Stockholm University

Statistical aspects on clinical trials with covariate adaptive randomisation and with ordinal response data

Anna Stoltenberg

Licentiate Thesis in Mathematical Statistics at Stockholm University, Sweden 2009

Statistical aspects on clinical trials with covariate adaptive randomisation and with ordinal response data

Anna Stoltenberg

2009

ABSTRACT
Before a clinical trial can be performed there are a number of considerations that need to be done. A study protocol is needed, where it is described how the trial will be conducted. The population of patients to be studied must be well defined. The protocol also need to include the objectives of the trial, the response variables chosen to measure efficacy and safety, and the methods to be used for statistical analyses. Preferably, trials are controlled, randomised and double-blind. In this licentiate thesis, we consider clinical trials with parallel treatment groups, where a so called covariate-adaptive randomisation procedure has been used to allocate patients to treatment. We discuss how the statistical analysis is affected by such randomisation and how the treatment effect can be measured when the primary efficacy outcome is an ordinal categorical variable.

When prognostic factors, suspected to influence the response variable, are identified, it is desirable to use a randomisation procedure that achieves balance between treatment groups with respect to these prognostic factors. Covariate-adaptive randomisation is a treatment allocation procedure where the allocation for a new patient depends on the prognostic factors of patients already recruited. A question is how covariate-adaptive randomisation affects a following standard statistical test. Of particular interest is if the significance level is maintained when the covariate-adaptive randomisation is not taken into account or if a rerandomisation test is needed. There is a discussion in paper I, "A study of p-values in clinical trials with covariate adaptive randomisation", regarding covariate-adaptive randomisation and rerandomisation tests. Simulations were carried out in order to study the possible effects of the p-values of two standard tests used in connection with ordinal data, the Wald test from a logistic regression model and the Cochran-Mantel-Haenszel (CMH) test. The simulations show that covariate-adaptive randomisation, if not taken into account properly in the analysis, may lead to incorrect type I error. The conclusion is that the gain with covariate-adaptive randomisation is limited and a rerandomisation test is needed.

In a clinical trial with an ordinal categorical response variable, a logistic regression can be applied to data under the assumption of proportional odds. When there are prognostic factors that need to be taken into account in the statistical analysis, they can easily be incorporated into the logistic regression model as covariates, and the odds ratio can be used as a measure of effect. However, with covariates the risk that the model assumption of proportional odds is violated increases. This is a reason to prefer a non-parametric method, such as the CMH test. When the CMH test is used, we recommend to chose an effect measures corresponding to this test; Mann-Whitney's U, Somers' D (equivalent with Mann-Whitney's U), or the Number Needed to Treat (NNT, Somers' D reciprocal). In paper II "Effect measures in clinical trials with ordinal data", these effect measures are discussed, in particular in the presence of prognostic factors.

Paper I A study of p-values in clinical trials with covariate adaptive randomisation

Paper II Effect measures in clinical trials with ordinal data

Clinical trials

'On the 20th of May 1747, I took twelve patients in the scurvy, on board the Salisbury at sea. The cases were as similar as I could have them. They all in general had putrid gums, the spots and lassitude, with weakness of the knees. They lay together in one place... and had one diet common to all. Two of these were ordered each a quart of cider a day, two others took 25 gutts of elixir vitriol 3 times a day upon an empty stomach. Two others took two spoonfuls of vinegar 3 times a day... Two of the worst patients were put under a course of seawater. Two others had each two oranges and one lemon given them every day... they continued but six days under this course, having consumed the quantity that could be spared. The two remaining patients took... an electuary recommended by a hospital surgeon. The consequence was that the most sudden and visible good effects were perceived from the use of oranges and lemons; one of those who had taken them being at the end of six days fit for duty, the other... was appointed to nurse the rest of the sick.' [1]

Robert A. Thon, *A History of Medicine in Pictures*, Parke, Davis and Co., 1957

Dr. James Lind tested several scurvy treatments on crew members of the British naval ship Salisbury and discovered that lemons and oranges were most effective in treating the dreaded affliction. This experiment is probably the first well-documented clinical trial. Regarding clinical trials much has happened since 1747, mostly during the second half of last century. Today, a clinical trial is defined as a study conducted by researchers on human subjects to test a medical treatment or prevention strategy. The medical treatment under examination could be a drug, a surgical procedure, a medical device or a therapy. [2]

Nowadays an application need to be sent to regulatory authorities, such as the Food and Drug Administration (FDA) in USA and the European Agency for the Evaluation of Medical Products (EMEA), before a drug can enter the market and be used by patients. An application includes a number of clinical trials and there are many requirements for these trials. For each clinical trial, a study protocol is needed, describing how the trial will be conducted and how the population of patients, that will be studied, is defined.

A clinical trial requires a precise definition of which patients are eligible for inclusion. This is to ensure that patients in the trial may be identified as representative of some future class of patients to whom the findings in the trial can be applicable. In focus is the type of patients considered most likely to benefit from the new treatment. The disease state of investigation must be established, and this often requires quite detailed inclusion and exclusion criteria in the study protocol. Duration of a study is limited and when the requirements are too stringent it will be difficult to find enough patients in time.

If possible, a trial is controlled with standard treatment or with placebo. The word placebo appeared in medical literature in the early 1800s. *Hooper's Medical Dictionary* of 1811 defined it as "an epithet given to any medicine more to please than benefit the patient." Placebo means *to please* and the antonym is nocebo, which means *to harm*.

To be able to say that a trial is reliable, it is preferable that the study is blind. The reason for performing a blind clinical trial is to avoid biased outcome. The comparison of treatments may be distorted if the patient herself or those responsible for treatment and evaluation know which treatment is being used. This problem is avoided by a double-blind study design, where the patient, physician and the evaluator are not aware of which treatment the patient actually receives.

Patients are best allocated to treatment by use of randomisation, which is one of the central characteristics of a clinical trial. Today most clinical trials are randomised and are usually utilized to evaluate the efficacy of treatment. One reason to use randomisation is to maintain approximate balance across treatment groups of prognostic factors that are affecting the response variable. This will eliminate selection bias between groups. An other reason is the justification of randomisation based tests. Randomisation is also a tool to keep the study blind.

There are a number of different algorithms that can be used when patients are allocated to treatment. The simplest form of randomisation is to flip a coin. However, this is never done in practice. Randomisation schemes are generated by computerised systems, more or less done in a complex way. A quite complex randomisation process that has achieved popularity lately is a so called covariate-adaptive randomisation. Covariate-adaptive randomisation is a treatment allocation procedure where the allocation for a new patient depends on prognostic factors of patients already recruited. The aim is to achieve balance between treatment groups with respect to these prognostic factors. The question is how covariate-adaptive randomisation affects a following standard statistical test that is randomisation based. Of particular interest is if the significance level is maintained.

The statistical aspects on clinical trials with covariate-adaptive randomisation with ordinal response data that are considered in this licentiate thesis appeared in three clinical studies performed in the stroke area. All three studies used a covariate-adaptive randomisation procedure proposed by Pocock and Simon [3]. The response variable in these studies was an ordinal categorical response, which was analysed by use of the Wald test from a logistic

regression model and the Cochran Mantel Haenszel (CMH) test. We have used these three stroke studies as basis for simulations. The simulations were done to investigate the properties of standard tests performed after a covariate-adaptive randomisation proposed by Pocock and Simon [3]. How these tests perform after covariate-adaptive randomisation is discussed in paper I "A study of p-values in clinical trials with covariate-adaptive randomisation". The conclusion is that the significance level is not always maintained and a rerandomisation test is needed.

This licentiate thesis concentrates on clinical trials with parallel treatment groups where we have a response outcome on an ordinal scale with categories described in words such as "none", "mild", "moderate" or "severe". To be specific, suppose we use a scale with m different categories. The frequencies in each row have a multinomial distribution and the probabilities of falling into different categories within each treatment are described in Table 1.

Table 1 Category probabilities for ordinal response in patients receiving test or control treatment

	Category			
Treatment	**1**	**2**	**…**	**m**
Test	π_{1T}	π_{2T}	…	π_{mT}
Control	π_{1C}	π_{2C}	…	π_{mC}

There are two main approaches to analyse data, parametric or non-parametric analyses [4]. For ordinal data, a preferred parametric approach is an ordinal regression method. In the statistical assessment of ordinal outcomes in comparative studies [5] it is essential that the ordinality of the ranked data is fully exploited. If the response is treated as a categorical variable on a nominal scale much of the information is wasted. In other word, it is not advisable to evaluate the data by calculating the proportions for each category of outcome and perform a chi-square test of association. Nor is it advisable to analyse ordinal outcome data with binary logistic regression. Information is lost when the variable is reduced by dichotomisation. There are several regression models that can be used for ordinal data, but for most of them there is no single effect measure that captures the impact of the treatment. We will concentrate on the proportional odds model. The proportional odds model is an extension of binary logistic regression and is sometimes referred to as the "ordinal logistic" model. It is also referred to as a "cumulative odds" model. The model is linear and additive on the logit scale, and use maximum likelihood methods to estimate a summary odds ratio. In this model the data is dichotomised across the scale, also refer to as "cut-points". As an example, the cut-points for an ordinal variable with the outcomes "none", "mild", "moderate" or "severe" can be found in Table 2.

Table 2 **Cut-points for a proportional odds model based on an ordinal outcome variable with 4 categories**

Cut-points	Proportional odds model: successive incremental cut-points
1	None vs. mild, moderate and severe
2	None and mild vs. moderate and severe
3	None, mild, and moderate vs. severe

With m categories, see Table 1, there are (m-1) cut points and the cumulative probability for a patient to fall into category k or better (a lower category) will be denoted Q_{kT} for test treatment and Q_{kC} for control treatment:

$$Q_{kT} = \pi_{1T} + \pi_{2T} + \ldots + \pi_{kT} \; ; \; Q_{kC} = \pi_{1C} + \pi_{2C} + \ldots + \pi_{kC}, \quad k = 1, \ldots, m \; .$$

Notice that $Q_{mT} = Q_{mC} = 1$. At each cut point an odds ratio can be calculated:

$$OR_k = \left\{ \frac{Q_{kT}/(1-Q_{kT})}{Q_{kC}/(1-Q_{kC})} \right\}$$

When all these odds ratios have a common value, i.e. $OR = OR_1 = OR_2 = \ldots = OR_{m-1}$, we have proportional odds, and the common value OR is a natural measure of treatment effect in the population.

When prognostic factors influence the response variable it is possible to include the prognostic factors and calculate an adjusted estimate of the odds ratio under the assumption of proportional odds. This is easily done by use of the PROC LOGISTIC procedure provided by the SAS® system [6]. Problems arise when the assumption of proportional odds is violated and the risk of violating the assumption increases with the number of prognostic variables.

The Cochran-Mantel-Haenszel (CMH) test is a non-parametric test, also called stratified Wilcoxon, since it is possible to include prognostic factors. Continuous prognostic factors need to be categorized before they can be included in the statistical analysis. A stratum is a given combination of levels from different factors, so e.g. with 2 prognostic factors of 2 levels each, there are four strata. The frequency tables from different strata are assumed to be independent given the row sums, and the null hypothesis is that in each stratum, the multinomial distributions in the two rows have common probability parameters over the response categories.

An ideal measure of treatment effect should exhibit good interpretability and good statistical properties. We would like the effect measure to tell the clinicians how likely it is that patients benefit from the test treatment. A good effect measure can communicate information useful to assess the clinical significance of any result found in a clinical trial. The odds ratio has been criticised for not having these properties. Besides, the assumption of proportional odds can be violated, especially when covariates are included in the logistic regression model. In this case

we prefer Somers' D (Somers' rank correlation index D) as effect measure. Somers' D is a modification of the Kendall tau rank correlation coefficient for the association between treatments and the response variable [7]. Somers' D can be adjusted for prognostic variables and there is no need for proportional odds. Somers' D is also called the expanded success rate difference, since in the case of binary response Somers' D reduces to success rate difference. More regarding Somers' D and corresponding effect measures, Mann-Whitney's U and the Number Needed to Treat (NNT, Somers' D reciprocal), in particular when adjustment is needed for prognostic factors, can be found in paper II "Effect measures in clinical trials with ordinal data".

References:

1. Hampton J.R. Size isn't everything. *Statistics in Medicine*. 2002;21:2807-2814

2. Collier R. Legumes, lemons and streptomycin: a short history of the clinical trial. *CMAJ*. 2009;180:23-24

3. Pocock S.J., Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trials. *Biometrics.* 1975;31:103-115.

4. Agresti A. *Analysis of ordinal categorical data*. Wiley 1984.

5. Scott S.C., Goldberg M.S., Mayo N.E. Statistical assessment of ordinal outcomes in comparative studies. *J Clin Epidemiol*. 1997;50:45-55.

6. Gameroff M.J. Using the proportional odds model for health-related outcomes: why, when, and how with various SAS® procedures. *SUGI 30 Statistics and Data Analysis.* Paper 205-30.

7. Somers, R. A new asymptotic measure of association for ordinal variables. *American Sociological review*. 1962;27:799-811.

Acknowledgements

I would like to thank AstraZeneca for financial support.

I wish to give Rolf Sundberg, my supervisor at Stockholm University, my sincere thanks. Rolf never gave up and was always able to squeeze in a meeting with me, even in times of his life when he was really occupied. I have never met anyone so committed and dedicated to statistics and he shared his expert knowledge with me. I do appreciate all valuable and constructive comments during my work with this licentiate thesis, which has improved my writing skills. I thank Rolf for all I have learned from him.

I would like to thank Olivier Guildbaud, my supervisor at AstraZeneca, who has, with his humble personality, kept my self esteem at a level that has made it possible to continue to the end. I admire Olivier for this broad experience and knowledge. I also highly appreciate his ability to explain things in a way that makes it understandable. Thank you Olivier, for always giving me of your time.

I want to express my sincere gratitude for the support from my managers. They have been the drive that kept me going when I gave up. First Lennart Claesson, who made me believe that it was possible, then Stig-Johan Wiklund who made it possible and now Jonas Häggström who is an inexhaustible source of inspiration and encourage at a daily basis.

I would like to thank all of my friends and colleagues at AstraZeneca who never showed any disbelief that I would do anything else than eventually finalise my licentiate thesis. When I doubted myself, they had confidence in me.

I want to thank my children, Erik and Ella, for always reminding me of my priorities and I am grateful to my husband, who never questioned my reasons for spending time, money and effort in the work with this licentiate thesis.

A study of p-values in clinical trials with covariate-adaptive randomisation

ABSTRACT

Covariate-adaptive randomisation is a treatment allocation procedure where the allocation for a new patient depends on prognostic factors of patients already recruited. The question is how covariate-adaptive randomisation affects a following standard statistical test. Of particular interest is if the type I error may be incorrect. Because of this concern, a rerandomisation test may be needed instead of a statistical test that does not take the covariate-adaptive randomisation into account.

The primary reason for using the covariate-adaptive randomisation is to achieve balance between treatment groups with respect to prognostic factors. The aspect of balance is not trivial and in studies with good balance the result will get higher credibility. Another reason is that with good balance one may intuitively expect that the power of a test is higher.

There are many aspects that need to be considered when covariate-adaptive randomisation is used. An important aspect is that regulatory authorities are conservative regarding covariate-adaptive randomisation and want the reason for using covariate-adaptive randomisation to be based on solid clinical and statistical ground. The programming of the covariate-adaptive randomisation procedure, as well as the rerandomisation test, is not straightforward in practice and errors may occur. Another consideration is the increased cost if a computerised randomisation system is needed.

In this paper, studies are simulated to investigate the properties of standard tests performed after a covariate-adaptive randomisation proposed by Pocock and Simon. The simulations are based on actual clinical data from three trials conducted in the stroke area. We have looked at the Wald test from a logistic regression model and the Cochran Mantel Haenszel (CMH) test.

The simulations show that covariate-adaptive randomisation, if not taken into account properly in the analysis, may lead to incorrect type I error. A recommended approach is to perform a rerandomisation test. The conclusion is that covariate-adaptive randomisation does no harm, if properly taken into account, but that the gain has its limitation.

Keywords: Covariate-adaptive randomisation; Taves' minimisation; simulation

1. INTRODUCTION

Randomisation is one of the central characteristics of a clinical trial and today most clinical trials are randomised. One aspect of randomisation is the maintenance of approximate balance across treatment groups of prognostic factors that are affecting the response variable. These prognostic factors can be known or unknown in the planning phase of the clinical trial. In combination with blinding, randomisation helps to minimise possible bias. Randomisation ensures an unbiased

treatment comparison, i.e. makes sure that any difference between the treatment groups is attributed to the treatment effect rather than to influence from prognostic factors.

General description of randomisation procedures will be found in Section 2. The covariate-adaptive randomisation method that has been used in this paper, proposed by Pocock and Simon, is described in Section 3. In Section 4 a rerandomisation test will be described. Issues regarding covariate-adaptive randomisation can be found in Section 5. In Section 6, there is an explanation of how the simulations are performed and the results from the simulations are also included. In Section 7 there is a discussion and in Section 8 the conclusions can be found.

2. RANDOMISATION PROCEDURES

Four classes of randomisation procedures can be distinguished: complete randomisation, restricted randomisation, covariate-adaptive randomisation and response-adaptive randomisation. We will restrict the discussions to the simple case where we have two parallel treatment groups. More general information regarding randomisation can be found in *Randomization in Clinical Trials* by William F Rosenberger [1].

2.1 Complete randomisation

Complete randomisation is simply toss of a fair coin, where the treatment assignments between patients are independent. This is only of theoretical interest, and in the worst case, all patients can be assigned to the same treatment.

2.2 Restricted randomisation

We refer to restricted randomisation when the treatment allocations to patients are mutually dependent in some way. To have equal numbers of patients assigned to each treatment group is one of the most common restrictions. A randomisation list can be generated prior to the inclusion of patients. The randomisation list includes randomisation numbers and each number is accompanied with a treatment code. Patients entering the study will be allocated to treatments in accordance with the randomisation numbers on the randomisation list in a consecutive order.

2.2.1 Balance between treatment groups

2.2.1.1 Truncated binomial randomisation

A restriction can be that we want to randomise a study with 2n patients such that we achieve equal number of patients in each treatment group. A way of assigning exactly n patients to each of two treatment groups is to use coin tossing until one treatment has been assigned to exactly n patients and then allocate the other treatment to the rest of the patients. This is referred to as the truncated binomial randomisation.

