an apparent effect of the risk factor or lack of the expected effect is achieved. In clinical trials it is well accepted to stop the trial early in favour of the alternative hypothesis. In epidemiological studies, in general, the need is not felt to stop early in case of a clear exposure effect. Little attention has been paid, however, to early stopping and accepting the null hypothesis. In metabolic epidemiological studies where analysis destroys the biological material, the question of efficient use of samples, for example, those stored in a biobank, becomes crucial. Also a slow accrual of cases or the costs of follow-up of a cohort nested study can make it desirable to stop a study early when it becomes clear that no relevant exposure effect will be found. Matching can further reduce the amount of information necessary to reach a conclusion. We derived test statistics *Z* (efficient score) and *V* (Fisher's information) for the sequential analysis of studies with dichotomous data where each case can be matched to one or more controls. A variable matching ratio is allowed. These test statistics can be entered into the software PEST to monitor the course of the study. The double sequential probability ratio test and the double triangular test were evaluated with simulated data for odds ratios equal to 1.5, 2.0 and 2.5 and various type I and type II error probabilities both under H_0 and under H_1. Our simulations resulted in average and median values for the amount of information (*V*), that are far less than those for a fixed sample size study. Efficiency gain can range from 32 per cent to 60 per cent. The proposed sequential analysis was applied in an investigation on the possible relationship between the polymorfism of the MTHFR-gene and rectal cancer in a cohort of women with cases matched by age to one and to three controls. A sequential analysis of matched data can lead to early stopping in favour of H_0 or H_1, thus conserving valuable resources for future testing. A sequentially designed study can be more economical and less arbitrary than a study that makes use of conditional power or conditional coverage probability calculations to decide early stopping.

4.1 Introduction

In randomized controlled clinical trials and epidemiological (prospective) case-control studies it can be desirable to have at one’s disposal a statistical analysis that uses the least possible number of observations to come to a decision.

In a clinical trial it may be unethical to subject more patients than necessary to a treatment that turns out to be inferior. For example, Newman *et al*[1] describe how a sequential analysis showed a poorer survival under radiotherapy plus razoxane than under radiotherapy alone in patients with inoperable lung cancer. Thus the trial could be stopped early, saving patients from an inferior treatment. Montaner *et al*[2] showed that oral corticosteroids can prevent early deterioration in patients with moderately severe AIDS-related pneumonia. After 37 patients were analysed sequentially the null hypothesis could be rejected. This result meant that 47 per cent (that is 33 out of 70) of the foreseen number of patients did not have to be included in the trial. Moss *et al*[3] used a sequential design to demonstrate the superiority of an implanted defibrillator over conventional medical treatment of patients at high risk for ventricular arrhythmia. All clinical trials described[1-3] were analysed using a so-called triangular test[4]. In an epidemiological prospective (cohort nested) case-control study, cases are often scarce or accrue slowly and continuation of the follow-up of the cohort is frequently costly. Especially in the emerging field of metabolic and genetic epidemiology with biosamples from cohorts stored in a biobank, analysis of these biosamples is frequently destructive, unlike analysis of questionnaire data. This introduces the need to be selective with regard to the hypotheses that can be tested using these biosamples. Contrary to the samples of the cases, control material is mostly abundantly available in cohort nested studies. Thus a first step in the analysis of such a study is to increase the number of controls per case. A second option is to analyse the data sequentially[4-6].

On average, a sequential analysis requires fewer observations than a fixed sample analysis under the same design specifications[4,5]. When, in an epidemiological study, each case can be matched to a control, and in particular to multiple controls, a sequential analysis of the matched data requires fewer case-control sets than a sequential analysis of unmatched data with the same design specifications.

In an earlier paper we developed and compared one-sample two-sided sequential *t*-tests for epidemiological studies with more than one control matched per case and a normally distributed exposure variable[6]. In the present paper we derive the test statistics for a sequential test with matched dichotomous data. A fixed and a variable matching ratio are considered. Two sequential tests are compared: the sequential probability ratio test and the triangular test. Comparisons are made by simulations.

The proposed sequential test was applied in a genetic epidemiological study investigating the possible relationship between the polymorfism of the MTHFR-gene and rectal cancer in a cohort of women with cases matched by age to one and to three controls.

When a sequential test is concluded, fixed sample estimation procedures can not be applied, because the maximum likelihood estimate of the parameter is biased. Valid estimation procedures lead to a median unbiased parameter estimate and corresponding confidence interval[4].

4.2 Example

We were confronted with the need for efficient use of biosamples in investigating data from the DOM cohort[7]. Participants in this cohort were 50 to 69 years old. Between 1974 and 1984, women in this population-based breast cancer screening cohort volunteered to provide overnight urine samples. For each of more than 16,532 women 100 cm^3 aliquots are stored at a temperature of -20°C in a biobank. It turned out that in these stored urine samples enough cells were available in the sediments to allow for PCR DNA probe analysis. Since the DNA was fragmented, it will not be possible to amplify full DNA, which otherwise could have solved the problem of limited biological material available for analysis.

Several hypotheses with respect to the roles of genetic polymorfism have been put forward in the literature that, when tested, will compete for the limited material. Since in particular the material of cases is limited (the number of potential controls is less of a problem given the cohort size), the question of efficient management of the samples emerged. We had earlier developed some strategies for handling this problem for a normally distributed exposure variable[6,8]. For a dichotomous exposure new tests had to be developed.

Colorectal cancer was among the first tumours in this cohort where the problem of efficient management became pressing, given problems in the continuation of a complete follow-up of the entire cohort. A role for the *methylene-tetra-hydrofolate-reductase* (MTHFR) gene was claimed in the literature. Mutations in this gene occur in up to 13 per cent of the population. The wild polymorfism was contrasted to the homozygote and heterozygote polymorfisms to explore the hypothesis that being a homozygote or heterozygote carrier of the MTHFR-gene mutation (677 C > T mutation) increases the risk of developing rectal cancer. Through follow-up by the regional cancer registration (IKMN) and a mortality register established with the general practitioners in the Utrecht region, incidence of or mortality due to rectal cancer was traced prospectively. Based on the date of incidence or mortality (whatever became first known), urine samples were thawed and tested. A total number of 72 cases were reported. For three of these cases no genetic information was available. Control women were matched to each case by age. We matched one and three controls per case in different analyses to study the efficiency aspect of multiple controls per case. So for 69 cases and 207 controls urine samples were retrieved from the biobank. Urines were treated according to a fixed protocol to obtain urine sediments and PCR gene product. All data were analysed sequentially in the chronological order in which the cases were notified.

4.3 The sequential tests

4.3.1 The test statistics Z and V

Consider an epidemiological study where each case can be matched to one or more controls and the possible relation with an exposure variable is to be tested. Let π_0 be the probability of exposure for the controls and π_1 the probability of exposure for the cases. Only discordant matched sets provide information and thus are used in the analysis. Let π be the probability that the case is exposed in a discordant pair of uncorrelated observations with

$$\pi = \frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)+\pi_1(1-\pi_0)}.$$

Testing H$_0$: π = 0.5 against H$_1$: $\pi = \pi_R$ can be reparameterized as testing H$_0$: θ = 0 against H$_1$: $\theta = \theta_R$. Then $\theta = \ln\{\pi/(1\text{-}\pi)\} = \ln\{(\pi_1(1\text{-}\pi_0))/(\pi_0(1\text{-}\pi_1))\} = \ln(\psi)$, where ψ stands for the odds ratio (OR).

When in the ith matched set or stratum a case is matched to M_i controls the conditional likelihood can be used to derive the test statistics Z_i and V_i per matched set (see Appendix I). The odds ratio ψ is assumed to be the same in all matched sets or strata.
For the ith set

$$Z_i = X_{1i} - \frac{1}{M_i+1}\sum_{m=1}^{M_i+1} X_{mi} = \frac{M_i}{M_i+1}\left[X_{i,case} - \overline{X}_{i,controls}\right]$$

$$V_i = \frac{1}{M_i+1}\sum_{m=1}^{M_i+1} X_{mi}^2 - \left(\frac{1}{M_i+1}\sum_{m=1}^{M_i+1} X_{mi}\right)^2 = \frac{1}{M_i+1}\sum_{m=1}^{M_i+1}\left(X_{mi} - \overline{X}_i\right)^2 \qquad (1)$$

where $X_{1i} = X_{i,case}$ is the value for the case in the ith set and $X_{2i}\ldots X_{M_i+1,i}$ are the values for the matched controls in the ith set (1 when exposed, 0 when not exposed) with $\overline{X}_{i,controls}$ as the average value for the controls in the ith set and $\overline{X}_i$ as the overall average value of the ith set. The test statistics Z and V are equal to ΣZ_i and ΣV_i, respectively, over the successive sets i = 1,2,...,n, where n is the number of sets observed so far.
For the ith matched set with all M_i equal to 1

$$Z_i = \frac{1}{2}\left(X_{1i} - X_{2i}\right) \qquad \text{and} \qquad V_i = \frac{1}{4}\left(X_{1i} - X_{2i}\right)^2$$

(that is, for each i, in fact $Z_i = \pm 0.5$ and $V_i = 0.25$) and Z and V become

$$Z = \frac{1}{2}\sum_{i=1}^{n}\left(X_{1i} - X_{2i}\right) \qquad \text{and} \qquad V = \frac{1}{4}\sum_{i=1}^{n}\left(X_{1i} - X_{2i}\right)^2 \qquad (2)$$

The test statistics in (2) can also be written as $Z = S_n - n/2$ and $V = n/4$, with S_n as the number of exposed cases and n the number of discordant matched pairs so far observed (see also equations (3.47) and (3.48) in Whitehead[4] with p_0=0.5).

In general, for C_i cases matched to M_i controls the results in the *i*th matched set can also be tabulated as follows:

*i*th matched set	cases	controls	total
Number of persons	C_i	M_i	N_i
Number exposed	S_i	T_i	E_i
Number not exposed	F_i	G_i	$\overline{E_1}$

where S_i stands for the number of exposed cases, T_i for the number of exposed controls, E_i for the total number of exposed persons and $\overline{E_i}$ for the total number of unexposed persons. The test statistic Z_i can be expressed as the difference between the observed number of exposed cases and the expected number under the null hypothesis, and the statistic V_i as the variance under the null hypothesis. Using this tabular notation,

$$Z_i = S_i - \frac{E_i C_i}{N_i} = \frac{S_i M_i - T_i C_i}{M_i + C_i} \qquad \text{and} \qquad V_i = \frac{C_i M_i E_i \overline{E_i}}{N_i^2 (N_i - 1)}$$

For $C_i = 1$, these equations become

$$Z_i = \frac{M_i}{M_i + 1}\left[S_i - \frac{T_i}{M_i}\right] \qquad \text{and} \qquad V_i = \frac{E_i \overline{E_i}}{N_i^2} \tag{3}$$

The above equations (3) for Z_i and V_i are equal to equations (1). (The variance formula as given here can be compared to formula (3.5) in Whitehead[4]. Our denominator is equal to $N_i^2(N_i - 1)$ due to the use of the conditional likelihood in our derivation, while that in formula (3.5) is N^3.)

The test statistic Z is the efficient score for θ under the null hypothesis H_0: $\theta = 0$; V stands for Fisher's information about θ contained in Z. Z follows approximately a Normal distribution with expectation θV and variance V when samples are large and θ is small[4,9]. For 1 : 1 matching Z and V are equal to the numerator and the square root of the denominator, respectively, of McNemar's (fixed sample size) test for matched pairs without continuity correction. The test statistics Z and V for a sequential test with 1 : M (M_i equal to M for all i) matching are equal to the numerator and the square root of the denominator, respectively, of the test statistic proposed by Miettinen[10]. They are related in the same way to the Mantel-Haenszel test statistic for matched data[11] (without continuity correction) and to the logrank test statistic[12].

When M_i is equal to M for all i, the ratio of the expectation of V_i under the null hypothesis for M = 2 to the expectation of V_i under the null hypothesis for M = 1 is equal to 1.33. Likewise, the ratio of the expectation of V_i under the null hypothesis for M = 3 to the expectation of V_i under the null hypothesis for M = 1 is equal to 1.50 (see Appendix II). This means that the amount of information for two controls matched per case is 1.33 times the amount of information obtained with matched pairs. For three controls matched per case, 1.50 times the amount of information contained in matched pairs is obtained.

In epidemiological studies a variable matching ratio can occur when not enough controls can be found to match to the same case, or when biological or genetic material stored in a biobank is either limited or not available (any more). Suppose a control is missing with probability π_m (i.e. the number of controls matched to a case is now variable for the ith set). Under the null hypothesis and, for example, $\pi_m = 0.25$ the ratio of $E(V_i \mid M_i = 2; \pi_m)$ to $E(V_i \mid M = 1)$ is equal to 1.125, and the ratio of $E(V_i \mid M_i = 3; \pi_m)$ to $E(V_i \mid M = 1)$ is equal to 1.336 (see Appendix III). (When M = 1 and the control is missing, the case-control set is uninformative; likewise when the case is missing).

4.3.2 Two sequential tests

We compare the behaviour of the test statistics Z and V in two types of sequential tests: the sequential probability ratio test (SPRT) and the triangular test (TT). Both tests require critical boundaries to be specified in advance. Characteristics of both tests are described at length by Whitehead[4]. For an open double SPRT the critical boundaries are as follows

$$Z = a + bV \qquad (u_1)$$

$$Z = \text{-a} + bV \qquad (u_2)$$

$$Z = a - bV \qquad (l_2)$$

$$Z = -a - bV \qquad (l_1)$$

The slopes ±b and intercepts ±a of these boundaries are functions of the error probabilities 2α (the two-sided type I error) and β (the type II error), as well as of the choice of a minimal relevant standardized exposure difference, θ_R, that defines the (two-sided) alternative hypothesis $\mathrm{H}_1\colon |\theta| = \theta_R$:

$$b = \frac{\theta_R}{\ln\left(\frac{1-\beta}{\beta}\right) + \ln\left(\frac{1-\alpha}{\alpha}\right)} \ln\left(\frac{1-\alpha}{\alpha}\right) \qquad \text{and} \qquad a = \frac{\ln\left(\frac{1-\beta}{\beta}\right) + \ln\left(\frac{1-\alpha}{\alpha}\right)}{2\theta_R}$$

For a double TT the critical boundaries are as follows

$$Z = a + cV \qquad (u_1)$$

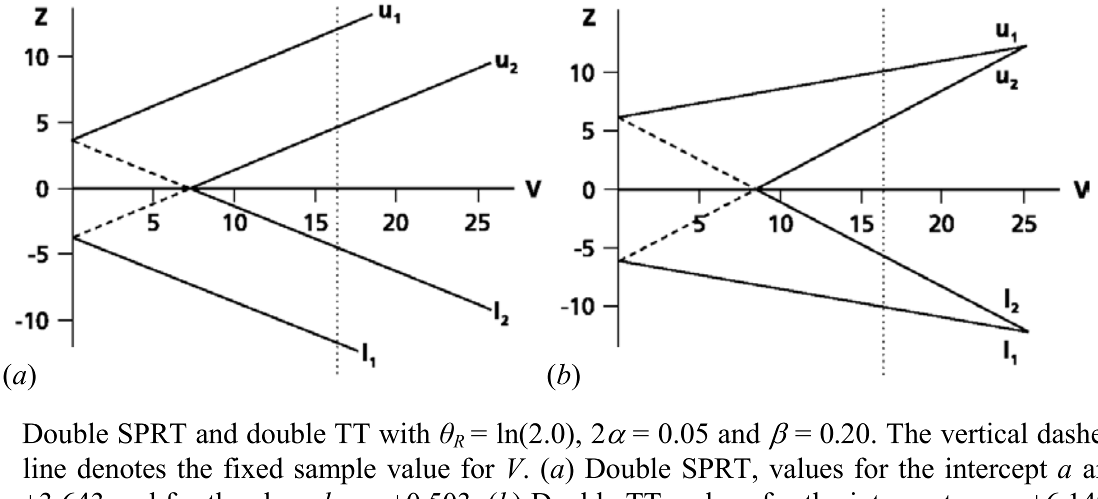
$$Z = -a + 3cV \qquad (u_2)$$

$$Z = a - 3cV \qquad (l_2)$$

$$Z = -a - cV \qquad (l_1).$$

The boundaries u_1 and u_2, and l_1 and l_2, cross at $V_{\max} = a/c$ and $Z = \pm 2a$.
For $\alpha \neq \beta$ no simple expressions can be given for a and c.

Throughout this paper, it is assumed that the boundaries of the tests are placed symmetrically with respect to the horizontal axis. Both sequential tests are conveniently presented in the form of a graph, plotting Z against V (see Figures 1(a) and (b) for illustration). The testing process continues as long as the sample path formed by successive Z-values plotted versus corresponding V-values remains between the boundaries u_1 and u_2 or between l_1 and l_2. The sampling is stopped and H_0 rejected when the sample path crosses u_1 or l_1. The test is stopped with acceptance of H_0 when the sample path crosses u_2 or l_2. The test is also stopped with the acceptance of H_0 when the sample path has crossed the first parts of both u_2 and l_2. To adjust for the discrete monitoring of a strictly spoken continuous process, Whitehead recommends to replace the intercept a by $a - 0.583\sqrt{(V_i - V_{i-1})}$ for $i = 1,2,...$ with $V_0 = 0$ (this adjustment is called the 'Christmas tree' correction).
Both tests are implemented in the computer program PEST[13]. Although PEST does not contain features for the design and analysis of matched data, it can be used by choosing the 'DEFAULT-response' option and entering precalculated values for Z and V.



(*a*) (*b*)

Figure 1 Double SPRT and double TT with $\theta_R = \ln(2.0)$, $2\alpha = 0.05$ and $\beta = 0.20$. The vertical dashed line denotes the fixed sample value for V. (*a*) Double SPRT, values for the intercept a are ±3.643 and for the slope b are ±0.503. (*b*) Double TT, values for the intercept a are ±6.149, the slope of the boundaries u_1 and l_1 is ± 0.244 and the slope of the boundaries u_2 and l_2 is ±0.731.

4.3.3 Simulations

The double SPRT and the double TT were evaluated with simulated data both under H_0: $\theta = 0$ and under H_1: $|\theta| = \theta_R$ with $\theta_R = \ln(\psi)$ for $\psi = 1.5$, 2.0 and 2.5 with $2\alpha = 0.10$, $\beta = 0.05$ and with $2\alpha = 0.05$, $\beta = 0.20$. Each evaluation consisted of 2500 simulation runs with $M = 1$, 2 and 3 controls matched per case. To evaluate the effect of a variable number M_i of controls matched per case, data were simulated and analyzed by the double SPRT for $\psi = 1.5$ with $2\alpha = 0.10$, $\beta = 0.05$ and with $2\alpha = 0.05$, $\beta = 0.20$. For $M_i = 2$ and $M_i = 3$ the probability of a missing control (π_m) was chosen equal to 0.25. All evaluations of both the double SPRT and the double TT were performed with the so-called 'Christmas tree' correction of the boundaries.

4.3.4 Estimation of sample size

For a fixed sample design with matched case-control pairs the sample size can be estimated using the relation $V_{fixed} = (u_\alpha + u_\beta)^2 / \theta_R^2$ (Whitehead[4]), where u_x denotes the standardized normal deviate exceeded with probability x. The total sample size can be estimated by dividing the number of discordant pairs $n = V_{fixed} \times 4$ (see equation (2)) by the probability of a discordant pair of uncorrelated observations, $\pi_{\text{disc}} = \pi_0(1 - \pi_1) + \pi_1(1 - \pi_0)$. For our simulated data evaluated sequentially π_{disc} can be approximated by

$$p_{disc} = \frac{4}{9} \times \frac{\psi^2 + \psi + 1}{(\psi + 1)^2}$$

(see Appendix IV). This probability is about 1/3 for an odds ratio ψ in the range 1.0 to 2.5.

In practice, for a sequential study the expected average or median value for V (as reported by the program PEST when designing the study) multiplied by 4 and divided by the probability of a discordant pair of observations estimates the average or median total number of case-control sets necessary for a study with 1 : 1 matching. For 1 : M matching this number can be reduced by a factor $(M + 1)/2M$ (Ury[14]). (Note the reciprocal relation with the expectation of V_i for M = 2 and M = 3.)

4.4 Results of the example

Homozygote or heterozygote carriers of the MTHFR-gene mutation were expected to have at least a twofold risk of developing rectal cancer compared to the carriers of the wild polymorphism. A double SPRT was designed with an OR equal to 2 as the alternative hypothesis ($\theta_R = \ln(2) = 0.69315$), a two-sided type I error $2\alpha = 0.05$ and a power $1\text{-}\beta = 0.80$. The expected average value for V under H_0 is equal to 10.1; the expected median value for V is equal to 8.8. The matched-pairs analysis ended without a decision after 69 matched pairs and 32 discordant pairs ($Z = -2.0$ and $V = 8.0$) (see Figure

2(*a*)). For three controls per case the sequential analysis could be terminated with accepting the null hypothesis after 41 matched sets and 35 discordant sets (Z = -0.25 and V = 7.3) (see Figure 2(*b*)). After stopping the sequential test, a median unbiased estimate[4] can be given for the odds ratio: OR = 0.87 with its 95 per cent confidence interval (0.39 ; 1.85).

The number of discordant pairs necessary for a fixed sample design with the design specifications as above is at least 66 ($V_{\text{fixed}} = 16.34$; $n \geq 16.34 \times 4$). Thus a saving of 47 per cent was reached in this study by a sequential analysis with three controls matched per case compared to a fixed sample design.

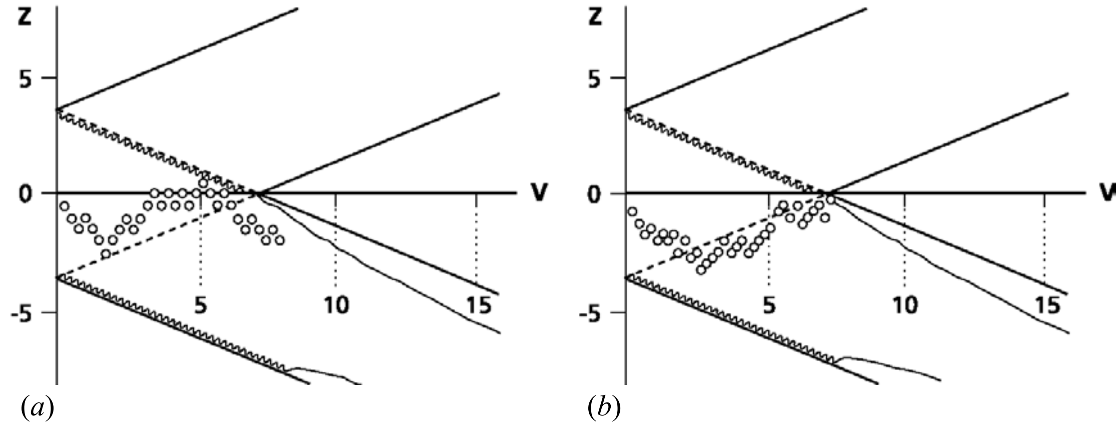


Figure 2 Sample path of successive ($Z;V$)-values for (*a*) 1:1 matching, and for (*b*) 1:3 matching in a double SPRT with θ_R = ln(2.0), 2α = 0.05 and β = 0.20. Data come from a cohort-nested case-control study relating exposure to the MTHFR-gene to the occurrence of rectal cancer in women (see text). The 'curved' lines indicate the 'Christmas tree' correction (see text and ref. 4).

4.5 Results of simulations

4.5.1 Type I and type II errors

For all evaluated OR's the resulting type I errors were not significantly different from their theoretical values. Simulations with a theoretical β equal to 0.05 resulted in acceptable type II errors. Only for $\psi \geq 2.0$ both the SPRT and the TT resulted in somewhat larger type II errors than the theoretical $\beta = 0.20$ for $M = 1$ or $M = 2$. (See Figures 3 and 4)

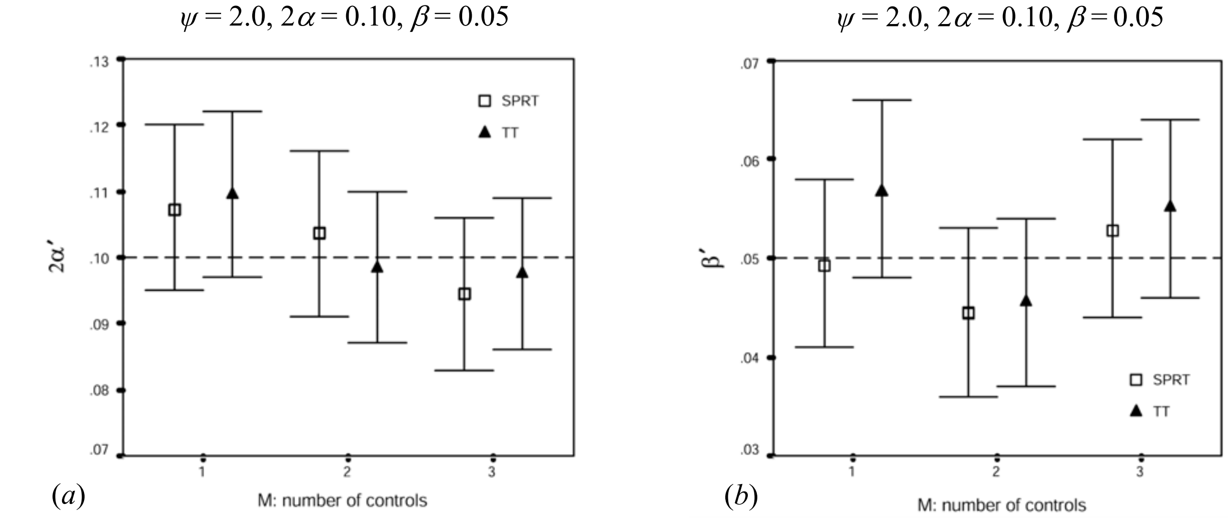


Figure 3 Achieved type I and type II error probabilities (*a*) $2\alpha'$ and (*b*) β' versus the number of controls matched per case (*M*) for $\theta_R = \ln(2.0)$, $2\alpha = 0.10$ and $\beta = 0.05$. The 0.95-confidence intervals are calculated using the normal approximation to the binomial distribution.