2.2.2 Permuted blocks

Randomisation in permuted blocks [2] helps to avoid imbalance and to increase the comparability of the treatment groups. For example, if two treatments, A and B, are to be compared, a permuted block of four patients can be AABB, ABAB, ABBA, BABA, BBAA, or BAAB. In a study with a total sample size of 4n, with permuted blocks of size four, n blocks are needed and the number of

patients assigned to each treatment can never differ by more than two, whenever patient assignment ends. With small block, we will get close to perfect balance.

2.2.3 Stratified randomisation

Prognostic factors are variables that are potentially correlated with response variable. A stratum is a given combination of levels from different factors, so e.g. with 2 prognostic factors of 2 levels each, there are four strata. With stratified randomisation, separate randomisations are performed for each stratum. Random permuted blocks can be used within each stratum, with varying block length as an option. The idea behind stratified randomisation is to keep the balance between the treatment groups within each prognostic factor level and still be able to claim that the study is a randomised clinical trial. Stratified block designs can, in some cases, lead to imbalance if not all randomisation numbers are used within the last blocks in many strata.

2.3 Covariate-adaptive randomisation

Covariate-adaptive randomisation is employed when there are known prognostic factors and there is a desire to balance between treatment arms with respect to these prognostic factors. This is a method in which the allocation of patients is determined by the current balance of the treatment groups, meaning that we use information from the previously allocated patients to assign treatment to the next patient to be included. When we refer to adaptive randomisation, we will mean the covariate-adaptive randomisation procedure proposed by Simon and Pocock, which is described in Section 3.

2.4 Response-adaptive randomisation

In response-adaptive randomisation, the treatment assignment to a new patient depends upon the treatment responses of previously included patients. Response-adaptive randomisation can be used when it is desirable to randomise patients to the most promising treatment. Response-adaptive randomisation is outside the scope of this paper.

3. THE ADAPTIVE RANDOMISATION METHOD OF POCOCK AND SIMON

The adaptive randomisation procedure considered in this paper is a technique developed by Pocock and Simon [3] and is the most commonly used adaptive randomisation procedure. It is based on an entirely deterministic allocation method originally proposed by Taves [4], which we will refer to as Taves' minimisation. Pocock and Simon introduced a random element to the minimisation procedure. The basic idea behind the method of this adaptive randomisation is that imbalance is measured marginally for each level of the prognostic factors and summed over the factors. The method balances treatments at each prognostic factor level marginally, but approximate balance within each stratum is also achieved, which is desirable. The assignment of treatment group to the next patient in line is determined to minimise the imbalance and depends on the values of the prognostic factors for the patients already included in the study.

Suppose that we have p prognostic factors for which treatment balance is desired. The numbers of levels of these factors are $r_1, r_2, \ldots, r_p$. The procedure is best described at an arbitrary time

point in the study when some of the patients have already been included in the study and a new patient is to be allocated to treatment. The imbalance measure, G_k, assuming that the new patient is allocated to treatment k (where k is A or B), is calculated as the total sum of absolute differences between number of patients (including the new patient) in each treatment group for each prognostic factor within each level of the prognostic factors:

$G_k = \sum_{i=1}^{p} \sum_{j=1}^{r_i} | n_{ijA} - n_{ijB} |$, where i represent a prognostic factor (i = 1, 2, …, p), and j a level within a factor (j = 1, 2, …, r_j). The imbalance measure G_A is calculated assuming that the new patient receives treatment A, and G_B is calculated assuming that the new patient is allocated to treatment B. If $G_A < G_B$, the new patient is assigned to treatment A with a higher probability than B, and to treatment B with a higher probability than A when $G_A > G_B$. When $G_A = G_B$ the new patient is equally likely to receive treatment A as treatment B.

Table 1 illustrates the calculation of the imbalance measures G_A and G_B. Imagine a clinical trial with two parallel treatment groups and with three important prognostic factors (condition at baseline, gender, and already on treatment or not) that we want to balance for and use in the adaptive randomisation procedure. Forty patients have already been allocated to treatment, 20 in each treatment group, and the new patient has a mild condition at baseline, is female and is not already on treatment. Assigning the new patient to treatment A would lead to a total imbalance of $G_A = |3 – 5| + |(9+1) – 7| + |6 – 5| + |2 – 3| + |10 – 13| + |(10+1) – 7| + |9 – 14| + |(11+1) – 6| = 25$, whereas assigning the new patient to treatment B would lead to a total imbalance of $G_B = |3 – 5| + |9 – (7+1)| + |6 – 5| + |2 – 3| + |10 – 13| + |10 – (7+1)| + |9 – 14| + |11 – (6+1)| = 19$. The total imbalance would be minimised if the new patient receives treatment B, since $G_A > G_B$ (25 > 19), and therefore the new patient will be assigned to treatment B with a higher probability than to treatment A.

Table 1 Calculation of imbalance measures G_A and G_B for the new 41th patient

	Prognostic factor	**Level**	**Treatment A**	**Treatment B**	**Imbalance**
Patients already allocated to treatment	Condition at baseline	None	3	5	
		Mild	**9**	**7**	
		Moderate	6	5	
		Severe	2	3	
	Gender	Male	10	13	
		Female	**10**	**7**	
	Already on treatment	Yes	9	14	
		No	**11**	**6**	
New patient allocated to treatment A	Condition at baseline	None	3	5	2
		Mild	**10**	**7**	**3**
		Moderate	6	5	1
		Severe	2	3	1
	Gender	Male	10	13	3
		Female	**11**	**7**	**4**
	Already on treatment	Yes	9	14	5
		No	**12**	**6**	**6**
					G_A=25
New patient allocated to treatment B	Condition at baseline	None	3	5	2
		Mild	**9**	**8**	**1**
		Moderate	6	5	1
		Severe	2	3	1
	Gender	Male	10	13	3
		Female	**10**	**8**	**2**
	Already on treatment	Yes	9	14	5
		No	**11**	**7**	**4**
					G_B=19

4. RERANDOMISATION TEST

A rerandomisation test can be applied to all types of response data, i.e. continuous, ordered or categorical data. Such a test is performed as follows:

Specify the problem. Assume that we have a clinical trial with 2 parallel treatment groups and want to test the null hypothesis that the two treatments are equivalent, i.e. the patient's response is the same regardless of treatment.

Choose a test statistic. If the two treatments are not equivalent, we want to be able to detect the difference and choose a test statistic that is sensitive to such a difference. Under the null hypothesis the test statistic has a certain null distribution under the used randomisation procedure, where the responses are considered as fixed.

Compute the test statistic for the observed responses using the treatments assigned to patients in the original randomisation. This test statistic will be referred to as the raw test statistic.

Estimate the null distribution of the test statistic by generating a large random sample from this distribution as follows:
(a) reallocate treatments to patients in accordance with the chosen randomisation procedure;
(b) compute the test statistic for this reallocation; and
(c) do independent repetitions of (a) and (b) a large number of times. The empirical distribution of the values from the generated test statistic estimates the null distribution. A large enough sample will accurately estimate this null distribution. In some situations it is possible to derive the true null distributions through complete enumeration of the possible allocations and their probability.

Accept or reject the hypothesis using the estimated null distribution. Determine the relevant critical region (one-tailed or two-tailed) from this null distribution, and reject the null hypothesis if the raw test statistic computed for the original randomisation falls into this region.

It is important to note that the responses (and corresponding covariate values) of patients are considered as given constants in the rerandomisation test described above. More information regarding rerandomisation and permutation tests can be found in a book written by Good [5]. In this book, Good use "permutation" and "rerandomisation" interchangeably. However, for a permutation test, all permutations of treatment assignment sequence are equally likely. After adaptive randomisation some treatment sequences are more likely than others and some are highly unlikely, so we prefer "rerandomisation" to "permutation" in the present context.

5. ASPECTS OF ADAPTIVE RANDOMISATION

5.1 Balance

The aspect of balance is not trivial. It is unnecessary to be exposed to potential criticism about the credibility of the result. This can be the case when there is severe imbalance between treatment groups regarding important prognostic factors. Achieving balance within strata becomes more

complex when there are many prognostic factors and when some of these factors have more than 2 levels. In several papers the balancing properties of adaptive randomisation have been compared to other allocation methods [6, 7, 8, 9]. The general view is that adaptive randomisation yields tight balance and outperforms stratified randomisation as the number of prognostic factors increases. Adaptive randomisation can cope with more factors than permuted blocks within strata, which becomes unworkable as the number of prognostic factors increases and the number of strata required quickly exceeds the number of patients in the trial [10]. However, if the number of prognostic factors with several levels is too large to be handled by randomisation with permuted blocks within strata, adaptive randomisation is perhaps not the only solution. In this case it might instead be better to decrease the number of prognostic factors.

McEntegart [11] believes that the primary reason to achieve good balance in a study, for instance by use of adaptive randomisation, is to get higher credibility among journal readers and regulatory authorities. It is appealing to see that there is treatment balance when data is tabulated by important prognostic factors, but that alone cannot justify the use of adaptive randomisation.

5.2 Power

Increasing the power could be a reason to use adaptive randomisation. However, Peto et al. [12] conclude that the gain in efficacy and balance relative to more simple randomisation methods are negligible. Weir and Lees [6] made simulations of trials including 1000 subjects per trial. They simulated normally distributed response variables and performed analyses of covariance (ANCOVA), where the prognostic factors used in the randomisation process were included as covariates. They compared adaptive randomisation to stratified random permuted block and could conclude that adaptive randomisation only resulted in slight improvement in power. In a paper by Lachin [13], where the statistical properties of randomisation in clinical trial are discussed, he concludes: “Although treatment imbalances affect power, the effects are trivial unless the imbalances are substantial. Therefore, for large trials, the susceptibility of a randomisation procedure to such imbalances is not a statistical concern.”

5.3 Regulatory guidance

In a trial using Taves’ minimisation as treatment allocation method, the Food and Drug Administration (FDA) requested that a non-deterministic element should be included in the randomisation procedure [14]. In addition, FDA requested that a rerandomisation test (see Section 4) should be performed to confirm the conclusion from the primary statistical analysis that did not take the randomisation procedure into account.

In *Statistical Principles for Clinical Trials* guideline, ICH E9 of the International Conference on Harmonisation [15], we can read: “Stratification by important prognostic factors measured at baseline (e.g. severity of disease, age, sex, etc.) may sometimes be valuable in order to promote balanced allocation within strata; this has greater potential benefit in small trials.” We can also find the following advice regarding adaptive randomisation: “Deterministic dynamic allocation procedures should be avoided and an appropriate element of randomness should be incorporated for each treatment allocation.” In other words, applicants are not recommended to use Taves’ minimisation, but as long as a random element is included there is no strong position against adaptive randomisation.

In “Points to consider on adjustment for baseline covariates” [16] (below referred to as the CPMP document), published by the Committee for Proprietary Medical Products (CPMP) of the European Agency for the Evaluation of Medical Products (EMEA) more advise is given: “… stratification for more than a few prognostic factors is not always possible, especially for small trials. In this situation, techniques of dynamic allocation such as minimisation are sometimes used to achieve balance across several factors simultaneously. Even if deterministic schemes are avoided, such methods remain controversial. Thus, applicants are advised to avoid such methods. If they are used, the reasons should be justified on solid clinical and statistical grounds.” And further, “Dynamic allocation is strongly discouraged. However, if it is used, then it is imperative that all factors used in the allocation scheme be included as covariates in the analysis. Even with this requirement, it remains controversial whether the analysis adequately reflects the randomisation scheme. Applicants will be required to describe the sensitivity analyses they intend to perform to support the conclusions from the primary analysis. Without adequate and appropriate supporting/ sensitivity analyses, an applicant is unlikely to be successful.”

The CPMP document has been criticised. Roe [17] thinks that “The almost dogmatic position in this document regarding dynamic allocation seems out of tune, …” and hopes that this will not stimulate a premature dismissal of a range of allocation methods for which there is empirical evidence that they can be beneficial in the practice of clinical trial design. Other critics to the view expressed in the CPMP document regarding adaptive randomisation are Buyse and McEntegart [18]. They believe that CPMP’s position is unfair, since it ignores recent methodological literature that encourages wider use of minimisation, and because it does not cite any references supporting its own view. Buyse and McEntegart consider CPMP’s position as unfounded because it endorses static balancing methods and rejects dynamic methods, when in fact there is no essential difference in the two, according to them. They also believe that CPMP’s position is unwise, because it favours use of randomisation methods that expose clinicians and the medical community to the risk of accidental bias, when this risk could have been limited, especially when the trials are small and cannot easily be repeated.

An answer to this criticism was written by Simon Day, Jean-Marie Grouin and John A. Lewis [19], all three authors involved in drafting the CPMP document. They have often observed, when reviewing applications from sponsor companies, that the use of adaptive randomisation results in more harm than good. They have rarely seen the need to use any allocation procedure more complex than simple stratified randomisation with permuted blocks. They have also seen trials where the choice of factors included in the randomisation algorithm has been poorly thought out, where the programming algorithm has been incorrect, and where the telephone system or web-based system for adaptive randomisation was unreliable.

5.4 Prognostic factors

In the CPMP document [16], it is a minimum requirement that the factors used in the randomisation procedure have to be included in the statistical model used for analysis. Our own view is that if a factor is so important that we use it in the randomisation procedure, it needs to be included in the statistical model used for statistical analysis. We want to point out that the CPMP document discourages the use of too many covariates and models that become too complex. Dynamic allocation procedures are said to be beneficial when the number of strata becomes larger [10]. However, since the guideline argues for including stratification factors in the

statistical model, and it discourage a large number of factors in that model, only a few stratification factors should be used. Statistical models with many covariates lead to complex models, which are complicated to interpret. With few stratification factors, other treatment allocation procedures than adaptive randomisation can be considered.

5.5 Computerised randomisation system

Most likely, any study allocating patients to treatment by use of an adaptive randomisation procedure becomes so complicated that there is a need of a computerised randomisation system. A telephone based interactive voice response (IVR) system, or a web-based system, see Cai et al [20], are examples of such. In the three stroke studies described in Section 6.1, an IVR system was used. The system had to be reliable and available all round the clock, since the patients arrived at hospital in an emergency situation, requiring immediate treatment allocation. For various reasons the physician occasionally never called the IVR system and the patient received a package of drug labelled with the lowest available randomisation number stored at the clinic. In this case, or if any of the prognostic variables were incorrect at the first call, it was always possible to update the IVR system later. The message heard when the IVR system was called, was available in 19 different languages. There was also a helpdesk facility available. In this example, many centres with few patients were included and there was an obvious risk for severe imbalance between treatment groups in small centres. It is not possible to answer if it was beneficial to use an IVR system, and what would have happened if it was not used. It can only be established that the desire of achieving balance required a lot of time for planning and was very expensive. Balance between treatments for each prognostic factor at all levels was achieved, but only one of the prognostic factors used at randomisation was correlated to the primary response variable.

6. SIMULATION

A number of authors have used computer simulations to show that better balance is achieved with adaptive randomisation compared to other stratification procedures [6, 7, 8, 9]. Better balance is most probably achieved, but in this paper we are more interested in how the adaptive randomisation procedure influences the statistical inference. Adaptive randomisation introduces a deterministic element to the randomisation, with the consequence that some sequences become more likely than others and some become highly unlikely. This may lead to better balance at the expense of a biased estimate of the variability in the test statistic, and there is a concern that the statistical test does not maintain its significance level after performing an adaptive randomisation. We have made simulations to see if the level of the type I error is maintained in tests that does not take the adaptive randomisation into account or if a rerandomisation test is needed. In order to make these simulations potentially relevant for applications in clinical trials, we based the simulations on three actual clinical studies, to be described below.

6.1 Description of the stroke studies used as basis for simulation

The clinical data that constitutes the basis for the simulations comes from three acute stroke studies, here referred to as studies 1, 2 and 3. They were double blind, multicenter, and placebo controlled phase IIb/III studies with two parallel treatment groups, active and placebo. The

efficacy response variable was disability at end of the 3 months study, as measured according to a 6-category ordered scale, the modified Rankin Scale (mRS), ranging from 0 (no symptoms) to 5 (severe disability), where death is merged with the latter category.

In study 1, 595 patients were included. The aim was to evaluate safety and no sample size calculation concerning efficacy was done. In the two other studies the sample size was based on an anticipated difference in efficacy. In study 2, 1699 evaluable patients were analysed for efficacy. The sample size calculation for study 3 was based on the result in study 2. This led to a larger sample size and study 3 was almost twice as large as study 2, with 3196 patients included in the efficacy analysis.

In study 1, two prognostic factors were used in the randomisation procedure, severity of stroke at baseline and country. Severity of stroke at baseline is a factor with four levels, none, mild, moderate and severe. In studies 2 and 3, two more two-level factors were identified as potentially prognostic and used at randomisation: whether alternative treatment was received or not, and side of brain where stroke occurred. The prognostic factors used in the randomisation were included as covariates in the statistical analyses.

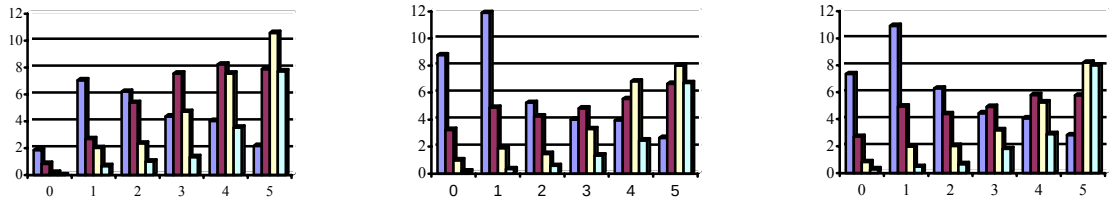
The patients in studies 1, 2 and 3 were randomised in accordance with an adaptive randomisation, as described in Section 3. When the imbalance measure for active treatment, G_{active}, was greater than the imbalance measure for the placebo group, $G_{placebo}$, the next patient was assigned to placebo with a higher probability than active, and to treatment active with a higher probability than placebo when $G_{active} < G_{placebo}$. When $G_{active} = G_{placebo}$ the new patient was equally likely to receive active treatment as placebo. The probability used in the three studies was chosen to be 0.75. However, this is not an obvious choice. Pocock [21] recommended a probability of 0.66 to 0.75. Weir and Lees [6] investigated the imbalance for various values of allocation probabilities, 0.85, 0.9, 0.95 and 1. McEntegart [11] used 0.75 as an example and the influential CONSORT group statement [22] gave a probability of 0.8 as an example.

6.2 Creation of pseudo studies

When the three studies had been conducted it was found that severity of stroke at baseline was the only prognostic factor used in the randomisation procedure that was actually correlated to the response variable, modified Rankin Scale (mRS), so only this prognostic factor is used in the simulation.