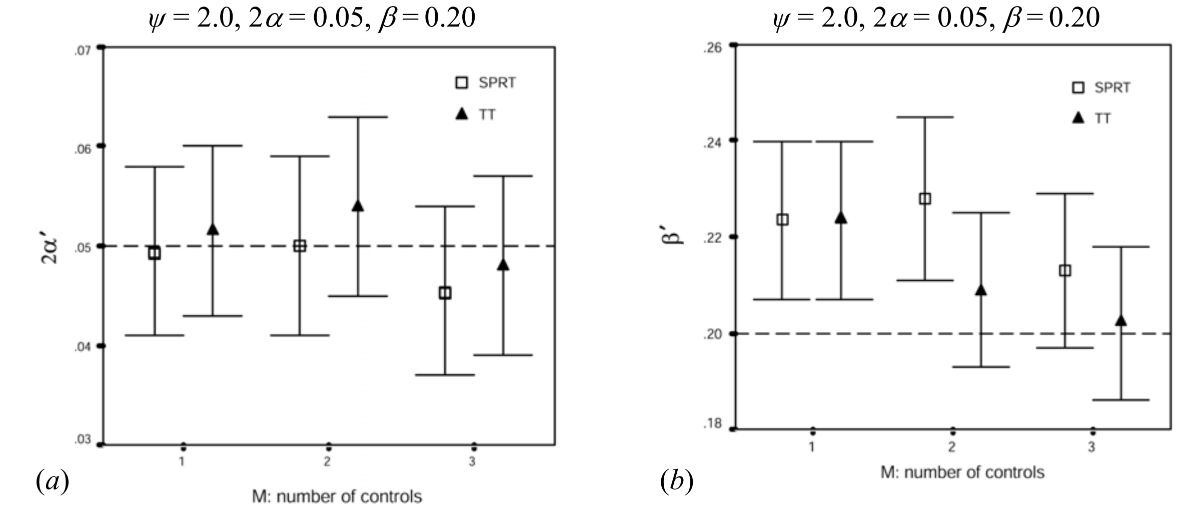


Figure 4 Achieved type I and type II error probabilities (*a*) $2\alpha'$ and (*b*) β' versus the number of controls matched per case (*M*) for $\theta_R = \ln(2.0)$, $2\alpha = 0.05$ and $\beta = 0.20$. The 0.95-confidence intervals are calculated using the normal approximation to the binomial distribution.

4.5.2 Sample size with a fixed number of controls per set

In Tables 1 and 2 observed and expected average and median values for *V* and the observed total number of (concordant and discordant) case-control sets (*N*) are given. Values tabulated are for 1 : 1 matching. Comparing the observed median values for *V* to the expected fixed sample size value (V_{fixed}) shows that savings in the amount of information used by a sequential study can range from 32 per cent to 46 per cent when the null hypothesis is true and from 36 per cent to 60 per cent when the alternative hypothesis is true, irrespective of the value for the odds ratio.

For 1 : M matching the expected reduction in total sample size N by a factor $(M+1)/2M$ was found for most simulations. Small deviations were seen when the type II error was larger than $\beta = 0.20$.

4.5.3 SPRT with a variable number of controls per set

The effect of a variable number of controls per case on the results of the double SPRT is shown in Table 3. To achieve the same amount of 'information' with a variable number of controls per case as with a fixed number more case-control sets are needed. The type I and type II errors were not significantly different from their theoretical values.

Table 1 Observed and expected (italic*)* values for the amount of information (V) and the observed total number of case-control sets (N) for 1 : 1 matching, ψ is the odds ratio, $2\alpha = 0.10$, $\beta = 0.05$ (V_{fixed} is the expected fixed sample size value for V)

ψ	H_0				H_1				V_{fixed}
	V		N		V		N		
	average	median	average	median	average	median	average	median	
(a) SPRT									
1.5	49.55 *47.82*	42.75 *41.59*	594.4	512.5	32.33 *32.57*	25.25 *26.47*	384.9	299.0	*65.84*
2.0	16.71 *16.36*	14.50 *14.23*	200.6	173.0	11.40 *11.14*	9.25 *9.06*	134.4	111.0	*22.53*
2.5	9.57 *9.36*	8.25 *8.15*	115.0	100.0	6.92 *6.38*	5.25 *5.18*	80.0	63.0	*12.89*
(b) TT									
1.5	47.29 *47.50*	44.50 *44.69*	568.3	535.0	34.56 *35.33*	31.75 *32.53*	412.0	375.5	*65.84*
2.0	16.30 *16.25*	15.25 *15.29*	195.7	185.0	12.30 *12.09*	11.25 *11.13*	144.7	134.0	*22.53*
2.5	9.11 *9.30*	8.50 *8.75*	109.6	103.0	7.12 *6.92*	6.75 *6.37*	81.8	74.0	*12.89*

Table 2 Observed and expected (italic) values for the amount of information (V) and the observed total number of case-control sets (N) for 1 : 1 matching, ψ is the odds ratio, $2\alpha = 0.05$, $\beta = 0.20$ (V_{fixed} is the expected fixed sample size value for V)

ψ	H_0				H_1				V_{fixed}
	V		N		V		N		
	average	median	average	median	average	median	average	median	
(a) SPRT									
1.5	30.11 *29.53*	26.25 *25.73*	361.3	312.5	33.99 *32.83*	26.00 *25.99*	405.2	309.0	*47.76*
2.0	10.16 *10.10*	9.00 *8.80*	122.5	107.0	12.18 *11.23*	9.75 *8.89*	142.5	110.0	*16.34*
2.5	5.90 *5.78*	5.25 *5.04*	70.7	63.0	7.79 *6.43*	5.75 *5.09*	90.7	72.0	*9.35*
(b) TT									
1.5	31.76 *31.81*	30.00 *29.79*	381.5	357.0	31.51 *32.05*	29.75 *30.67*	373.9	358.0	*47.76*
2.0	10.95 *10.89*	10.25 *10.19*	132.0	124.0	10.98 *10.97*	10.25 *10.50*	128.5	121.0	*16.34*
2.5	6.27 *6.23*	6.00 *5.83*	74.8	71.0	6.45 *6.28*	6.25 *6.01*	74.5	71.0	*9.35*

Table 3 The ratio of the average and the median number of case-control sets for a variable number of controls (M_i) (with probability of a missing control $\pi_m = 0.25$) to those for a fixed number of controls for the double SPRT

	H_0		H_1	
	average	median	average	median
(a) $\psi = 1.5$, $2\alpha = 0.10$, $\beta = 0.05$				
$M_i = 2$	1.168	1.171	1.190	1.196
$M_i = 3$	1.113	1.125	1.125	1.169
(b) $\psi = 1.5$, $2\alpha = 0.05$, $\beta = 0.20$				
$M_i = 2$	1.173	1.174	1.188	1.202
$M_i = 3$	1.126	1.128	1.067	1.037

4.6 Discussion

Sequential analysis of epidemiological studies can have advantages above a fixed sample size approach. It offers the possibility to stop early when enough evidence for an apparent effect of the risk factor or lack of the expected effect is achieved. Control for possible confounders or a more precise effect estimate could be an argument for not stopping early to reject the null hypothesis. On the other hand, studies on the relevance of certain gene mutations for the development of tumours focus primarily on the exploration of hypotheses, using a limited number of biological samples, as proposed in this paper. Such studies are meant to build evidence and create prior information for further affirmative studies. Once a role for a gene mutation is detected in a sequential analysis a more elaborate (sequential) design including possible effect modifiers and/or confounders can be warranted. However, a possible lack of an exposure effect certainly is an argument in favour of early stopping and accepting the null hypothesis. Destructive or expensive laboratory tests can demand a minimization of the number of those tests performed. Strömberg[15], for example, calls for development of methods for early stopping of inconclusive epidemiological studies and further discussion.

In epidemiological studies, and especially in cohort nested studies, cases are often scarce while controls are abundant. Then, matching of controls and cases can increase the efficiency of an epidemiological study and multiple controls per case can improve the power of the study (Ury[14]).

In clinical trials a matched data structure can arise when patients are subjected to two successive treatments in a random sequence (the so-called cross-over trials). A similar situation arises when two drugs or treatments are applied to different sides of the mouth (Fertig *et al*[16]) or two equivalent parts of the body. Most clinical cross-over trials have to do with the within-subject comparison of two treatments and thus matched pairs of observations.

In this paper, the test statistics Z and V were derived for the sequential analysis of studies with a dichotomous outcome where each case can be matched to one or more than one control. Using the computer program PEST, designed to perform sequential analysis, the test statistics can be entered into this computer program to monitor the course of the study. (It is thus possible to analyse stratified data, like the case-control study described by Strömberg[15] (four strata) and matched case-control studies (many strata), sequentially using PEST[4,13] by calculating the test statistics for each stratum and combining them.)

Our simulations resulted in acceptable type I errors for the evaluated OR's. The resulting type II errors were slightly larger than their theoretical values, especially for an OR = 2.5. Bellisant *et al*[17] performed some simulations to evaluate the small sample properties of an open SPRT and a TT applied to non-comparative phase II clinical trials. These non-comparative studies can be compared statistically to the results of our matched pair simulations. They performed simulations only with $\beta = 0.05$. Although they do not mention this explicitly, their simulations show higher type II errors for large θ_R ($= \ln(3.86) = 1.35$) especially with the TT.

The test statistics Z and V are derived assuming that the likelihood for the parameter θ resembles the normal likelihood. Sprott[18] showed that among various reparametrizations of the binomial distribution the transformation

$$\eta(\pi) = \int_0^\pi \frac{dt}{[t(1-t)]^{2/3}}$$

completely removes the component of asymmetry. Therefore Whitehead[4] suggests the use of $\theta' = 4^{1/3}[\eta(\pi) - \eta(0.5)]$ instead of θ for a sequential test with dichotomous data. The use of θ' leads to the same test statistics under H_0 as the use of θ (equations (1), (2) and (3)). Only the critical boundaries of the sequential tests are adjusted. In theory, the use of θ' instead of θ should lead to type I and type II errors that are closer to their theoretical values 2α and β. In fact, our simulations showed only negligible differences.

The statistic V is based on the 'observed' information about θ. For a dichotomous response the 'observed' and 'expected' information are equal. We considered the use of Cox's test as a sequential analogue of Wald's W_e test[19]. This test has the disadvantage that for every new case-control set not only Z and V but also the ML-estimate for θ has to be calculated. For a sequential test, the ML-estimate is biased[4]. Although Cox's test is asymptotically equivalent to the sequential tests described in this paper, some simulations for small samples (i.e. large ψ, ψ=2.5) show type II errors larger than their theoretical values, but in addition far too small type I errors.

Two types of sequential tests were compared: an open sequential probability ratio test (SPRT) and a triangular test (TT). The TT is, by the shape of its boundaries, a 'closed' test. Theoretically, the SPRT can carry on infinitely without reaching a decision. The PEST-program also provides a possibility for a closed SPRT, the so-called 'truncated SPRT'. The number of observations is then set to a limit to prevent this carrying on infinitely. Slopes and intercepts of the critical boundaries are adjusted to maintain the same significance level and power. In general, an SPRT is more efficient, in terms of the amount of information used, than a TT when the true parameter value θ is equal to 0 (the null hypothesis) or when $|\theta| \geq \theta_R$ (the alternative hypothesis). This is confirmed by our simulations. Median numbers of case-control sets were smaller for the double SPRT than for the double TT both under the null hypothesis and under the alternative hypothesis, when the same parameterization and type I and type II errors were applied.

Pasternack and Shore[20] mentioned the possibility of applying a group sequential test to matched data from case-control studies. Laboratory or logistic conditions can require analysis of the data in groups or batches. Group sequential designs as well as designs that use α-spending functions for repeated testing can be easily implemented in PEST[4,13].

A number of (recent) publications have paid attention to the early stopping of a trial or study due to lack of treatment difference or exposure effect (Lan and Wittes[21]; Hunsberger *et al*[22]; Betensky[23]; Strömberg[15]; Ware *et al*[24]). Lan and Wittes[21] and Hunsberger *et al*[22] provide conditional power (CP) calculations for studies that were not sequential by design. Betensky[23] developed lower boundaries for CP calculations in repeated significance and O'Brien-Fleming (group-sequential) designs. Strömberg[15] considers the conditional coverage probability (CCP) instead of CP. (CP is the probability of rejecting the null hypothesis at the end of the study (when only part of the total amount of planned information is observed), conditional on the observed data and assuming an alternative hypothesis for the remainder of the study. CCP is the probability that a two-sided confidence interval around the final estimate given the observed data and an assumed alternative hypothesis for the remainder of the study, includes the parameter value under the null hypothesis). Betensky[23] and Strömberg[15] both observe that less attention is paid to designs for early acceptance of H_0 than to designs that allow early termination due to a clearly apparent result. Strömberg even calls for 'further discussion concerning early stopping of epidemiologic studies'. We agree with both authors that early stopping in favour of H_0 can conserve valuable (time, financial, genetic, biological) resources for future testing (see for example[6,8,25,26]). We emphasize, however, that CP and CCP calculations require extrapolation of the observed data and are based on rather arbitrary assumptions[27]. We take the view that a (group-)sequential analysis can be more economical and more objective than an analysis based on CP or CCP calculations.

An alternative sequential approach that focuses on effect estimation more than on hypothesis testing is the repeated confidence interval (RCI) approach, as described by Jennison and Turnbull[28]. RCIs can be constructed by inverting a group sequential test. Group sequential tests perform a number of interim analyses on accumulating data. Each interim analysis is performed using a nominal significance level that is smaller than the desired overall type I error 2α to guarantee that the overall significance level is not inflated by the multiple testing procedure. The critical values for such an interim analysis are used to construct the corresponding RCI for that interim inspection.

In general, we think that a clear distinction must be made between hypothesis testing and effect estimation. Repeated testing of the null hypothesis can lead to early stopping of a study either because there is evidence for a significant effect and the null hypothesis is rejected or because there is no such evidence given the data thus far and the null hypothesis can be accepted. Advantages of early stopping with acceptance of the null hypothesis lie in making the most efficient use of finite resources. Effect estimation can be carried out repeatedly during a sequential study or once at termination of the study when a final estimate with its confidence interval is desired. For example, in an epidemiological study monitoring the effects of environmental hazards, early stopping is mostly not relevant, but RCIs can be useful to contemplate the effect size in accumulated data. At termination of the study point and interval estimates should be adjusted according to the sequential stopping rule. Jennison and Turnbull do not indicate any adjustments, however. A limitation of the RCI approach is the lack of a power aspect when RCIs are used to stop a study early. The power guarantee is an essential part of the sequential design as suggested by Whitehead, which makes this design more appropriate for hypothesis testing.

In case of early stopping and accepting the null hypothesis emphasis lies more on hypothesis testing than on effect estimation. Nevertheless, a median unbiased estimate for the parameter θ, and thus for the odds ratio, and a confidence interval can be obtained[4,13].

Average or median sample size for a study with 1 : 1 matching can be estimated from the average or median value for V multiplied by 4 and divided by the probability of a discordant pair or set (p_{disc}). As p_{disc} is often unknown, its expected value (see Appendix IV) can be used to get an approximate idea of the total number of matched sets necessary.

Our simulations confirm that a sequential analysis requires on average less case-control sets than a fixed sample analysis. For 1 : 1 matching savings in the amount of information used can range from 32 per cent to 60 per cent. Where efficiency is the aim, optimizing the case-control ratio in epidemiological studies is, in our opinion, an additional essential strategy whether a classical case-control or a cohort-nested study is concerned, since in epidemiological studies the collection or accrual of cases in a cohort or population may not be as simple or inexpensive as finding controls. When more than one control can be matched to each case, the amount of information per case-control set becomes larger. Also the probability of a discordant, informative set is larger. A variable matching ratio due to missing control information can easily be dealt with.

We further suggest that authors claiming efficiency gain give additional estimations of such gain (for example in per cent) to allow comparisons of different strategies for early stopping. We hope that this paper and further discussion may stimulate other investigators to consider these sequential designs as strategies for ethical or efficiency aspects in epidemiological studies, in order to pursue only the most promising hypotheses.

4.7 Acknowledgements

The study was supported by the Praeventiefonds (the Netherlands), grant nr 28-1560. We thank Charles Gimbrère, head of the Comprehensive Cancer Registration (IKMN), for providing the information about cases and controls, Olga van der Hel for laboratory examinations on the MTHFR-gene, Jan van den Broek for advice and stimulating discussions and Maria Schipper for carefully reading and commenting on the manuscript. Last, but certainly not least, we thank both referees for their valuable and helpful comments.

4.8 References

1. Newman CE, Cox R, et al. Reduced survival with radiotherapy and razoxane compared with radiotherapy alone for inoperable lung cancer in a randomised double-blind trial. Br J Cancer 1985; 51: 731-2.
2. Montaner JSG Lawson LM, et al. Corticosteroids prevent early deterioration in patients with moderately severe Pneumocystis carinii pneumonia and the acquired immunodeficiency syndrome (AIDS). Ann Int Med 1990; 113: 14-20.
3. Moss AJ, Hall WJ, et al for the multicenter automatic defibrillator implantation trial investigators. Improved survival with an implanted defibrillator in patients with coronary disease at high risk for ventricular arrhythmia. N Engl J Med 1996; 335: 1933-40.
4. Whitehead J. The design and analysis of sequential clinical trials, rev. 2nd ed. Chichester: John Wiley & Sons Ltd, 1997.
5. Armitage P. Sequential medical trials. 2nd ed. Oxford: Blackwell Scientific Publications, 1975.
6. Van der Tweel I, Kaaks R, Noord PAH van. Comparison of one-sample two-sided sequential t-tests for application in epidemiological studies. Stat Med 1996; 15: 2781-95.
7. De Waard F, Collette HJA, Rombach JJ, Baanders-van Halewijn EA, Honig C. The DOM project for the early detection of breast cancer, Utrecht, the Netherlands. J Chronic Dis 1984: 37: 1-44.
8. Van Noord PAH. Selenium and human cancer risk (nail keratin as a tool in metabolic epidemiology). [thesis] Utrecht: University of Utrecht, the Netherlands, 1992.
9. Cox DR, Hinkley DV. Theoretical statistics. London: Chapman and Hall, 1974.
10. Miettinen OS. Individual matching with multiple controls in the case of all-or-none responses Biometrics 1969; 25: 339-55.
11. Breslow NE, Day NE. Statistical methods in cancer research, volume I. Lyon: IARC, 1980.
12. Armitage P, Berry G. Statistical methods in medical research, 3d ed. Oxford: Blackwell Science, 1994.
13. Brunier H, Whitehead J. PEST3 Operating manual. Reading: University of Reading, U.K., 1993.
14. Ury HK. Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. Biometrics 1975; 31: 643-9.
15. Strömberg U. A method for deciding early stopping of inconclusive case-control studies in settings where data are stratified. Stat Med 1997; 16: 2327-37.
16. Fertig JW, Chilton NW, Varma AO. Studies in the design and analysis of dental experiments. 9. Sequential analysis (Sign test). J Oral Ther Pharm 1964; 1: 45-56.
17. Bellisant E, Benichou J, Chastang C. Application of the triangular test to phase II cancer clinical trials. Stat Med 1990; 9: 907-17.
18. Sprott DA. Normal likelihoods and their relation to large sample theory of estimation. Biometrika 1973; 60: 457-65.
19. Cox DR. Large sample sequential tests for composite hypotheses. Sankhyā 1963;25:5-12.
20. Pasternack BS, Shore RE. Sample sizes for individually matched case-control studies: a group sequential approach. Am J Epidemiol 1982; 115: 778-84.
21. Lan KKG, Wittes J. The B-value: a tool for monitoring data. Biometrics 1988; 44: 579-85.
22. Hunsberger S, Sorlie P, Geller NL. Stochastic curtailing and conditional power in matched case-control studies. Stat Med 1994; 13: 663-70.
23. Betensky RA. Conditional power calculations for early acceptance of H0 embedded in sequential tests. Stat Med 1997; 16: 465-77.
24. Ware JH, Muller JE, Braunwald E. The futility index. An approach to the cost-effective termination of randomized clinical trials. Am J Med 1985; 78: 635-43.
25. Van der Tweel I, Noord PAH van, Kaaks R. Application of a sequential t-test in a cohort nested case-control study with multiple controls per case, J Clin Epidemiol 1993; 46: 253-9.
26. Kaaks R, Tweel I van der, Noord PAH van. Efficient use of biological banks for biochemical epidemiology: exploratory hypothesis testing by means of a sequential t-test. Epidemiology 1994; 5: 429-38.

27. Van der Tweel I, Noord PAH van. A method for deciding early stopping of inconclusive case-control studies in settings where data are stratified. by Strömberg U. [letter] Stat Med 1999; 18: 361-3.
28. Jennison C, Turnbull BW. Interim analyses: the repeated confidence interval approach' (with discussion). J Roy Stat Soc B 1989; 51: 305-61.
29. Hosmer DW, Lemeshow S. Applied logistic regression, New York: Wiley 1989.
30. Mood AM, Graybill FA, Boes DC.Introduction to the theory of statistics, 3d ed.McGraw Hill 1974.

APPENDIX I

In case of $1 : M_i$ matching the conditional likelihood for the *i*th matched set or stratum with M_i controls matched to each case is equal to[29]

$$L_i(\theta) = \frac{e^{\theta X_{1i}}}{e^{\theta X_{1i}} + \sum_{m=2}^{M_i+1} e^{\theta X_{mi}}}$$

The logarithm of $L_i(\theta)$ is equal to.

$$l_i(\theta) = \theta X_{1i} - \ln\left(\sum_{m=1}^{M_i+1} e^{\theta X_{mi}} \right)$$

Taking the first derivative with respect to θ and substituting $\theta = 0$ leads to

$$Z_i = X_{1i} - \frac{1}{M_i + 1} \sum_{m=1}^{M_i+1} X_{mi}$$

The negative of the second derivative with respect to θ and substitution of $\theta = 0$ leads to

$$V_i = \frac{1}{M_i + 1} \sum_{m=1}^{M_i+1} X_{mi}^2 - \left(\frac{1}{M_i + 1} \sum_{m=1}^{M_i+1} X_{mi} \right)^2$$

APPENDIX II

Let η be the probability of exposure under the null hypothesis with $\eta = \pi_1 = \pi_0$. For C_i cases matched to M_i controls the results in the ith matched set can be tabulated as follows:

ith matched set	Cases	Controls	Total
Number of persons	C_i	M_i	N_i
Number exposed	S_i	T_i	E_i
Number not exposed	F_i	G_i	$\overline{E}_i$

For the ith matched set, V_i is equal to $\dfrac{C_i M_i E_i \overline{E}_i}{N_i^2 (N_i - 1)}$.

Because E_i can be assumed to follow a binomial distribution with parameters N_i and η, the expectation of V_i is equal to

$$\mathscr{E}(V_i) = \frac{C_i M_i}{N_i^2 (N_i - 1)} \mathscr{E}(E_i \overline{E}_i) = \frac{C_i M_i}{N_i^2 (N_i - 1)} N_i (N_i - 1) \eta (1 - \eta) = \frac{C_i M_i}{N_i} \eta (1 - \eta)$$

For $C_i = 1$ and for $M_i = 1$, the expectation of V_i is equal to $\mathscr{E}(V_i) = \eta(1-\eta)/2$, for $M_i = 2$ the expectation of V_i is equal to $\mathscr{E}(V_i) = 2\eta(1-\eta)/3$, and for $M_i = 3$ the expectation of V_i is equal to $\mathscr{E}(V_i) = 3\eta(1-\eta)/4$.

APPENDIX III

Suppose, the probability of a missing control in a matched set is equal to π_m. For 1 case and M_i controls in the ith matched set, the expectation of V_i given M_i becomes

$$\mathscr{E}\left(V_i \middle| M_i\right) = \sum_{k=1}^{M_i} \binom{M_i}{k} \left(1 - \pi_m\right)^k \pi_m^{M_i - k} \ \mathscr{E}(V_i \mid k \text{ controls present in the } i\text{th set})$$

$\mathscr{E}(V_i \mid k$ controls present in the ith set) is given in appendix II with $k = M_i$, $C_i = 1$ and $N_i = k+1$.

APPENDIX IV

The probability of a discordant pair of uncorrelated observations π_{disc} is equal to $\pi_0(1\text{-}\pi_1) + \pi_1(1\text{-}\pi_0)$. The probability π_1 can be substituted by a function of π_0 and ψ. The probability π_0 is replaced in our simulations by U, a random variable with a uniform distribution on [0,1] and thus with expectation $\mathscr{E}(U) = 1/2$, variance $\text{var}(U) = 1/12$ and $\mathscr{E}(U^2) = 1/3$.

The probability of a discordant pair of uncorrelated observations π_{disc} can then be rewritten as

$$\frac{\psi U(1-U) + U(1-U)}{\psi U + (1-U)} = (\psi + 1)\frac{U(1-U)}{\psi U + (1-U)} = (\psi + 1)\frac{NUM}{DEN}$$

The expectation of π_{disc} is equal to $(\psi+1)\ \mathscr{E}(\text{NUM/DEN})$. The expected value of NUM/DEN can be approximated by[30]

$$\frac{\mathscr{E}(NUM)}{\mathscr{E}(DEN)} - \frac{\text{cov}(NUM,DEN)}{(\mathscr{E}(DEN))^2} + \frac{\mathscr{E}(NUM)\text{var}(DEN)}{(\mathscr{E}(DEN))^3}$$

with $\mathscr{E}(\text{NUM}) = 1/6$, $\mathscr{E}(\text{DEN}) = (\psi+1)/2$, $\text{var(DEN)} = (\psi\text{-}1)^2/12$ and

$$\begin{aligned}\text{cov(NUM,DEN)} &= \mathscr{E}\{U(1\text{-}U)[\psi U+(1\text{-}U)]\} - \mathscr{E}(\text{NUM})\,\mathscr{E}(\text{DEN})\\ &= \mathscr{E}\{\psi U^2(1\text{-}U)+U(1\text{-}U)^2\} - (\psi+1)/12\\ &= \psi\mathscr{E}(U^2\text{-}U^3)+\mathscr{E}(U\text{-}2U^2+U^3) - (\psi+1)/12\\ &= \psi(1/3\text{-}1/4)+(1/2\text{-}2/3+1/4) - (\psi+1)/12=0\end{aligned}$$

and thus becomes $\frac{4}{9} \times \frac{\psi^2 + \psi + 1}{(\psi + 1)^3}$.

Then the expected value of π_{disc} becomes $\frac{4}{9} \times \frac{\psi^2 + \psi + 1}{(\psi + 1)^2}$.

CHAPTER 5

Early stopping in clinical trials and epidemiological studies for futility: conditional power versus sequential analysis

I. van der Tweel*, P.A.H. van Noord#

**Centre for Biostatistics, Utrecht University, Utrecht, the Netherlands*
#Julius Center for Health Sciences and Primary Care, Utrecht University, Utrecht, the Netherlands

(An abridged version of this chapter was published in the
Journal of Clinical Epidemiology 2003; 56:610-617)

Abstract

Early stopping of a clinical trial is well accepted when there is enough evidence for a significant effect. However, during the course of a trial there can be reasons to consider early termination for 'futility'. In epidemiological studies costly or destructive laboratory tests or slow case accrual can make it desirable to stop a study early for reasons of efficiency.

Estimation of the conditional power (CP) is proposed as a decision tool to stop a study early or to continue it. We consider the disadvantages of this method. We propose (group) sequential continuation of the trial or study as a less arbitrary strategy. We re-analysed two data sets from the literature to illustrate the advantages of a sequential approach.

We conclude that (group) sequential analyses have several advantages as compared to the use of CP. We therefore plea that more studies should consider a sequential design and analysis, where possible, to enable early stopping when enough evidence has accumulated to conclude a lack of the expected effect. Such a strategy can save valuable resources for more promising hypotheses.

Keywords: sequential tests, conditional power, early stopping, randomized clinical trials, epidemiological studies

5.1 Introduction

In clinical trials early stopping for ethical or economical reasons is well accepted. For that purpose, often one or more interim analyses or a (group) sequential analysis are planned at the design phase of the trial, mostly to allow the trial to stop as soon as enough evidence for a significant effect is available. Nevertheless for some trials no interim analysis was foreseen[1], but in the course of the trial the need for it is felt. Reasons include a much slower than anticipated patient accrual, intervening results of comparable trials, etc. In these cases the question behind an interim analysis often is whether the trial can be stopped because of lack of a relevant difference between the treatment groups.

In observational epidemiological studies, in general, no need is felt to stop early in case of a clear exposure effect. However, in cohort studies, slow case accrual, costly laboratory tests or tests that require destruction of unique biological samples can make it desirable to stop early when the data obtained so far indicate no relevant effect.

Thus in the course of some clinical trials or cohort studies the question can arise whether to stop early and accept the null hypothesis (no difference between treatments or no exposure effect) or to continue. This paper starts with a review of two examples from the literature where this question was raised. Next we discuss the use of conditional power as proposed by the authors of these examples as a strategy for early stopping and its disadvantages. Then we present the results of re-analyses of both examples using a sequential approach to illustrate its advantages.

5.2 Two examples from the literature

5.2.1 Example I:

The Lupus Nephritis Collaborative Study (LNCS) was a multicenter clinical trial to evaluate standard drug therapy S (prednison plus cyclophosphamide) versus standard drug therapy plus plasmapheresis P (plasma exchange) in the treatment of nephritis associated with systemic lupus erythematosis (SLE).[2]

The principal outcome was renal failure or death. The study was designed to detect a 50% reduction in the hazard rate (from a value of 0.30 for S to a value of 0.15 for P) with a one-sided type I error α = 0.05 and a power 1-β equal to 0.88. The fixed sample size was targeted to be 125 patients. During the trial it had to be decided whether a renewal application for funding should be submitted, so an external advisory committee reviewed the emerging results. Life table analysis of the primary study outcome showed no meaningful difference between the two groups. The cumulative survival probability was even slightly higher for the standard treatment S. By that time 46 patients were randomized to S and 40 patients to P. The average follow-up duration was 97 weeks, ranging from 1 to 225 weeks.

Because of this lack of difference in the principal outcome measure and, to a lesser extent, the slow recruitment of patients, the need for further continuation of the trial was questioned.

5.2.2 Example II:

The Atherosclerosis Study in Communities (ASC) was designed as a nested case-control study to look at the association of atherosclerosis with the presence of three viral antibodies: cytomegalovirus (CMV), herpes simplex virus 1 (HSV1) and herpes simplex virus 2 (HSV2).[3] The presence of antibodies to each virus in a blood sample indicated exposure to the virus. The hypothesis to be tested was that exposure to these viruses was associated with early atherosclerosis. The cases were persons without clinical evidence of atherosclerosis who were found to have abnormal arterial wall thickening of their carotid arteries as determined by non-invasive B-mode ultrasound. The controls were matched, amongst others, on age and gender.

The study was designed to detect an odds ratio (OR) of 2 in 300 matched pairs for occurrence of the antibodies in cases when compared to controls, using McNemar's test with a two-sided type I error 2α of 0.05 and a power 1-β of 0.80.

Based on the data from the first 100 matched pairs, the authors questioned whether the study should be continued. The laboratory blood analysis of antibodies was costly, so analysis of the blood of all cases and controls was not desirable if there would be little chance to detect the hypothesized relationships.

5.3 Methods

5.3.1 Conditional Power

Calculation of the conditional power (CP) was proposed and used as a decision tool for either stopping a study early or continuing it. The CP is defined, at a given information fraction k (i.e. given the data processed so far), as the probability $p_k(\theta)$ that a statistical test will reject the null hypothesis H_0 at the end of the study and assuming θ as the parameter value of interest for the remainder of the study.[4] The information fraction k is the observed proportion of the total amount of planned information, i.e. for example the planned fixed sample size. Under the null hypothesis H_0 the parameter θ is assumed to be equal to 0; under the alternative hypothesis H_1 (the absolute value of) θ is assumed to be equal to θ_R (depending on a one-sided or a two-sided test), θ_R reflecting the treatment difference or effect size relevant to detect. When $k = 0$, $p_k(\theta)$ is the unconditional power function. The null hypothesis H_0 can be accepted at an information fraction k if the CP assuming a certain value of θ is smaller than some value γ.

In this paper we will focus on early stopping of a trial or a study to accept H_0. With this in view several authors have proposed different values for the parameter θ to calculate the CP. Hunsberger *et al*[3] calculated the CP, amongst other values, for $\theta = \theta_R$ and for a parameter value θ estimated from the first part of the data. Pepe and Anderson[5] proposed for a two-stage experimental design an 'optimistic but plausible alternative given the initial data' under which the CP might be calculated: $\hat{\mu}_1 + 1.0\, s.e.(\hat{\mu}_1)$, where $\hat{\mu}_1$ is the mean of the first n_1 observations. This parameter value is the limit of the (one-sided) 84%-confidence interval (CI) for the parameter estimate based on the data processed so far. Here we denote this parameter value as θ_2. Strömberg[6] used, amongst other values, the limits of the (one-sided) 75%- or 90%-CI for the parameter estimate as parameter values, i.e. $\hat{\mu}_1 + 0.674\, s.e.(\hat{\mu}_1)$ or $\hat{\mu}_1 + 1.282\, s.e.(\hat{\mu}_1)$, respectively. We denote these parameter values as θ_1 and θ_3, respectively.

The decision to stop the study and accept the null hypothesis is based on a CP, calculated under a parameter value θ, falling below some prespecified value γ, say 0.1 (conservative) to 0.3 (non-conservative).[7]

5.3.2 Disadvantages of the Conditional Power

CP calculations require extrapolation of the study results obtained so far and are based upon rather arbitrary choices. First, a choice must be made for a plausible parameter value θ to calculate the CP e.g. a) the parameter value θ_R as specified in the design phase under H_1 or b) the parameter value based on the data obtained so far or c) the parameter value based on a limit of the CI for the parameter estimate (θ_1, θ_2 or θ_3). Second, a choice must be made for the critical value γ for the CP to decide for early stopping or continuing. Third, a choice must be made for the 'optimal' information fraction *k* to estimate the CP.

In addition, early stopping rules affect the type I and type II error probabilities of the hypothesis testing procedure as well as the validity of the final parameter estimates.

5.3.3 A sequential approach as an alternative for CP

A study that is sequential by design does not depend on the above-mentioned, rather arbitrary choices (the parameter value θ, the value of γ and the 'optimal' information fraction *k*) necessary for the calculation of the CP. In a sequential approach the data can be analysed under the same (*a priori*) specifications as were made in the fixed sample design.[8] The general approach for a sequential analysis[9] is as follows. A null hypothesis H_0 and an alternative hypothesis H_1 are formulated for a suitable measure θ of treatment difference or exposure effect. Two test statistics *Z* and *V* can be derived depending on the type of response variable: *Z* is a measure of the treatment difference or the exposure effect, *V* reflects the amount of information about θ contained in *Z*. For each new patient or group of patients values for *Z* and *V* can be calculated and presented graphically by plotting *Z* against *V* (see Figures 1a and 1b for illustration of one-sided sequential tests). When *Z* and *V* are calculated after each new patient, the analysis is called a continuous sequential test; when *Z* and *V* are calculated after each new group of patients it is called a group sequential test.

After each calculation of Z and V one of three decisions is made (Figures 1a and 1b):

a) the analysis is stopped and H_0 is rejected (the upper boundary is crossed),

b) the analysis is stopped and H_0 is accepted (the lower boundary is crossed),

c) the process is continued with one or more new patients (the new (*Z*,*V*)-point is still within the two boundaries).

Various types of sequential tests are described[9] of which we used the (truncated) sequential probability ratio test (SPRT) and the triangular test (TT) as illustration. All tests require critical boundaries to be specified in advance. These boundaries depend on θ_R, the (one- or two-sided) type I error (α or 2α) and the type II error (β).
Characteristics of the sequential tests are described at length by Whitehead.[9]

So, when in the course of a study the question arises whether it can be stopped early, we propose to analyse the available data sequentially using the same values for θ_R, type I error and type II error as specified in the original fixed sample design. If the sequential analysis based on the data from the first part of the study leads to the decision to continue the study, it can easily be continued sequentially. With this approach patients, money, time, biological samples, etc can be saved by an average amount of 35 to 65%, particularly when H_0 or H_1 is true[9], as we will show by sequential re-analysis of the data from the LNCS-trial[2] and from the ASC-study.[3]

When a sequential analysis is stopped, a median unbiased estimate for the treatment or exposure effect can be calculated with its CI. This estimate is adjusted for the multiple testing due to the repeated looks at the cumulative data.

The computer program PEST[10] was used for the sequential analyses.

5.4 Sequential re-analysis of the two examples from the literature

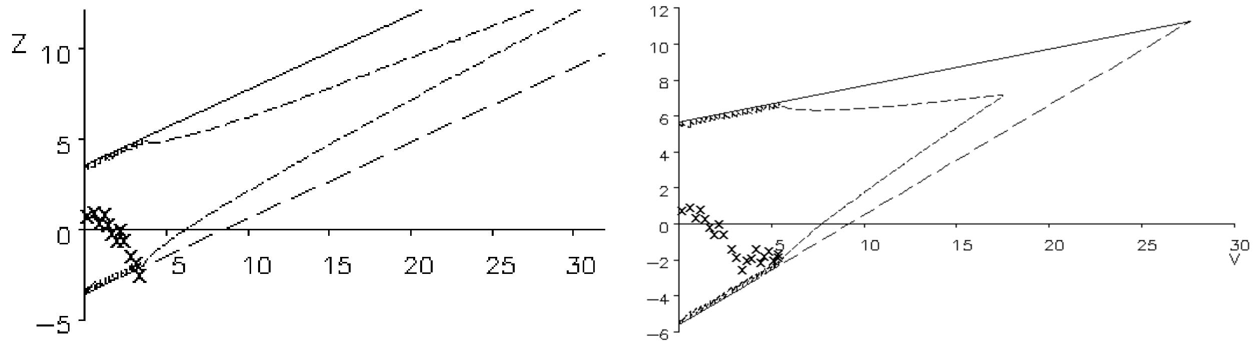
5.4.1 Example I:

In the LNCS-trial[2] unconditional and conditional power calculations were based on data observed in the first 86 patients. The actual incidence of renal failure was much lower than anticipated, so the unconditional power was less than 0.5 to detect a 50% reduction in hazard rate. Conditional Power values were calculated for various assumed values of the incidences of renal failure under P and under S. These CP values were less than 0.15 for a 50% reduction in hazard rate. Thus it was considered highly unlikely that plasmapheresis would add significant benefit to standard drug therapy and the trial was terminated.

We re-analysed the LNCS-data sequentially with a (truncated) SPRT and a TT. The specifications for the sequential tests were the same as the ones used in the design phase for the fixed sample size calculations.[2] As measure of treatment difference θ was taken equal to the negative logarithm of the hazard ratio (HR). Thus the null hypothesis can be formulated as H_0: $\theta = 0$ with a (one-sided) type I error $\alpha = 0.05$ and a type II error $\beta = 0.12$. The HR was powered to be 0.5, which translated into an alternative hypothesis H_1: $\theta = \theta_R = -\ln(0.5) = 0.693$. The events and their time-to-event and the censored follow-up times were entered chronologically into the sequential analysis. Note that the test statistic Z is the log-rank statistic and Z^2/V corresponds to the well-known log-rank test.

Using an SPRT (Figure 1a), the trial could have been terminated with acceptance of the null hypothesis after inclusion of the results of the first 58 patients with 4 events in 29 patients in group S and 10 events in 29 patients in group P. The median unbiased estimate of the HR was equal to 1.73 with a 95% CI (0.56 ; 5.22).

Using a TT (Figure 1b), results of all 86 patients with 10 events in group S and 12 events in group P were entered sequentially. The sample path touched the boundary for acceptance of the null hypothesis, but just did not cross it. The median unbiased estimate of the HR was equal to 1.37 with a 95% CI (0.58 ; 3.21). (NB Lachin reported HR = 1.41 (0.61 ; 3.26) as the final estimate for the HR for the 86 patients.)



STOP the study – a boundary has been crossed

CONTINUE the study – no boundary has yet been crossed

Figure 1a
LNCS-data analysed in a one-sided truncated SPRT with θ_R=-ln(0.5), $\alpha = 0.05$ and $\beta = 0.12$. (The truncation point was cut off.)

1b
LNCS-data analysed in a one-sided TT with θ_R=-ln(0.5), $\alpha = 0.05$ and $\beta = 0.12$.

5.4.2 Example II:

In the ASC-study[3] CP values were calculated based on data from the first third of the study (100 matched pairs) and assuming the estimated current trend as the parameter value for the rest of the study. For HSV1 and HSV2 CP values of about 0.39 were found, whereas the CP for CMV was 0.90. The authors decided to complete the study till the planned end. They also calculated CP values assuming the parameter value θ_R as specified originally under H_1. These CP values were 0.72 for HSV2, 0.86 for HSV1 and 0.94 for CMV.

To obtain an idea of the 'sensitivity' of our approach we created ten random permutations of all 340 matched pairs resulting from both the HSV1 and the HSV2 data at the end of the study (Table 1).

Table 1 Results at the end of the ASC study[3] (after 340 matched pairs).

		cases			
		HSV1		HSV2	
		neg.	*pos.*	*neg.*	*pos.*
controls	*neg.*	34	65	195	51
	pos.	46	195	56	38
OR		1.41		0.91	
exact p-value		0.09		0.70	
exact 95%-CI		(0.95 ; 2.11)		(0.74 ; 1.64)	

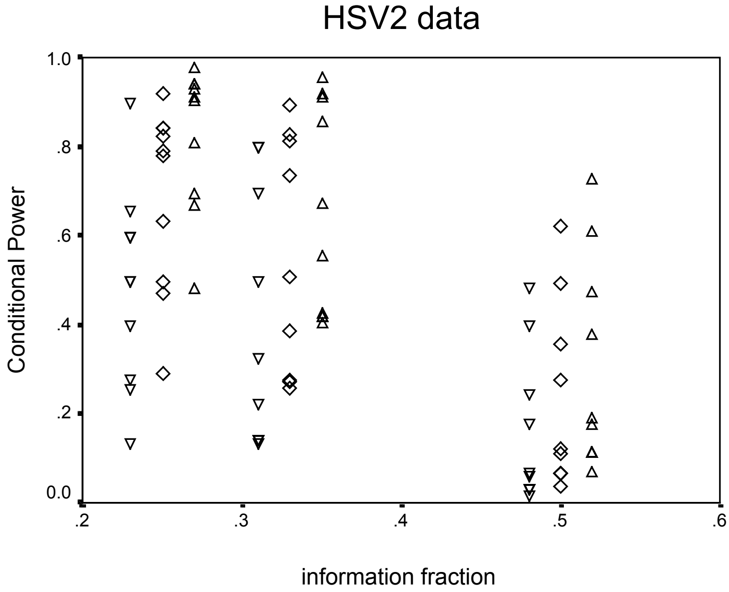


Figure 2 Conditional power values for information fraction k = 1/4, 1/3 and 1/2 for parameter values θ equal to θ_1 (∇), θ_2 (◊) and θ_3 (Δ) for 10 permutations of the HSV2 data. (For illustration purposes k was diminished and increased by a small amount.)

For each permuted data set we adapted the method as described[3] for the calculation of the CP for McNemar's test to allow its use for two-sided tests. Figure 2 shows the CP values calculated when 75, 100 or 150 matched pairs were analysed (i.e. at k = 1/4, 1/3 or 1/2 respectively) for three parameter values θ_1, θ_2 and θ_3.

When 1/4 or 1/3 of the planned case-control sets were analysed, CP values were in general too high (i.e. higher than 0.3) to terminate the study. When half of the fixed sample size was analysed, CP values were lower but still varied considerably. When the current trend in the data was extrapolated, CP values were less than 0.2 for 6 of 10 and 9 of 10 permutations at information fractions of 1/3 and 1/2, respectively (not shown). In general, using larger parameter values θ led to higher CP values.

We also analysed each permutation of the data by a group sequential test and a continuous sequential test, both with a double SPRT and a double TT. We analyzed the data group sequentially after inclusion of 20%, 40%, 60%, 80% and 100% of the matched pairs respectively. All sequential tests were designed[9,11] for 1 : 1 matched dichotomous data with the same specifications as for the original fixed sample case-control study (H$_0$: $\theta = 0$ versus H$_1$: $|\theta| = \theta_R = \ln(2)$ with $2\alpha = 0.05$ and $\beta = 0.20$). The parameter θ was chosen equal to the logarithm of the OR, that is $\ln[\pi(1-\pi)]$ with π as the probability of a discordant pair with case 'antibody positive' and control 'antibody negative'. The test statistics are $Z = S_n$ - n/2 and $V = n$/4, where n is the number of discordant pairs observed, S_n is the *observed* number of discordant pairs with case 'antibody positive' and control 'antibody negative' and n/2 is the *expected* number of discordant pairs under the assumption that the probability of a discordant pair π is equal to 1/2. Note that here Z^2/V corresponds to McNemar's test for matched pairs without continuity correction.

The group sequential and continuous sequential tests for the 10 permutations of the HSV2 data all ended with the acceptance of H_0. Group sequential analysis with an SPRT or TT led more often to early termination of the HSV2 data than calculation of the CP with θ_2 as the parameter value ($CP < 0.3$). When results of continuous and group sequential analyses are compared, in general, more data sets could be saved using a continuous sequential test. The percentage saving using a continuous SPRT was greater than or at least equal to the amount of saving using a continous TT for all permutated data sets (Table 2).

Table 2 The percentage saving in matched sets after continuous or group sequential analysis for ten permutations of HSV2 data. All sequential analyses accepted H_0.

	continuous sequential		group sequential	
permutation	SPRT	TT	SPRT	TT
1	67	64	60	60
2	50	50	60	40
3	69	49	60	60
4	71	61	60	60
5	76	68	60	60
6	61	59	60	40
7	75	65	60	40
8	71	69	80	60
9	54	54	40	40
10	61	61	60	60

Using the SPRT, sequential evaluation of matched sets would have saved 50% to 76% as compared to the analysis of the planned fixed sample size. The median number of matched sets needed was 96.5. The range of values for V in the ten simulations was 5.5 to 11.0 with a median value of 7.75. With respect to the fixed sample size V-value (=16.34) the range for V was 0.34 to 0.67 with a median value equal to 0.47. (Note that V is in terms of ‘information’ i.e. discordant pairs, thus saving 33% to 66% of these.) With the TT 49% to 69% of the fixed sample number of matched sets could have been saved. The median number of matched sets needed was 118. The range of values for V in the ten simulations was 7.75 to 11.75 with a median value of 9.125. With respect to the fixed sample size V-value (=16.34) the range for V was 0.47 to 0.72 with a median value equal to 0.56. (In terms of discordant pairs this means savings of 28% to 53%.) Group sequential SPRT and TT resulted in a median saving of 60% of the matched sets.

In Figure 3 the median unbiased estimates for the ORs and their 95% CIs are depicted for each of the ten permuted data sets. All confidence intervals include 1, the value of the OR under H_0.

Results of continuous sequential analyses using a (truncated) SPRT or a TT for one of the simulated HSV2 data sets are presented as an example in Figures 4a and 4b. The (truncated) SPRT (Figure 4a) could be stopped after 87 matched sets with 36 discordant sets (a saving in terms of the number of matched sets of 71%); the TT (Figure 4b) could be stopped after 118 matched sets with 43 discordant sets (a saving of 61%).

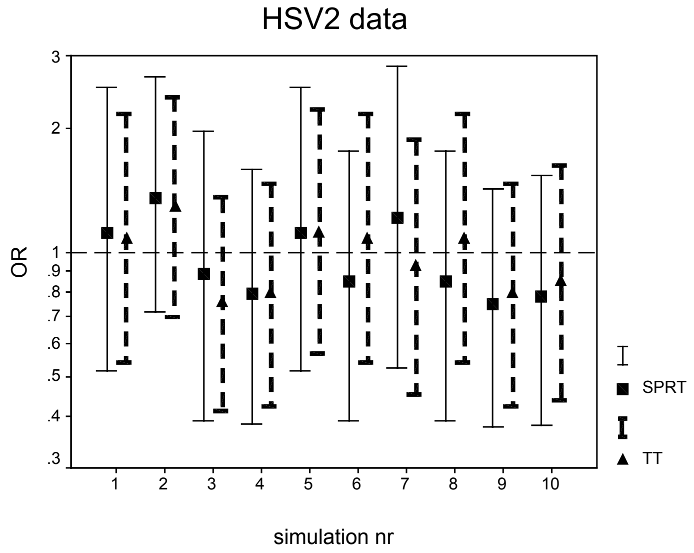
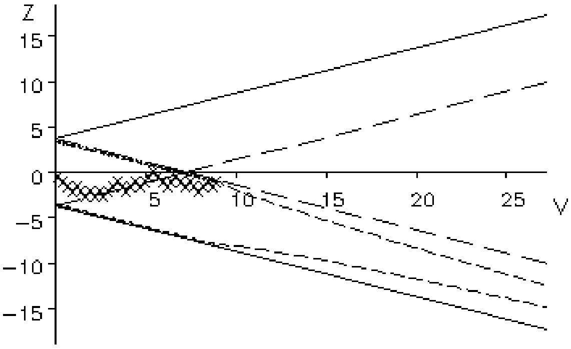
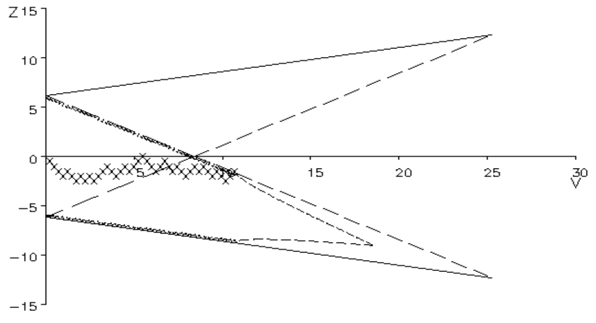


Figure 3 Median unbiased estimates and 95% confidence intervals for the Odds Ratio (OR) for 10 permutations of the HSV2 data analysed by the SPRT and TT.



STOP the study – a boundary has been crossed



STOP the study – a boundary has been crossed

Figure 4a
One of the permutations of the HSV2 data analysed in a two-sided truncated SPRT with $\theta_R = \ln(2)$, $2\alpha = 0.05$ and $\beta = 0.20$. (The truncation point was cut off.)

4b
One of the permutations of the HSV2 data analysed in a two-sided TT with $\theta_R = \ln(2)$, $2\alpha = 0.05$ and $\beta = 0.20$.

In only one of the ten permuted HSV1 data sets a CP value based on θ_2 between 0.1 and 0.3 could have led to early termination after 1/3 of the data sets were analysed.

When the current trend was extrapolated, CP values for this same permutation were less than 0.1 after 1/4, 1/3 and 1/2 of the data was observed. Sequential analyses of the ten permuted data sets led to savings ranging from 0 to 91% (Table 3).

Table 3 The percentage saving in matched sets after continuous or group sequential analysis for ten permutations of HSV1 data.
Most analyses accepted H_0, except for those marked by a * for which H_0 was rejected; and those marked by $ for which H_0 was accepted after the planned 300 matched sets, but within the actually analysed 340 matched sets.

	continuous sequential		group sequential	
permutation	SPRT	TT	SPRT	TT
1	25	53*	0	20
2	44	44	40	40
3	$0^{\$}$	22	0	20
4	9	20*	0	20*
5	91*	84*	80*	80*
6	$0^{\$}$	58*	60*	60*
7	$0^{\$}$	35*	$0^{\$}$	20*
8	43	46	40	40
9	71	65	60	60
10	66	66	60	60

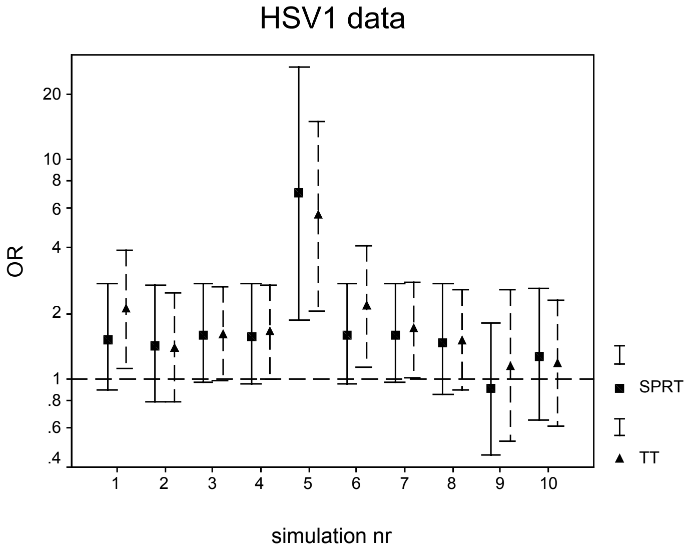


Figure 5 Median unbiased estimates and 95% confidence intervals for the Odds Ratio (OR) for 10 permutations of the HSV1 data analysed by the SPRT and TT.

Again, for most of the permutations, a continuous sequential analysis saved more data sets (median saving of 44%) than a group sequential analysis (median saving of 40%). One of the permuted data sets ended very early with the rejection of H_0. For the other data sets the parameter estimates and their CIs were similar (Figure 5).

5.5 Discussion

During the course of a clinical trial or prospective epidemiological study the need can be felt to consider early stopping for 'futility', because accrual is much slower than anticipated or because the hypothesized effect size does not seem to exist. For AIDS clinical trials a strong relationship between early patient enrollment and the eventual ability of the trial to attain its target sample size was noticed.[12] For these trials the feasibility or futility can be gleaned from the early enrollment patterns to a considerable extent. When patient enrollment in a trial lags behind and it looks unlikely that the target sample size will be reached, investigators should consider whether to stop the trial for futility with unconclusive results or change to a sequential continuation of the trial hoping this will lead to more conclusive evidence.

Stochastic curtailment procedures, calculation of the conditional coverage probability (CCP) and calculation of the CP have been proposed as decision tools for early stopping of a trial or study. Stochastic curtailment procedures were developed to enable early stopping by evaluating the CP at multiple times during a study.[13] In case early stopping is considered at just a single point in time these procedures may be conservative in terms of their type I and type II error probabilities. The CCP is defined[6] as the probability that a two-sided (95%) confidence interval around the final estimate, given the observed data and an assumed alternative hypothesis H_1 for the remainder of the study, includes the parameter value under the null hypothesis H_0 (e.g. the value 1 when H_0: OR = 1 is assumed). If the CCP is high for plausible values of the assumed parameter, the study is likely to be inconclusive.