In Figure 1, for each study 1-3, the joint distribution of severity at baseline (none, mild, moderate, severe) and disability at end of study (mRS score 0 to 5) is visualised. No consideration of treatment is done here. The first four bars in Figure 1 represent no disability at end of study where the first bar is the percentage of patients with no severity of stroke at baseline and the following bars represent mild, moderate and severe stroke at baseline. The next four bars represent a mRS score of 1, and so on. The sum of the 24 bars is 100%. As we can see, patients in study 1 had more severe strokes at baseline and were more disabled at end of study, compared to patients in the other two studies. Only a small proportion of the patients have no symptoms after 3 months of receiving active treatment or placebo. Study 2 and 3 were similar, with large proportion of patients with the outcome 0 (no symptoms) or 1 (mild symptoms) on the mRS scale and the condition at baseline were none or mild stroke for more patients than in study 1.

Figure 1 **The joint distribution of severity of stroke at baseline (none, mild, moderate, severe) and disability at end of study, mRS score (0 to 5), for each study 1, 2 and 3**



Study 1 Study 2 Study 3

New pseudo studies are generated by multinomial sampling with 24 categories and probability parameters given by the observed distributions in Figure 1. Each hypothetical patient in a pseudo study is randomly assigned a number between 1 and 24. For example, if a patient is assigned the number 6, this corresponds to the sixth bar in one of the distributions in Figure 1. This means that the patient is assumed to have a mild condition for the prognostic factor 'severity of stroke at baseline' and an outcome of score 1 on the mRS. When all patients in a pseudo study randomly have been assigned a number between 1 and 24, numbers are translated to a value for the prognostic factor and a mRS score as response.

Pseudo studies of sizes 50, 595 and 1699 were sampled from each of the three studies in Figure 1. The size of 50 was chosen to represent a small study. A study with only 50 patients is not likely to be large enough to detect any treatment effect, and the choice may seem unrealistic. However, if there are any patterns to be seen, we hope to see them more clearly in many small studies, than in few larger ones. The other study sizes, 595 and 1699, were the actual sizes in studies 1 and 2. They represent a more realistic choice of sizes, even though the study of 595 patients was planned to only evaluate safety and not large enough to show any treatment effect. The reason for not simulating studies with the same size as in study 3, a size of 3196, is both that this is time consuming and that we will probably not be able to see anything that has not already been seen in the pseudo studies of smaller sizes.

For the larger sample sizes of 595 and 1699, eight pseudo studies for each of the three studies were created. Pseudo studies of size 50 were faster to sample than studies of larger sizes and 99 pseudo studies for each of the three studies were generated. In Appendix 1 a selection of underlying distributions of the pseudo studies can be found. When the sample size is as large as 1699 there are small variations between the distributions of the eight pseudo studies, and the distributions of severity of stroke at baseline by mRS score are similar to those of the original studies. Totally, 297 + 24 + 24 = 345 pseudo studies were generated, see Table 2.

Table 2 Number of pseudo studies generated for each study in Figure 1

	Study			
Pseudo studies of size (N)	**1**	**2**	**3**	**Total**
50	99	99	99	297
595	8	8	8	24
1699	8	8	8	24

6.3 Statistical tests used in the studies

The simulations of the pseudo studies are based on three stroke studies. In these stroke studies the primary effect variable (mRS) was analysed by use of two tests, the non-parametric Cochran-Mantel-Haenszel (CMH) test and the Wald test in a logistic regression model. Since these two tests were used in the three stroke studies, we have chosen to use these test in the pseudo studies as well. With no covariates the CMH test reduces to the Wilcoxon-Mann-Whitney (WMW) test. In simple comparisons of two treatments, without any prognostic factors, the Wald test in a logistic regression model is equivalent to the WMW test and provides identical p-values when the hypothesis that the treatments are equally effective is tested [23]. This means that if there is a difference between the p-values from the Wald test in a logistic regression model and the CMH test, this is probably due to the use of prognostic factors in these tests. The standard tests do not take the adaptive randomisation into account, and therefore we want to evaluate them to see if they will maintain their significance level under adaptive randomisation.

6.3.1 Cochran-Mantel-Haenszel test

The Cochran-Mantel-Haenszel (CMH) test is a test of conditional independence. In a clinical trial with binary response (0 or 1), the CMH test tests the null hypothesis of conditional independence in S 2x2 tables, where S is the number of strata. A 2x2 frequency table for the response in stratum h (h = 1,… ,S) can be seen in Table 3.

Table 3 Frequency table for binary response in patients treated with active or placebo in stratum h

	0	**1**
Active	n_{A0h}	n_{A1h}
Placebo	n_{P0h}	n_{P1h}

In each stratum h, given the row sums of the 2x2 tables, the first cell counts n_{A0h} and n_{P0h} in the two rows are assumed to have a binomial distribution. The 2x2 tables from different strata are assumed to be independent given the row sums. The null hypothesis is that in each stratum h, the two binomial distributions have equal probability parameters. Under this null hypothesis, given the row and column marginals of each 2x2 table, the first cell n_{A0h} in each table is then

(conditionally) distributed according to a known hypergeometric distribution with mean μ_{A0h} and variance var(n_{A0h}), and the test statistic,

$$z_{CMH} = \frac{\left[\sum_{h=1}^{S}(n_{A0h} - \mu_{A0h})\right]^2}{\sum_{h=1}^{S}\mathrm{var}(n_{A0h})},$$

has approximately a chi-squared null distribution with one degree of freedom.

However, in our situation, the response variable is the modified Rankin Scale (mRS), ranging from 0 (no symptoms) to 5 (severe disability), where death is merged with the latter category. In each stratum h, the frequencies in each row of the resulting 2x6 frequency table are assumed to have a multinomial distribution given the row sums, and the frequencies in different categories within each treatment are denoted as in Table 4.

Table 4 Frequency table for modified Rankin Scale score in patients receiving with active or placebo in stratum h

	0	**1**	**2**	**3**	**4**	**5/death**
Active	n_{A0h}	n_{A1h}	n_{A2h}	n_{A3h}	n_{A4h}	n_{A5h}
Placebo	n_{P0h}	n_{P1h}	n_{P2h}	n_{P3h}	n_{P4h}	n_{P5h}

The frequency tables from different strata are assumed to be independent given the row sums, and the null hypothesis is that in each stratum, the multinomial distributions in the two rows have common probability parameters over the response categories.

In this more general case, the CMH test is defined in terms of the mean scores

$$\bar{x}_{Ah} = \sum_{k=0}^{5} n_{Akh}x_{kh} \Big/ \sum_{k=0}^{5} n_{Akh},$$

$$\bar{x}_{Ph} = \sum_{k=0}^{5} n_{Pkh}x_{kh} \Big/ \sum_{k=0}^{5} n_{Pkh},$$

within stratum h, where $x_{0h}, x_{1h}, \ldots, x_{5h}$ are scores assigned to the outcomes 0, 1, …,k, respectively. The CMH statistic, given these scores, is

$$z_{CMH} = \left[\sum_{h=1}^{S}(\bar{x}_{Ah} - \bar{x}_{Ph})\right]^2 \Big/ \sum_{h=1}^{S}\mathrm{var}(\bar{x}_{Ah} - \bar{x}_{Ph}).$$

In each stratum, the frequencies in the two marginals of the 2x6 table are considered fixed, and the cell-frequencies within the table are then (conditionally) distributed according to a known multiple hypergeometric distribution under the null hypothesis. This means that under the null hypothesis, the (conditional) variances $\mathrm{var}(\bar{x}_{Ah} - \bar{x}_{Ph})$ are known, and the statistic z_{CMH} is approximately chi-square distributed with one degree of freedom. The null hypothesis is rejected if z_{CMH} is too large. Different choices of the scores x_{0h}, x_{1h}, …, and x_{5h} lead to different versions of the CMH test. The scores chosen are so called Modified Ridit scores, which correspond (within each table) to Wilcoxon midrank scores. The SAS procedure PROC FREQ was used with the option "modified ridit" to perform this CMH test in the stroke studies. Unfortunately, only a 2-sided test is performed, and only a 2-sided p-value is provided by PROQ FREQ, i.e. the p-value does not indicate whether a difference is in favour of the active treatment or placebo.

The CMH statistics have low power for detecting an association in which the patterns of association for some of the strata are in the opposite direction of the patterns displayed by other strata. Thus, a non-significant CMH statistic suggests either that there is no association or that no pattern of association has enough strength or consistency to dominate any other pattern.

6.3.2 Wald test

A logistic regression model can be applied to ordinal response data and an associated Wald test can be performed by use of PROC LOGISTIC from SAS®. Wald test is a standard way to use the likelihood function to perform large-sample inference. Wald's test, the likelihood ratio (LR) test and the score test usually give similar results for studies with large samples. In studies with few observations it is preferable to choose the LR test instead of a Wald test, since the Wald test is only an approximation of the LR test. The LR test incorporates the log-likelihood at H_0 as well as at $\hat{\beta}$. Since the Wald test was used in the three stroke studies and since no LR test and only Wald test is given with PROC LOGISTIC, we have chosen to evaluate the Wald test for the influence of adaptive randomisation.

Consider a binary response applied to a logistic regression model with only treatment as factor, logit = $\alpha + \beta x$, where β is the log odds ratio. The significance test focus on H_0: $\beta = 0$, the hypothesis of independence. The Wald test uses the test statistic $z = \hat{\beta}/SE$, where $\hat{\beta}$ is the maximum likelihood estimate of β and SE the standard error of $\hat{\beta}$. The test statistic $z = \hat{\beta}/SE$ has an approximate standard normal distribution under H_0. One refers z to the standard normal table to obtain one- or two-sided p-values. Equivalently, for the two-sided alternative, z^2 has a chi-squared probability above the observed value. z^2 is asymptotically chi-squared with 1 degree of freedom for large samples.

In our situation, the response is an ordered variable with six categories. With a six category scale, there are five cut-off points and the frequencies in the five 2x2 tables at each cut-point can be seen in Table 5, for a given strata.

Table 5 Frequency tables at each cut-point for the modified Rankin Scale score in patients treated with active or placebo

	0	≥ 1
Active	n_{A11}	n_{A21}
Placebo	n_{P11}	n_{P21}

	≤ 1	≥ 2
Active	n_{A12}	n_{A22}
Placebo	n_{P12}	n_{P22}

…

	≤ 4	5/death
Active	n_{A15}	n_{A25}
Placebo	n_{P15}	n_{P25}

For each 2x2 table an odds ratio can be calculated as, $OR_k = \frac{n_{A1k}/n_{A2k}}{n_{P1k}/n_{P2k}}$, k = 1, …, 5. In a cumulative logit model, originally proposed by Walker and Duncan [24] and later called the proportional odds model by McCullagh [25], it is assumed that all these odds ratios have a common value, i.e. $OR = OR_1 = \ldots = OR_5$, not only within each stratum, but also over all strata. With this approach we can apply the same model to data when we have an ordered response variable as in the binary case described above. In this situation β is the log odds ratio, log OR, in the model, $\text{logit} = \alpha_k + \beta x$, k = 1, …, 5.

In the statistical program package SAS®, the Wald chi-square statistic given by PROC LOGISTIC is computed by squaring the estimated log odds ratio divided by its standard error estimate, and the p-value of the Wald chi-square statistic with one degree of freedom is given. This means that a p-values corresponding to a 2-sided test is provided, i.e. the p-value does not indicate whether a difference is in favour of the active treatment or placebo.

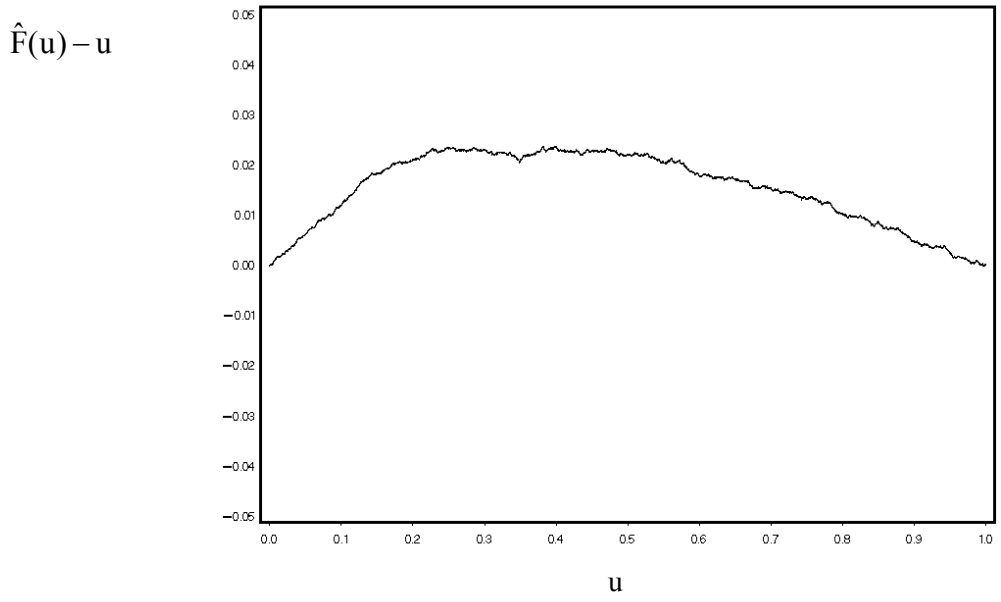
6.4 Estimation of null distribution

In the pseudo studies we use the adaptive randomisation suggested by Pocock and Simon. For each pseudo study, new treatment allocations are generated using the allocation probabilities described in Section 6.1. The new treatment reallocations are done conditional on the observed sequence of patient entries, and the values of the response and the prognostic factor, severity of stroke at baseline, are also considered as given. After each generated sequence of treatment reallocations, the value of the CMH test statistic and the values of the Wald test statistic from a logistic regression model are calculated. See Section 6.3 for details regarding these tests. Both tests adjust for the prognostic factor, severity of stroke at baseline, which was used in the randomisation procedure.

For each such pseudo study, 40 000 sequences of treatment reallocations are done, leading to the calculation of 40 000 pairs of test statistic values (one from the CMH test, and one from the Wald test). For each type of test, under the assumption that the null hypothesis of equal treatment effect is true, these test statistic values come from an underlying null distribution function (one for CMH, and one for Wald) that depends on the pseudo study, sample size and treatment allocation method. The empirical distribution function, based on 40 000 test statistic values, is a good estimate of the null distribution function. An alternative to these test statistics is the corresponding p-values, here obtained by assuming that the test statistic is χ^2 distributed. After 40 000 repetitions of the treatment allocations , the p-values are sorted in increasing order from $p_{(1)}$, $p_{(2)}$, …, to $p_{(40\,000)}$, and their empirical distribution function is used to estimate the underlying true null distribution function of p-values. We will denote such a null distribution function of p-values by F, and the corresponding estimate, the empirical distribution function, by $\hat{F}$. A test that

perfectly satisfies the distributional assumption on it, and in particular is not affected by the adaptive randomization or by finite sample deviation from its asymptotic distribution, will have a null distribution function F that is the uniform distribution, satisfying F(u)=u, $0 < u < 1$. The difference between the actual F and the "ideal" uniform distribution function can theoretically be described through a p-p-plot (u, F(u)), $0 < u < 1$; but since F is unknown, we must instead study the p-p-plot (u, $\hat{F}(u)$), $0 < u < 1$, and evaluate by how much this curve differs from the diagonal line (u, u), $0 < u < 1$, of equality. Equivalently, we can study the vertical difference $\hat{F}(u) - u$ plotted against u, as illustrated in Figure 2.

Figure 2 Difference between empirical distribution function, $\hat{F}$, and u, $\hat{F}(u) - u$ plotted against u



In Figure 2, we can see an example of the difference $\hat{F}(u)$ - u for one pseudo study based on 40 000 generated p-values. When F(u) = u it is possible to analyse the trial with a test that ignores that an adaptive randomisation is used. That is, it is desirable that the difference $\hat{F}(u) - u$ is as close to zero as possible. A curve above 0 reflects a test that overestimates the p-values. A test that underestimates the p-values will have a curve $\hat{F}(u) - u$ that lies below 0. Overestimated p-values are conservative. From a regulatory point of view, it is better to overestimate a p-value, since this means that there is a lower risk of claiming a statistically significant treatment effect, when there is no treatment effect. In other words, the actual type I error of the test is then less than specified. In Figure 2 we can see an example of overestimated p-values, since $\hat{F}(u) - u$ is positive for all $0 < u < 1$.

When a test performs poorly in the simulation, i.e. the difference $\hat{F}(u) - u$ is far from 0, it is not necessarily only due to the adaptive randomisation. It can also be finite sample deviation from its asymptotic distribution for the test statistics. In an attempt to separate the source of error due to poor χ^2 distribution approximation from the error that arises because the adaptive randomisation is ignored in the statistical analysis, a new set of treatment allocations is carried out, without

adaptive randomisation. We use the same pseudo studies of size 50, with the same prognostic factors and responses as we used for the adaptive randomisation. Instead of an adaptive randomisation we perform unstratified treatment allocations with a single permuted block of size 50, with 25 patients in each treatment group. Patients are reallocated to treatment and 40 000 pairs of p-values, one from the CMH test, and one from the Wald test from a logistic regression model are calculated. After 40 000 repetitions of the treatment allocation, the p-values are sorted and new null distributions are estimated, $\hat{F}'(u)$.

6.5 Results

Pseudo studies with 50 patients are fast to generate, and we created 99 pseudo studies with 50 patients each. Instead of presenting results from all these 99 pseudo studies, we selected 8 of them as follows. For each pseudo study the value $u_{0.05}$, such that $\hat{F}(u_{0.05}) = 0.05$, for the Wald test from a logistic regression model was determined. The $u_{0.05}$-values were sorted and ranked from 1 to 99 and we chose to present the results selected from the eight pseudo studies with ranks 1 (minimum), 15, 29, 43, 57, 71, 85, and 99 (maximum). The same eight pseudo studies were used to illustrate the behaviour of the CMH test.

In Appendix 2, we can find the curves of the difference $\hat{F}(u) - u$ in pseudo studies generated from study 1, 2 and 3, with sample size of 50, 595 and 1699 patients, both for the CMH test and the Wald test in a logistic regression model, after adaptive randomisation. Totally, there are 144 curves evaluated.

We can see in Appendix 2, that for the CMH test the p-values can be underestimated as well as overestimated. The absolute differences $\left|\hat{F}(u) - u\right|$ are more pronounced for pseudo studies with only 50 patients and the curves get closer to 0 when the number of patients increases. With 595 patients, curves seem to be close to 0 and with 1699 patients the CMH test performs quite well, with curves very close to 0.The curves from the small pseudo studies with 50 patients cannot be directly compared to curves from pseudo studies with more patients, particularly not the extreme lower and upper curves. The reason for this is that the curves for pseudo studies with 50 patients are based on 99 underlying pseudo studies, and that only a selection of these, including the extremes (minimum and maximum), is presented.