Calculation of the CP and the CCP is not straightforward; sometimes approximations to the exact CP have to be made and/or ad-hoc computer programs have to be written. The calculation requires extrapolation of the data obtained so far and is based upon rather arbitrary choices. First, a parameter value θ must be chosen. Ware *et al*[14] calculated their 'futility index' (defined as the conditional probability that a trial will fail to demonstrate an effect given the results already observed) using θ_R. Pepe and Anderson[5] showed that stopping rules based on θ_R may be overconservative in many cases. Therefore, they proposed to base the CP calculations on θ_2, the limit of the (one-sided) 84%-CI for the parameter estimate. Strömberg[6] used the limits of the (one-sided) 75%- or 90%-CI as the parameter values (θ_1 and θ_3 respectively) for his CCP calculations. Thus a variety of possible parameter values θ can be thought of and the CP value can be quite different depending upon the value chosen.[5] Second, a threshold must be determined for γ, the CP value below which the decision to stop the study and accept H_0 can be made. Ware *et al*[14] used a value equivalent to a γ of 0.33. Pepe and Anderson[5] recommend values less than or equal to 0.3. Betensky[15] works with threshold values of 0.1 (conservative) and 0.3 (non-conservative). Conservative values for γ do not affect the power much while non-conservative values will lead to power loss. Strömberg[6] gives no clear choice for the critical value for the CCP. He admits that his approach permits a flexible definition of an 'inconclusive study'. Third, the 'optimal' information fraction *k* must be chosen. Pepe and Anderson[5] find it difficult to provide a general recommendation, but think that values between 1/4 and 1/2 have intuitive appeal. In practice, the moment for calculating the CP arises more from the trial or study progress than from a possible 'ideal' information fraction.

Therefore, we propose to analyse the data sequentially as an alternative to the estimation of CP or CCP. Either a group sequential or a continuous sequential test procedure can be used in such situations, where the choice depends on pragmatic considerations. We illustrated our proposal with re-analyses of two examples from the literature.[2,3] In our re-analyses we chose the same specifications for the sequential test (θ_R, α and β) as for the original fixed sample size design. Lachin *et al*[2] concluded, based on calculations of, amongst others, the (un)conditional power that it was highly unlikely that there would be any benefit of additional plasmapheresis. For these calculations they specified a range of plausible values for the true incidences or parameter values. Our sequential re-analysis of the LNCS-data showed that savings in the number of patients compared to the corresponding fixed sample size could be none (the TT) or 33% (the SPRT). Hunsberger *et al*[3] decided to complete their association studies based on CP values larger than 0.3 for all three antibodies. CP values calculated under the originally specified parameter value θ_R were highest (always higher than 0.3) and thus indeed very conservative. Our results confirm that a decision based on the CP to stop the study or to continue is rather arbitrary. We studied the 'sensitivity' of our sequential analyses by the creation and analysis of 10 random permutations of the final HSV1 and HSV2 data.[3] Sequential analyses more often led to early termination of the studies and in a more objective way.

Sequential analyses have several important advantages as compared to the use of CP:

- no extrapolation of the observed data is necessary;
- no choice needs to be made for the threshold value γ for the CP;
- no choice is required for the timing of the calculation of the CP, i.e. for the information fraction k;
- stopping guarantees the type I and type II error probabilities, so there is no power loss;
- median unbiased parameter estimates and their CIs can be calculated after stopping;

- standard software is commercially available[10], so no ad-hoc computer programs have to be written;
- different kinds of outcome variables (binary, ordinal, survival, normal) can be analysed with this software, while the calculation of the CP is complicated e.g. for censored survival data[15] and demands simplifying assumptions;
- adjustment for prognostic factors or confounding variables is possible[9,10] ;
- sequential analysis using the boundaries approach as in PEST can be performed either continuously or in a group-sequential way without adaptations, contrary to the adaptations proposed for conditional power calculations[7,15] at each time point.

On average, a sequential analysis requires less patients to come to a decision than a fixed sample size analysis.[9]

Although in theory these advantages are clear, we acknowledge that pragmatic considerations can hamper the practical realization of sequential analysis. At the same time we emphasize that only the primary outcome variable is monitored sequentially and thus needs to be up-to-date. Furthermore, when continuous sequential monitoring is not easy or convenient, group sequential analysis could be considered.

Both for fixed sample size tests and for sequential tests the (expected) sample size is based on the type I and type II errors and on θ_R. The choice for θ_R is in most situations conducted by practical considerations of the feasibility of the trial. Although some authors therefore consider the use of θ_R as arbitrary, we think that it is the most ‘objective’ choice, especially when a trial or study is already underway and transformed from a fixed sample size test to a sequential test.

Additional results of simulations[11] performed to determine the characteristics of sequential tests on matched data are given in Table 4.

Table 4 The minimal amount of information $V_{\min}$ needed before H_0 can be accepted in a 1 : 1 matched case-control study or clinical trial with paired dichotomous data with type I error 2α and type II error β (results from simulations).[11] For comparison the corresponding fixed sample size information V_{fixed} is given. OR: Odds Ratio.

		$2\alpha = 0.10$, $\beta = 0.05$		$2\alpha = 0.05$, $\beta = 0.20$	
	OR	$V_{\min}$	V_{fixed}	$V_{\min}$	V_{fixed}
SPRT	1.5	17.75	65.83	11.00	47.76
	2.0	6.00	22.53	4.25	16.34
	2.5	3.25	12.89	2.50	9.35
TT	1.5	25.25	65.83	18.75	47.76
	2.0	9.00	22.53	6.75	16.34
	2.5	4.75	12.89	4.00	9.35

From these results we can derive the minimal information fraction. For example, for the ASC-study[3] (OR=2.0, $2\alpha = 0.05$, $\beta = 0.20$) the minimal information fraction using a SPRT is equal to 4.25/16.34=0.26 and using a TT it is equal to 6.75/16.34=0.41. Before this amount of information is gathered in the study, a sequential analysis cannot be stopped early and accept H_0. This minimal information fraction derived in terms of V is the same as the fraction in terms of the number of discordant pairs.

Recently, Whitehead introduced a 'futility design' in the PEST program (version 4.1).[10] This 'futility design' also allows early termination of a trial when cumulating results look disappointing. But, interim analyses using the 'futility design' cannot result in early stopping to reject H_0. For the 'futility design' either the slope of the boundary must be specified or a maximum value for V, the amount of information to be used. Although a guideline is given for the choice of the value for the slope of the boundary, specifications for the 'futility design' are also arbitrary.

Betensky[15] and Strömberg[6] both observed that little attention is paid to designs of epidemiological studies with a possible early acceptance of H_0. Strömberg even calls for 'further discussion concerning early stopping of epidemiologic studies'. Our results confirm their statements that early stopping for 'futility' can conserve valuable resources.[11,16]

We share the view that one should terminate a clinical trial earlier than planned only in exceptional circumstances.[17] When a study is stopped early, savings in the number of patients, biological samples, in time, costs, etc. must be weighed against the decreased precision of the (adjusted) parameter estimate. This is probably of more concern when the null hypothesis is rejected than when it is accepted for 'futility' with the conclusion of 'no relevant effect'. When a trial or study is stopped early, then the effect estimates (although perhaps non-significant) and their CIs should be published. This is especially valuable for contribution to future overviews or meta-analyses.

Quoting Jennison and Turnbull[4, p.219]: '… it is wise to define a study protocol as unambiguously as possible at the outset. If this is done thoroughly and an interim analysis schedule is also defined, the full range of group sequential tests are available for use and one of these tests may be preferred to stochastic curtailment.'

Our results show that, when a study was not designed sequentially, (group) sequential continuation of the study both provides a more objective strategy requiring no arbitrary assumptions and turns out to be at least as and often even more efficient than the calculation of CP .

5.6 Acknowledgement

We gratefully thank prof. John Lachin and co-workers for taking the trouble to retrieve for us the LNCS data, that were described in their publication.[2]

5.7 References

1. Pocock SJ, Hughes MD. Practical problems in interim analyses, with particular regard to estimation. Contr Clin Trials 1989; 10: 209S-21S.
2. Lachin JM, Lan S-P and the Lupus Nephritis Collaborative Study Group. Termination of a clinical trial with no treatment group difference: The Lupus Nephritis Collaborative study. Control Clin Trials 1992; 13:62-79.
3. Hunsberger S, Sorlie P, Geller NL. Stochastic curtailing and conditional power in matched case-control studies. Stat Med 1994; 13: 663-70.
4. Jennison C, Turnbull BW. Group sequential methods with applications to clinical trials. Boca Raton: Chapman & Hall/CRC, 2000.
5. Pepe MS, Anderson GL. Two-stage experimental designs: early stopping with a negative result. Appl Stat 1992; 41: 181-90.
6. Strömberg U. A method for deciding early stopping of inconclusive case-control studies in settings where data are stratified. Stat Med 1997; 16: 2327-37.
7. Betensky RA. Early stopping to accept H0 based on conditional power: approximations and comparisons. Biometrics 1997; 53: 794-806.
8. Van der Tweel I, Noord PAH van. A method for deciding early stopping of inconclusive case-control studies in settings where data are stratified. by Strömberg U. [letter] Stat Med 1999; 18: 361-3.
9. Whitehead J. The design and analysis of sequential clinical trials, rev. 2nd ed. Chichester: John Wiley & Sons Ltd, 1997.
10. MPS Research Unit. PEST4: Operating manual. Reading: University of Reading, U.K., 2000.
11. Van der Tweel I, Noord PAH van. Sequential analysis of matched dichotomous data from prospective case-control studies. Stat Med 2000; 19: 3449-64.
12. Haidich A-B, Ioannidis JPA. Patterns of patient enrollment in randomized controlled trials. J Clin Epidemiol 2001, 54: 877-83.
13. Lan KKG, Simon R, Halperin M. Stochastically curtailed tests in long-term clinical trials. Comm Statistics-Sequential Analysis 1982; 1: 207-19.
14. Ware JH, Muller JE, Braunwald E. The futility index. An approach to the cost-effective termination of randomized clinical trials. Am J Med 1985; 78: 635-43.
15. Betensky RA. Conditional power calculations for early acceptance of H0 embedded in sequential tests. Stat Med 1997; 16: 465-77.
16. Van der Tweel I, Kaaks R, Noord PAH van. Comparison of one-sample two-sided sequential t-tests for application in epidemiological studies. Stat Med 1996; 15: 2781-95.
17. Simon R. Some practical aspects of the interim monitoring of clinical trials. Stat Med 1994; 13:1401-9.

CHAPTER 6

Sequential tests for gene-environment interactions in matched case-control studies

I. van der Tweel*, M. Schipper*

**Centre for Biostatistics, Utrecht University, Utrecht, the Netherlands*

submitted for publication

Summary

The sample size necessary to detect a significant Gene x Environment interaction in an observational study can be large. For reasons of cost-effectiveness and efficient use of available biological samples we investigated the properties of sequential designs in matched case-control studies to test for both non-hierarchical and hierarchical interactions. We derived the test statistics *Z* and *V* and their characteristics when applied in a two-sided triangular test.

Results of simulations show good agreement with theoretical values for *V* and the type I error. Power values were larger than their theoretical values for very large sample sizes.

Median gain in efficiency was about 27%. For a 'rare' phenotype gain in efficiency was larger when the alternative hypothesis was true than under the null hypothesis.

Sequential designs lead to substantial efficiency gains in tests for interaction in matched case-control studies.

Keywords: sequential tests, gene-environment interactions, matching, case-control studies, sample size

6.1 Introduction

To study the association between a disease, for example cancer, and a genetic risk factor or an environmental factor case-control studies can be designed and analysed. To adjust for possible confounding factors a case can be matched to one or more controls by, for example, ethnicity or age. Both genetic and environmental factors may contribute to the susceptibility of disease and these factors may interact in their influence on the risk of disease.[1] Then a gene-environment interaction can be interesting to explore. Interactions can be tested both hierarchically by first including the main effects (for example G and E for a genetic and an environmental factor) in the model and subsequently testing for the interaction between these main effects (GxE), and non-hierarchically, thus modelling the interaction but not the associated main effects. A non-hierarchical model can be of interest when both the genetic type and the exposure are required to increase the risk of disease.[1,2] This kind of association is discussed as model D.[1]

Sample sizes necessary to detect significant GxE interactions can be large when the occurrence of the genetic factor is rare.[3] Furthermore, when an increased risk does exist for a genetic factor, the interaction is likely to be very moderate[4] and thus, again, large sample sizes will be needed. Besides, epidemiological studies often use biological samples that are limited in quantity or costly to obtain. These considerations make it essential to search for study designs that make very efficient use of available resources.[5-8]

Matched case-control studies already require a smaller sample size than unmatched studies under the same model specifications to study an association with disease.[7,9] While matching already leads to more efficient designs in terms of fixed sample size, sequential methods may reduce the average sample size even further.[6,7] A sequential analysis tests the cumulative data every time new information becomes available.

In the following we describe the test statistics we derived for the sequential analysis of both hierarchical and non-hierarchical GxE interactions in matched case-control studies. We investigated the efficiency of these sequential analyses by simulations.

6.2 Models and hypotheses

In this paper we study a recessive bi-allelic disease with prevalence of the recessive allele a equal to q_a. Then a dichotomous disease phenotype can be defined from the three genotypes AA, Aa and aa. The probability of the disease phenotype, P(G=1), equals q_a^2 for aa and the probability of no disease phenotype, P(G=0), equals $1–q_a^2$ for AA or Aa. The probability distribution for the dichotomous exposure variable is p_e for exposure, i.e. P(E=1), and $1–p_e$ for no exposure, i.e. P(E=0).

a) GxE interaction in a non-hierarchical model:
The likelihood for a non-hierarchical interaction in a conditional logistic regression model, i.e. without main effects G and E, can be written as

$$L(\beta_{GE}) = \prod_{i=1}^{n} \frac{e^{\beta_{GE}G_{i1}E_{i1}}}{e^{\beta_{GE}G_{i1}E_{i1}} + e^{\beta_{GE}G_{i2}E_{i2}}} \tag{1}$$

where G_{i1} and E_{i1} denote the case information for genetic and environmental factors, respectively, and G_{i2} and E_{i2} the control information, in the ith matched set.

b) GxE interaction in a hierarchical model:
The likelihood for a hierarchical interaction in a conditional logistic regression model, i.e. with main effects G and E, can be written as

$$L(\beta_G, \beta_E, \beta_{GE}) = \prod_{i=1}^{n} \frac{e^{\beta_G G_{i1} + \beta_E E_{i1} + \beta_{GE}G_{i1}E_{i1}}}{e^{\beta_G G_{i1} + \beta_E E_{i1} + \beta_{GE}G_{i1}E_{i1}} + e^{\beta_G G_{i2} + \beta_E E_{i2} + \beta_{GE}G_{i2}E_{i2}}} \tag{2}$$

where G_{i1} and E_{i1} denote the case information for genetic and environmental factors, respectively, and G_{i2} and E_{i2} the control information, in the ith matched set.

6.3. The sequential tests

For a sequential test as developed by Whitehead[10] to test the null hypothesis H_0: $\theta = 0$ versus H_1: $\theta \neq 0$ two test statistics are needed, the score statistic Z and Fisher's information V. The parameter of interest θ is standardized such that under the null hypothesis it always equals 0. An example is the logarithm of the odds ratio for the interaction. When the information of the genetic and environmental factors for a new case-control set is obtained, new values for Z and V can be calculated. This is called a continuous sequential analysis. Z and V can also be calculated when information on more than one matched case-control set becomes available, a so-called group sequential analysis. Based on Z and V a decision with respect to H_0 is made. Therefore cumulative (Z,V)-values are plotted in a graph forming a pathway. Critical boundaries that enable to make the decision are fixed beforehand. These critical boundaries depend on the (two-sided) type I error 2α, the type II error β and the parameter θ_R. θ_R is the expected value of the standardized parameter under the alternative hypothesis H_1. Each new (Z,V)-point is compared to the predefined boundaries. This leads to one of three decisions:

i) enough evidence is obtained to reject the null hypothesis when the upper or lower boundary is crossed, no more observations are necessary;

ii) enough evidence is obtained to accept the null hypothesis when the inner wedge is reached, no more observations are necessary;

iii) more evidence is needed to come to a decision when the new point is still within the boundaries, so more observations are necessary.

We investigated the behaviour of the double Triangular Test (TT)[10] as a sequential test for GxE interaction (see Figures 1a and 1b for illustration of a double TT).

The test statistics Z and V for a sequential test on interaction in a non-hierarchical model on matched case-control data can be derived in a rather straightforward way.[7,10] This derivation is given in Appendix I.

For a sequential test on interaction in a hierarchical model on matched-case-control data the test statistics Z and V are derived by first estimating the main effects in a conditional logistic regression model without an interaction term on the available cumulative data.[10] The estimated main effects are substituted as nuisance parameters so that Z and V can subsequently be derived for the sequential test on interaction (see Appendix II).

In a matched case-control design only discordant sets are informative for the relation to test. This means, for example, that when we study the effect of exposure, only matched pairs with the case exposed and the control not, or vice versa, are informative for a possible association between an exposure factor and a disease. This aspect applies also to tests on GxE interaction (see Table 1). The fact that only discordant sets are informative influences the total number of matched sets necessary for a test on interaction. Based on the conditional logistic regression model, the probability of a discordant pair equals

$$\pi_{disc} = \frac{p_e q_a^2 (1 - p_e q_a^2)(\mathrm{OR} + 1)}{1 + p_e q_a^2 (\mathrm{OR} - 1)} \text{ (see Appendix III),}$$

where OR is the Odds Ratio for the interaction term in the model, i.e. OR=exp(β_{GE}). We assume that the odds ratios for the main effects of G and E are equal to 1.

Table 1 Gene (G) x environment (E) interaction in a matched case-control set (G=0: genetically not at risk, G=1: genetically at risk, E=0: not exposed, E=1: exposed)

		Control			
		G = 0	G = 0	G = 1	G = 1
		E = 0	E = 1	E = 0	E = 1
Case	G = 0 E = 0				@
	G = 0 E = 1				@
	G = 1 E = 0				@
	G = 1 E = 1	@	@	@	

@: case-control pairs informative for a test on interaction

6.4 Fixed sample size determination

Gauderman describes how the fixed sample size can be estimated for a test on GxE interaction in a hierarchical model.[3] His calculations are based on a likelihood ratio test statistic for a conditional logistic regression analysis of matched case-control data.

First the expected log-likelihood $\ell^1 = \mathscr{E}(\ln(L(\beta_G, \beta_E, \beta_{GE}))$, with L as defined in Equation (2), is maximized with respect to the observable phenotype and exposure data. This results in expected MLEs $\hat{\beta}_G^1$, $\hat{\beta}_E^1$, $\hat{\beta}_{GE}^1$ and an expected log-likelihood $\hat{\ell}^1$. Then, the expected log-likelihood $\ell^0 = \mathscr{E}(\ln(L(\beta_G, \beta_E)))$ under the null hypothesis (i.e. H_0: $\beta_{GE} = 0$) is maximized. This leads to expected MLEs $\hat{\beta}_G^0$, $\hat{\beta}_E^0$ and an expected log-likelihood $\hat{\ell}^0$. The likelihood ratio test statistic is defined as $\Lambda = 2(\hat{\ell}^1 - \hat{\ell}^0)$. For N matched sets NΛ is the non-centrality parameter of the χ^2-distribution under the alternative hypothesis H_1. When both the genetic factor and the environmental factor are dichotomous, the test on interaction has one degree of freedom and N can be computed as $N = (z_\alpha + z_\beta)^2 / \Lambda$ with 2α as the (two-sided) type I error, β as the type II error and z_x as the standardized normal deviate exceeded with probability x. For these calculations user-friendly software is provided.[11]

We followed the same approach for a non-hierarchical model. Now the expected log-likelihoods $\ell^1 = \mathscr{E}(\ln(\mathrm{L}(\beta_{GE})))$ and $\ell^0 = \mathscr{E}(\ln(\mathrm{L}(\emptyset)))$ (the 'null' model), with L as defined in Equation (1), are maximized leading to the MLE $\hat{\beta}^1_{GE}$ and expected log-likelihoods $\hat{\ell}^1$ and $\hat{\ell}^0$. Again the test on interaction has one degree of freedom and N can be calculated as before.

The necessary fixed sample size for a test on interaction in a non-hierarchical model can also be estimated using Whitehead's test statistic V (Fisher's information), $V = (z_\alpha + z_\beta)^2 / (\beta_{GE})^2$ with $\beta_{GE} = \theta_R = \ln(\mathrm{OR})$ under H_1.[10] The number of discordant matched case-control sets is equal to $N_{\text{disc}} = 4V$ (see Appendix I). The total number of matched case-control sets can be estimated by $N_{\text{tot}} = 4V / \pi_{\text{disc}}$ with π_{disc} as the probability of a discordant matched set.

6.5 Sample size determination for a sequential test

When a sequential test is used, sample size is a stochastic variable and therefore cannot be determined beforehand. Only an average or median estimate or other characteristics of its distribution can be given. This estimate can be derived by multiplying the average or median value for V by 4 and dividing it by the probability of a discordant set. (The computer program PEST version 4 provides the average and median values for V).[12] A GxE interaction can be tested continuously or group sequentially. In genetic laboratory determinations a 96-wells plate is common for PCR-based genotyping. For a GxE interaction 2 wells are required for each matched case-control set, 1 to determine the phenotype of the case and 1 to determine the phenotype of the control, leading to a group size of, e.g., 44, when space is also reserved for blank and control samples. A group sequential analysis will in general be less efficient in terms of sample size than a continuous sequential analysis. However, it may be more efficient in terms of cost or logistic (laboratory) considerations.

For a 'rare' gene a group size of 44 often contains no information to be able to estimate the main effect of G in a hierarchical model. In these situations the group size was increased to 3*44.

6.6 Data generation and simulations

To investigate the performance of the test statistics simulation studies were carried out both under H_0: OR = 1.0 and under H_1 with an OR of 1.5, 2.0 and 3.0, where OR is the Odds Ratio for the interaction term in the model. The odds ratios for the main effects G and E are assumed to be equal to 1. The two-sided type I error 2α was set equal to 0.05, the power $1\text{-}\beta$ to 0.80 or 0.90, $p_e = P(E{=}1)$ to 0.25, and $q_a^2 = P(G{=}1)$ to 0.01 for a 'rare' phenotype or 0.40 for a 'common' phenotype. For each combination of an OR and P(G=1), a large population

Table 2a Results of 2500 simulations of sequential analyses on non-hierarchical interactions using a double triangular test with P(G=1) = 0.40 and P(E=1) = 0.25

Notation:
G fixed: fixed sample size estimate according to Gauderman
W fixed: fixed sample size estimate according to Whitehead
Nav: average number of case-control sets
Nmed: median number of case-control sets
N90: 90th percentile of the number of case-control sets
Vav: average value for V
Vmed: median value for V
V90: 90th percentile value for V
rej. fr.: fraction of simulations that rejected H_0

OR = 1.5

	2α	1-β	G fixed	W fixed		Nav	Nmed	N90	Vav	Vmed	V90	rej. fr.
H_0	0.05	0.80			cont.seq. grp seq. *theor.*	698.2 713.8 *707.6*	656 660 *662*	970 968 *975*	31.50 32.17 *31.84*	29.50 30.25 *29.80*	43.50 44.00 *43.89*	0.040 0.046
		0.90			cont.seq. grp seq. *theor.*	948.9 960.3 *947.8*	884 924 *887*	1315 1364 *1306*	42.67 43.21 *42.65*	39.75 40.25 *39.91*	58.75 59.75 *58.78*	0.046 0.056
H_1	0.05	0.80	910	891	cont.seq. grp seq. *theor.*	595.3 599.8 *598.2*	568 572 *572*	916 924 *926*	31.67 32.67 *32.05*	30.50 31.00 *30.67*	48.00 50.03 *49.61*	0.806 0.823
		0.90	1218	1193	cont.seq. grp seq. *theor.*	737.1 722.2 *724.8*	692 660 *681*	1180 1144 *1154*	39.23 39.32 *38.83*	36.75 36.75 *36.46*	62.00 62.25 *61.82*	0.900 0.901

Table 2a (continued)

OR = 2.0

	2α	$1\text{-}\beta$	G fixed	W fixed		Nav	Nmed	N90	Vav	Vmed	V90	rej. fr.
H_0	0.05	0.80			cont.seq.	242.0	229	338	10.91	10.25	14.75	0.050
					grp seq.	255.4	220	352	11.55	11.00	15.75	0.044
					theor.	*242.2*	*227*	*334*	*10.90*	*10.20*	*15.02*	
		0.90			cont.seq.	325.2	307	457	14.62	13.75	20.25	0.046
					grp seq.	339.8	308	484	15.24	14.25	20.75	0.048
					theor.	*324.2*	*304*	*447*	*14.59*	*13.66*	*20.11*	
H_1	0.05	0.80	282	266	cont.seq.	184.5	175	284	11.32	10.75	17.00	0.802
					grp seq.	193.2	176	308	11.97	11.50	18.00	0.820
					theor.	*178.8*	*171*	*277*	*10.97*	*10.50*	*16.98*	
		0.90	378	357	cont.seq.	222.7	211	350	13.68	13.00	21.25	0.909
					grp seq.	293.1	220	352	14.70	14.00	22.50	0.907
					theor.	*216.6*	*203*	*345*	*13.29*	*12.48*	*21.15*	

OR = 3.0

	2α	$1\text{-}\beta$	G fixed	W fixed		Nav	Nmed	N90	Vav	Vmed	V90	rej. fr.
H_0	0.05	0.80			cont.seq.	97.8	92	143	4.43	4.00	6.00	0.048
					grp seq.	112.7	88	176	5.09	5.00	6.75	0.053
					theor.	*96.4*	*90*	*133*	*4.34*	*4.06*	*5.98*	
		0.90			cont.seq.	129.5	122	189	5.79	5.50	7.75	0.046
					grp seq.	144.4	132	220	6.49	6.25	9.00	0.050
					theor.	*129.1*	*121*	*178*	*5.81*	*5.44*	*8.01*	
H_1	0.05	0.80	100	87	cont.seq.	63.9	61	99	4.49	4.75	7.25	0.756
					grp seq.	73.6	88	88	5.60	5.75	8.00	0.774
					theor.	*58.3*	*56*	*90*	*4.37*	*4.18*	*6.76*	
		0.90	134	116	cont.seq.	79.7	76	124	5.96	5.75	8.75	0.858
					grp seq.	89.9	88	132	6.81	6.75	10.25	0.882
					theor.	*70.5*	*66*	*112*	*5.29*	*4.97*	*8.42*	

Table 2b Results of 2500 simulations of sequential analyses on non-hierarchical interactions using a double triangular test with $P(G=1) = 0.01$, $P(E=1) = 0.25$

Notation:
G fixed: fixed sample size estimate according to Gauderman
W fixed: fixed sample size estimate according to Whitehead
Nav: average number of case-control sets
Nmed: median number of case-control sets
N90: 90th percentile of the number of case-control sets
Vav: average value for V
Vmed: median value for V
V90: 90th percentile value for V
rej. fr.: fraction of simulations that rejected H_0

OR = 1.5

	2α	1-β	G fixed	W fixed		Nav	Nmed	N90	Vav	Vmed	V90	rej. fr.
H_0	0.05	0.80			cont.seq.	27034.2	25564	37799	31.97	30.25	44.75	0.060
					grp seq.	26392.1	24684	37536	32.20	30.00	45.00	0.062
					theor.	*25472.0*	*23840*	*35112*	*31.84*	*29.80*	*43.89*	
		0.90			cont.seq.	36413.8	34522	50558	42.99	40.50	59.28	0.048
					grp seq.	35158.2	32846	49148	42.79	49.75	60.00	0.064
					theor.	*34120.0*	*31928*	*47024*	*42.65*	*39.91*	*58.78*	
H_1	0.05	0.80	31312	30679	cont.seq.	19983.3	18950	31333	31.46	29.75	49.25	0.852
					grp seq.	21614.6	20856	32736	33.64	32.50	50.75	0.748
					theor.	*20677.4*	*19787*	*32006*	*32.05*	*30.67*	*49.61*	
		0.90	41918	41070	cont.seq.	23771.7	22095	37519	37.41	35.25	58.30	0.942
					grp seq.	27208.9	25828	42284	42.54	40.50	65.75	0.846
					theor.	*25051.6*	*23523*	*39884*	*38.83*	*36.46*	*61.82*	

Table 2b (continued)

OR = 2.0

	2α	1-β	G fixed	W fixed		Nav	Nmed	N90	Vav	Vmed	V90	rej. fr.
H_0	0.05	0.80			cont.seq.	9303.1	8833	13152	10.98	10.25	15.00	0.058
					grp seq.	8776.5	8338	12408	11.01	10.25	15.00	0.046
					theor.	*8720.0*	*8160*	*12016*	*10.90*	*10.20*	*15.02*	
		0.90			cont.seq.	12485.8	11838	17584	14.70	13.75	20.25	0.052
					grp seq.	11689.6	11000	16192	14.58	13.38	20.25	0.050
					theor.	*11672.0*	*10928*	*16088*	*14.59*	*13.66*	*20.11*	
H_1	0.05	0.80	9289	8759	cont.seq.	5908.6	5655	9215	11.08	10.25	17.00	0.839
					grp seq.	6377.6	6116	9724	11.55	11.00	17.00	0.782
					theor.	*5850.7*	*5600*	*9056*	*10.97*	*10.50*	*16.98*	
		0.90	12436	11726	cont.seq.	7065.8	6662	11113	13.21	12.25	20.75	0.932
					grp seq.	7702.5	7260	11968	13.93	13.25	21.53	0.886
					theor.	*7088.0*	*6656*	*11280*	*13.29*	*12.48*	*21.15*	

OR = 3.0

	2α	1-β	G fixed	W fixed		Nav	Nmed	N90	Vav	Vmed	V90	rej. fr.
H_0	0.05	0.80			cont.seq.	3741.1	3573	5426	4.40	4.00	6.00	0.047
					grp seq.	3524.3	3256	5324	4.42	4.00	6.25	0.054
					theor.	*3472.0*	*3248*	*4784*	*4.34*	*4.06*	*5.98*	
		0.90			cont.seq.	4838.0	4586	7020	5.75	5.50	7.50	0.052
					grp seq.	4616.7	4400	6644	5.84	5.50	8.00	0.054
					theor.	*4648.0*	*4352*	*6408*	*5.81*	*5.44*	*8.01*	
H_1	0.05	0.80	3024	2622	cont.seq.	1860.9	1758	2961	4.62	4.25	7.25	0.818
					grp seq.	1913.8	1804	2992	4.75	4.75	7.00	0.802
					theor.	*1765.7*	*1689*	*2731*	*4.37*	*4.18*	*6.76*	
		0.90	4048	3510	cont.seq.	2272.7	2120	3589	5.58	5.00	8.75	0.919
					grp seq.	2337.6	2200	3652	5.80	5.50	8.75	0.901
					theor.	*2137.4*	*2008*	*3402*	*5.29*	*4.97*	*8.42*	

Table 3a Results of 2500 simulations of sequential analyses on hierarchical interactions using a double triangular test with $P(G=1) = 0.40$, $P(E=1) = 0.25$

Notation: G fixed: fixed sample size estimate according to Gauderman
Nav: average number of case-control sets
Nmed: median number of case-control sets
N90: 90th percentile of the number of case-control sets
Vav: average value for V
Vmed: median value for V
V90: 90th percentile value for V
rej. fr.: fraction of simulations that rejected H_0

OR = 1.5

	2α	1-β	G fixed		Nav	Nmed	N90	Vav	Vmed	V90	rej. fr.
H_0	0.05	0.80		grp seq. *Theor.*	1431.4	1364	1980	32.00 *31.84*	30.05 *29.80*	44.28 *43.89*	0.052
		0.90		Grp seq. *Theor.*	1933.6	1804	2684	43.29 *42.65*	40.30 *39.91*	60.01 *58.78*	0.061
H_1	0.05	0.80	2007	Grp seq. *Theor.*	1283.8	1232	2024	30.68 *32.05*	29.21 *30.67*	48.15 *49.61*	0.868
		0.90	2687	Grp seq. *Theor.*	1505.9	1408	2376	36.02 *38.83*	33.77 *36.46*	57.58 *61.82*	0.945

Table 3a (continued)

OR = 2.0

	2α	1-β	G fixed		Nav	Nmed	N90	Vav	Vmed	V90	rej. fr.
H_0	0.05	0.80		grp seq. *theor.*	496.8	484	660	11.10 *10.90*	10.40 *10.20*	15.24 *15.02*	0.053
		0.90		grp seq. *theor.*	666.8	616	924	14.92 *14.59*	14.04 *13.66*	20.23 *20.11*	0.054
H_1	0.05	0.80	674	grp seq. *theor.*	461.0	440	704	11.29 *10.97*	10.81 *10.50*	17.33 *16.98*	0.802
		0.90	902	grp seq. *theor.*	538.7	484	880	13.20 *13.29*	12.45 *12.48*	20.94 *21.15*	0.912

OR = 3.0

	2α	1-β	G fixed		Nav	Nmed	N90	Vav	Vmed	V90	rej. fr.
H_0	0.05	0.80		grp seq. *theor.*	213.0	220	308	4.66 *4.34*	4.44 *4.06*	6.31 *5.98*	0.051
		0.90		grp seq. *theor.*	277.7	264	396	6.11 *5.81*	5.72 *5.44*	8.40 *8.01*	0.045
H_1	0.05	0.80	270	grp seq. *theor.*	191.5	176	308	4.69 *4.37*	4.48 *4.18*	7.12 *6.76*	0.815
		0.90	361	grp seq. *theor.*	230.3	220	352	5.66 *5.29*	5.35 *4.97*	8.89 *8.42*	0.908

Table 3b Results of 2500 simulations of sequential analyses on hierarchical interactions using a double triangular test with $P(G=1) = 0.01$, $P(E=1) = 0.25$

Notation: G fixed: fixed sample size estimate according to Gauderman
Nav: average number of case-control sets
Nmed: median number of case-control sets
N90: 90th percentile of the number of case-control sets
Vav: average value for V
Vmed: median value for V
V90: 90th percentile value for V
rej. fr.: fraction of simulations that rejected H_0

OR = 1.5

	2α	$1-\beta$	G fixed		Nav	Nmed	N90	Vav	Vmed	V90	rej. fr.
H_0	0.05	0.80		grp seq. *theor.*	36336.5	34320	51084	32.80 *31.84*	30.89 *29.80*	46.08 *43.89*	0.066
		0.90		grp seq. *theor.*	48906.2	45672	69300	44.24 *42.65*	41.34 *39.91*	62.48 *58.78*	0.074
H_1	0.05	0.80	44614	grp seq. *theor.*	28704.1	27324	44484	30.97 *32.05*	29.22 *30.67*	48.14 *49.61*	0.888
		0.90	59726	grp seq. *theor.*	33674.7	32076	52549	36.35 *38.83*	34.65 *36.46*	56.48 *61.82*	0.968

Table 3b (continued)

OR = 2.0

	2α	1-β	G fixed		Nav	Nmed	N90	Vav	Vmed	V90	rej. fr.
H_0	0.05	0.80		grp seq. *theor.*	12380.0	11748	17292	11.14 *10.90*	10.50 *10.20*	15.33 *15.02*	0.057
		0.90		grp seq. *theor.*	16478.5	15444	22968	14.86 *14.59*	13.83 *13.66*	20.64 *20.11*	0.064
H_1	0.05	0.80	13942	grp seq. *theor.*	8513.5	8052	13332	10.13 *10.97*	9.59 *10.50*	15.89 *16.98*	0.915
		0.90	18664	grp seq. *theor.*	11545.6	11088	17952	14.07 *13.29*	13.48 *12.48*	21.87 *21.15*	0.887

OR = 3.0

	2α	1-β	G fixed		Nav	Nmed	N90	Vav	Vmed	V90	rej. fr.
H_0	0.05	0.80		grp seq. *theor.*	4869.7	4620	6864	4.38 *4.34*	4.09 *4.06*	5.99 *5.98*	0.039
		0.90		grp seq. *theor.*	6487.0	6204	8976	5.88 *5.81*	5.53 *5.44*	8.10 *8.01*	0.043
H_1	0.05	0.80	4965	grp seq. *theor.*	3355.4	3168	5148	4.46 *4.37*	4.28 *4.18*	6.86 *6.76*	0.850
		0.90	6646	grp seq. *theor.*	3914.3	3564	6204	5.23 *5.29*	4.85 *4.97*	8.30 *8.42*	0.935

of matched cases and controls was simulated. The probability of occurrence of the several combinations of phenotype and exposure for both cases and controls was based on the law of Hardy-Weinberg and the conditional logistic regression model.

For each combination of the power, the OR and P(G=1) 2500 simulations were run on random samples from the corresponding large population both under H_0 and under H_1. Every simulation run resulted in an average, median and 90th percentile for the information statistic V and for the total number of case-control sets N. The fraction of simulations that resulted in the rejection of H_0 is an estimate of the type I error and the power of the test.

6.7 Results of simulations

For the non-hierarchical interaction model both continuous and group sequential analyses were simulated (Table 2a and 2b). For data with a 'common' phenotype (P(G=1)=0.40) and an OR equal to 1.5 or 2.0 the simulation results agreed very well with the theoretical values for 2α and $1\text{-}\beta$. For an OR equal to 3.0 resulting significance levels were about 0.05, but power values were lower than the theoretical values. The median gain in number of matched case-control sets necessary for a sequential test compared to Gauderman's fixed sample size estimate was about 25%. For data with a 'rare' phenotype (P(G=1)=0.01) power values were mostly larger than their theoretical values, especially for an OR = 1.5. Median efficiency gain was about 22%.

For the hierarchical interaction model only group sequential analyses were simulated (Table 3a and 3b). For the data with a 'common' phenotype simulation results for all ORs resembled their theoretical values. Only for an OR = 1.5 power values were larger than the theoretical values. Median efficiency gain as compared to Gauderman's fixed sample size estimate was about 34%. For data with a 'rare' phenotype power values were larger than theoretical values, but significance levels were about the theoretical value of 0.05. Median efficiency gain was about 30%.

In general, for data generated under H_0 the median efficiency gain was smaller than that under H_1.

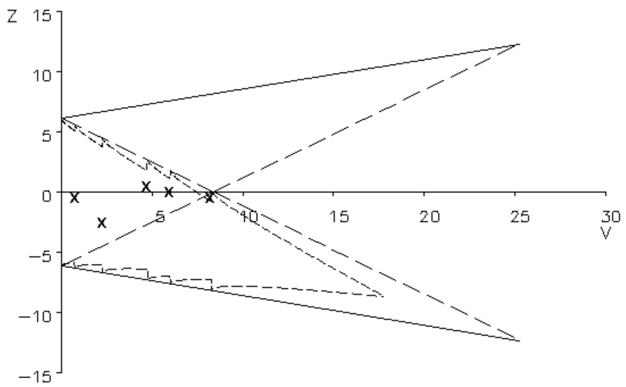
6.8 An example

Breast cancer is caused by genetic factors, environmental factors or a combination of these two in most cases. Van der Hel (submitted) investigated the combined effects of smoking and genetic polymorphisms in relevant metabolic genes. She also looked at the cumulative effect of putative at risk phenotypes on breast cancer risk. N-acetyltransferase 1 and 2 (NAT1, NAT2), glutathione S-Transferase M1 (GSTM1) and T1 (GSTT1) are enzymes, involved in carcinogen metabolisms. The genes coding for the NAT enzymes contain polymorphic sites, which cause variable enzymatic activity. GSTM1 or GSTT1 null phenotype results in a complete lack of enzymatic activity.

Follow-up of a population-based screening program for early detection of breast cancer in the Netherlands (DOM) revealed 942 women with incident breast cancer. One thousand control women were randomly selected from the DOM-cohorts. As environmental factor the smoking status of the women was assessed at baseline by a self-administered questionnaire. Women were classified as 'never' smokers or 'ever' smokers. The probability of being an 'ever' smoker was estimated as 0.30. When a woman had three or four putative phenotypes at risk she was considered 'susceptible', otherwise she was not. The probability to be 'susceptible' was estimated as 0.30 (Van der Hel submitted).

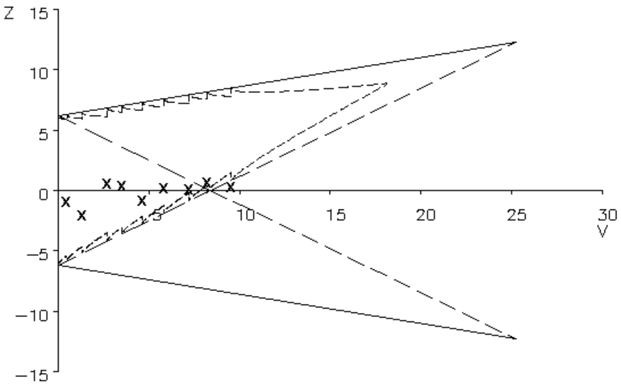
A total of 579 cases and controls could be matched on age. To detect an OR equal to 2 with a hierarchical test on interaction, a two-sided significance level $2\alpha = 0.05$ and a power $1\text{-}\beta = 0.80$ a fixed sample size of at least 674 matched sets (according to Gauderman) would be required. For a non-hierarchical test on interaction with the same specifications at least 290 (according to Whitehead) to 308 matched sets (according to Gauderman) would be required. Cumulative data on cases and controls were analysed in the chronological order the cases became apparent. Data were analysed sequentially with group sizes of 44 case-control sets.

The non-hierarchical test on interaction led to the acceptance of the null hypothesis (i.e. no interaction between G and E) after 5 groups of data (= 220 case-control sets of which 33 informative sets) were analysed (Figure 1a). Z was equal to –0.5 and V was equal to 8.25. The median unbiased estimate for OR was 0.92 (95% C.I. 0.46 ; 1.84).



STOP the study – a boundary has been crossed

Figure 1a
Results of a sequential test on non-hierarchical interaction using a double triangular test with $2\alpha = 0.05$ and $1\text{-}\beta = 0.80$ to detect an OR = 2



STOP the study – a boundary has been crossed

1b
Results of a sequential test on hierarchical interaction using a double triangular test with $2\alpha = 0.05$ and $1\text{-}\beta = 0.80$ to detect an OR = 2

Thus an efficiency gain of about 29% with respect to the fixed sample size was reached by using a group sequential test.

The hierarchical test on interaction also concluded that there was no evidence for an interaction between G and E after the analysis of 9 groups of data (= 396 case-control sets of which 73 were informative) (Figure 1b). Z was equal to 0.24 and V was equal to 9.56. The median unbiased estimate for OR was 1.11 (95% C.I. 0.57 ; 2.19). The main effects were estimated as OR_G = 1.173 and OR_E = 1.167. Here an efficiency gain of 41% was reached by using a group sequential test instead of a fixed sample test.

6.9 Discussion

The sample size necessary to detect a significant GxE interaction in an observational study can be (very) large. If an increased risk does exist for the combination of a genetic and an environmental factor, its size is more likely to be moderate (OR of 1.2 to 1.6) than large, thus requiring a large number of observations.[4] As an alternative to a large study, many small studies can be pooled. But pooling of small studies may be hampered by publication bias (i.e. positive findings are more likely to be published than negative findings).[4] Therefore a large, conclusive study to detect a moderate interaction would be preferred to a meta-analysis of small studies.

For reasons of cost-effectiveness and efficient use of available resources, like biological samples, we investigated the properties of sequential designs in matched case-control studies to test for interaction. Matched study designs already require smaller sample sizes than unmatched designs. Sequential tests require, on average, a smaller sample size than their fixed sample size counterparts.

We derived the test statistics for sequential tests on hierarchical and non-hierarchical interactions in matched case-control studies. For non-hierarchical interactions we compared results of continuous and group sequential tests. The continuous sequential analyses reflect the theoretical properties of the tests, while the group sequential tests reflect more the way laboratory analyses will be performed in practice.

For the non-hierarchical models we estimated fixed sample sizes according to Gauderman and by using Whitehead's expression for V. Differences between the two estimates arise because Gauderman bases his calculations on the likelihood ratio test, while Whitehead uses the score test. For the hierarchical models we could only estimate sample size following Gauderman's formula. Efficiency gains for the sequential tests were calculated with respect to Gauderman's fixed sample size estimate.

Results of our simulations for sequential tests showed a good agreement with theoretical values for both types of interaction when a 'common' phenotype was assumed. Efficiency gains ranged from 19 to 48% for ORs less than or equal to 2. Only for very small studies (OR = 3) the gain was obviously less. For the 'rare' phenotype the gain in efficiency was largest for data generated under the alternative hypothesis. The probability of a discordant, and thus informative, pair depends on the OR and, especially when the phenotype is ‘rare’, this leads to smaller values for π_{disc} under H_0 (when OR=1) than under H_1 (when OR>1). This results sometimes in median estimates of the number of matched sets necessary that are only slightly smaller or even somewhat larger than the fixed sample size estimate. Because π_{disc} is smaller under H_0 than under H_1, more case-control sets are needed under H_0 than under H_1 to get the same amount of information V.

In general, the larger studies in terms of sample size led to power values that were larger than their theoretical values. Resulting significance levels agreed well with their theoretical values.

The ability of a sequential test to accept the null hypothesis was illustrated by the example of the breast cancer data. A possible interaction with an OR=2 between genetic susceptibility and smoking was tested non-hierarchically and hierarchically. Both sequential tests accepted the null hypothesis without using all the available data. Gains of 29% and 41% in the number of matched sets (as compared to the fixed sample size) necessary to come to a decision were reached for the non-hierarchical and hierarchical test, respectively.

Case-only designs are mentioned as alternative to matched case-control designs.[2,13,14] Gauderman shows that case-only designs can be more efficient than matched case-control designs to study (gene x gene) interaction.[14] Case-only designs require no selection of controls, but they are only useful to test an interaction in the cases. They depend strongly on the assumption that the genetic and the environmental factor are independent in the large population. If that association has still to be examined, a (matched) case-control study yields more information. If genetic and environmental main effects are also of interest or have to be adjusted for, case-only designs are no option. Further work will show efficiency gains of the use of sequential designs in case-only studies and compare these efficiency gains to those from non-hierarchical models in matched case-control studies.

When biological samples are scarce or laboratory examinations are costly savings in samples, labour, and/or costs can be very valuable. Sequential tests can be very useful to handle the available data efficiently and can lead to considerable savings.

When biological samples for controls are abundant, but those for cases are scarce, still more efficiency can be obtained by matching more than one control to a case.[7,15]

6.10 Acknowledgements

We acknowledge the help from professor Theo Stijnen in the derivation of the test statistics for the hierarchical model and thank him for critically reading the manuscript. We thank Olga van der Hel for providing the data and the information on the research questions for the example.

6.11 References

1. Ottman R. Theoretical Epidemiology: Gene-environment interaction: Definitions and study designs. Prev Med 1996; 25: 764-70.
2. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 1994; 13: 153-62.
3. Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. Stat Med 2002; 21: 35-50.
4. Brennan P. Gene-environment interaction and aetiology of cancer: what does it mean and how can we measure it? Carcinogenesis 2002; 23: 381-7.
5. Van der Tweel I, Noord PAH van, Kaaks R. Application of a sequential t-test in a cohort nested case-control study with multiple controls per case, J Clin Epidemiol 1993; 46: 253-9.
6. Van der Tweel I, Kaaks R, Noord PAH van. Comparison of one-sample two-sided sequential t-tests for application in epidemiological studies. Stat Med 1996; 15: 2781-95.
7. Van der Tweel I, Noord PAH van. Sequential analysis of matched dichotomous data from prospective case-control studies. Stat Med 2000; 19: 3449-64.
8. Aplenc R, Zhao H, Rebbeck TR, Propert KJ. Group sequential methods and sample size savings in biomarker-disease association studies. Genetics 2003; 163: 1215-9.
9. Schork NJ, Fallin D, Tiwari HK, Schork MA. Handbook of Statistical Genetics, ed. DJ Balding et al. Chichester: Wiley, 2001; pp. 741-64.
10. Whitehead J. The design and analysis of sequential clinical trials, rev. 2nd ed. Chichester: John Wiley & Sons Ltd, 1997.
11. Morrison J, Gauderman WJ. Quanto, version 0.4. University of Southern California, 2002.
12. MPS Research Unit Reading. PEST4, Operating manual. Reading: University of Reading, 2000.
13. Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! Am J Epidemiol 1996; 144: 207-13.
14. Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol 2002; 155: 478-84.
15. Ury HK. Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. Biometrics 1975; 31: 643-9.

APPENDIX I A non-hierarchical interaction model

The likelihood for a non-hierarchical interaction in a conditional logistic regression model has the form

$$L(\beta_{GE}) = \prod_{i=1}^{n} \frac{e^{\beta_{GE}G_{i1}E_{i1}}}{e^{\beta_{GE}G_{i1}E_{i1}} + e^{\beta_{GE}G_{i2}E_{i2}}}$$

where G$_{i1}$ and E$_{i1}$ denote the case information for genetic and environmental factors, respectively, and G$_{i2}$ and E$_{i2}$ the control information, in the *i*th matched set.
The logarithm of L(β_{GE}) is equal to

$$\ell(\beta_{GE}) = \sum_{i=1}^{n} \{\beta_{GE}G_{i1}\, E_{i1} - \ln(e^{\beta_{GE}G_{i1}E_{i1}} + e^{\beta_{GE}G_{i2}E_{i2}})\}$$

Taking the first derivative with respect to β_{GE} leads to

$$\partial\ell \,/\, \partial\beta_{GE} = \sum_{i=1}^{n} \{G_{i1}E_{i1} - \frac{G_{i1}E_{i1}e^{\beta_{GE}G_{i1}E_{i1}} + G_{i2}E_{i2}e^{\beta_{GE}G_{i2}E_{i2}}}{e^{\beta_{GE}G_{i1}E_{i1}} + e^{\beta_{GE}G_{i2}E_{i2}}}\}$$

Substituting $\beta_{GE} = 0$ gives

$$\partial\ell \,/\, \partial\beta_{GE}(\beta_{GE} = 0) = \sum_{i=1}^{n} \{G_{i1}E_{i1} - \frac{G_{i1}E_{i1} + G_{i2}E_{i2}}{2}\} = Z$$

Note that for 1 case and 1 control in the *i*th matched set

$$Z_i = G_{i1}E_{i1} - \frac{1}{2}(G_{i1}E_{i1} + G_{i2}E_{i2}) = \frac{1}{2}(G_{i1}E_{i1} - G_{i2}E_{i2}) = \pm\, 1/2$$

The negative of the second derivative with respect to β_{GE} is equal to

$$\partial^2\ell \,/\, \partial\beta_{GE}^2 = \sum_{i=1}^{n} \{\frac{\sum_{j=1,2}(G_{ij}E_{ij})^2\, e^{\beta_{GE}G_{ij}E_{ij}} \,.\, \sum_{j=1,2} e^{\beta_{GE}G_{ij}E_{ij}} - (\sum_{j=1,2} G_{ij}E_{ij}\, e^{\beta_{GE}G_{ij}E_{ij}})^2}{(\sum_{j=1,2} e^{\beta_{GE}G_{ij}E_{ij}})^2}\}.$$

Substituting $\beta_{GE} = 0$ gives

$$V = \sum_{i=1}^{n} \{\frac{\sum_j (G_{ij}E_{ij})^2 \cdot 2 - (\sum_j G_{ij}E_{ij})^2}{4}\} =$$

$$= \frac{1}{4}\sum_{i=1}^{n}(G_{i1}E_{i1} + G_{i2}E_{i2})(2 - (G_{i1}E_{i1} + G_{i2}E_{i2})) = \mathrm{n}/4,$$

because (G$_{i1}$E$_{i1}$ = 1 and G$_{i2}$E$_{i2}$ = 0) or (G$_{i1}$E$_{i1}$ = 0 and G$_{i2}$E$_{i2}$ = 1).
For the *i*th matched set V_{i} = 1/4.

APPENDIX II A hierarchical interaction model

The likelihood for a hierarchical interaction in a conditional logistic regression model has the form

$$L(\beta_G, \beta_E, \beta_{GE}) = \prod_{i=1}^{n} \frac{e^{\beta_G G_{i1} + \beta_E E_{i1} + \beta_{GE} G_{i1} E_{i1}}}{e^{\beta_G G_{i1} + \beta_E E_{i1} + \beta_{GE} G_{i1} E_{i1}} + e^{\beta_G G_{i2} + \beta_E E_{i2} + \beta_{GE} G_{i2} E_{i2}}}$$

where G_{i1} and E_{i1} denote the case information for genetic and environmental factors, respectively, and G_{i2} and E_{i2} the control information, in the *i*th matched set.

This likelihood is identical to the likelihood of fitting a logistic regression model to a set of data with constant response, no intercept and differences between the corresponding values for case and control for G, E and GE:

$$L(\beta_G, \beta_E, \beta_{GE}) = \prod_{i=1}^{n} \frac{e^{\beta_G \Delta G_i + \beta_E \Delta E_i + \beta_{GE} \Delta GE_i}}{1 + e^{\beta_G \Delta G_i + \beta_E \Delta E_i + \beta_{GE} \Delta GE_i}}$$

where $\Delta G_i = G_{i1} - G_{i2}$ (the difference between the case value G_{i1} and the control value G_{i2}),

$\Delta E_i = E_{i1} - E_{i2}$ (the difference between the case value E_{i1} and the control value E_{i2}) and

$\Delta GE_i = G_{i1}E_{i1} - G_{i2}E_{i2}$ (the difference between the case value $G_{i1}E_{i1}$ and the control value $G_{i2}E_{i2}$).

The logarithm of $L(\beta_G , \beta_E , \beta_{GE})$ is equal to

$$\ell(\beta_G, \beta_E, \beta_{GE}) = \sum_{i=1}^{n} \{\beta_G \Delta G_i + \beta_E \Delta E_i + \beta_{GE}\ \Delta GE_i - \ln(1 + e^{\beta_G \Delta G_i + \beta_E \Delta E_i + \beta_{GE} \Delta GE_i})\}$$

The first-order derivative with respect to β_{GE} is

$$\partial\ell / \partial\beta_{GE} = \sum_{i=1}^{n} \{\Delta GE_i - \frac{\Delta GE_i e^{\beta_G \Delta G_i + \beta_E \Delta E_i + \beta_{GE} \Delta GE_i}}{1 + e^{\beta_G \Delta G_i + \beta_E \Delta E_i + \beta_{GE} \Delta GE_i}}\} \qquad (3)$$

We derived estimates for β_G and β_E by fitting the reduced model (i.e. a model without the interaction term) in a logistic regression analysis. Subsequently the null hypothesis H_0: $\beta_{GE} = 0$ can be tested in a sequential analysis.