We can see that the Wald test from a logistic regression model performs poorly in small studies, where Wald test have a tendency to underestimate the p-values. When p-values are under-estimated there is a risk for rejecting the null hypotheses at a higher significance level than planned for. For larger sample sizes, 595 and 1699 patients, the behaviour of Wald's test is remarkably similar to that of the CMH test. The curves do not get closer to 0 when the sample size increase from 595 to 1699, especially not for pseudo studies generated from study 3. For sample sizes 595 or 1699, the p-values may as well be overestimated as underestimated.

In Appendix 3, we can in total see eighteen graphical presentations, each for 50 patients and with eight curves representing the selected eight pseudo studies. There are six graphs for each of the studies 1, 2, and 3, where three of these curves are obtained after performing the CMH test and the other three after the Wald test has been used. The graph at the top is taken from Appendix 2

and illustrates the curves $\hat{F}(u) - u$ versus u after adaptive randomisation. The graph in the middle represents the use of single permuted block randomisation, $\hat{F}'(u)$ -u versus u. At the bottom is a graph with eight difference curves, $\hat{F}(u) - \hat{F}'(u)$.

In Appendix 3, we can see that the CMH test performs extremely well after randomisation with single permuted block for pseudo studies with 50 patients. This means that most of the error in the statistical evaluation in Appendix 2 comes from ignoring that adaptive randomisation was used. The curves after adaptive randomisation change little when the single permuted block curves are subtracted from them.

When single permuted block randomisation is used, the estimated null distribution of p-values from the Wald test is similar to the estimation after adaptive randomisation. For small sample sizes the approximation of the χ^2 distribution for the Wald test statistic is poor regardless of randomisation procedure. When the curve from the unstratified randomisation is removed from the curve of the adaptive randomisation, the behaviour of Wald test is remarkably similar to that of the CMH test. See Appendix 3.

In the pseudo studies with sample size 1699, there are only small variations in the underlying distributions of the prognostic factors and the response, compared to the distributions in smaller pseudostudies (see Appendix 1 for a selection of underlying distributions). When the underlying distributions for the repetitions of treatment allocation are similar, the estimation of the null distributions can be expected to be alike, leading to curves of difference close to each other. In Appendix 2, we can see that the curves are close to each other, even if it is more pronounced for the CMH test than for the Wald test. On the other hand, for the smaller sample sizes, 50 or 595 patients, the empirical null distribution function is sensitive to the underlying distribution of the prognostic factor and response. Even when there is only small variation between the distributions of the prognostic factor and of the response, the p-values can be underestimated or overestimated. The outcome is unpredictable. We are not aware if any pattern in the underlying distribution of the prognostic factor and the response can predict the direction of the result, that is if the p-value will be under- or overestimated.

In a rerandomisation test (see Section 4) we choose a test statistic or the corresponding p-value, as was done in Section 6.4. This choice will not influence the outcome of the rerandomisation test, since the p-value is a monotone function of the original test statistic and can be used in the same way. After the null distribution has been estimated, we use the raw p-value, p_{raw}, (which does not take the adaptive randomisation procedure into account) as follows. If the raw p-value is within the defined critical region of the estimated null distribution, the null hypothesis is rejected and there is a statistically significant treatment effect. The critical region for the rerandomisation test at significance level α (approximately) is based on the estimated null distribution function, $\hat{F}$, as follows: (a) determine the critical value u_α for which $\hat{F}(u_\alpha) = \alpha$; and (b) reject the null hypothesis (same effect in the two treatment groups) if $p_{raw} \leq u_\alpha$. Thus, u_α is the empirical α-quantile based on the empirical distribution function $\hat{F}$. The critical values for rerandomisation tests at a significance level of 5%, after the performance of the Wald test and the CMH test after adaptive randomisation, $\hat{F}(u_{0.05}) = 0.05$, and after single permuted block randomisation, $\hat{F}'(u_{0.05}) = 0.05$, in pseudo studies of size 50, can be found in Table 6. As mentioned in the first paragraph

of this section, the pseudo studies have been sorted by the $u_{0.05}$ values from the Wald test for each stroke study, which can be seen in Figure 6.

Table 6 Critical values $u_{0.05}$, empirically based on 40 000 p-values generated through rerandomisation for 8 (of 99) selected pseudo studies of size 50 from each of stroke studies 1 – 3

Study 1				**Study 2**				**Study 3**			
Adaptive randomisation		**Single permuted block**		**Adaptive randomisation**		**Single permuted block**		**Adaptive randomisation**		**Single permuted block**	
CMH	**Wald**	**CMH**	**Wald**	**CMH**	**Wald**	**CMH**	**Wald**	**CMH**	**Wald**	**CMH**	**Wald**
0.039	0.026	0.051	0.035	0.040	0.023	0.050	0.032	0.044	0.028	0.051	0.034
0.047	0.031	0.051	0.034	0.047	0.034	0.050	0.039	0.047	0.033	0.052	0.039
0.048	0.035	0.049	0.037	0.047	0.036	0.050	0.038	0.046	0.036	0.051	0.039
0.051	0.037	0.051	0.037	0.053	0.039	0.052	0.039	0.042	0.039	0.051	0.047
0.054	0.041	0.050	0.040	0.054	0.042	0.051	0.038	0.051	0.041	0.052	0.041
0.058	0.044	0.052	0.042	0.050	0.045	0.051	0.044	0.057	0.044	0.052	0.039
0.057	0.049	0.049	0.039	0.055	0.049	0.052	0.046	0.056	0.048	0.050	0.042
0.061	0.059	0.050	0.051	0.084	0.077	0.048	0.044	0.059	0.068	0.051	0.055

The results in Table 6 confirm the findings based on the curves in Appendix 3. We can see that after adaptive randomisation, the significance level is not maintained for small sample sizes and the critical values for the rerandomisation tests can be far from 0.05, especially for the Wald test, but also for the CMH test. After randomisation with single permuted block, the CMH test performs amazingly well. This is not the case for the Wald test from a logistic regression model, where the $\hat{F}'(u_{0.05})$-values are far from 0.05.

Critical values $u_{0.05}$ for the rerandomisation tests at the 5% significance level after adaptive randomisation for pseudo studies with 595 and 1699 patients, are found in Table 7.

Table 7 **Critical values $u_{0.05}$, empirically based on 40 000 p-values generated through rerandomisation for pseudo studies of size 595 and 1699 from each of stroke studies 1 – 3**

595 patients						1699 patient					
Study 1		Study 2		Study 3		Study 1		Study 2		Study 3	
CMH	Wald	CMH	Wald	CMH	Wald	CMH	Wald	CMH	Wald	CMH	Wald
0.048	0.045	0.045	0.042	0.049	0.046	0.049	0.048	0.049	0.049	0.049	0.046
0.050	0.049	0.050	0.050	0.050	0.048	0.049	0.049	0.047	0.049	0.048	0.047
0.049	0.049	0.049	0.050	0.048	0.048	0.051	0.050	0.049	0.050	0.049	0.047
0.050	0.049	0.051	0.051	0.050	0.048	0.051	0.050	0.049	0.051	0.048	0.050
0.048	0.049	0.052	0.051	0.049	0.049	0.052	0.051	0.050	0.051	0.049	0.051
0.051	0.050	0.051	0.052	0.050	0.049	0.050	0.051	0.049	0.052	0.051	0.051
0.054	0.052	0.053	0.053	0.053	0.050	0.052	0.051	0.053	0.053	0.050	0.052
0.050	0.052	0.057	0.058	0.052	0.050	0.051	0.052	0.052	0.054	0.052	0.053

In pseudo studies with large sample size (595 or 1699 patients) the critical values are closer to the significance level of 0.05, but still it is not possible to say that the type I error is not affected when the adaptive randomisation is ignored in the statistical analysis, see Table 7.

An alternative to the use of critical regions for rejecting the null hypothesis (same effect regardless of treatment received) is an adjusted p-value. An adjusted p-value, p_{adj}, can be calculated as $p_{adj} = \hat{F}(p_{raw})$, which is the proportion of p-values in the empirical distribution function equal to or lower than p_{raw}. The adjusted p-value can be compared directly with the chosen significance level α, and the resulting test will then approximately have a correct significance level. For the adjusted p-value, as well as the critical region, to be as correct as possible, the number of repetitions, n, needs to be large enough. An asymptotic interval that with 95% probability will cover the true p-value $F(p_{raw})$, can be calculated as $p_{adj} \pm 1.96\sqrt{p_{adj}(1-p_{adj})/n}$. With n = 40 000, which was the number of repetitions done for each pseudo study and $p_{adj} = 0.05$, the width of the interval, $I_{p_{adj}}$, is approximately 0.004, i.e. $I_{p_{adj}} = 0.05 +/- 0.002$. For $p_{adj} = 0.01$ and $p_{adj} = 0.1$ the widths are 0.002 and 0.006 representatively after 40 000 repetitions.

7. DISCUSSION

In a clinical trial, where the statistical test does not take the adaptive randomisation into account, the p-value can be questioned. This is confirmed in the simulation of the pseudo studies, based on actual clinical data, evaluated above. The p-value may be overestimated as well as underestimation. Not surprisingly, when it comes to maintaining the type I error, studies with large sample size perform better than smaller trials. However, regardless of sample size, there is no

guarantee that the significance level is maintained if a test, which does not take the adaptive randomisation into account, is used.

When a test performs poorly it is not necessarily only due to the adaptive randomisation. It can also be due to a poor rate of convergence towards the asymptotic distribution of the test statistics. In the simulations, in studies with small sample sizes, we also looked at the maintenance of significance level after randomisation in single permuted block. The CMH test performs well after single permuted block randomisation, even in small studies. For the Wald test from the logistic regression model, this is not the case. In small studies, the Wald test statistic is not approximately χ^2 distributed. If it can be suspected that the test statistic deviates from its asymptotic distribution, which can be the case in a study with small sample size, it can be beneficial to consider a rerandomisation test (or permutation test) also in studies where adaptive randomisation has not been used.

Adaptive randomisation is complicated in various ways. Even programming errors have occurred and an example of this is a trial that had to rerecruit over 1000 women when a mistake in the randomisation algorithm caused serious imbalance [26]. In the same way as the execution of an adaptive randomisation procedure can be complicated, a rerandomisation test can be difficult to perform. This can be an excuse for not performing a rerandomisation test. However, Green et al [27] refer to a case where U.S. regulatory body requested that a trial that had used Taves' minimisation procedure had to be reanalysed using rerandomisation test. Regulatory authorities are conservative regarding adaptive randomisation. In ICH E9 [15] adaptive randomisation is not encourage, and CPMP discourages its use. The reasons for using adaptive randomisation need to be based on solid clinical and statistical grounds to be supported by the European authorities [16]. Especially in a trial with large sample size it can be questioned if there are any solid and statistical grounds.

When a rerandomisation test is performed, it is usually only to support the primary statistical analysis, which does not take the adaptive randomisation into account. For the study to be conclusive, the rerandomisation test needs to confirm the result. A problem arises when the raw p-value indicates there is statistically significant result but the rerandomisation test does not, or vice versa. We suggest that the adjusted p-value from the rerandomisation test should be used as the main basis for the evaluation of the primary objective in a confirmatory clinical trial, and that this should be stated already in the planning phase. To be trustworthy, it is essential to state á priori that the primary p-value will come from a rerandomisation test. In a recent paper by Hasegawa and Tango [28], simulations have been made to compare the rerandomisation test to standard ANCOVA. Hasegawa and Tango come to the same conclusion when they state "In conclusion, we suggest that 1) Pocock–Simon's procedure could be used as a method of randomization, and 2) a permutation test could be used as the primary statistical analysis in a confirmatory randomized controlled trial with important prognostic factors."

8. CONCLUSION

We can conclude from the simulations that the p-values from the CMH test and the Wald test can be over- or underestimated when performed after adaptive randomisation. The use of adaptive randomisation cannot be ignored in the statistical analysis. We suggest that an adjusted p-value

from a rerandomisation test should be used for the evaluation of a clinical trial when adaptive randomisation has been used. The rerandomisation test should be the primary analysis and for the trial to be reliable this need to be stated in the planning phase.

When adaptive randomisation is to be used, Taves' minimisation should be avoided, since it does not include an element of randomness. In the statistical evaluation, the prognostic variables used in the randomisation process should be included as covariates in the statistical model used for analysis. If a proper analysis that takes adaptive randomisation into account is used, adaptive randomisation doesn't make much harm. However, the gain has its limitations. The decision for using adaptive randomisation should be based on solid clinical and statistical grounds, which is also required by the regulatory authorities. During the planning phase of a clinical trial, the various costs (e.g. computerised systems), including logistic difficulties and potential programming and calculation errors, should be taken into account when considering whether adaptive randomisation should be used or not. We conclude that adaptive randomisation needs to be used with caution and after other possible randomisation procedures have been thoroughly considered. There are alternatives to adaptive randomisation, i.e. stratified randomisation. Simulation is a great tool to investigate if adaptive randomisation will gain in balance and power compared to what can be expected from other randomisation methods.

References

1. Rosenberger, W. F., Lachin, J. M. *Randomization in Clinical Trials, Theory and Practice.* New York: John Wiley & Sons 2002.

2. Friedman LM, Furberg CD, DeMets DL. *Fundamentals of Clinical Trials.* 3rd Ed. St. Louis: Mosby-Year Book, Inc. 1996.

3. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trials. *Biometrics*, 1975;31:103-115.

4. Taves DR. Minimization: a new method of assigning patients to treatment and control groups. *Clinical Pharmacology and Therapeutics*, 1974;15:443-453.

5. Good P. *Permutation tests*. Springer-Verlag New York 1994.

6. Weir CJ, Lees KR. Comparison of stratification and adaptive methods for treatment allocation in an acute stroke clinical trial. *Statistics in medicine*, 2003;22:705-726.

7. Begg CB, Iglewicz B. A treatment allocation procedure for sequential clinical trials. *Biometrics*, 1980;36:81-90.

8. Therneau TM. How many stratification factors are 'too many' to use in a randomization plan. *Controlled Clinical Trials*, 1993;14:98-108.

9. Zielhuis GA et al. The choice of a balanced allocation method for a clinical trial in otitis media with effusion. *Statistics in Medicine*, 1990;9:237-246.

10. Pocock SJ. Allocation of patients to treatment in clinical trials. *Biometrics*, 1979;35:183-197.

11. McEntegart, DJ. The pursuit of balance using stratified and dynamic randomization techniques: an overview. *Drug Information Journal*, 2003;37:293-309.

12. Peto R, et al. Design and analysis of randomized clinical trials required prolonged observation of each patient. I: Introduction and design. *Br J Cancer* , 1976;34:585-612.

13. Lachin JM. Statistical properties of randomization in clinical trials. *Controlled Clinical Trials*, 1988; 9:289-311.

14. Ebbutt A, Kay R, McNamara J, Engler J. The analysis of trials using a minimization algorithm. In: 20th Anniversery Conference Report. Macclesfield England: PSI; 1998:12-14

15. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. ICH Harmonised Triplicate Guideline. E9: Statistical Principles for Clinical Trials. *Statistics in medicine*, 1999;18:1905-1942.

16. 'CPMP. Points to consider on adjustment for baseline covariates (CPMP/EWP/908/99). EMEA: London, 2002

17. Roes KCB. Regulatory perspectives; Dynamic allocation as a balancing act. *Pharmaceutical statistics*, 2004;3:187-191.

18. Buyse M. McEntegart. Achieving balance in clinical trials: An Unbalanced view from EU regulators. *Applied Clinical Trials*, 2004;13:36-40.

19. Day S, Grouin J-M, Lewis JA. Achieving balance in clinical trials. *Applied Clinical Trials*, 2005;14:24-26.

20. Cai H, Xia J, Xu D, Gao D, Yan Y. A generic minimization random allocation and blinding system on web. *Journal of Biomedical Informatics*, 2008;39:706-719.

21. Pocock SJ. *Clinical Trials. A Practical Approach.* Wiley: Chichester, 1983.

22. Altman DG, Schulz KF, Moher D et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Int Med*, 2001;134:663-694.

23. Whitehead J. Sample size calculations for ordered categorical data. *Statistics in Medicine,* 1993;12:2257-2271.

24. Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. *Biometrika*. 1967; 54:167-179.

25. McCullagh P. Regression models for ordinal data (with discussion). *J. R. Statist. Soc. B.* 1980;42:2:109-142.

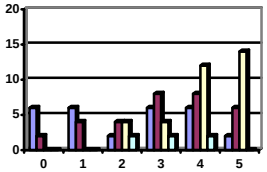
26. Comparative Obstetric Mobile Epidural Trial (COMET) Study Group UK. Effect on low-dose mobile versus traditional epidural techniques on mode of delivery: a randomised controlled trial. *Lancet*, 2001;358:19-23.

27. Green H, McEntegart DJ, Byrom B, Ghani S, Shepard S. Minimization in crossover trials with non-prognostic strata: theory and practical application. *J Clin Pharm Ther*, 2001;26:121-128.

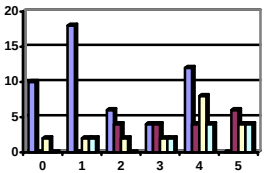
28. Hasegawa T, Tango T. Permutation test following covariate-adaptive randomisation in randomized controlled trials. *Journal of Biopharmaceutical Statistics*, 2009;19:106-119.