Substituting the derived estimates b_G and b_E for β_G and β_E and $\beta_{GE} = 0$ in (3) leads to

$$Z = \sum_{i=1}^{n} \{\Delta GE_i - \frac{\Delta GE_i e^{b_G \Delta G_i + b_E \Delta E_i}}{1 + e^{b_G \Delta G_i + b_E \Delta E_i}}\} =$$

$$= \sum_{i=1}^{n} \{\Delta GE_i(1 - \frac{e^{b_G \Delta G_i + b_E \Delta E_i}}{1+e^{b_G \Delta G_i + b_E \Delta E_i}})\}.$$

Note that Z equals the sum over all non-zero values for the variable of interest (ΔGE$_i$) times the difference between the constant response (=1) and the fitted value for a logistic regression model with covariates ΔG$_i$ and ΔE$_i$, a constant response and no intercept term.

To derive Fisher's information V we have to work out the second derivatives of the log-likelhood with respect to the parameter of interest β_{GE}.
We simplify the notation by denoting η_i for exp($\beta_G\Delta$G$_i$ + $\beta_E\Delta$E$_i$ + β_{GE} ΔGE$_i$).

$$\partial^2 \ell / \partial\beta_{GE}^2 = \sum_{i=1}^{n} \{-\frac{(\Delta GE_i)^2\,\eta_i\,(1+\eta_i) - (\Delta GE_i \eta_i)^2}{(1+\eta_i)^2}\} =$$

$$= \sum_{i=1}^{n} \{(\Delta GE_i)^2 \frac{\eta_i^2 - \eta_i\,(1+\eta_i)}{(1+\eta_i)^2}\} = \sum_{i=1}^{n} \{(\Delta GE_i)^2 \frac{-\eta_i}{(1+\eta_i)^2}\} \qquad (4)$$

$$\partial^2 \ell / \partial\beta_{GE}\partial\beta_G = \sum_{i=1}^{n} \{-\frac{(\Delta GE_i)\ \Delta G_i\,\eta_i\,(1+\eta_i) - (\Delta GE_i \eta_i)\ (\Delta G_i \eta_i)}{(1+\eta_i)^2}\} =$$

$$= \sum_{i=1}^{n} \{\Delta GE_i\ \Delta G_i \frac{\eta_i^2 - \eta_i\,(1+\eta_i)}{(1+\eta_i)^2}\} = \sum_{i=1}^{n} \{\Delta GE_i\ \Delta G_i \frac{-\eta_i}{(1+\eta_i)^2}\} \qquad (5a)$$

$$\partial^2 \ell / \partial\beta_{GE}\partial\beta_E = \sum_{i=1}^{n} \{-\frac{(\Delta GE_i)\ \Delta E_i\,\eta_i\,(1+\eta_i) - (\Delta GE_i \eta_i)\ (\Delta E_i \eta_i)}{(1+\eta_i)^2}\} =$$

$$= \sum_{i=1}^{n} \{\Delta GE_i\ \Delta E_i \frac{\eta_i^2 - \eta_i\,(1+\eta_i)}{(1+\eta_i)^2}\} = \sum_{i=1}^{n} \{\Delta GE_i\ \Delta E_i \frac{-\eta_i}{(1+\eta_i)^2}\} \qquad (5b)$$

The information V is equal to $-[l_{\theta\theta}(\theta,\varphi) - \{l_{\theta\varphi}(\theta,\varphi)\}'.\{(l_{\varphi\varphi}(\theta,\varphi)\}^{-1}.\{l_{\theta\varphi}(\theta,\varphi)\}]$ [10], where $\theta = \beta_{GE}$ is the parameter of interest and $\varphi = (\beta_G, \beta_E)'$ is the vector of nuisance parameters.
Thus

$$l_{\theta\theta}(\theta,\varphi) = \partial^2 \ell / \partial\beta_{GE}^2$$

$$\{l_{\theta\varphi}(\theta,\varphi)\}' = \{\partial^2 \ell / \partial\beta_{GE}\partial\beta_G \;;\; \partial^2 \ell / \partial\beta_{GE}\partial\beta_E\}'$$

$$l_{\varphi\varphi}(\theta,\varphi) = \begin{pmatrix} \partial^2 \ell / \partial\beta_G^2 & \partial^2 \ell / \partial\beta_G\partial\beta_E \\ \partial^2 \ell / \partial\beta_E\partial\beta_G & \partial^2 \ell / \partial\beta_E^2 \end{pmatrix} \qquad (6)$$

Substituting $\theta = 0$ and $\varphi^* = (b_G, b_E)'$ into the expression for V leads for the separate parts of this equation to:
(4) this second derivative is equal to the negative of the sum over all non-zero values for ΔGE$_i$ times p$_i$(1-p$_i$) with p$_i$ = η_i / (1 + η_i);
(5a), (5b) these second derivatives are the negative of the sum over all non-zero values for ΔGE$_i$ times the covariates ΔG$_i$ and ΔE$_i$, respectively, times p$_i$(1-p$_i$);
(6) this matrix is the covariance matrix of the parameter estimates for the fit of a logistic regression model with covariates ΔG$_i$ and ΔE$_i$, a constant response and no intercept term. (See also Whitehead par. 7.6.2[10] for the derivation of *Z* and *V*).

(N.B. The same values for *Z* and *V* are derived for the conditional logistic regression model when a Cox PH regression model with strata is used to estimate the nuisance parameters.)

APPENDIX III

We define p_e = P(E=1), q_a^2 = P(G=1) and OR = exp(β_{GE}).

For the cases the following holds:

$$P_1(G=0 \wedge E=0) = (1-p_e)(1-q_a^2)e^{b_0}$$
$$P_1(G=0 \wedge E=1) = p_e(1-q_a^2)e^{b_0}$$
$$P_1(G=1 \wedge E=0) = (1-p_e)q_a^2 e^{b_0}$$
$$P_1(G=1 \wedge E=1) = p_e q_a^2 \mathrm{OR} e^{b_0}$$

with

$$e^{b_0} = \frac{1}{1+p_e q_a^2(\mathrm{OR}-1)}$$

For the controls the following holds:

$$P_2(G=0 \wedge E=0) = (1-p_e)(1-q_a^2)$$
$$P_2(G=0 \wedge E=1) = p_e(1-q_a^2)$$
$$P_2(G=1 \wedge E=0) = (1-p_e)q_a^2$$
$$P_2(G=1 \wedge E=1) = p_e q_a^2$$

resulting in

$$\begin{aligned}\pi_{disc} &= P_1(G=1 \wedge E=1)\left[P_2(G=0 \wedge E=0) + P_2(G=0 \wedge E=1) + P_2(G=1 \wedge E=0)\right] \\ &+ P_2(G=1 \wedge E=1)\left[P_1(G=0 \wedge E=0) + P_1(G=0 \wedge E=0) + P_1(G=0 \wedge E=0)\right] \\ &= \frac{p_e q_a^2(1-p_e q_a^2)(\mathrm{OR}+1)}{1+p_e q_a^2(\mathrm{OR}-1)}\end{aligned}$$

CHAPTER 7

Repeated looks at accumulating data: to correct or not to correct?

The history of the development of statistical theory has shown two main schools: the Bayesian and the frequentist school. Fundamental differences between these two schools have divided statisticians in the past into Bayesians and frequentists. In the preceding chapters of this thesis we applied sequential testing theory following the frequentist approach. Recent developments and publications have drawn attention again to the less known likelihood approach and its applications in sequential testing. In this chapter I describe the three approaches and their main features and differences. The differences culminate when sequential methodology is applied. Therefore, especially this aspect will be discussed.

7.1 The frequentist approach: Neyman-Pearson theory

Sequential testing theory was developed by Abraham Wald during the Second World War to try to minimize cost of industrial experiments. Later sequential testing was adopted in clinical trial settings for ethical reasons: we want to stop a trial early if a new drug or treatment is especially beneficial or harmful. Wald (1902-1950) and also Neyman (1894-1981) were the most influential exponents of the frequentist philosophy.[1] Central in the Neyman-Pearson theory is the *likelihood ratio* (LR), that is defined as LR = $P(x \mid \theta_1) / P(x \mid \theta_2)$, the ratio of the probability distribution or likelihood of the observed data, summarized by x, given that hypothesis H_1: $\theta = \theta_1$ is true and the probability distribution or likelihood of the observed data given hypothesis H_2: $\theta = \theta_2$ is true, where θ is the parameter of interest. If the LR is large, the observed data contain evidence favouring θ_1, if the LR is small, the observed data contain evidence favouring θ_2 and if the LR equals 1, there is no evidence for either θ_1 or θ_2. Wald's Sequential Probability Ratio Test (SPRT) is based on the Neyman-Pearson theory. It says that one should continue collecting data as long as $B < LR < A$, to stop data collection and decide for H_2 as soon as $LR \leq B$ and to stop data collection and decide for H_1 as soon as $LR \geq A$.[2] A and B are functions of the type I error α and the type II error β. Thus based on the LR, Wald's SPRT chooses between the two hypotheses using a *stopping rule* and a *decision rule*.

7.2 The frequentist approach: significance testing

While Neyman-Pearson theory is concentrated on hypothesis testing and decision-making, another frequentist approach (usually ascribed to Fisher)[3] concentrates more on significance testing using critical values or p-value procedures. A statistical test is performed under the assumption that the null hypothesis is true. Based on the observed data, one rejects or accepts the null hypothesis. There is no explicit role for an alternative hypothesis. Repeated significance testing procedures as introduced and fully explained by Armitage[4] can be viewed in this light. The statistical test is repeated after each new group of observations. To maintain an overall type I error α of, say, 0.05 and to avoid 'chance capitalization' or 'inflation of the error rate', each interim analysis is performed at a lower nominal value for α. Pocock and O'Brien and Fleming, amongst others, have developed different ways to 'divide' α for a fixed number of interim analyses.[5,6] DeMets and Lan describe how the overall α can be 'spent' more flexible according to the amount of information (time) used.[7]

Whitehead further elaborated Wald's SPRT in his 'boundaries approach'.[8] He developed continuous stopping boundaries such that the type I error is maintained and power requirements are satisfied. The use of these boundaries is a very flexible way of performing interim analyses. Test statistics are the efficient score statistic Z and Fisher's information V. Z is a cumulative measure for the effect size, V is a measure for the amount of information about the parameter θ contained in Z. The parameter θ that is to be tested can be standardized such that it is always equal to zero under the null hypothesis. Under the null hypothesis the distribution of Z is Normal with mean θV and variance V. Whitehead's approach is thus close to Fisher's: no choice between two hypotheses is made, but the null hypothesis is rejected or accepted based on the cumulative observed data. When the sequential test leads to the decision to stop further data collection, the p-value has to be adjusted for the multiple looks at the data.

7.3 The Bayesian approach

Already in the 18th century, Bayes (1701-1761) developed his theory on probability which was published (posthumously) in 1763. Bayesians express their prior knowledge, ideas, theories, … in a *prior* distribution function for the parameter of interest. Subsequently they observe data as result of an experiment. The product of the prior distribution function and the information about this parameter contained in the data and expressed in the likelihood, leads to the *posterior* distribution function. This posterior distribution can thus be viewed as an update of the prior information or the way belief is altered by data. If the two hypotheses H_1 and H_2 are to be distinguished, the posterior probability ratio can be expressed as the product of the LR and the prior probability ratio: $P(\theta_1 \mid x) / P(\theta_2 \mid x) = LR \,.\, P(\theta_1) / P(\theta_2)$.

When cumulative data from an experiment are analysed sequentially following the Bayesian approach, the posterior distribution describes the currently available information about the parameter of interest. This information can be used to decide whether to stop the experiment because enough evidence is already gathered or whether additional evidence is needed. In a Bayesian sequential setting no adjustment is necessary for interim looks at accumulating data.[8] The fact that test results following the Bayesian approach depend to a large extent on the choice of the prior distribution makes the approach less attractive.

7.4 The likelihood approach

Over decades statisticians have divided themselves into two, often controversial, groups: the frequentists and the Bayesians. The frequentists try to answer the question: 'What should I do?', while the Bayesians ask: 'What should I believe?'. Neither of these approaches explicitly answers the question: 'What do the data say?'[3,9] Perhaps a '*third way'*, the likelihood approach, deserves more attention than it has got until now. The concept of likelihood can be ascribed to Fisher (1890-1962). Fisher was against the use of prior probability distributions, but also rejected the idea that probability can only be interpreted in a long-run frequency way. For example: one can state, that, if the null hypothesis is true, the probability that we observe a specific test result, is smaller than, say, 0.05. We mean to say that, if we would repeat our experiment a very large number of times, we would observe this test result or one more extreme in less than 5% of the experiments. Fisher's ideas are formulated as

- whenever possible to get exact results we should base inference on probability statements, otherwise it should be based on the likelihood;
- the likelihood can be interpreted subjectively as a rational degree of belief, but it is weaker than probability, since it does not allow an external verification, and
- in large samples there is a strengthening of likelihood statements where it becomes possible to attach some probabilistic properties ('asymptotic approach to a higher status').[1]

Fisher's view differs, however, from the 'pure likelihood' view as supported by, amongst others, Royall[3] and Blume.[9] This 'pure likelihood' view, or 'evidentialism' as Vieland and Hodge called it[10], tries to answer the question 'What do the data say?' by the use of a methodology based only on the likelihood function. The Likelihood Principle states that the likelihood function contains all of the information in an experiment relevant for statistical inference about the parameter θ.[11] According to the Law of Likelihood, as formulated by Hacking, the observed data are evidence supporting one hypothesis *over* another hypothesis and the LR measures the strength of that evidence.[3] Note that no choice is made is for one or the other hypothesis.

7.5 (Mis)interpretation of the p-value

Controversies arise between the frequentist and the likelihood approach when it comes to statistical inference. The controversies arise because of the way p-values are used and interpreted in the frequentist approach. A p-value is the probability that the null hypothesis is rejected erroneously. It is, however, also interpreted as a measure of strength of evidence against the null hypothesis: 'the smaller the p-value, the stronger the evidence'. Several authors show that data from different experiments can have the same likelihood, but do not necessarily lead to the same p-value.[1,3,12] As an example, suppose we observe 8 successes in 10 experiments with probability of success for each experiment equal to 0.5, a typical example of a binomial experiment. The one-sided p-value corresponding to 8 or more successes in 10 experiments is equal to 0.055. If we had not planned beforehand to do exactly 10 experiments, but to continue until 2 failures (8 successes) were observed, the sampling scheme is a different one. We then would have a negative binomial experiment. In that case the one-sided p-value corresponding to 8 or more successes is equal to 0.0195. While these two experiments have the same likelihood and thus the same evidence about the parameter of interest, the p-values are different and even lead to different conclusions with regard to the rejection of the null hypothesis. The (strong) likelihood principle states that two data sets that produce proportional likelihoods should lead to identical conclusions and thus should also carry the same evidence about the parameter of interest. People are inclined to give different interpretations to the p-value depending on the sample size of the experiment. If two experiments that are identical except for their sample sizes produce results with the same p-value, these results do not represent equally strong evidence against the null hypothesis. Some statisticians will argue that the evidence is stronger in the smaller experiment, while others will state that the results of the larger experiment give stronger evidence.[3] Thus, the α-postulate as formulated by Cornfield: ‘All hypotheses rejected at the same critical level have equal amounts of evidence against them’ or in other words ‘Equal p-values represent, at least approximately, equal amounts of evidence’ is wrong.[3]

7.6 Correction of the type I error

The discrepancy between frequentist and likelihood inference culminates in the use of sequential methodology. When, for example, accumulating results of a clinical trial are monitored applying a frequentist method, at each interim analysis part of the overall α is spent. Smaller nominal values for α must be used at each interim analysis to guarantee that the overall value of α is not inflated at the end of the trial. The consequence is that an experiment cannot be extended beyond its planned sample size, because the preset level of α is already ‘spent’.[12] The decision to continue or to stop further data collection depends not only on the information obtained so far but also on a stopping rule. This stopping rule is a function of the type I error α. The type I error is not part of the likelihood approach. Here the evidence in the data is entirely independent of the type of sampling, be it sequential or fixed. It has been shown that the likelihood function in sequential experimentation ignores the stopping rule and thus that the evidence from an experiment is independent of the stopping rule.[1,13] So, in a sequential experiment multiple looks at the data do not affect the likelihood function. (Note that also Armitage remarks in passing: ‘In fact, the likelihood function is unaffected, apart from a constant multiplier, by the stopping rule under which the data were collected.’)[4]

7.7 Strength of the evidence

Pawitan, however, also emphasizes that the likelihood principle states something about evidence, but not about any particular course of action.[1] Then the question can arise when 'enough' evidence is obtained in favour of one of two hypotheses. One could argue that 'enough' evidence is a very subjective matter. Nevertheless, several likelihood supporters have searched how to quantify the amount of evidence. The LR measures the strength of the evidence. Let us denote the value of the LR by k. Royall suggested benchmark values of k=8 and k=32 to distinguish between weak, moderate or fairly strong and strong evidence.[3] If the LR $\geq k$, the data show *(fairly) strong* evidence in favour of H_1, if the LR $\leq 1/k$, the data show *(fairly) strong* evidence in favour of H_2 and if $1/k <$ LR $< k$, the data show *weak* evidence. (Of course this is just a crude categorisation of a continuous measure.) We can see what the data tell us by graphing the likelihood function and by calculating $1/k$ likelihood intervals. An $1/k$ likelihood interval encloses all values for the parameter of interest θ for which $L(\theta)/L(\hat{\theta}) \geq 1/k$, where $L(\theta)/L(\hat{\theta})$ is the normalized or standardized likelihood function and $\hat{\theta}$ is the maximum likelihood estimate (MLE) for θ. Or, in other words, it consists of all values that are 'consistent with the observed data'. Any θ within the likelihood interval is supported by the data because the best-supported value, the MLE $\hat{\theta}$, is only better supported by a factor k or less. One could notice a similarity between likelihood intervals and confidence intervals. An 1/8 likelihood interval corresponds to a 95.9%-confidence interval and an 1/32 likelihood interval corresponds to a 99.1%-confidence interval. (An 1/6.67 likelihood interval corresponds to a 95%-confidence interval.) Nevertheless, likelihood intervals should not be interpreted as identical to confidence intervals. Furthermore, a (frequentist) confidence interval depends also on the number of interim looks at the data, while likelihood intervals depend only on the data itself.

7.8 Misleading evidence

Strong evidence, however, can be misleading evidence. Observations can hold strong evidence supporting H_1 over H_2, while in fact H_2 is true. However, although evidence can be misleading, the probability of observing strong misleading evidence is small and limited by a *universal upper bound* P(LR $\geq$ k) $\leq 1/k$, when the *true* distribution of the data is according to H_2.[14] This important fact implies that it is difficult to collect, deliberately or not, strong misleading evidence.[9] As a devil's advocate, one could plan to continue sampling until enough evidence is gathered for one's favourite hypothesis although it is an erroneous one compared to the rival hypothesis. The probability that one will be successful in the end is always smaller than $1/k$ even if the number of observations is unlimited.

The probability of misleading evidence can be compared with the type I error α. Both point into the direction of H_1, when in fact H_2 is true. The type II error β can be compared to the probability of failing to find strong evidence in favour of H_1, i.e. the probability of finding only weak evidence plus the probability of finding misleading evidence in favour of H_2.

Under a sequential design the probability of observing misleading evidence is greater than that under a fixed sample size design. Although this probability increases with each look at the data, it remains bounded because the amount by which it increases converges to zero as the sample size grows.[15]

7.9 An example

As an example, let us look at a simple sequential experimental design.[11,16] A clinical trial compared remission times for two treatments for acute leukaemia: 6-mercaptopurine (treatment A) and placebo (treatment B). The original study was stopped after the analysis of 21 pairs of patients. For each pair of patients a preference was recorded for A or B according to which therapy resulted in a longer remission time. Under the null hypothesis H_0 the probability of a preference for A was 0.5, under the alternative hypothesis H_A this probability was thought equal to 0.75. A triangular test was designed with $\alpha = 0.05$ and power $1\text{-}\beta = 0.95$. In the first two columns of Table 1 the data for the

Table 1 Data for the clinical trial as described[11,16]

number of pairs n	preference	Z	V	LR
1	A	0.5	0.25	1.50
2	B	0	0.50	0.75
3	A	0.5	0.75	1.12
4	A	1.0	1.00	1.69
5	A	1.5	1.25	2.53
6	B	1.0	1.50	1.27
7	A	1.5	1.75	1.90
8	A	2.0	2.00	2.85
9	A	2.5	2.25	4.27
10	A	3.0	2.50	6.41
11	A	3.5	2.75	9.61
12	A	4.0	3.00	14.42
13	A	4.5	3.25	21.62
14	B	4.0	3.50	10.81
15	A	4.5	3.75	16.22
16	A	5.0	4.00	24.33
17	A	5.5	4.25	36.40
18	A	6.0	4.50	54.75
19	A	6.5	4.75	82.11
20	A	7.0	5.00	123.16
21	A	7.5	5.25	184.74

21 pairs of patients are shown. In the next two columns the test statistics *Z* and *V* for the sequential triangular test as described by Whitehead are given.[8] *Z* is equal to the difference of the observed total number of preferences for A and the expected number (=n*0.5); *V* is equal to n/4 (=n*0.5*(1-0.5)). In the last column the LR is given for the accumulating data, i.e. the likelihood function for the data under H_A: $\theta = 0.75$ divided by the likelihood function for the data under H_0: $\theta = 0.50$.

In Figure 1 *Z* and *V* are plotted in a triangular design. After the 16th pair the upper boundary of the triangular test was crossed, which led to the conclusion that the null hypothesis could be rejected. The 90%-confidence interval for θ is (0.59 ; 0.88).

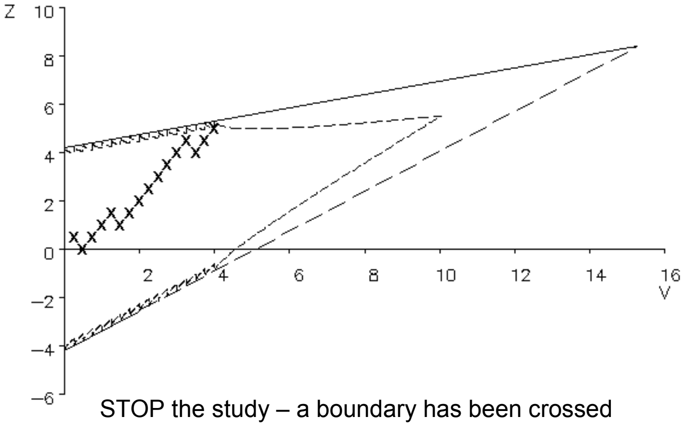


Figure 1 Results of the trial plotted as test statistics *Z* versus *V* in a triangular test

In Figure 2 the LR = P(x | θ= 0.75) / P(x | θ= 0.5), is plotted against the number of pairs. The LR was greater than 8 after 11 pairs and greater than 32 after 17 pairs of patients.

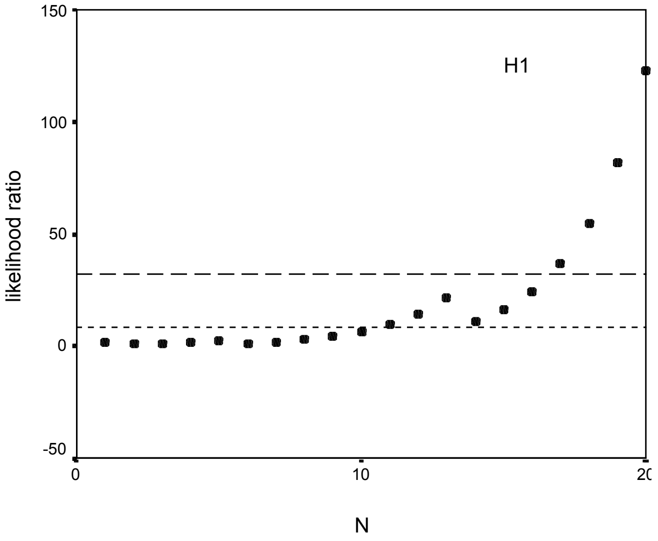


Figure 2 Results of the trial plotted as the LR versus the number of pairs N. The dotted line corresponds to a LR of 8, the dashed line to a LR of 32.

In Figure 3 the standardized likelihood function is plotted together with the 1/8 and 1/32 likelihood intervals for the data from the trial (14 preferences for A in 17 pairs of patients). The value $\theta = 0.5$ is not consistent with the data, the value $\theta = 0.75$ is included in both likelihood intervals.

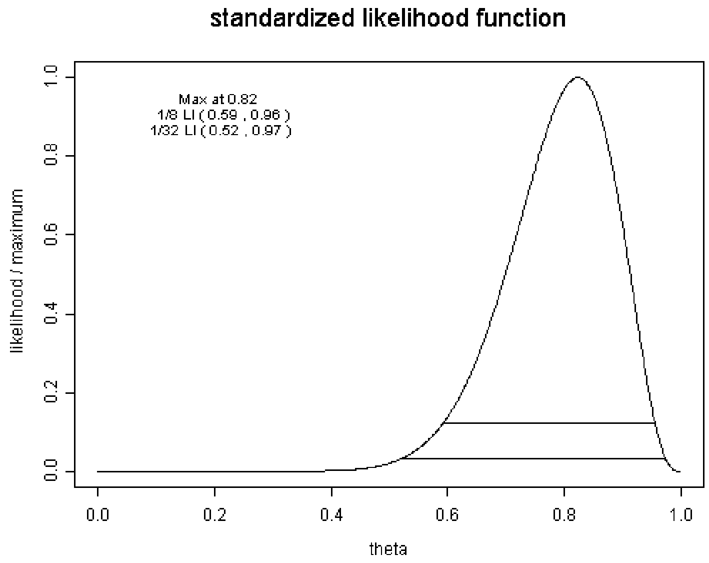


Figure 3 Results of the trial plotted as the standardized likelihood function versus θ, together with the 1/8 and 1/32 likelihood intervals (LI)

To compare results from a trial that rejected the null hypothesis with those from an experiment that led to the acceptance of the null hypothesis I simulated preference data under the null hypothesis H_0: $\theta = 0.50$. Results of the simulated data are presented in Figures 4, 5 and 6. After the 17th pair the lower boundary of the triangular test was crossed, which led to the conclusion that the null hypothesis could be accepted (Figure 4). The 90%-confidence interval for θ is (0.30 ; 0.69).