Appendix 1 Distributions

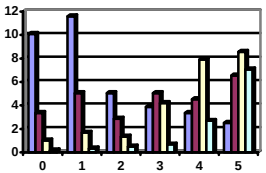
50 patients, study 1, population 77



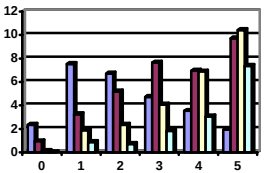
50 patients, study 3, population 64



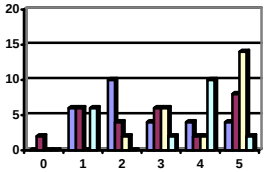
595 patients, study 2, population 5



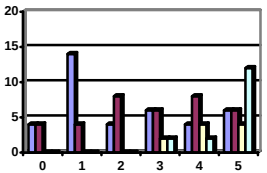
1699 patients, study 1, population 8



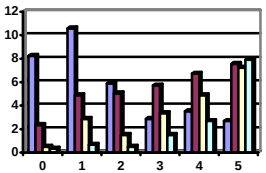
50 patients, study 1, population 12



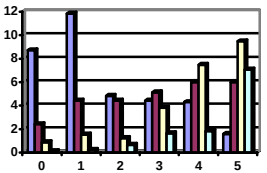
50 patients, study 3, population 14



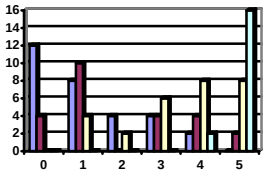
595 patients, study 3, population 4



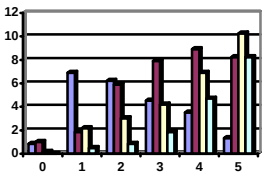
1699 patients, study 2, population 1



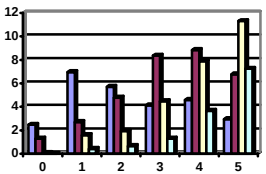
50 patients, study 2, population 93



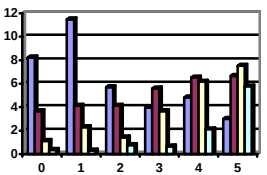
595 patients, study 1, population 8



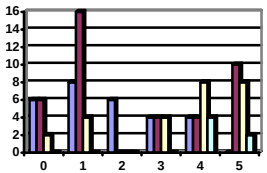
1699 patients, study 1, population 1



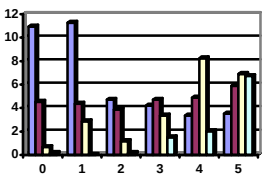
1699 patients, study 2, population 5



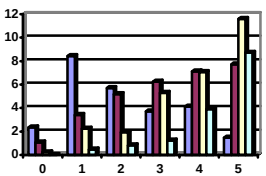
50 patients, study 2, population 64



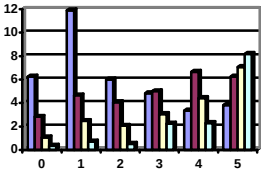
595 patients, study 2, population 1



1699 patients, study 1, population 4

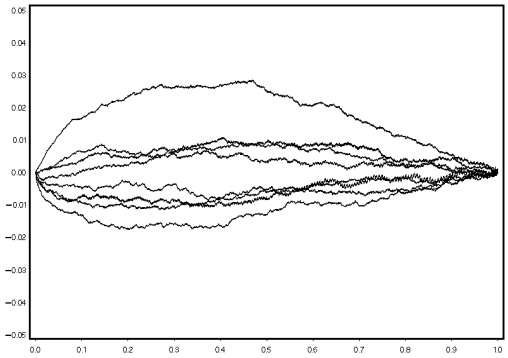


1699 patients, study 3, population 1

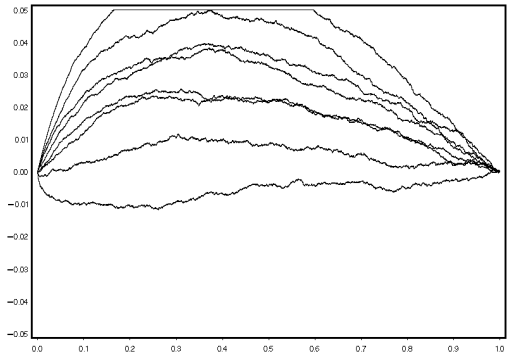


Appendix 2 Graphical presentation of $\hat{F}(u) - u$ vs. u

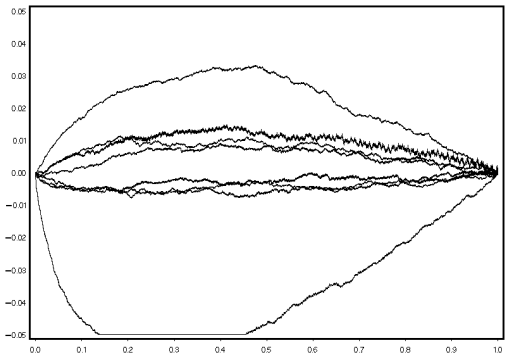
CMH test, 50 patients, study 1



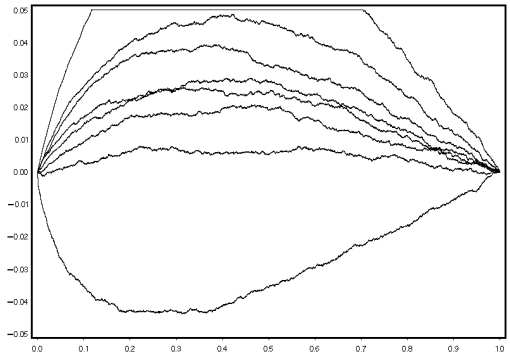
Wald test, 50 patients, study 1



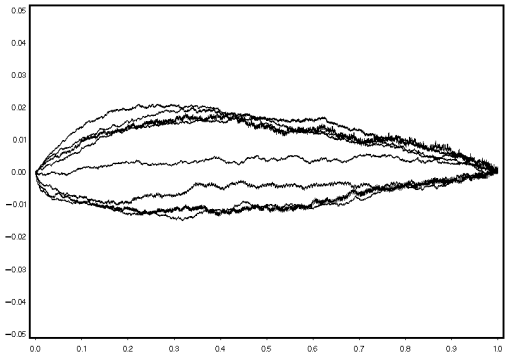
CMH test, 50 patients, study 2



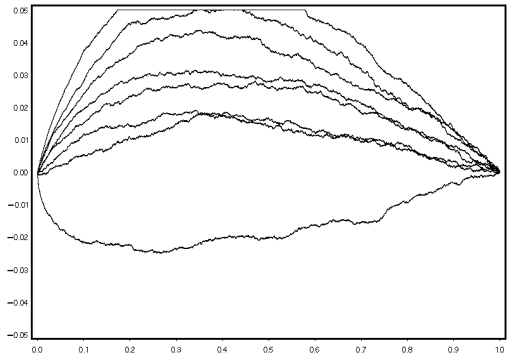
Wald test, 50 patients, study 2



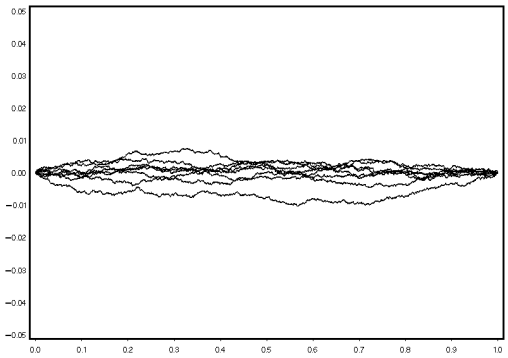
CMH test, 50 patients, study 3



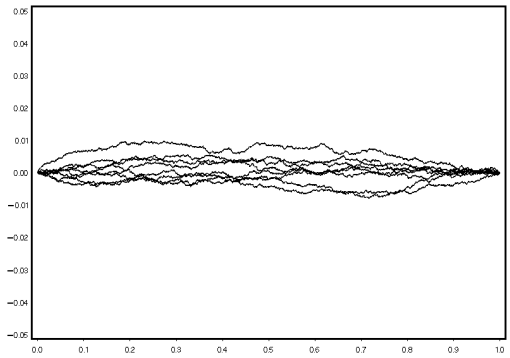
Wald test, 50 patients, study 3



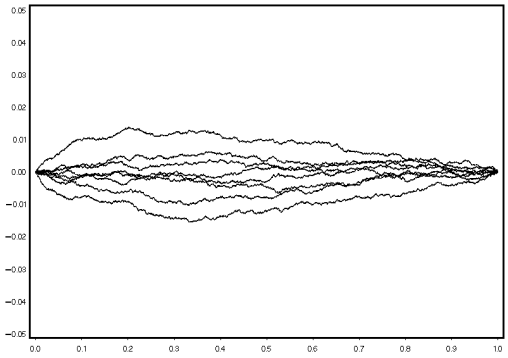
CMH test, 595 patients, study 1



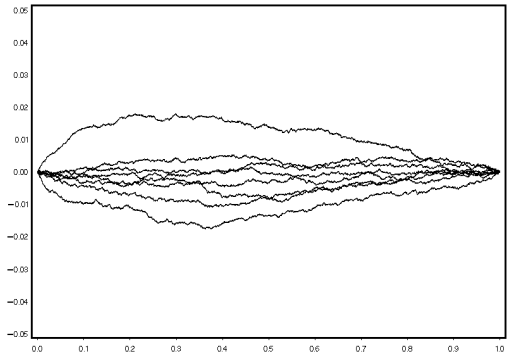
Wald test, 595 patients, study 1



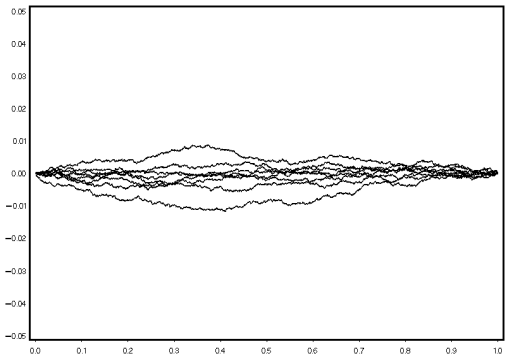
CMH test, 595 patients, study 2



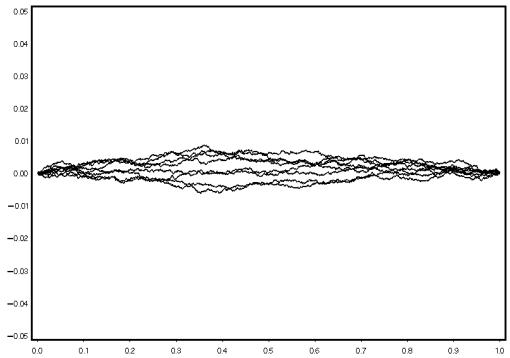
Wald test, 595 patients, study 2



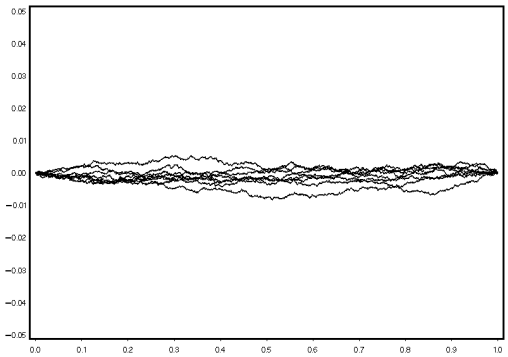
CMH test, 595 patients, study 3



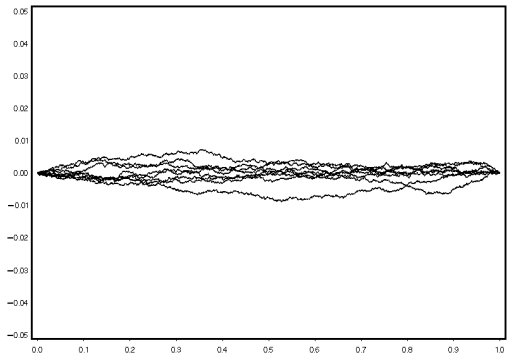
Wald test, 595 patients, study 3



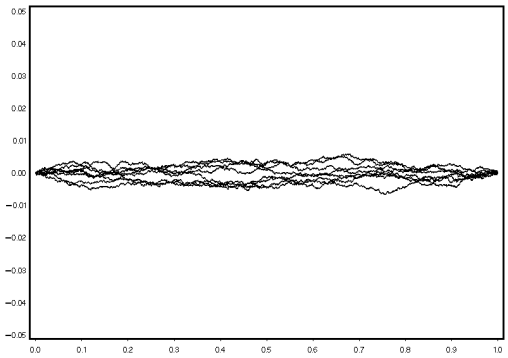
CMH test, 1699 patients, study 1



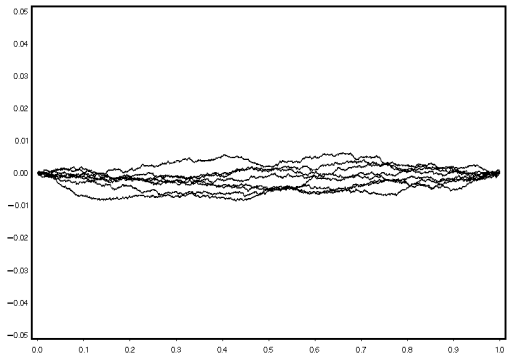
Wald test, 1699 patients, study 1



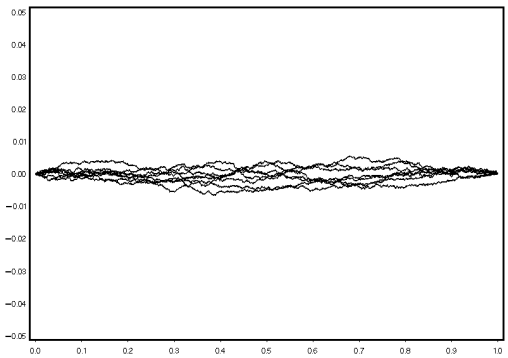
CMH test, 1699 patients, study 2



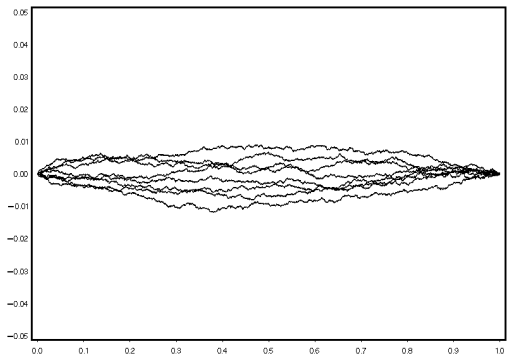
Wald test, 1699 patients, study 2



CMH test, 1699 patients, study 3

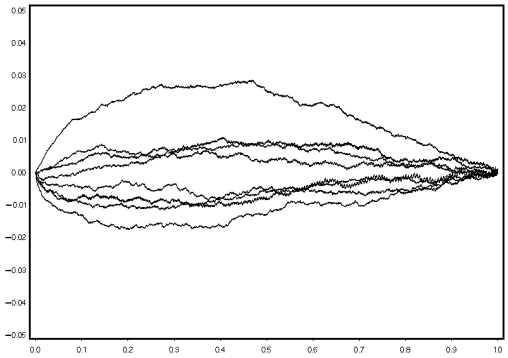


Wald test, 1699 patients, study 3

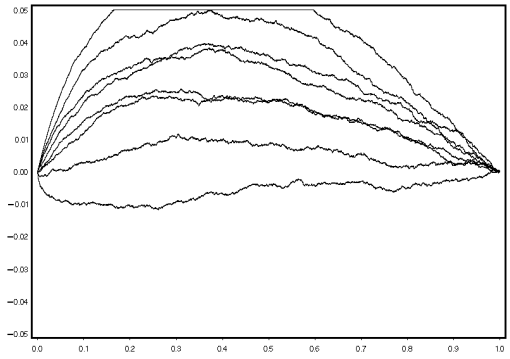


Appendix 3 Graphical presentation of $\hat{F}(u) - u$ vs. u

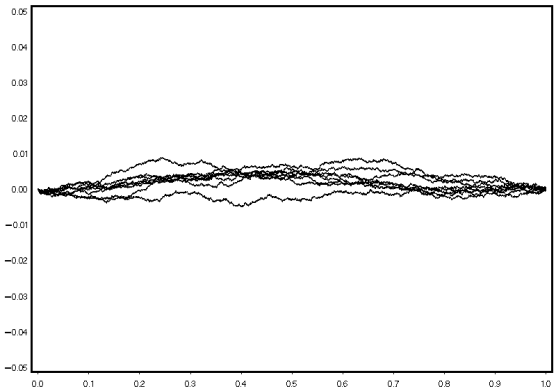
Adaptive randomisation, CMH test, 50 patients, study 1



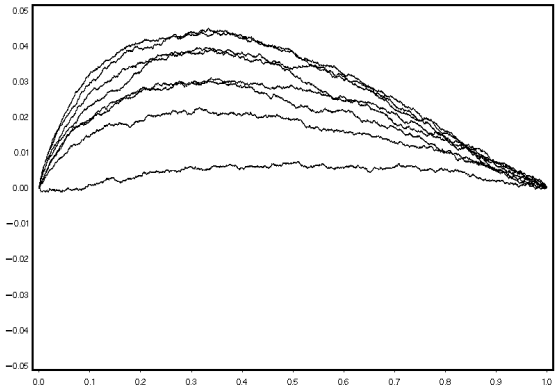
Adaptive randomisation, Wald test, 50 patients, study 1



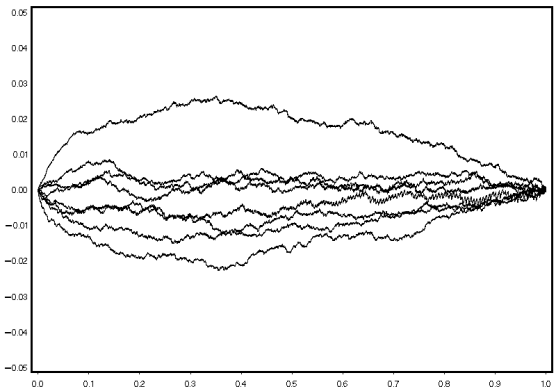
Single permuted block, CMH test, 50 patients, study 1



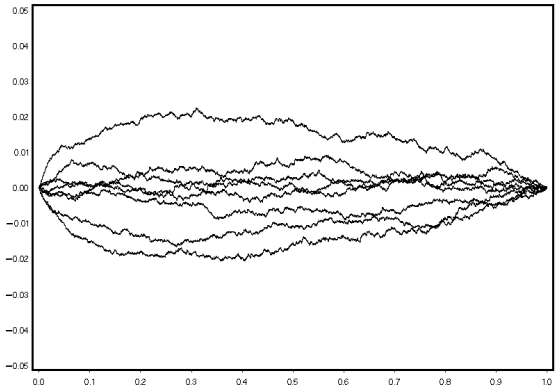
Single permuted block, Wald test, 50 patients, study 1



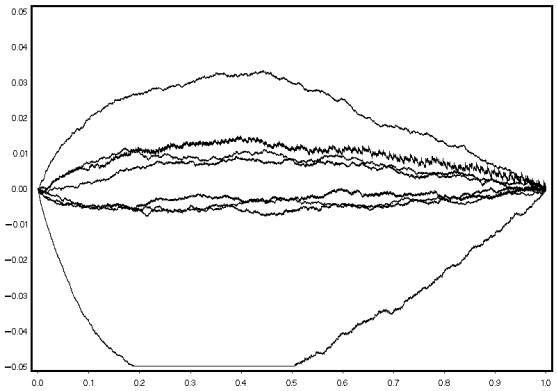
Difference between adaptive randomisation and single permuted block
CMH test, 50 patients, study 1