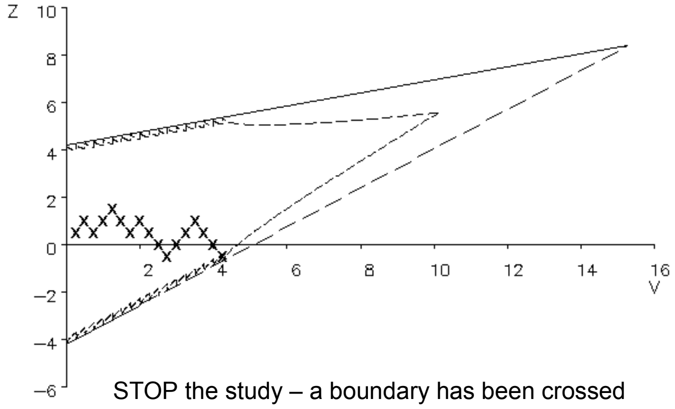


Figure 4 Results of the simulated data plotted as test statistics Z versus V in a triangular test

The LR = P(x | θ = 0.75) / P(x | θ = 0.5) was smaller than 1/8 after 16 pairs and smaller than 1/32 after 18 pairs of patients (Figure 5).

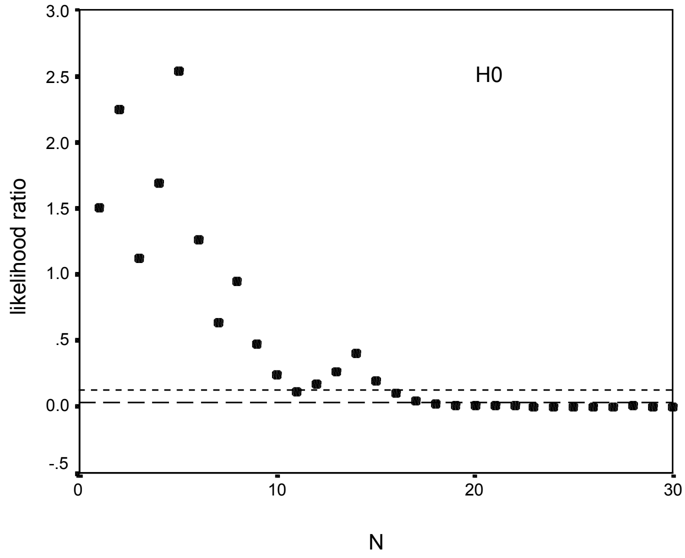


Figure 5 Results of simulated data plotted as the LR versus the number of pairs N. The dotted line corresponds to a LR of 1/8, the dashed line to a LR of 1/32.

In Figure 6 the standardized likelihood function is plotted together with the 1/8 and 1/32 likelihood intervals for the simulated data (8 preferences for A in 18 pairs of patients). The value θ = 0.5 is consistent with the data, the value θ = 0.75 is not consistent with the data.

(Note that the LR is invariant to the choice of the parameter i.e. it makes no difference whether θ is used or ψ = log(OR) = log(θ/(1-θ)))

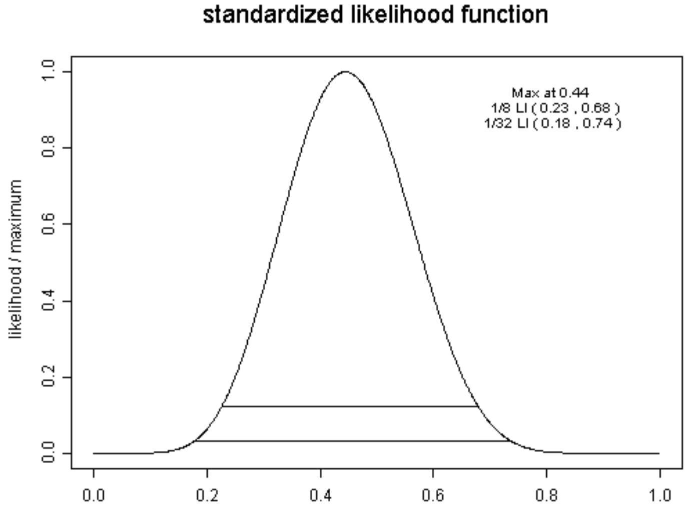


Figure 6 Results of the simulated data plotted as the standardized likelihood function versus θ, together with the 1/8 and 1/32 likelihood intervals (LI)

Blume gives an approximation to the probability P(M) that a sequential design will generate misleading evidence.[15] For Bernoulli-data, like the example above, $P(M\,|\,k=8) \approx 0.0922$, while $P(M\,|\,k=32) \approx 0.0230$. A 10,000 simulations for these data under H_0 resulted in $P(M\,|\,k=8) = 0.1002$ and $P(M\,|\,k=32) = 0.0238$; a 10,000 simulations under H_A resulted in $P(M\,|\,k=8) = 0.0874$ and $P(M\,|\,k=32) = 0.0224$. (The universal upper bound for P(M) is equal to 0.125 for $k = 8$ and equal to 0.03125 for $k = 32$.) As this probability is greater for a sequential design than for a fixed sample design, a recommendation could be to compare the LR for a sequential design with a threshold $k = 32$ and not with $k = 8$ when looking for (clear) evidence.

7.10 Summary

In a Bayesian sequential setting, as in the likelihood approach, no adjustment is necessary for repeated looks at accumulating data. The practical problem in the Bayesian approach of statistical testing lies in the choice of an appropriate prior distribution and the amount of (subjective) belief that is assigned to it. In large sample problems the data will dominate the prior distribution and thus determine the posterior distribution so that the Bayesian approach becomes the likelihood approach. This is also the case when a non-informative prior distribution is used. Furthermore, the invariance property that holds for the likelihood approach does not hold in a Bayesian setting.

The Neyman-Pearson approach indeed makes use of the LR, but its numerical value is not interpreted as a measure of the strength of evidence. Only its extremeness is compared to critical boundaries to make a decision. The p-value was added to have a measure of the strength of the evidence after all. This use of the p-value comes into conflict with the likelihood principle. (Data sets with the same (or proportional) likelihood carry the same evidence and should thus lead to the same p-value.)

The likelihood approach is a simple and elegant '*third way*' to deal with evidence in experimental data. It makes a clear distinction between the degree of uncertainty and the strength of the evidence. Other favourable qualities of the likelihood approach are:

- Two hypotheses of equal importance are compared instead of focusing on the acceptance or rejection of the null hypothesis.
- No correction for interim looks at accumulating data is necessary, so there is also no problem in extending an already obtained sample.
- The MLE can be used for the parameter of interest without adjustment, while following a frequentist sequential test it is biased.[8]
- For a sequence of observations the universal upper bound applies, i.e. the probability of finding strong misleading evidence of strength k or greater cannot exceed, and often is much less than, the value $1/k$.[14]

Of course there are topics that call for further investigation:

- The Law of Likelihood is restricted to the comparison of simple hypotheses and does not apply to most composite hypotheses[3], although Blume suggests a transformation such that the Law can be applied.[9]
- In multi-parameter models there is no general way to eliminate nuisance parameters. Royall and Blume suggest some *ad hoc* methods, of which the use of profile likelihoods looks the most satisfactory.[3,9] This has to be further investigated, especially in the context of sequential likelihood testing.
- Simulations of sequential designs using the LR will have to show their characteristics, like the efficiency, the average sample size to come to a decision, the probabilities of weak and of misleading evidence, … for different outcome variables.

7.11 Conclusion

Recent developments and publications on the likelihood approach and especially its application in sequential designs[3,9,14,15] prompted me to go into this '*third way*' and compare it with the frequentist and the Bayesian approach. Although there are still topics to investigate further before a definite recommendation can be made to turn into this way, I would like to end with the following conclusion.

Because the number of interim looks at accumulating data does not affect the LR, sequential designs based on the LR are a very natural way of monitoring the strength of evidence in the observations. A sample of observed data can be enlarged without worrying about the effect on the type I error, and thus without any adjustments. For (simple) sequential testing problems in observational, epidemiological studies on matched case-control data, where the goal is to achieve evidence for one hypothesis over another and where it is important to use up available biological specimen as efficiently as possible, the likelihood approach is an objective answer to the question: 'What do the data say?'

7.12 References

1. Pawitan Y. In all likelihood: Statistical modelling and inference using likelihood. Oxford: Clarendon Press, 2001.
2. Wetherill GB, Glazebrook KD. Sequential methods in statistics. 3rd ed. London: Chapman and Hall, 1986.
3. Royall R. Statistical evidence. A likelihood paradigm. London: Chapman & Hall, 1997.
4. Armitage P. Sequential medical trials. 2nd ed. Oxford: Blackwell Scientific Publications, 1975.
5. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977; 64:191-9.
6. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979; 35: 549-56.
7. DeMets DL, Lan KKG. Interim analysis: the alpha spending function approach. Stat Med 1994; 13: 1341-52.
8. Whitehead J. The design and analysis of sequential clinical trials, rev. 2nd ed. Chichester: John Wiley & Sons Ltd, 1997.
9. Blume JD. Likelihood methods for measuring statistical evidence. Stat Med 2002; 21: 2563-99.
10. Vieland VJ, Hodge SE. Statistical Evidence: A likelihood paradigm (Book review). Am J Hum Genet 1998; 63: 283-9.
11. Berry DA. Interim analysis in clinical trials. The role of the likelihood principle. Am Stat 1987; 41: 117-22.
12. Cornfield J. Sequential trials, sequential analysis and the likelihood principle. Am Stat 1966; 20: 18-23.
13. Anscombe FJ. Sequential medical trials. JASA 1963; 58: 365-83.
14. Royall R. On the probability of observing misleading statistical evidence (with discussion). JASA 2000; 95: 760-80.
15. Blume JD. On observing misleading evidence in sequential trials. http://alexander.stat.brown.edu/~jblume/slides/, 2003.
16. Freireich EJ, Gehan EA, Frei E, Schroeder L, et al. The effect of 6-Mercaptopurine on the duration of steroid-induced remission in acute leukemia: A model for evaluation of other potentially useful therapy. Blood 1963; 21: 699-716.

Summary

Sequential statistical tests on matched case-control data are an efficient way of handling the available biological material in an observational epidemiological study. In a cohort or a biological bank controls are mostly abundantly present, while cases are (much) more limited. One or more controls can be matched to a case to control for confounding factors thus enhancing the efficiency of the analysis.

In chapter 1 two versions of a one-sample sequential *t*-test were compared. The exact calculation of the log-likelihood ratio using Kummer's function was compared to Rushton's approximation in Wald's open Sequential Probability Ratio Test (SPRT). The average sample sizes for the two sequential *t*-tests with one control per case were already smaller than the sample size required for the corresponding non-sequential paired *t*-test. Matching multiple controls per case increased the efficiency. Simulations showed that the gain in efficiency was largest when two controls were matched to each case compared to 1 control per case, when the null hypothesis could be rejected. The exact calculation using Kummer's function was somewhat more conservative than Rushton's approximation, i.e. it showed higher power values and lower type I error values. The use of the one-sample sequential *t*-test was illustrated with data from the DOM cohort. Cases were pre-menopausal women with breast cancer. The research hypothesis was whether the selenium content in toenail clippings from these women is already decreased before tumour occurrence. One to 5 controls were matched on age to each of 64 cases. Most of the sequential tests to detect standardized differences of 0.3, 0.4 or 0.5 led to the acceptance of the null hypothesis. For some tests to detect a standardized difference of 0.3 no conclusion could be reached, because the number of available case-control sets was too small.

In chapter 2 we investigated how a standardized difference between the mean of the exposure distribution of the cases and that of the controls can be expressed as an expected odds ratio (OR). This relation was elaborated for both a one-sample (paired) and a two-sample (unpaired) sequential *t*-test. An example was given of a two-sample sequential *t*-test with 5 controls per case to test the null hypothesis of equal selenium content in toenails of cases and of controls. Cases were pre-menopausal women with breast cancer. An expected standardized difference of 0.25 between the average exposure values for cases and for unmatched controls of the selenium distribution in the cohort corresponding to a minimum expected OR of 2 between the highest and the lowest quintile of the exposure distribution was tested sequentially. After the analysis of the information of 31 cases and 155 controls the null hypothesis could be accepted. Compared to the 61 cases detected in the cohort, a saving of 49% was obtained.

In chapter 3 again two versions of a one-sample sequential *t*-test were compared, Rushton's approximation to the log-likelihood ratio and the use of a profile log-likelihood function, as described by Whitehead, in Wald's open SPRT. Simulations showed a smaller average sample size for Whitehead's procedure when the null hypothesis was rejected. Rushton's approximation led to a smaller average sample size when the null hypothesis could be accepted. Especially for larger sample sizes the power resulting from Rushton's version was too small. For small sample sizes the significance level resulting from Whitehead's version was somewhat too large. In this chapter a possible relation between the selenium content in toenail clippings and colorectal cancer in pre-menopausal women was investigated. Three controls were matched on age to each case. The null hypothesis ('no difference in selenium content between cases and controls') could be accepted after the 8th matched case-control set was analysed. Thus a saving of 75% of available case-control sets was obtained compared to the fixed sample size necessary to detect a standardized difference of 0.5.

In chapter 4 we described the sequential test we developed for dichotomous exposure variables in matched case-control data. This test relates to the (fixed sample size) McNemar test and Mantel-Haenszel test for matched data. Results from simulations using the SPRT were compared to results using a triangular test (TT) for various odds ratios, type I and type II errors. The resulting type I errors were acceptable in general. Type II errors were somewhat larger than their theoretical values for larger odds ratios, i.e. for smaller samples. In general, an SPRT requires less information to come to a decision than a TT when the null hypothesis or the alternative hypothesis is true and thus in these situations is a more efficient test. Our simulations confirmed that sequential analyses require on average fewer case-control sets than fixed sample size analyses. Savings ranged from 32% to 60%. We illustrated this sequential test investigating the relation between a mutation of the MTHFR-gene and the occurrence of rectal cancer. With one control matched by age to each of the 69 available cases no decision could be made. With three controls matched to each case the null hypothesis could be accepted after 35 discordant sets. For a fixed sample size analysis at least 66 discordant sets would have been needed to detect an OR equal to 2. Thus, using a sequential test with 3 controls matched to each case, 47% of the matched sets could be saved.

During the course of a clinical trial or a prospective epidemiological study the need can be felt to stop early for 'futility'. One of the reasons is that there seems to be no indication for a relevant effect thus far. Calculation of the conditional power (CP) is proposed and used in the literature as a decision tool for early stopping of a trial or study. In chapter 5 we discussed the disadvantages of CP. As an alternative we proposed to perform a (group) sequential test on the already available data under the same specifications (effect size, α, β) as the original trial or study design. When the (group) sequential test does not lead to a decision yet, new data can be collected and analysed until enough evidence is obtained for a decision. We illustrated the use of a sequential test instead of CP by re-analysing two examples from the literature. It turns out that a (group) sequential test is at least as and often even more efficient than the estimation of CP. Use of a (group) sequential test is a more objective strategy to decide for early stopping than calculation of CP requiring no arbitrary assumptions.

Case-control studies can be designed to study an association between a genetic risk factor or an environmental factor and the occurrence of a disease. Both genetic and environmental factors may contribute to the susceptibility of the disease and it can be interesting to explore the gene-environment (GxE) interaction. GxE interactions can be tested both hierarchically and non-hierarchically. In chapter 6 we described the properties of sequential designs in matched case-control studies to test for GxE interactions. Results of simulations showed a good agreement with theoretical values for V, the necessary amount of information, and the type I error. Power values were larger than their theoretical values for very large sample sizes. The median gain in efficiency was about 27%. For a ‘rare’ phenotype gain in efficiency was larger when the alternative hypothesis was true than under the null hypothesis. The probability of a discordant, and thus informative, set of data is always smaller under the null hypothesis than under the alternative hypothesis. We illustrated the developed sequential tests by an example using data from the DOM project again. Cases were women with incident breast cancer. Each case was matched by age to one control. A possible interaction with an OR of 2 between genetic susceptibility and smoking was tested non-hierarchically and hierarchically. Both sequential tests accepted the null hypothesis. Gains in the necessary number of matched sets of 29% and 41% were reached. So sequential tests on GxE interaction in matched case-control data also show gains in efficiency.

In the last chapter of this thesis I described two approaches to statistical testing theory, the frequentist and the Bayesian approach. The frequentist approach tries to answer the question: ‘What should I do?’; the Bayesian approach wonders: ‘What should I believe?’ The sequential tests as described in this thesis follow the frequentist approach. There are, however, some fundamental problems to the frequentist approach. One is that the type I error has to be adjusted when accumulating data are tested repeatedly. Another problem is the two roles the p-value plays. The p-value or significance level is the probability of obtaining the observed results or more extreme ones if the null hypothesis is true. One could say that the p-value is the probability of obtaining misleading evidence. The p-value is, however, often also wrongly used as a measure for the strength of the evidence against the null hypothesis. The practical problem in the Bayesian approach lies in the subjective choice of an appropriate prior distribution for the parameter of interest.

There is a 'third way' that has drawn little attention until now: the likelihood approach. This approach poses the question: 'What do the data say?' en tries to answer this question by looking at the likelihood of the data under hypothesis H_1 compared to the likelihood of the data under hypothesis H_2. The likelihood ratio measures the strength of the evidence in the observations in favour of H_1 over the strength of the evidence in favour of H_2. Contrary to the frequentist approach, the likelihood approach makes a clear distinction between a measure for the strength of the evidence and the probability of misleading evidence. Under the likelihood approach cumulative data can be tested repeatedly without adjustments. In a simple example data are analysed sequentially using both the frequentist and the likelihood approach. The likelihood approach shows clear advantages compared to the frequentist and the Bayesian approach. Nevertheless, some topics will have to be investigated further before the likelihood approach can be recommended in general as a more efficient and objective way of repeated testing of accumulating data.

In case of simple sequential tests of epidemiological, observational matched case-control data with the purpose to collect evidence for one of two hypotheses using up the available biological material as efficiently as possible, the likelihood approach can give an objective answer to the question: 'What do the data say?'

Nederlandse inleiding en samenvatting

Een epidemiologische observationele studie

Halverwege de jaren zeventig van de 20^{e} eeuw startte de toenmalige vakgroep Epidemiologie van de Rijksuniversiteit Utrecht onder leiding van prof.dr F. de Waard het DOM project. Dit project was bedoeld om vrouwen in de stad Utrecht en omstreken periodiek te screenen op borstkanker. Het DOM project is één van de vele voorbeelden van een epidemiologische, observationele studie. In een dergelijke studie worden geen behandelingen met elkaar vergeleken zoals in een klinisch vergelijkend onderzoek. In deze observationele studie werd een grote groep vrouwen op vrijwillige basis regelmatig geobserveerd om vast te kunnen stellen of op zeker moment in de tijd borstkanker ontstaat. In sommige screeningsrondes van het DOM project werd aan vrouwen gevraagd om hun urine van de voorafgaande nacht mee te willen nemen of een afgeknipte teennagel. Deze monsters werden voor latere analyse opgeslagen in een biologische bank. Combinatie van de karakteristieken van de deelneemsters aan het project met informatie van de regionale kankerregistratie leidde, na kortere of langere tijd, tot de identificatie van de *cases*, vrouwen met een ziekte, bijvoorbeeld borstkanker. Deze cases kunnen worden vergeleken met controles, deelneemsters die gedurende dezelfde studieperiode de ziekte niet ontwikkelen.

Biologisch materiaal

Voortdurend worden nieuwe laboratoriumtechnieken ontwikkeld. Hiermee kunnen een groot aantal (nieuwe) hypotheses omtrent een mogelijk verband tussen blootstelling aan een stof en een ziekte getest worden op het opgeslagen biologische materiaal van cases en controles. Als de ziekte weinig voorkomt, zullen er weinig cases zijn en is de hoeveelheid opgeslagen biologisch materiaal beperkt. Als biologisch materiaal eenmaal verwerkt is in laboratoriumonderzoek kan het, in het algemeen, niet meer opnieuw gebruikt worden. Om het grote aantal interessante hypotheses te combineren met de beperkte hoeveelheid biologisch materiaal zijn statistische methoden nodig die veelbelovende hypotheses kunnen onderscheiden van minder veelbelovende, daarbij gebruik makend van zo weinig mogelijk biologisch materiaal.

Sequentiële analyse

Halverwege de vorige eeuw werd een nieuwe statistische techniek ontwikkeld: de sequentiële analyse. Deze werd vooral om economische redenen toegepast in de industriële kwaliteitscontrole. Om ethische redenen wordt sequentiële analyse de laatste jaren steeds meer toegepast in klinisch onderzoek ter vergelijking van medicijnen of behandelingen. Sequentiële methoden bieden een onderzoeker de mogelijkheid om een onderzoek te beëindigen zodra voldoende ‘bewijs’ is vergaard om de nulhypothese (‘er is geen verband tussen blootstelling en ziekte') te kunnen aannemen of te kunnen verwerpen ten gunste van de alternatieve hypothese ('er is wel een verband tussen blootstelling en ziekte'). Na iedere nieuwe waarneming of groep waarnemingen worden de cumulatieve gegevens getest. Op basis van deze cumulatieve resultaten kan besloten worden om de studie te stoppen of om meer gegevens te vergaren en te testen. Als na iedere nieuwe waarneming wordt getest spreken we over een continue sequentiële analyse; als na een groep van nieuwe waarnemingen wordt getest spreken we over een groepssequentiële analyse. Een sequentiële analyse heeft, gemiddeld genomen, minder waarnemingen nodig om tot een beslissing te komen dan een vergelijkbare studie met een vooraf vastgestelde grootte. Sequentiële methoden zijn dus bruikbaar om efficiënter met beschikbare gegevens om te gaan.

Matchen

Soms zijn er factoren die zowel met de blootstelling aan een stof als met de ziekte samenhangen. Dergelijke factoren kunnen de mogelijke relatie tussen de blootstelling en de ziekte verstoren. Voorbeelden van mogelijk verstorende factoren zijn leeftijd, etniciteit en (bij vrouwen) menopauzale status. *Matchen* is een techniek die wordt gebruikt om te corrigeren voor dergelijke verstorende factoren. Op basis van de waarde(n) van de verstorende factor(en) kan iedere case gematcht worden aan één of meer controles. Als een ziekte vrij veel voorkomt en er dus in de loop van de studie veel cases bekend worden, is het voldoende om één case aan één controle te matchen. Als een ziekte echter zeldzaam is, zijn er weinig cases, maar daarentegen veel controles beschikbaar. Om toch in dergelijke situaties voldoende zeggingskracht te hebben, kan meer dan één controle aan een case worden gematcht. Voor een proefopzet waarbij gebruik gemaakt wordt van matchen zijn over het algemeen minder cases en controles nodig dan voor een proefopzet waarin niet gematcht kan worden. Matchen is een manier om efficiënter om te gaan met de beschikbare gegevens.

Omvang van een studie

Het is een kwestie van *Good Statistical Practice* om in de ontwerpfase van een vergelijkend klinisch onderzoek het benodigde aantal patiënten te bepalen, dus nog vóór de gegevens worden verzameld. We spreken dan over een studie met een van tevoren vastgestelde omvang. De omvang van een dergelijke studie wordt bepaald door de grootte van het verschil in uitkomst tussen de behandelingen dat men zou willen detecteren als het bestaat, door de gewenste type I fout van de studie en door de gewenste *power* van de studie. De type I fout van een studie is de kans op een vals-positief resultaat of de kans dat de nulhypothese ten onrechte wordt verworpen. Een studie kan bijvoorbeeld <u>ten onrechte</u> concluderen dat er een verband bestaat tussen blootstelling aan een stof en een ziekte. Deze type I fout wordt meestal op 5% gesteld. De *power* van een studie is de zeggingskracht oftewel de kans op een terecht-positief resultaat oftewel de kans dat de nulhypothese terecht wordt verworpen. In dit geval is de conclusie van een studie terecht dat er een verband bestaat tussen blootstelling en ziekte. De *power* wordt meestal op 80% of 90% gesteld.

In de meeste epidemiologische, observationele studies wordt de omvang van de studie bepaald door praktische aspecten als tijd, kosten, beschikbaarheid van personen, enz. Als een epidemiologische studie sequentieel geanalyseerd zal gaan worden, kan het aantal waarnemingen, dat nodig is om een beslissing te kunnen nemen, niet van tevoren worden vastgesteld. Wel kan, op basis van de type I fout, de *power* en het te detecteren effect, een schatting worden gemaakt van het gemiddelde of mediane aantal waarnemingen dat nodig zal zijn.

Continue en dichotome variabelen

Als de blootstellingsvariabele een continue grootheid is, dat wil zeggen binnen zekere grenzen alle mogelijke waarden kan aannemen zoals bijvoorbeeld het seleniumgehalte van afgeknipte teennagels, dan kunnen alle waarden gebruikt worden voor de analyse. Als daarentegen de blootstellingsvariabele maar twee mogelijke waarden kan aannemen (een zogenaamde dichotome variabele), zoals wél of géén genetische mutatie, dan is het mogelijk dat een set van een case en één of meer gematchte controles geen bruikbare informatie bevat. Dit kan gebeuren als zowel de case als de controle(s) allen wél blootgesteld waren aan de stof of juist allen níet blootgesteld waren, de zogenaamde concordante sets. Alleen de zogenaamde discordante case-controle sets bevatten informatie voor de statistische analyse. In een discordante set is de case wél blootgesteld aan een stof en de controle níet of omgekeerd. De totale omvang van de studie (concordante + discordante sets) zal afhangen van de kans op een discordante set. Als deze kans klein is, zal een groot aantal case-controle sets geanalyseerd moeten worden om genoeg informatie te kunnen vergaren om tot een beslissing te kunnen komen. Meer controles matchen per case, als dat mogelijk is, vergroot de kans op een discordante set. Dit is dus ook een manier om efficiënter met de beschikbare data om te gaan.

Dit proefschrift beschrijft hoe sequentiële analyse van gematchte case-controle sets op een efficiënte manier met waardevol biologisch materiaal kan omgaan, waardoor een groot aantal interessante hypotheses getoetst kan worden. De ontwikkelde sequentiële toetsen worden geïllustreerd met voorbeelden die data uit het DOM project gebruiken.