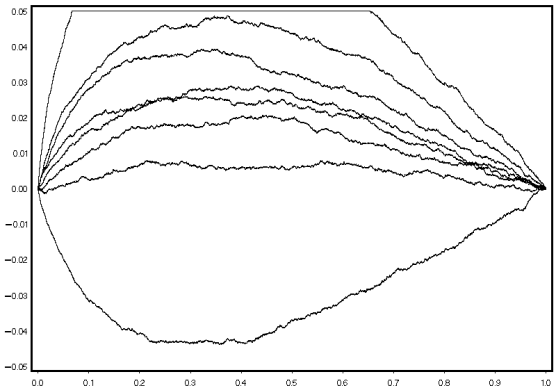
Difference between adaptive randomisation and single permuted block
Wald test, 50 patients, study 1



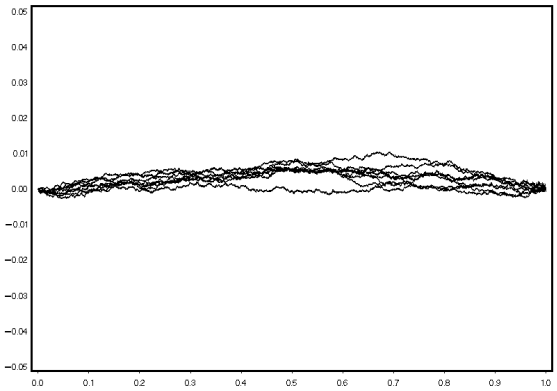
Adaptive randomisation, CMH test, 50 patients, study 2



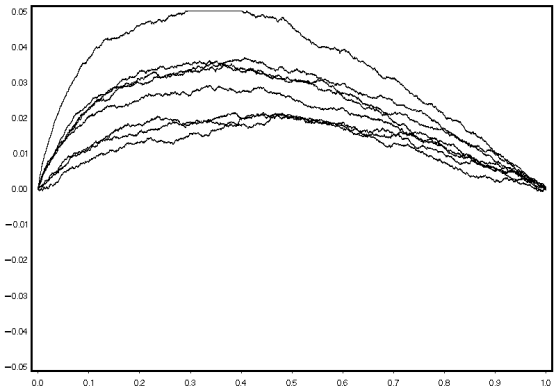
Adaptive randomisation, Wald test, 50 patients, study 2



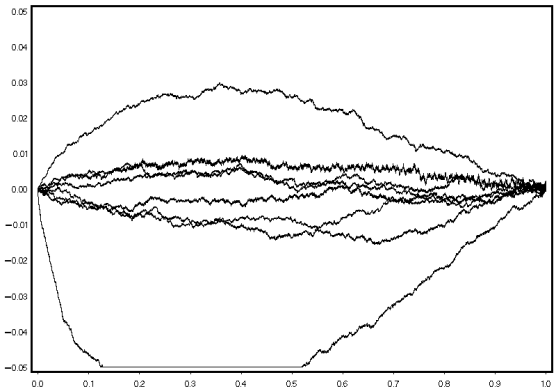
Single permuted block, CMH test, 50 patients, study 2



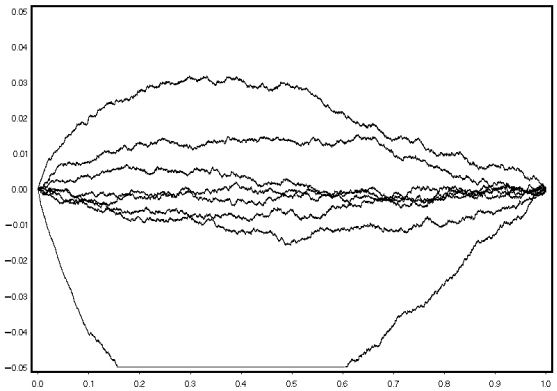
Single permuted block, Wald test, 50 patients, study 2



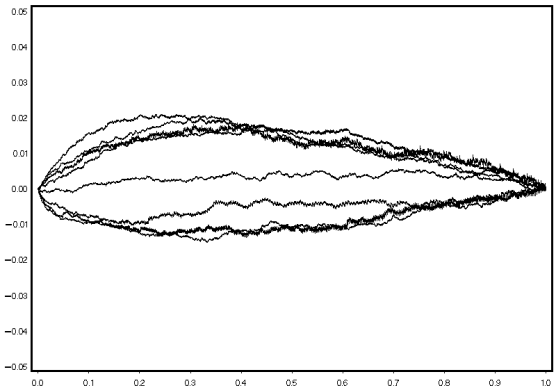
Difference between adaptive randomisation and single permuted block
CMH test, 50 patients, study 2



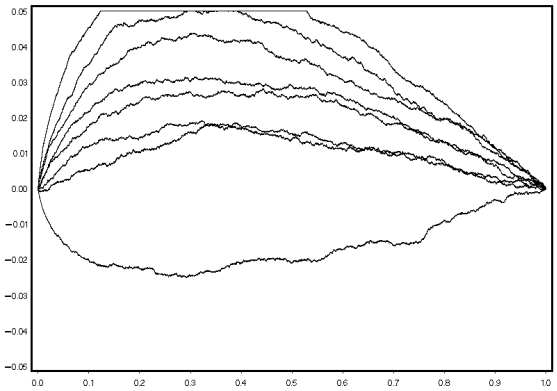
Difference between adaptive randomisation and single permuted block
Wald test, 50 patients, study 2



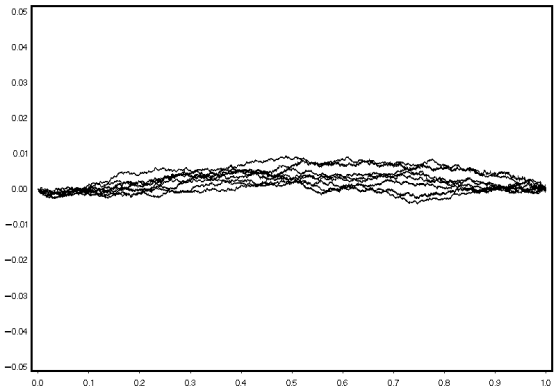
Adaptive randomisation, CMH test, 50 patients, study 3



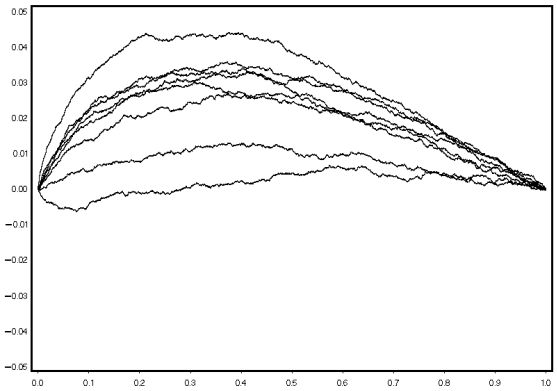
Adaptive randomisation, Wald test, 50 patients, study 3



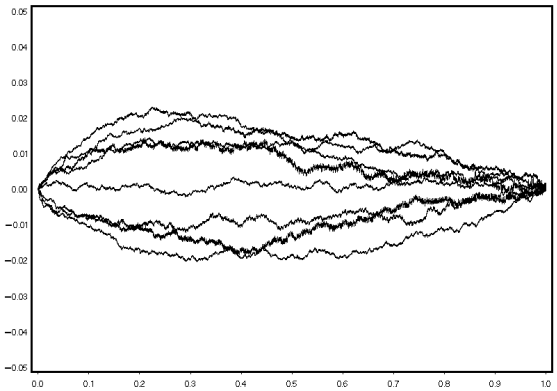
Single permuted block, CMH test, 50 patients, study 3



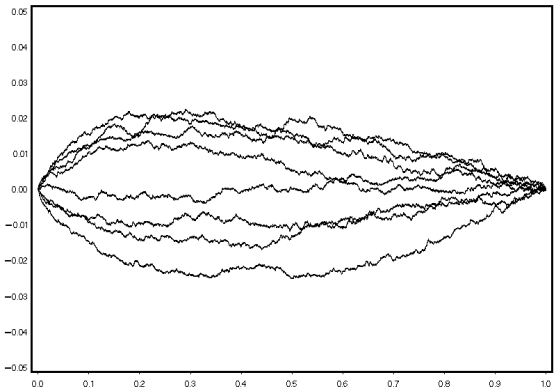
Single permuted block, Wald test, 50 patients, study 3



Difference between adaptive randomisation and single permuted block
CMH test, 50 patients, study 3



Difference between adaptive randomisation and single permuted block
Wald test, 50 patients, study 3



Effect measures in clinical trials with ordinal data

SUMMARY

In a clinical trial with an ordinal categorical response variable, a logistic regression can be applied to data under the assumption of proportional odds. An estimate of the odds ratio can be used as a measure of treatment effect. Problems arise when the assumption of proportional odds is violated. Moreover, the odds ratio has been criticised for being difficult to interpret. Instead of an odds ratio as effect measure, we propose Somers' D, or its reciprocal, Number Needed to Treat (NNT), or Mann-Whitney's U (equivalent with Somers' D). These effect measures can be used even in situations when there are covariates that need to be adjusted for in the analysis. This will be illustrated with an example. We recommend an NNT defined via Somers' D, as its reciprocal, but will also discuss alternative definitions that have been proposed.

KEY WORDS: proportional odds; Mann-Whitney's U; Somers' D; number needed to treat; covariates; prognostic factors; stratification

1 INTRODUCTION

In a clinical trial with two parallel treatment groups, the effect can be measured in a variety of ways. One approach is to choose a random pair of patients, one from each treatment group, compare the outcome of the two random patients and decide who has a preferred clinical outcome (or declare a tie). This approach leads to effect measures such as Somers' D or Mann-Whitney's U normalized by the product of the sample sizes. The Number Needed to Treat (NNT) has been suggested as another option. NNT has received considerable attention in recent years, particularly among clinicians. Primarily, this article will handle ordinal categorical response variables, but we will see that the effect measures considered can also be applied to binary and continuous response variables.

Binary and ordinal categorical data can be analysed by use of a logistic regression model. Under the assumption of proportional odds (constant odds ratio), the odds ratio can be estimated and used as a measure of treatment effect. In simple comparisons of two treatments and no covariates, through a statistical test of equality, this approach is equivalent to the non-parametric Wilcoxon-Mann-Whitney test [1].

In clinical trials, it is not unusual to have prognostic factors influencing the response variable. In this situation we want an estimate of the treatment effect adjusted for the prognostic variables. In a logistic regression model it is possible to include the prognostic factors and calculate an adjusted estimate of the odds ratio under the assumption of proportional odds. However, an odds ratio is difficult to interpret and the risk of violating the assumption of proportional odds increases with the number of prognostic variables. Instead we recommend an effect measure that corresponds to the Cochran-Mantel-Haenszel (CMH) test that is not dependent on the assumption of proportionality. The comparison between test and control treatment in terms of rankings of the response variable can be adjusted, and it will be shown that a corresponding treatment effect estimate and associated confidence interval can be calculated.

In many leading health research journals the established policy is to prefer point and interval estimates of effect measures to p-values. To the clinician, the p-value is a probabilistic abstraction that is commonly misinterpreted, in particular when dichotomised at 0.05 or some other conventional significance level: "significant" is interpreted as "real" and "non-significant" as "null", see Newcombe [2]. A p-value is the answer to the question "Is it reasonable that the

two treatment effects are the same?" but the magnitude of the difference is more interesting. To be able to judge the clinical significance of an observed difference, a statistical test with its p-value should be supplemented by a point estimate and an associated confidence interval, when this is possible.

We will restrict the discussion to the simple case where we have two parallel independent groups for comparing test treatment (T) to control treatment (C). We assume that lower values of the clinical outcome measure correspond to favourable results.

2 EFFECT MEASURES

An ideal measure of treatment effect should exhibit good interpretability and good statistical properties. We would like the effect measure to tell the clinicians how likely it is that patients benefit from the test treatment. A good effect measure can communicate information useful to assess the clinical significance of any result found in a clinical trial, and also be the basis for sample size determination in the process of designing the study. We also want to be able to calculate confidence intervals in an accurate, but not too complicated way. It is desirable that the measure can be used to compare the results from several studies in the same therapeutic area. For ordinal categorical response variables it is common to use the odds ratio as an effect measure, but we will see that there are other measures where the assumption of proportional odds is not needed. These alternative effect measures are equivalent to each other, the difference is just a matter of scaling. They are also simpler to interpret for audiences not familiar with odds ratios, and are widely recommended.

2.1 Odds ratio

Consider a clinical trial with two parallel treatment groups where we have an outcome on an ordinal scale with categories described in words such as "mild", "moderate" or "severe". To be specific, suppose we use a scale with m different categories. The frequencies in each row have a multinomial distribution and the probabilities of falling into different categories within each treatment are described in Table 1.

Table 1 Category probabilities for ordinal response

	Category			
Treatment	**1**	**2**	**…**	**m**
Test	π_{1T}	π_{2T}	…	π_{mT}
Control	π_{1C}	π_{2C}	…	π_{mC}

The cumulative probability for a patient to fall into category k or better (a lower category) will be denoted Q_{kT} for test treatment and Q_{kC} for control treatment:

$$Q_{kT} = \pi_{1T} + \pi_{2T} + \ldots + \pi_{kT} \; ; \; Q_{kC} = \pi_{1C} + \pi_{2C} + \ldots + \pi_{kC} \; , \; k = 1, \ldots , m \; .$$

Notice that $Q_{mT} = Q_{mC} = 1$. With m categories, there are (m-1) cut points between the categories. At each cut point an odds ratio can be calculated:

$$OR_k = \left\{ \frac{Q_{kT} / (1 - Q_{kT})}{Q_{kC} / (1 - Q_{kC})} \right\}$$

When all these odds ratios have a common value, i.e. OR = OR_1 = OR_2 = … = OR_{m-1}, we have proportional odds, and the common value OR is a natural measure of treatment effect in the population.

In the special case where we have a binary success/failure - type response, this odds ratio reduces to ordinary odds ratio in a 2x2 table, i.e. OR = (π_T / (1-π_T)) / (π_C / (1-π_C)), where π_T is the success rate in the population treated with test treatment and π_C the corresponding rate in the control group.

The cumulative logit model was originally proposed by Walker and Duncan [3] and later called the proportional odds model by McCullagh [4]. With our notation this model can be written

$$\log\left\{\frac{Q_{kt}}{(1-Q_{kt})}\right\} = \alpha_k + \theta x_t .$$

Here t=T or C, with x_T=1 and x_C=0, and the unknown intercept parameters α_k (k = 1, …, m-1) satisfy the condition $\alpha_1 \le \alpha_2 \le \cdots \le \alpha_{m-1}$. We can see that θ = log OR, and that θ and OR=exp(θ) are independent of the cut points k = 1, …, m-1. The Newton-Raphson method or the iterative reweighted least squares method can be used for maximum likelihood estimation of the parameters. These can be obtained through standard statistical software such as the SAS® system.

The validity of the proportional odds model can be checked by testing the assumption that all odds ratios are equal. The standard test is a score test [5]. The estimated odds ratio can be used as a summary measure of efficacy, even when the assumption of proportionality is not fulfilled. However, we do not know what happens with the properties of this estimator and corresponding interval estimator, and that would need further investigation. Violation of the assumption of equal log odds could lead to misinterpretation of the outcome in a clinical trial.

2.2 Somers' D

Somers' rank correlation index D is a modification of the Kendall tau rank correlation coefficient for the association between treatments and the response variable [6]. There is no need for proportional odds. Consider a random pair of patients where one patient is randomly selected from the group treated with test drug, whereas the other is randomly (and independently) selected from the group of patients treated with the control drug. Let T be the outcome from the patient with test drug in the pair, and C the outcome from the patient receiving control treatment. When two patients are randomly picked in this way, there are three possibilities: T < C, T = C and T > C. P(T < C) is the probability that a random patient given the test treatment has a better (lower) outcome than another random patient receiving control, and analogously for P(T > C). Here, the outcome is not restricted to an ordinal or binary response, but when the response variable is discrete, there is a positive probability for T = C. The net gain P(T < C) - P(T > C) is Somers D, which in the case of binary response reduces to success rate difference, π_T - π_C.

For an unbiased estimation of Somers' D, we will consider all possible (test, control) pairs of patients. With n_T and n_C patients in the treatments groups, let T_i, i = 1, …, n_T and C_j j = 1, …, n_C represent the outcomes in the test and control treatment groups, respectively. Let $I_D(T_i, C_j)$ represent the comparison within pairs of patients (i, j),

$$I_D(T_i, C_j) = \begin{cases} 1 & \text{when } T_i < C_j \\ 0 & \text{when } T_i = C_j \\ -1 & \text{when } T_i > C_j \end{cases}$$

There are $n_T \cdot n_C$ possible pairs of patients and we let $\hat{D}$ be the total sum of all $I_D(T_i, C_j)$ divided by the number of all possible pairs, i.e.

$$\hat{D} = \sum_{i=1}^{n_T} \sum_{j=1}^{n_C} I_D(T_i, C_j) / (n_T \cdot n_C).$$

If n_T and n_C are large, $\hat{D}$ is approximately distributed according to a normal distribution with mean D and variance V_D. Goodman and Kruskal [7] proposed an estimate $\hat{V}_D$ of V_D that could be used for inferences. This estimate $\hat{V}_D$ is implemented in the PROC FREQ procedure provided by the SAS® system. It is this estimate $\hat{V}_D$ that is used to form approximate confidence intervals for D of the form $\hat{D} \pm z_{1-\alpha/2}\sqrt{\hat{V}_D}$. The variance estimate $\hat{V}_D$ does not require any assumption about proportional odds. In particular, it is appropriate not only under the null hypothesis that the ratios in each category are the same for both treatments; H_0: $\pi_{kT} = \pi_{kC}$, k = 1, … , m. A variance estimate under this H_0 is also provided by PROQ FREQ, but will not be considered here.

It should be mentioned that Goodman and Kruskal [7] derived their asymptotic results about $\hat{D}$, V_D, and $\hat{V}_D$ under a multinomial model for the 2m frequencies in a $2 \times m$ table, where also the row totals n_T and n_C are random. It can, however, be shown through standard asymptotic results and methods that the resulting $\hat{V}_D$ under their model can be used also under the present model where the rows totals, n_T and n_C, are fixed, i.e. $(\hat{D} - D)/\sqrt{\hat{V}_D}$ is approximately N(0, 1) distributed under the present model, if n_T and n_C are large. Details regarding this are outside the scope of this article, but briefly, the key to this lies in the conditional behaviour of the 2m-nomial distribution for the frequencies in a $2 \times m$ table given the rows totals n_T and n_C.

Kraemer and Kupfer [8] have made a review of effect measures used in clinical trials. They call Somers' D the expanded success rate difference. They propose that, in all randomised clinical trials, along with reporting the p-value comparing T with C, researchers should report Somers' D, together with its associated standard error, and confidence interval.

2.3 Number needed to treat

The number needed to treat, NNT, was introduced into the medical literature by Laupacis et al. [9] in 1988 as an easily understood and "clinically useful measure of the consequences of treatment", and has been widely used since then. Originally NNT was defined as $1/(\pi_T - \pi_C)$ for trials with two balanced parallel treatment groups and binary response variables, i.e. NNT was defined as the number $n=n_T=n_C$ such that $n(\pi_T - \pi_C)=1$. This means that in average, if NNT patients are treated with each treatment, one additional patient will benefit from being treated with test treatment compared to control treatment.

It is not obvious how NNT is best defined when we have more than two ordinal response categories or the response is continuous. However, as described in section 2.2, Somers' D is

the net gain P(T < C) – P(T > C), and its reciprocal can be viewed as a number needed to treat, defined in terms of random pairs (T, C), with success and failure corresponding to the outcome T < C and T > C, respectively. For ordinal categorical (and continuous) response variables we recommend that NNT is defined in this manner as the inverse of Somers' D, i.e. $NNT_D = 1/D$, which has also been proposed by Kraemer and Kupfer [8].

In therapeutic areas where ordinal data is common, such as pain, psychiatry, stroke, multiple sclerosis, and rheumatism, it is not unusual that an NNT estimate is given alone. However, an estimate of the precision should also be given, as for any other estimated effect measure.

2.3.1 Joint outcome table specification technique

One objection to effect measures in trials with parallel treatment groups, is that they reflect an effect on a population level, and not what is truly desired, namely the expected effect for an individual patient. However, we can never observe a single individual both with test and control treatment in a trial with parallel groups, so it is not possible to estimate a truly individual treatment effect. In spite of this fact, attempts have been made to translate a treatment effect at the population level into an expected effect in the individual patient. Saver [10] has recently developed a joint outcome table specification technique. He hypothetically assumes that each patient receives both treatments, and obtains a joint (T, C)-distribution, seeking to create a sort of crossover situation. The approach developed is to ask disease experts to complete "the biologically most reasonable" joint distribution table of individual patient outcomes, given the observed marginal distributions from the parallel treatment groups [11]. The table is completed by iterative redistribution of individual patients from their destined outcomes under control therapy to their destined outcomes under test treatment. Saver introduced a number needed to treat, NNT, and a number needed to harm, NNH, separately. Saver's NNT, NNT_{Saver}, is based on the bivariate distribution and defined as $1/P(T \leq C - d)$, where d is a selected non-negative integer. Similarly, Saver's NNH, NNH_{Saver} is defined as $1/P(T \geq C + d)$.

In addition to the biologically most plausible NNT, where he lets disease experts specify the distribution, he refers to the minimum and maximum possible NNT. Saver denotes his NNT as minimum possible when it is assumed that test treatment cannot harm the patient, i.e. $NNH_{Saver} = \infty$, and the potential response of a patient under test treatment can only be either equal to, or at most one score lower than, that under the control treatment, i.e. Saver's minimum possible NNT equals $1/P(T \leq C - 1) = 1/P(T = C - 1)$ under these restrictions on the (T, C)-distribution (see example in section 4.1). Saver's maximum possible NNT is derived by completing the joint outcome table under the assumption that every patient who improves does so by the largest number of steps compatible with the marginal distributions. In the same way as the biologically most plausible NNT, both the minimum possible NNT and the maximum possible NNT are derived given the observed marginal distributions from the test and control treatment, but the joint distributions will be different.

Saver's approach can be criticised. In a clinical trial with parallel treatment groups the comparison of the test and control treatment is at group level and the responses in the treatment groups are assumed to be independent. The crucial weakness is that the joint distribution depends on judgment of the participating experts. Results in clinical trials should be built on evidence from outcome data and not on personal judgements. Other experts will probably come to other conclusions and the reproducibility of the result can be questioned. The value of NNT_{Saver} will not only vary with the opinion of the disease experts, but also depends on the choice of d and whether or not to take NNH_{Saver} into account. This leads to more than one definition of NNT_{Saver} and more than one value of NNT can be calculated. With more than one definition, there is also more than one interpretation. This can be confusing. It would be interesting to get the view of the regulatory authorities on the joint

outcome table specification technique. Saver's estimate of the biologically most plausible NNT will not be the same as Somers' D. This is not surprising, because: (a) the estimated quantities are quite different in nature, so their values have different meanings and thus are not comparable; and (b) the methods used to estimate these quantities are quite different, with Saver's method being based on expert's judgment of the joint distribution of responses.

2.3.2 Reciprocal of mean score difference

An alternative definition of NNT is the reciprocal of the mean score difference (msd) between the two compared groups. This NNT (denoted NNT_{msd}) is equal to the number of patients needed in each group for the total sum of scores in the test group to be at least 1 unit less than the total sum of scores in the control group. This can be compared with the original definition of NNT; number of patients needed per group to get one additional patient with favourable outcome in the test group than in the control group. With binary outcome data, NNT_{msd} is equal to NNT_D, but not for ordinal data and with continuous response it would be unnatural to invert the difference of the mean values.

The mean score difference is related to Saver's NNT-measures. We will now show, in terms of true rates, that the minimum possible NNT is equal to the reciprocal of mean score difference. Under the assumption for the minimum NNT (see 2.3.1), the bivariate distribution for an ordinal scale with m different categories can be found in Table II. Here Δ_k is the difference $\pi_{kC} - \pi_{kT}$, where π_{kC} is the marginal rate of patients in category k in the control group and π_{kT} is the rate of patients in the same category for patients treated with test drug.

Table 2 **Bivariate distribution for an ordinal scale assuming outcome in test treatment group is equal or one score lower than in control group**

Control treatment	Test treatment 1	2	…	m-2	m-1	m	Control distribution
1	$\pi_{1T} - \sum_{k=2}^{m} \Delta_k$	0	…	0	0	0	π_{1C}
2	$\sum_{k=2}^{m} \Delta_k$	$\pi_{2T} - \sum_{k=3}^{m} \Delta_k$	…	0	0	0	π_{2C}
3	0	$\sum_{k=3}^{m} \Delta_k$	…	0	0	0	π_{3C}
…	…	…	…	…	…	…	…
m-2	0	0	…	$\pi_{(m-2)T} - \sum_{k=(m-1)}^{m} \Delta_k$	0	0	$\pi_{(m-2)C}$
m-1	0	0	…	$\sum_{k=(m-1)}^{m} \Delta_k$	$\pi_{(m-1)T} - \Delta_m$	0	$\pi_{(m-1)C}$
m	0	0	…	0	Δ_m	π_{mT}	π_{mC}
Test distribution	π_{1T}	π_{2T}	…	$\pi_{(m-2)T}$	$\pi_{(m-1)T}$	π_{mT}	

Now Saver's minimum possible NNT equals $1/P(T = C - 1)$ in this particular bivariate distribution, and

$$P(T = C - 1) = \sum_{k=2}^{m} \Delta_k + \sum_{k=3}^{m} \Delta_k + ... + \sum_{k=(m-1)}^{m} \Delta_k + \Delta_m = \sum_{k=2}^{m} (k-1)\Delta_k = \sum_{k=2}^{m} (k-1)(\pi_{kC} - \pi_{kT}).$$

Since $\sum_{k=1}^{m} (\pi_{kC} - \pi_{kT}) = 0$, $\sum_{k=2}^{m} (k-1)(\pi_{kC} - \pi_{kT}) = \sum_{k=1}^{m} k(\pi_{kC} - \pi_{kT}) =$ msd, it follows that Saver's minimum possible NNT, $1/P(T = C - 1)$, equals the reciprocal of mean score difference, 1/msd.

This means that Saver's minimum possible NNT can be obtained directly from the mean score difference, without having to go through the bivariate distribution in Table 2. However, due to the underlying assumptions, Saver's minimum possible NNT is not useful in practice. In addition to the problem of translating a study with two parallel independent treatment groups into a crossover situation, it is unrealistic to assume that the only alternatives are that the response of a patient after receiving test treatment is equal or one score lower than after receiving control treatment for that patient.

In the binary case NNT_{msd} equals the original definition of NNT. For other response variables, due to the definition of NNT_{msd}, the interpretation depends on the scale used. Different scales will lead to more than one interpretation, not comparable with the original definition. An example of a misinterpreted NNT_{msd} is discussed in section 4.1 below.

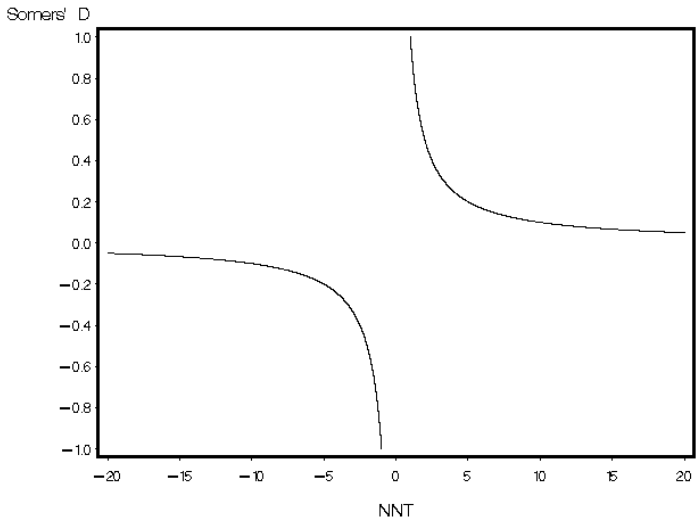
2.3.3 Dichotomisation of outcome variable before NNT estimation

In therapeutic areas where ordinal data is common, such as pain, psychiatry, stroke, multiple sclerosis, and rheumatism, it is common to dichotomise the outcome variable and perform a so-called responder analysis. The definition of a responder varies depending on the used outcome scale. A responding patient can be a patient achieving a pre-specified score on an improvement scale or having a reduction of, for example, 50% after treatment compared to baseline. The proportion of responders in each treatment group is then reported, together with the corresponding inverse of the success rate difference, i.e. NNT. What all these definitions of a responder have in common is that they need to be sanctioned by the clinicians working in the therapeutic area and the regulatory authorities. NNT is then used to compare studies and to compare competing drugs with each other. It can be questioned if its possible to compare NNTs based on different responder definitions. Originally NNT was suggested for binary outcome variables, and this may be one reason for the dichotomisation of the response variable before the NNT is calculated. Another possible explanation is that NNT is easier to derive from a dichotomised endpoint. A dichotomised outcome scale reduces the computational complexity, but discards substantial outcome information. There is no reason to waste outcome information by dichotomising before calculating NNT.

2.3.4 Criticism of NNT

A problem with NNT concerns the calculation of a confidence interval (CI), irrespectively of the distribution of the outcome variable. The CI for NNT based on Somers' D is calculated as the inverse of the CI for Somers' D. The null hypothesis of equal treatment effect, i.e. $P(T < C) - P(T > C) = 0$ (or $\pi_T = \pi_C$, in the binary case) corresponds to both $-\infty$ and ∞ on the NNT scale, see Figure 1.

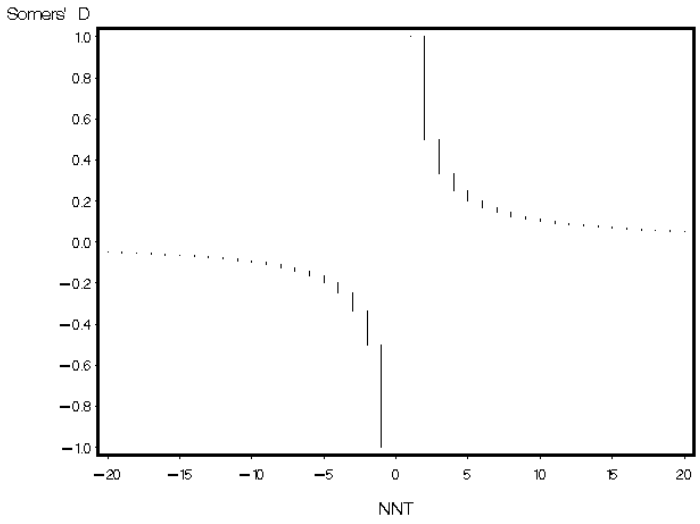
Figure 1 Somers' D versus number needed to treat



Statistically non-significant results correspond to confidence regions being a union of two disjoint intervals. This situation is easily misinterpreted. As an example, assume there is a success rate difference of 0.1 with the confidence interval –0.05 to 0.25. This leads to an NNT = 10 with the confidence limits –20 and 4. How can this be interpreted and what is the meaning of a negative NNT? As a means of resolving this, Altman [12] introduced NNTH, number of pairs of patients, one patient from each treatment group, needed for one additional patient to be harmed from test treatment, and NNTB, number of pairs needed for one additional patient to benefit. Using these descriptors, Altman suggests that the confidence interval is rewritten as NNTH 20 to ∞ to NNTB 4. This emphasises the continuity and combines the interval for NNTH from 20 to ∞ and the interval for NNTB from 4 to ∞. When the introduction of NNTB and NNTH is accepted, the confidence interval in the non-significant situation has a better chance to be understood and misinterpretations can hopefully be avoided. Altman's NNTH is not equivalent to Saver's NNH, NNH_{Saver} = 1/P(T ≥ C + d), whose value depends on the choice of d and the opinions of the disease experts. Sometimes Saver assumes that the test drug cannot harm the patients and just ignores the number needed to harm.

Usually, NNT is presented without decimals and is rounded upwards to the closest integer. This can make NNT too blunt or seriously biased. The classical advertise slogan: "9 out of 10 cover girls prefer …" cannot be described in terms of an integer NNT in a satisfactory way. Such an NNT will simply be 2. More precisely, any Somers' D of 0.50 to 0.99 will correspond to NNT=2, see Figure 2. There is no reason for NNT to be presented as an integer. NNT is an value in the same way as any mean value, and should be presented with decimals.

Figure 2 Somers' D versus number needed to treat, where NNT is an integer



The use of NNT as a clinically easily understood measure has been challenged, for example in a paper by Grieve [13]. He states: "The estimated NNT is biased, the estimate has no finite moments, the simplicity of the method of calculating a CI has unhelpful properties and doubts about the basic definition." The paper considers binary response data, but the criticism is valid more generally.

Senn's view [14] is that NNT is an extremely misleading way to summarize the results of individual trials. He argues that the collection of patients in a clinical trial is heterogeneous, and therefore it is highly unlikely that a single NNT applies to them. One could always expect to find subgroups for which the NNT was different. However, this argument is probably true also for other efficacy measures.

2.4 Mann-Whitney's U

Mann-Whitney's U or Mann-Whitney rank measure of association, allowing for ties, normalized by the product of the sample sizes, $U/n_Tn_C = P(T < C) + ½P(T = C)$, is the probability that the outcome for a random patient given test treatment (T) is better (lower) than the outcome for another randomly chosen patient receiving control (C), + half the probability that the two patients have equal outcomes. The last term takes care of possible ties, which in the continuous case is not needed, since $P(T = C) = 0$. We estimate Mann-Whitney's U/n_Tn_C as

$$\hat{U}/n_Tn_C = \sum_{i=1}^{n_T}\sum_{j=1}^{n_C} I_U(T_i,C_j)/(n_T \cdot n_C), \text{ where } I_U(T_i, C_j) = \begin{cases} 1 & \text{when } T_i < C_j \\ ½ & \text{when } T_i = C_j \\ 0 & \text{when } T_i > C_j \end{cases}$$

The relationship between Somers' D and Mann-Whitney's U/n_Tn_C is interesting. We write $D = P(T < C) - P(T > C)$ and $U/n_Tn_C = P(T< C) + ½P(T = C)$. Since $P(T < C) + P(T = C) + P(T > C) = 1$, U/n_Tn_C can be expressed as $(D + 1) / 2$. With the Wilcoxon-Mann-Whitney's test we have the null hypothesis H_0: $U/n_Tn_C = 0.5$, i.e. when 2 random patients (one from each treatment

group) are compared, none of them is more likely to have better outcome than the other. In other words, the treatment effect is equal in both groups and this is identical to the null hypothesis H_0: D = 0.

The Mann-Whitney's U statistic is widely recommended as an effect measure. Here are some examples:

- Ryu and Agresti [15] refer to the effect measure as a simple and useful way to describe the difference between two distributions of ordinal categorical variables.
- The measure is called *probability of superiority* in a paper by Grissom [16] and is referred to as "An intuitively appealing indicator of magnitude of effect in applied research is an estimate of the probability of the superior outcome of one treatment over another".
- Vargha and Delaney [17] call P(T < C) + ½P(T = C) a *measure of stochastic superiority* of C over T.
- The parameter P(T < C) is found to be "easily understood by our clinical colleagues" in a paper by Hauck et al. [18].
- Zhou [19] discusses P(T < C) in the situation where T and C are two independent normally distributed variables.
- The *common language effect size indicator* expresses how often a score sampled from one normal distribution will be greater than a score sampled from another normal distribution. The effect measure is described in a paper by McGraw and Wong [20].
- Acion et al. [21] compare *Cohen's d* for ordinal or continuous response measures with P(T < C). *Cohen's d* is the same as *standardized mean difference,* i.e. the difference between the T and C group means, divided by the within-group standard deviation. Acion et al. characterize P(T < C) "as a measure that presents good qualities of meaning, simplicity, and robustness" and provide examples with real data where its performance is contrasted with Cohen's d. Senn [22] does not agree that P(T < C) is a measure that presents good qualities of meaning, simplicity, and robustness, and gives examples when it is not.
- There is an immediate identity between Mann-Whitney's U/n_Tn_C and the *area under the curve (AUC)* measure in procedures for *receiver operator characteristic (ROC)* curve comparing responses of two treatments. If we sample a T patient and an independent C patient, AUC is the probability that the T patient has a treatment outcome preferable to the C patient (where we toss a coin to break any ties), symbolically: AUC = P(T < C) + ½P(T = C). This has been pointed out by Bamber [23], and more information can be found in a paper by Hanley and McNeil [24].