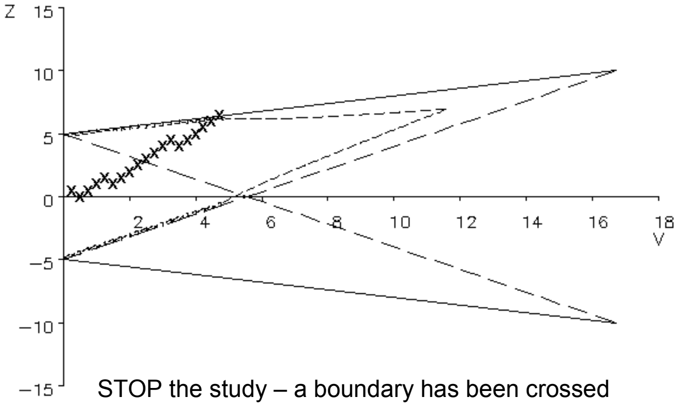
Twee eenvoudige voorbeelden

Aan de hand van twee eenvoudige, hypothetische voorbeelden wil ik het gebruik van de sequentiële analysemethode illustreren. Om het verband tussen een mogelijke genetische mutatie en het krijgen van borstkanker te onderzoeken is iedere vrouw met borstkanker (de *case*) in de chronologische volgorde waarin de kanker zich openbaarde gematcht aan een controlevrouw zonder kanker. Van alle vrouwen is bekend of ze de genetische mutatie hadden of niet. Alleen de discordante combinaties A (case met de mutatie, controle zonder de mutatie) en B (case zonder de mutatie, controle met de mutatie) bevatten informatie voor de analyse. In onderstaande tabel zijn de resultaten van de eerste 19 case-controle paren vermeld.

tabel 1 Resultaten van 19 case-controle paren in termen van combinatie A of B (zie tekst)

paar	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
combinatie	A	B	A	A	A	B	A	A	A	A	A	A	A	B	A	A	A	A	A

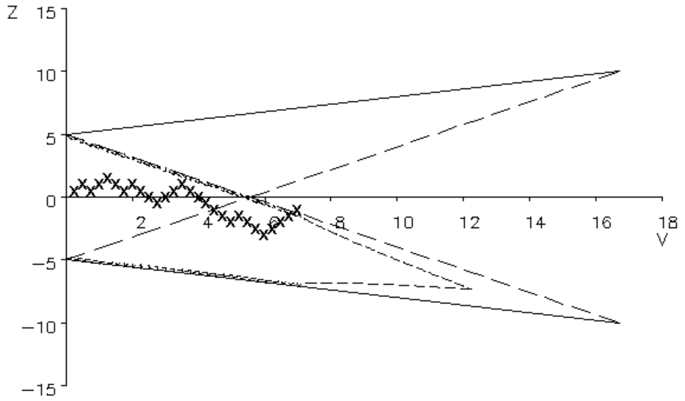
Na ieder nieuw paar worden nieuwe waarden voor grootheden *Z* en *V* berekend en uitgezet in een figuur (zie bijvoorbeeld figuur 1).



Figuur 1 Resultaten van de sequentiële analyse van de 19 case-controle paren uit tabel 1, die leiden tot de conclusie dat er een verband bestaat tussen het hebben van de genetische mutatie en het krijgen van de ziekte.

Z wordt verticaal uitgezet en is een maat voor het effect; *V* wordt horizontaal uitgezet en is een maat voor de cumulatieve hoeveelheid informatie. Als in een case-controle paar combinatie A optreedt, gaat *Z* een stapje naar boven in de figuur; als sprake is van combinatie B gaat *Z* een stapje naar beneden. *V* gaat bij ieder nieuw case-controle paar een stapje naar rechts. Bij de start van het onderzoek worden in de figuur kritieke grenslijnen getrokken, in dit voorbeeld in de vorm van twee driehoeken. Het verloop en de helling van de grenslijnen hangen af van de type I fout, de *power* en de grootte van het verwachte effect. Als het cumulatieve 'pad' van (Z,V)-waarden de bovengrens van de bovenste driehoek of de ondergrens van de onderste driehoek overschrijdt, dan kan de nulhypothese verworpen worden met als conclusie dat er een verband is tussen het hebben van de genetische mutatie en het krijgen van borstkanker. Als het cumulatieve 'pad' de binnenste wigvormige grenzen overschrijdt, dan kan de nulhypothese geaccepteerd worden. In dit voorbeeld is de nulhypothese: 'de kans op een genetische mutatie is voor cases en controles even groot, namelijk gelijk aan 0.5'. De alternatieve hypothese zou kunnen zijn: 'de kans op een genetische mutatie is voor cases gelijk aan 0.75'. Voor dit voorbeeld is V gelijk aan een kwart van het aantal (discordante) case-controle paren; Z is gelijk aan het aantal paren met combinatie A verminderd met het verwachte aantal als de nulhypothese waar is. Figuur 1 laat zien dat na de verwerking van de gegevens van 19 paren het 'pad' de bovengrens overschrijdt en de nulhypothese verworpen wordt. Van de 19 paren hadden 16 de combinatie A en maar 3 de combinatie B. Het verwachte aantal paren met combinatie A, als de nulhypothese waar is, is 19/2=9.5. Z is dan gelijk aan $16 - 9.5 = 6.5$ en $V = 19/4 = 4.75$. De conclusie is dat de kans op een genetische mutatie voor cases groter is dan voor controles en dat er dus een verband bestaat tussen het hebben van de genetische mutatie en het krijgen van borstkanker. Om dit verband te kunnen detecteren zouden ten minste 43 discordante sets nodig geweest zijn in een studie met een vaste grootte. Een sequentiële analyse levert in dit geval 56% besparing op.

Een andere set gegevens uit een vergelijkbaar onderzoek maar met een andere genetische mutatie zou tot figuur 2 kunnen leiden.



Figuur 2 Resultaten van de sequentiële analyse van een onderzoek waarin geen verband is tussen het hebben van de genetische mutatie en het krijgen van de ziekte.

Hier overschrijdt het 'pad' van cumulatieve (Z,V)-waarden één van de wigvormige binnengrenzen nadat de gegevens van 28 paren zijn verwerkt en wordt de nulhypothese geaccepteerd. In dit geval is de conclusie dat er geen verband bestaat tussen het hebben van de genetische mutatie en het krijgen van de ziekte.

Hoofdstuk 1

In hoofdstuk 1 werden een exacte en een benaderende versie voor een sequentiële toets beschreven voor de situatie dat de blootstelling aan een stof gemeten kan worden als een continue grootheid. Als voorbeeld kan hierbij het seleniumgehalte in de teennagels dienen. De onderzoeksvraag was of het seleniumgehalte in de afgeknipte teennagels al vóór het optreden van de tumor verlaagd was. Het verschil tussen het seleniumgehalte van een case en het gemiddelde seleniumgehalte van de gematchte controles werd gebruikt als maat voor een (mogelijk) verschil in blootstelling. Beide versies van de sequentiële toetsen hadden gemiddeld genomen minder waarnemingen nodig om tot een beslissing te komen dan een toets met een vooraf vastgesteld aantal waarnemingen. Het matchen van méér dan 1 controle per case verhoogde de efficiëntie. De winst in het aantal benodigde case-controle sets was het grootste als 2 controles gematcht werden aan 1 case in vergelijking met 1 controle per case, wanneer de nulhypothese (‘geen verband tussen seleniumgehalte en ziekte’) verworpen kon worden. De gevonden type I fout van de exacte versie was wat kleiner en de gevonden *power* wat groter dan die van de benaderende versie. Het gebruik van een dergelijke toets werd geïllustreerd met data uit het DOM project. De cases waren pre-menopauzale vrouwen die borstkanker kregen. Eén tot 5 controles werden gematcht op leeftijd aan ieder van de 64 beschikbare cases. De meeste sequentiële toetsen leidden tot het accepteren van de nulhypothese; soms waren echter te weinig gegevens beschikbaar om een beslissing te kunnen nemen.

Hoofdstuk 2

In hoofdstuk 2 is onderzocht hoe een verwacht verschil in blootstelling als continue grootheid tussen cases en controles uitgedrukt kan worden in een *Odds Ratio* (OR), een in de epidemiologie gebruikelijke maat. Als voorbeeld zijn weer de pre-menopauzale vrouwen met borstkanker uit het DOM project geanalyseerd. Een verwacht gestandaardiseerd verschil in seleniumgehalte van 0.25 tussen cases en ongematchte controles komt dan overeen met een verwachte OR van 2 voor de hoogste 20% ten opzichte van de laagste 20% van de blootstelling. Na analyse van 31 cases en 155 controles kon de nulhypothese worden geaccepteerd. Vergeleken met de 61 beschikbare cases, werd dus een besparing van 49% bereikt.

Hoofdstuk 3

Ook in hoofdstuk 3 werden twee versies van een sequentiële toets voor een continue blootstellingsvariabele vergeleken, de benaderende versie (R) uit hoofdstuk 1 en een nieuwere versie volgens Whitehead (W). Simulaties lieten zien dat het gemiddelde aantal case-controle sets voor de W-versie kleiner was dan voor de R-versie als de nulhypothese verworpen kon worden; met de R-versie waren minder case-controle sets nodig als de nulhypothese geaccepteerd kon worden. Voor grote aantallen case-controle sets was de *power* van de R-versie te laag. Voor kleine aantallen case-controle sets was de type I fout van de W-versie wat te groot. In dit hoofdstuk werd de mogelijke relatie onderzocht tussen het seleniumgehalte in de teennagels en het optreden van darmkanker bij pre-menopauzale vrouwen. Drie controles werden op leeftijd gematcht aan iedere case. De nulhypothese kon al na de achtste case-controle set geaccepteerd worden. Dat betekende een besparing van 75% in het aantal case-controle sets ten opzichte van de studieomvang zoals die van tevoren vastgesteld kon worden om een gestandaardiseerd verschil in seleniumgehalte van 0.5 aan te kunnen tonen als dat zou bestaan.

Hoofdstuk 4

Blootstelling is echter niet altijd in een continue waarde uit te drukken. Vaak is alleen maar bekend of iemand wel of niet blootgesteld is geweest, of dat wel of niet sprake is van een genetische mutatie. In hoofdstuk 4 werd de sequentiële toets beschreven die we ontwikkelden voor deze zogenaamde dichotome data. Met behulp van simulaties zijn de eigenschappen van deze toets bestudeerd voor twee versies van een sequentiële toets: de SPRT (Sequential Probability Ratio Test) en de TT (Triangular Test). De SPRT en TT verschillen in de (steilheid van de) grenzen die gehanteerd worden voor het beslissingsproces. In het algemeen heeft de SPRT-versie minder informatie nodig dan de TT-versie om tot een beslissing te komen als de nulhypothese (‘er is geen verband tussen blootstelling en ziekte’) of de alternatieve hypothese (‘het verwachte verband is er wel’) waar is. Dit werd door onze simulaties bevestigd. De type I fouten in de simulaties waren acceptabel; de gevonden waarden voor de *power* waren soms aan de lage kant, vooral wanneer het om kleine aantallen case-controle sets ging. We illustreerden het gebruik van deze sequentiële toets met onderzoek naar een mogelijke relatie tussen het wel of niet hebben van een bepaalde genetische mutatie en het optreden van rectumkanker. Vrouwen uit het DOM project die deze vorm van kanker hadden gekregen werden op leeftijd gematcht aan controlevrouwen zonder kanker. Met 1 controle gematcht per case kon na cumulatieve analyse van de beschikbare 69 cases nog geen beslissing genomen worden. Met 3 controles gematcht per case kon de nulhypothese na 35 discordante sets geaccepteerd worden. Voor een vergelijkbare toets met een van tevoren vastgesteld aantal waarnemingen zouden ten minste 66 discordante sets nodig zijn geweest om een

OR van 2 aan te kunnen tonen. Met een sequentiële toets met 3 gematchte controles kon in dit geval dus 47% bespaard worden.

Hoofdstuk 5

In de loop van een klinisch vergelijkend onderzoek of een epidemiologische, observationele studie wordt soms overwogen om voortijdig te stoppen. Een van de redenen hiervoor is dat het verwachte verschil tussen twee behandelingen of het verwachte verband tussen blootstelling en ziekte er niet lijkt te zijn. In de literatuur wordt berekening van de *Conditional Power* (CP) voorgesteld en gebruikt als instrument om te beslissen of een studie voortijdig gestopt kan worden. CP wordt berekend op een moment dat de studie nog gaande is. Het is de kans dat, gegeven de hoeveelheid informatie die op dat moment verzameld is, de statistische test de nulhypothese aan het geplande einde van de studie zal verwerpen. In hoofdstuk 5 bespraken we de nadelen van de CP. Als alternatief stelden we voor om een (groeps)sequentiële toets toe te passen op de al verzamelde gegevens. Als deze nog geen beslissing mogelijk maakt, dan kunnen nieuwe gegevens worden verzameld en geanalyseerd tot voldoende 'bewijs' is bereikt om wel een beslissing te kunnen nemen. We illustreerden dit voorstel door twee voorbeelden uit de literatuur opnieuw te analyseren. Een (groeps)sequentiële toets bleek minstens zo efficiënt te zijn als het berekenen van de CP. Bovendien zijn voor een (groeps)sequentiële toets geen arbitraire aannames nodig zoals voor berekening van de CP. Een (groeps)sequentiële toets komt op een meer objectieve manier tot de beslissing om een studie voortijdig te stoppen.

Hoofdstuk 6

Case-controle studies kunnen ontworpen worden om een verband tussen een genetische factor of een factor die de blootstelling aan een stof representeert en het optreden van een ziekte te bestuderen. Zowel de genetische factor als de blootstellingsfactor kunnen bijdragen aan de ontvankelijkheid voor een ziekte en het kan interessant zijn om hun zogenaamde interactie te bestuderen. Interactie betekent dat het hebben van de genetische factor én het blootgesteld zijn geweest aan de betreffende stof méér effect heeft op het ontstaan van de ziekte dan beide factoren afzonderlijk. Zowel hiërarchische als niet-hiërarchische interacties kunnen worden bestudeerd. Bij hiërarchische interacties worden de genetische factor, de blootstellingsfactor en een factor voor hun interactie in het model opgenomen. Bij niet-hiërarchische interacties wordt verondersteld dat zowel de genetische factor als de blootstelling nodig zijn voor het ontstaan van de ziekte. In hoofdstuk 6 beschreven wij de eigenschappen van sequentiële toetsen op interacties in gematchte case-controle studies. Onze simulaties lieten een goede overeenkomst zien tussen de theoretische en de gevonden benodigde hoeveelheid informatie om tot een beslissing te kunnen komen. Ook de gevonden type I fouten kwamen goed overeen met de theoretische waarden. Voor hele grote studies waren de gevonden waarden voor de *power* aan de hoge kant. De mediane winst in het aantal te analyseren case-controle sets was ongeveer 27%. Voor weinig voorkomende genetische afwijkingen was de winst groter als de alternatieve hypothese waar was dan als de nulhypothese waar was. Dit houdt verband met de kans op een discordante set, die klein is voor weinig voorkomende genetische afwijkingen. Het voorbeeld dat de ontwikkelde sequentiële toetsen illustreerde maakte weer gebruik van gegevens uit het DOM project. Cases waren vrouwen die borstkanker gekregen hadden. Iedere case werd op leeftijd gematcht aan een controle zonder borstkanker. Getoetst werd of er sprake was van interactie tussen een genetische afwijking en het rookgedrag. De sequentiële analyses leidden zowel voor de toets op hiërarchische interactie als voor de toets op niet-hiërarchische interactie tot het accepteren van de nulhypothese. Dat wil zeggen dat het veronderstelde interactie-effect tussen een genetische afwijking en roken niet bevestigd kon worden. De winst in het benodigde aantal case-controle sets was 29% en 41%. Sequentiële analyse kan dus ook in geval van toetsen op interactie in gematchte case-controle sets efficiënter omgaan met de beschikbare gegevens.

Hoofdstuk 7

In het laatste hoofdstuk worden twee stromingen in de statistiek beschreven, de frequentistische en de Bayesiaanse stroming. De frequentistische stroming probeert de vraag te beantwoorden: ‘Wat zou ik moeten doen?’; de Bayesiaanse stroming vraagt zich af: ‘Wat zou ik moeten geloven?’ De sequentiële toetsen die in dit proefschrift beschreven worden, zijn frequentistisch van aard. Een sequentiële toets houdt in dat de cumulatieve data herhaaldelijk worden geanalyseerd tot voldoende ‘bewijs’ is verkregen om een beslissing te kunnen nemen. De frequentistische benadering vereist echter dat de type I fout die per tussentijdse analyse gehanteerd wordt, kleiner gekozen wordt dan de 5% die als acceptabel beschouwd wordt voor de analyse van een studie met een van tevoren vastgestelde omvang. De beslissing om een studie te stoppen of te continueren hangt af van de inmiddels verzamelde informatie, maar ook van een stopregel die weer afhangt van de type I fout. Behalve het feit dat aanpassing van de type I fout nodig is bij herhaald toetsen, heeft de frequentistische benadering een veel fundamenteler probleem. De p-waarde of overschrijdingskans, die als resultaat van een statistische toets vermeld wordt, is de kans op de verkregen resultaten of extremere <u>áls</u> de nulhypothese waar is. Men kan dus zeggen dat de p-waarde de kans is dat de verkregen resultaten misleidend ‘bewijs’ zijn. De p-waarde wordt echter ook vaak, ten onrechte, gebruikt als maat voor de sterkte van het ‘bewijs’ tegen de nulhypothese: ‘hoe kleiner de p-waarde, des te sterker het ‘bewijs’’.

De Bayesiaanse stroming combineert ‘a priori’ kennis of ideeën over de verdeling van de te toetsen gegevens met de gegevens zelf tot een ‘a posteriori’ verdeling die dus beschouwd kan worden als een *update* van de ‘a priori’ kennis of, met andere woorden, de manier waarop een zeker geloof door de gegevens veranderd kan worden. Het grote probleem van de Bayesiaanse benadering is echter het subjectieve karakter van de ‘a priori’ verdeling.

Er is nog een derde stroming die tot nu toe weinig aandacht in de statistiek heeft gekregen: de *likelihood* stroming. Deze stroming stelt de vraag: ‘Wat vertellen de gegevens?’ en probeert die vraag te beantwoorden door te kijken naar de aannemelijkheid van de gegevens onder de ene hypothese (H_1) ten opzichte van de aannemelijkheid van de gegevens onder een andere hypothese (H_2). Het quotiënt van de twee aannemelijkheden is een maat voor de sterkte van het ‘bewijs’ in de gegevens ten gunste van ófwel H_1 ófwel H_2. In de *likelihood* stroming wordt de maat voor de sterkte van het ‘bewijs’ losgekoppeld van de kans op een misleidend ‘bewijs’. Terugkomend op het verschil tussen de frequentistische en de *likelihood* stroming: de interpretatie van de p-waarde hangt voor veel mensen af van de grootte van een studie. Twee experimenten die, behalve de grootte, volkomen identiek zijn en resultaten met dezelfde p-waarde produceren, hoeven niet evenveel ‘bewijs’ tegen de nulhypothese te bevatten. Ook kunnen gegevens uit verschillende experimenten precies dezelfde aannemelijkheid hebben, maar toch leiden tot verschillende p-waarden.

De beslissing om een studie te stoppen of te continueren op basis van het aannemelijkheidsquotiënt is niet afhankelijk van de type I fout of van een stopregel. Cumulatieve data kunnen dus zonder aanpassing van de type I fout herhaald geanalyseerd worden. In een eenvoudig voorbeeld werden data sequentieel geanalyseerd met behulp van een frequentistische toets en met een toets die gebruik maakt van de aannemelijkheidsratio. De *likelihood* benadering laat duidelijke voordelen zien vergeleken met de frequentistische en de Bayesiaanse benadering. Er blijven echter nog diverse vragen die verder uitgezocht moeten worden, voordat de benadering in het algemeen aanbevolen zal kunnen worden als een efficiënte en objectieve manier voor het herhaald toetsen van cumulatieve data.

Samenvattend komt dit hoofdstuk tot de volgende conclusie. In het geval van eenvoudige sequentiële toetsen van epidemiologische, observationele gematchte case-controle data, waarbij het doel is om ‘bewijs’ te vergaren voor één van twee hypotheses en waarbij de beschikbare hoeveelheid biologisch materiaal zo efficiënt mogelijk moet worden gebruikt, kan de *likelihood* stroming een objectief antwoord geven op de vraag: ‘Wat vertellen de gegevens?’

Dankwoord

" ... the whole of life is sequential,
for our future actions are conditioned to some extent by our past experience ..."
(Wetherill and Glazebrook, 1986)

Zo ook de totstandkoming van dit proefschrift. Het is niet, zoals veelal gebruikelijk, het resultaat van onderzoek gedurende een periode als onderzoeker-in-opleiding. De eerste 4 hoofdstukken waren al gepubliceerd, toen Rick Grobbee mij vroeg of ik er iets in zou zien om er een proefschrift van te maken. Prof. dr D.E. Grobbee, beste Rick, ik wil je hartelijk bedanken dat jij dit in gang gezet hebt. Jij had zelfs al wat voorwerk verricht, toen je het mij vroeg, door te informeren naar de mogelijkheden van een promotie als een academische vooropleiding ontbreekt. Wij waren het er over eens dat er ook een biostatisticus als promotor gevraagd moest worden. Prof. dr Th. Stijnen, beste Theo, ik waardeer het erg dat jij je gelijk bereid toonde om ook promotor te willen zijn. Ik wil jou en prof. dr J.C. van Houwelingen bedanken voor de brief die jullie geschreven hebben en die mij toelating tot een academische promotie heeft verleend.

Dr P.A.H. van Noord, beste Paul, jij hebt, zonder het op dat moment te weten, aan de wieg van dit proefschrift gestaan. Eind jaren tachtig vroeg jij mij of ik mee wilde denken bij de toepassing van sequentiële technieken in gematchte case-controle studies. Het waardevolle biologische materiaal van vrouwen in de DOM-cohorten moest zo zuinig mogelijk gebruikt worden. Jij hebt mij bedolven onder literatuur en ideeën. Ik wil je bedanken voor de bijzonder plezierige samenwerking en ik hoop dat we die de komende jaren kunnen voortzetten.

Dr R. Kaaks, beste Rudolf, ook jij was duidelijk betrokken bij de totstandkoming van de eerste hoofdstukken van dit proefschrift. Bedankt voor de, soms urenlange, telefonische discussies.

Dr O.L. van der Hel, beste Olga, bedankt dat wij jouw data ter illustratie mochten gebruiken.

I would like to thank prof. J.R. Whitehead, mrs S. Todd and mr F. Baksh from the University of Reading (U.K.) for the stimulating discussions around Fazil's thesis. Dear John, I have appreciated our contacts very much. Thank you for taking place in the defence committee. Thank you also for inviting Paul van Noord and me to come to Reading to discuss Fazil's work. Dear Fazil, thank you, I have learned a lot from your thesis. Dear Sue, I enjoyed our meetings, thank you also.

Mijn collega's van het Centrum voor Biostatistiek wil ik heel erg bedanken voor de plezierige samenwerking. Ik ben blij dat we in goede harmonie een aantal taken hebben kunnen verdelen, waardoor ik het promoveren kon combineren met het waarnemend hoofd zijn. In het bijzonder wil ik Maria Schipper, Jan van den Broek en Cas Kruitwagen bedanken voor het lezen, becommentariëren en bediscussiëren van verschillende hoofdstukken. Wiebe Pestman, bedankt voor je hulp bij het opnieuw maken van een aantal figuren.
Beste Maria, ik vind het heel fijn dat je op deze dag als paranimf naast mij wilt staan. Dank je wel voor het meedenken. Ik verheug me al op de voortzetting van ons onderzoek en de uitwerking van al onze onderzoeksplannen.

De KLEPpers van het Julius Centrum wil ik bedanken voor de goede contacten. Michael Edlinger, bedankt voor alle kopjes koffie en gezelligheid.

Monique den Hartog wil ik hartelijk bedanken voor het opnieuw invoeren van de eerste hoofdstukken en de lay-out van het hele proefschrift. Je hebt mij veel werk uit handen genomen.

Femke Bulten van de afdeling Vormgeving van de faculteit Biologie wil ik hartelijk bedanken voor het ontwerpen van de omslag.

Beste mede-(ex-)diakenen van de Ark in Maarssenbroek, dank jullie wel voor het anderhalf jaar lang overnemen van mijn werkzaamheden.

Lieve vriendenclub, Annie en Martijn, Ingrid en Huib, Marina en Henk, Ineke en Dirk, bedankt voor jullie belangstelling.

Lieve schoonfamilie, dank jullie wel voor jullie meeleven.

Lieve Liane en Rob, zus en broer, dank jullie wel voor jullie liefde en belangstelling. Liane, ik waardeer het heel erg dat je op deze dag als paranimf naast mij wilt staan. Ik hoop dat het minder 'eng' zal zijn dan je dacht.

Lieve mama en papa, jullie vonden het vanzelfsprekend dat ik op de HBS 'exact' zou kiezen in een tijd dat dat helemaal niet zo vanzelfsprekend was voor een meisje. Ook na de HBS hebben jullie mij gestimuleerd om in een 'exacte' richting verder te gaan. Dank jullie wel voor jullie liefde, belangstelling en betrokkenheid.

Allerliefste Gerard, dank je wel voor alles waarmee jij het mogelijk hebt gemaakt dat ik dit proefschrift af kon maken. Dank voor je luisterend oor en je wijze raad; je geduld, humor en relativeringsvermogen hebben mij heel veel goed gedaan. Bedankt ook voor het lezen en herlezen van de Nederlandse inleiding en samenvatting. Deze is door jouw op- en aanmerkingen en vragen om uitleg hopelijk begrijpelijk geworden voor familie, vrienden en kennissen.

Curriculum vitae

Ingeborg van der Tweel werd op 12 januari 1955 in Bussum geboren. Zij behaalde in 1971 het HBS-B diploma aan het bijzonder Willem de Zwijger-lyceum in Bussum.

In het najaar van 1971 trad zij in dienst van de stichting Mathematisch Centrum in Amsterdam als computerprogrammeur. Daar volgde zij de opleiding Wetenschappelijk Rekenen A (Wiskundig Genootschap), waarvoor zij in 1975 het diploma kreeg. In 1976 behaalde zij het diploma Statistisch Assistent VVS (Vereniging Voor Statistiek) en in 1977 het diploma Statistisch Analist VVS.

Op 1 september 1978 kwam zij in dienst bij de computerafdeling van de vakgroep Cardiologie (faculteit Geneeskunde) van de Rijksuniversiteit Utrecht.

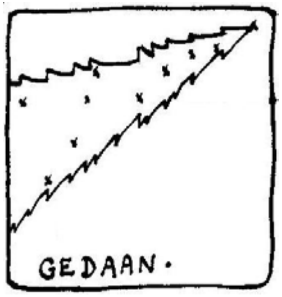
Zij behaalde het diploma Statisticus VVS in 1987 en het tentamen voor een aanvullend caput ‘Analyse van Overlevingsduren’ in 1989. De computerafdeling van de vakgroep Cardiologie werd opgenomen in de facultaire computerdienst van de faculteit Geneeskunde en in 1988 werd zij door de faculteit Geneeskunde gedetacheerd bij het interfacultaire Centrum voor Biostatistiek i.o..

Het Centrum voor Biostatistiek is een samenwerkingsverband van de faculteiten Biologie, Farmaceutische Wetenschappen, Diergeneeskunde en het Universitair Medisch Centrum van de Universiteit Utrecht. Het Centrum voor Biostatistiek werd op 1 januari 1990 formeel opgericht en stond tot 1 juni 2002 onder leiding van dr ir J.A.J. Faber. Na het samengaan van het Academisch Ziekenhuis Utrecht en de faculteit Geneeskunde in het Universitair Medisch Centrum Utrecht kwam zij op 1 januari 2000 in dienst van de faculteit Biologie.

Sinds april 1998 is zij vanuit het Centrum voor Biostatistiek twee dagen in de week gedetacheerd bij het Julius Centrum voor Gezondheidswetenschappen en Eerstelijns Geneeskunde.

In 2000 werd zij geregistreerd als Biostatisticus-VVS door de commissie Registratie Biostatistici van de Vereniging Voor Statistiek en Operationele Research.

Zij kreeg in 2002 de Basiskwalificatie Onderwijs toegekend door de faculteit Biologie. Sinds 1 juni 2002 is zij waarnemend hoofd van het Centrum voor Biostatistiek.


GEDAAN.