For the Mann-Whitney statistic, Newcombe [2] has written an extensive work where he compares eight different ways to calculate a confidence interval. The methods treat the distributions of T and C as continuous, but he states that they also apply to ordinal categorical responses. He recommends a pseudo-score-type confidence interval that assumes exponential distributions for T and C. Recently, Ryu and Agresti [15] compared Newcombe´s method with six other existing confidence interval methods for U/n_Tn_C for categorical outcome data. Since $U/n_Tn_C = (D + 1) / 2$ there is an unambiguous equivalence between the CI for normalized Mann-Whitney's U and the CI for Somers' D. To calculate a CI for P(T < C), when T and C are normally distributed, see Zhou [19]. The authors mentioned have only considered rather small sample sizes, less than 100 per treatment group, and unbalanced treatment groups. For non-parametric tests for proving non-inferiority in clinical trials with ordinal categorical data, see Munzel and Hauschke [25].

3 ADJUSTING FOR COVARIATES

In clinical trials covariates must be planned for and specified before the treatment code is broken. Here, we will not consider the choice of such covariates, but only assume that they

are pre-specified prognostic variables, presumably correlated with the response variable. In particular, we will consider prognostic factors used in the randomisation process when patients are allocated to treatment. If prognostic factors are important enough to be included in the randomisation process, these prognostic factors should also be included as covariates in the statistical model used for analysis. A statistical analysis can be adjusted for prognostic factors that were not used in the randomisation procedure, provided that they were predefined. The identification of prognostic factors can be done in a blind review of data. Whenever there are prognostic factors correlated with the response variable, the use of an effect measure adjusted for the prognostic factors should be considered. If the assumption of proportional odds may be violated, the adjusted Somers' D should be used.

When the odds ratio is estimated by use of logistic regression under the assumption of proportional odds, covariates can simply be added to the model described in section 2.1, whether they are binary, categorical or continuous. The model then becomes of the form

$$\log\left\{\frac{Q_{kt}}{(1-Q_{kt})}\right\} = \alpha_k + \theta x_t + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

For the test treatment group x_t=1, whereas x_t=0 in the control group, and θ = log OR. The regression coefficients β_1 to β_p correspond to the covariates x_1 to x_p. Since we assume proportional odds, θ and OR = exp(θ) do not depend on the choice of cut points, k=1, …, m-1, nor do the other regression coefficients. In this model it is assumed that the odds are proportional in all strata for categorical covariates and over the whole scale for a continuous variable. Hence, the risk for violating the assumption of proportional odds increases with the number of covariates. There are alternative regression models for ordinal responses, see Ananth and Kleinbaum [26] for a review of methods and applications. However, none of the suggested alternative models provides us with an overall measure of the treatment effect.

A method proposed when the assumption of proportional odds is not trusted is to form a Somers' D adjusted for stratification variables. This has been described by Stokes, Davis and Koch [27]. No assumption of proportional odds is needed. An adjusted Somers' D can be estimated by

$$\hat{D}_{adj} = \sum_{h=1}^{S} w_h \hat{D}_h \Big/ \sum_{h=1}^{S} w_h \text{ , with estimated variance } \hat{V}_D = \left\{ \sum_{h=1}^{S} \left(w_h s_h \right)^2 \Big/ \left(\sum_{h=1}^{S} w_h \right)^2 \right\}$$

where S is the number of strata, $\hat{D}_h$ is the estimated within-strata Somers' D value, and s_h is the within-strata standard error. The weights, w_h, have been chosen to be functions of the number of patients within strata and are calculated as $n_{hT} \cdot n_{hC} / (n_{hT}+n_{hC})$.

When the non-parametric Cochran-Mantel-Haenszel test is performed, the difference between the means in the two treatment groups is formed within strata and a weighted sum of these differences is used to combine information across strata. The weights are usually proportional to $n_{hT} \cdot n_{hC} / (n_{hT}+n_{hC})$ which is the same as the weights used in the calculation of the adjusted Somers' D. O´Gorman et al. [28] have compared two methods of estimating a risk difference in a stratified analysis. More to be read regarding stratified U can be found in the work of Koch et al. [29], where issues in clinical trials with ordinal categorical outcome are addressed.

A common value for the D_h s is not required. When the stratum-specific theoretical quantities $D_h = P(T_h < C_h) - P(T_h > C_h)$, estimated by the $\hat{D}_h$ s, have a common value, then $\hat{D}_{adj}$ estimates this common value; otherwise $\hat{D}_{adj}$ estimates a weighted mean of the stratum-specific theoretical quantities with weights. If the effect measures differ from stratum to stratum, subgroup analyses may be relevant. A homogeneity test, to test if the effect is constant over strata, can be performed. The null hypothesis that there is a common value in all strata, i.e. H_0: $D_1 = D_2 = \ldots = D_S$ can be tested with the test statistic $H = \sum_{h=1}^{S} \frac{\left(\hat{D}_h - \hat{D}'\right)^2}{s_h^2}$, where $\hat{D}' = \sum_{h=1}^{S} \hat{D}_h / s_h^2 \Big/ \sum_{h=1}^{S} 1/s_h^2$. The test statistic H is χ^2-distributed with S-1 degrees of freedom. For general information regarding this homogeneity test for large samples, see Rao [30].

When the sample size for each treatment within each stratum is sufficiently large (≥ 10) we will have an approximately normal distribution via the central limit theorem. A 95% confidence interval corresponding to $\hat{D}_{adj}$ is then $(\hat{D}_{adj} \pm 1.96\sqrt{\hat{V}_D})$.

After computation of the estimated D_{adj} with the confidence interval (D_L, D_U), a similarly adjusted Mann-Whitney's $U/n_T n_C$ is simply calculated as $(\hat{D}_{adj} + 1) / 2$, and the confidence interval as (U_L, U_U) where $U_L = (D_L + 1) / 2$ and $U_U = (D_U + 1) / 2$. Similarly, if the NNT is of interest it can simply be calculated as $NNT_{adj} = 1/D_{adj}$ with the confidence interval ($1/D_U$, $1/D_L$). Again, a non-significant effect leads to a union of two disjoint confidence intervals, here as in the unadjusted case, with analogous interpretation problems.

4 EXAMPLE

As an illustration we will use data from an acute stroke study, SAINT I [31], which was a randomised, double blind, placebo controlled, multi-centre study with two parallel treatment groups. The primary outcome in SAINT I was disability after 3 months treatment, as measured according to a 6-category ordinal scale, the modified Rankin Scale (mRS), ranging from 0 (no symptoms) to 5 (severe disability), where deaths are merged with the latter category. There were 1699 patients included to evaluate efficacy, 850 were randomised to received active drug (test) and 849 were allocated to placebo (control). We will first neglect that several prognostic factors were used when treatment was allocated to patients in the study.

4.1 Analysis without stratification

In Table 3 we can see the distribution of mRS after active and placebo treatment in SAINT I, and the odds ratio at each cut point.

Table 3 Distribution of mRS after active and placebo treatment in SAINT I

	mRS					
Treatment	**0**	**1**	**2**	**3**	**4**	**5**
Active	15.4%	18.0%	11.4%	14.2%	16.9%	24.0%
Placebo	11.0%	20.0%	11.7%	12.7%	20.6%	24.0%
Odds ratio	1.47	1.12	1.09	1.16	1	

Fitting a proportional odds model to the data resulted in an estimated odds ratio of 1.13 with a confidence interval of (0.96, 1.33). The odds ratios at each cut point range from 1 to 1.47 and the assumption of proportionality can be questioned. However, a score test for the proportional odds assumption is not quite statistically significant (p = 0.059) at a significance level of 5%.

For the calculation of Somers' D, the estimate of P(T < C) and P(T > C) is 0.432 and 0.393, respectively, resulting in Somers' $\hat{D}$ = 0.039, Mann-Whitney's $\hat{U}/n_T n_C$ = 0.52 and NNT_D = 25.6. The variance for Somers' D is obtained through the statistical software SAS® , see Goodman and Kruskal [7] for a description, which leads to $\hat{V}_D = 0.0159^2$ and a 95% confidence interval given by $0.039 \pm 1.96 \cdot 0.0159$, that is (0.008, 0 .070). After transformation we get the CI for Mann-Whitney's $U/n_T n_C$ (0.504, 0.535) and NNT_D (14.3, 127.6). Here, the validity of the inferences does not rely on the assumption of proportional odds.

In section 2.3.1 the joint outcome table specification technique was described, where Saver [10] simulates a crossover situation, given the marginal distributions under active and placebo. Saver has applied his technique to SAINT I data in Table 3. When Saver calculated what he refers to as the minimum possible NNT, he assumes that no patients were harmed by treatment with active drug and a patient receiving active substance will score equal or one unit lower, as compared to after receiving placebo. Saver obtains the estimated bivariate distribution in Table 4, and a minimum possible NNT of 7.9. Note that this value 7.9 can be obtained directly as an estimate of 1/msd without having to go through the bivariate distribution in Table 4; because of the relationship derived in section 2.3.2.

Table 4 Estimated bivariate distribution for the mRS scale assuming outcome in active group equal or one score lower than placebo, SAINT I data

	Active						
Placebo	**0**	**1**	**2**	**3**	**4**	**5/death**	**Placebo distribution**
0	0.110	0	0	0	0	0	0.110
1	0.044	0.156	0	0	0	0	0.200
2	0	0.024	0.093	0	0	0	0.117
3	0	0	0.021	0.106	0	0	0.127
4	0	0	0	0.036	0.169	0	0.206
5/death	0	0	0	0	0.001	0.240	0.240
Active distribution	0.154	0.180	0.114	0.142	0.169	0.240	

Saver's maximum possible NNT is derived by completing the joint outcome table under the assumption that every patient who improves does so by the largest number of steps

compatible with the marginal distributions. For SAINT I data this maximum possible NNT is 16.7. Saver's biologically most plausible NNT, having disease experts specify (in their opinion) the biologically most reasonable joint distribution of responses under active and control treatment. Saver estimates the biologically most plausible NNT in SAINT I to be 9.8. The method is described in the paper by Saver [11], but without confidence intervals. However, in an earlier published paper [32], the confidence interval for the biologically most plausible NNT is given as 8.7 to 10.9, based on the variability across 10 experts' estimates of the most plausible joint distribution of responses. In all of the three different NNTs calculated by Saver, he assumes that the active drug cannot harm so that NNH_{Saver} is infinity.

In a paper by Lees et al. describing the stroke study SAINT I [31], the reciprocal mean score difference (see section 2.3.2) was used as effect measure, with the following comment: "The benefit amounts to an average improvement of 0.13 points on the modified Rankin Scale per patient, which suggests that about eight patients would need to be treated to achieve improvement equal to 1 point on the scale for one patient." This is not completely true. It is the expected total sum of scores that will be one unit lower, and not one patient scoring one unit lower, when 8 patients are treated with active drug and compared to placebo.

4.2 Analysis with adjustment for covariates

The stratification variables used in SAINT I were three important prognostic variables identified at study planning:
- total NIHSS score at baseline
- side of infarct (right or left side)
- treatment with or intent to treat with a competing drug (rt-PA).

Total NIHSS score at baseline is a variable with 4 categories, and the two others are binary variables, leading to 16 strata. The within-stratum estimates of Somers' D_h and standard error s_h can be provided by the statistical software SAS® in the PROC FREQ procedure. The estimates in SAINT I are listed in Table 5.

Table 5 Somers' D within each stratum in SAINT I

NIHSS score at baseline, category	Side of infarction	Treated/ Intention to treat with rt-PA	D_h	s_h	No. of patients in active group	No. of patients in placebo group
1	Left side	No	-0.0366	0.0744	114	115
1	Left side	Yes	0.0143	0.1824	20	21
1	Right side	No	0.1035	0.0652	144	150
1	Right side	Yes	0.1252	0.1488	27	29
2	Left side	No	0.0686	0.0957	71	71
2	Left side	Yes	0.2526	0.1449	30	26
2	Right side	No	0.0710	0.0816	93	101
2	Right side	Yes	0.1028	0.1101	52	55
3	Left side	No	0.0983	0.1206	40	45
3	Left side	Yes	0.0043	0.1457	24	39
3	Right side	No	0.0758	0.0902	82	60
3	Right side	Yes	0.1198	0.1153	44	48
4	Left side	No	0.0761	0.1057	46	40
4	Left side	Yes	0.0542	0.1464	28	29
4	Right side	No	0.0033	0.1714	15	20
4	Right side	Yes	0.0600	0.2206	10	10

In some strata, some of the categories of the mRS scale were not observed in any treatment group. When that occurred, the scale was collapsed, i.e. the number of categories was reduced, in that stratum.

For these data we get $\hat{D}_{adj} = 0.058$, and the corresponding 95% confidence interval $\hat{D}_{adj} \pm 1.96\sqrt{\hat{V}_D}$ is found to be (0.005, 0.110). We can also compute a stratified Mann-Whitney's measure as a function of the Somers' D measure as $U_{adj} = (D_{adj} + 1) / 2$. With the data from SAINT I we get $U_{adj} = 0.529$ and the corresponding adjusted 95% confidence interval (0.502, 0.555). We can now use the reciprocal of the stratified Somers' D to calculate a stratified NNT for SAINT I and when that is done we directly get an estimated value of 17.4 patients with the 95% confidence interval (9.1, 212.8).

The results from the unadjusted and adjusted statistical analyses are the same, the treatment effect of active drug is small. Somers' D (CI) is 0.039 (0.008, 0.070) in the unadjusted case and 0.058 (0.005, 0.110) when we have adjusted for the three stratification variables. In the latter case the treatment effect is slightly greater than the estimate when we ignore the stratification variables. However, the uncertainty is greater and the confidence interval is wider. When we apply a logistic regression model to data under the assumption of proportional odds an estimate of the odds ratio is 1.13 with a confidence interval of (0.96, 1.33). When the model is expanded to include the 3 covariates the estimate is 1.20 for the odds ratio and (1.01, 1.42) for the confidence interval. Here the result is only just statistically significant, but the hypothesis of proportionality is rejected ($p < 0.05$). This means

that the use of a logistic regression adjusted for covariates can be questioned, due to violation of the assumption of proportional odds.

5 CONCLUSION

For ordinal categorical outcome variables, a logistic regression analysis is valid under the assumption of proportional odds. The requirement of proportional odds can never be completely fulfilled, but an odds ratio based on proportional odds can always be estimated and used as a measure of treatment effect, even when odds are not proportional. However, when the model can be questioned it is likely to give us unreliable results. When there are prognostic factors in the study that are included as covariates in the statistical model, the risk is higher that the assumption of proportionality is not satisfied. Odds ratios have been criticised as being difficult to interpret. Odds ratio gained popularity in an effort to salvage retrospective case-control studies, not primarily in randomised controlled trials, according to Kraemer and Kupfer [8]. Kraemer [33] points out that odds ratios often yield results that are puzzling or misleading and should not be considered as “gold standard”.

Instead of an odds ratio as effect measure, we recommend Somers’ D or an equivalent effect measure, such as Mann-Whitney’s U/n_Tn_C or the corresponding number needed to treat. These effect measures do not require an assumption of proportional odds. Somers’ D is a widely recommended effect measure that has good qualities of meaning and simplicity. In the binary case Somers’ D reduce to success rate difference, which is one of the most commonly used measures for binary response. Somers’ D can directly be adjusted for covariates and can be inverted to give a number needed to treat.

One should be cautious regarding number needed to treat. A problem with NNT concerns the associated confidence intervals. A statistically non-significant result leads to two disjoint intervals difficult to interpret. Also, there is disagreement on how to define NNT. Saver uses a joint outcome table specification technique for calculation of an NNT. The reciprocal of mean score differences is a related suggestion. Sometimes the response variable has been dichotomised before an NNT is calculated. That is not necessary and collapsing an ordinal or continuous variable into a binary variable reduces outcome information. Somers’ D represents the net gain, $P(T < C) – P(T > C)$, and its reciprocal is a natural definition of NNT, which reduces to the original definition of NNT in the binary case.

Standard statistical software, such as SAS®, makes the use of logistic regression easy. It is simple to get an estimate of the odds ratio, also when prognostic factors are added to the model. At the present time, when there are prognostic factors that need to be adjusted for, the estimation of Somers’ D and Mann-Whitney’s U/n_Tn_C is not as convenient. However, in this article a suggestion has been given for how an estimate of an adjusted Somers’ D and corresponding confidence interval can be obtained.

Wilcoxon-Mann-Whitney (WMW) and its extension Cochran–Mantel-Haenszel (CMH) tests should be supplemented by an effect measure. A natural measure, corresponding to the WMW test, is Somers’ D or Mann-Whitney’s U/n_Tn_C. Accordingly, the adjusted versions of Somers’ D or Mann-Whitney’s U/n_Tn_C should be used when a CMH test is used. In all randomised clinical trials with categorical ordinal outcome variables, along with the p-value comparing T with C, we recommend to report Somers’ D or Mann-Whitney’s U/n_Tn_C, as well as the corresponding standard error and a confidence interval. It can be optional to also report an NNT estimate (with decimals), but if reported, a corresponding confidence interval has to be given, and NNT should be defined as the reciprocal of Somers’ D.

References

1. Whitehead J. Sample size calculations for ordered categorical data. *Statistics in Medicine.* 1993;12:2257-2271.

2. Newcombe RG. Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 1: General issues and tail-area-based methods. *Statistics in Medicine.* 2006;25:543-557. Part 2: Asymptomatic methods and evaluation. *Statistics in Medicine.* 2006;25:559-573.

3. Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. *Biometrika*. 1967; 54:167-179.

4. McCullagh P. Regression models for ordinal data (with discussion). *J. R. Statist. Soc. B.* 1980;42:2:109-142.

5. Agresti A. *Analysis of ordinal categorical data*. Wiley. 1984.

6. Somers, R. A new asymptotic measure of association for ordinal variables. *American Sociological review*. 1962;27:799-811.

7. Goodman LA, Kruskal WH. Measures of association for cross classifications. IV: Simplification of asymptotic variances. *Journal of the American Statistical Association*. 1972;67:415-421.

8. Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry*. 2006;59:990-996.

9. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine.* 1988;318:1728-1733.

10. Saver JL. Number needed to treat estimates incorporating effects over the entire range of clinical outcomes. *Archives of Neurology*. 2004;61:1066-1070.

11. Saver JL. Novel end point analytic techniques and interpreting shifts across the entire range of outcome scales in acute stroke trials. *Stroke*. 2007;38:3055-3062.

12. Altman DG. Confidence intervals for the number needed to treat. *British Medical Journal.* 1998;317:1309-1312.

13. Grieve AP. The number needed to treat: a useful clinical measure or a case of the Emperor's new clothes? *Pharmaceutical Statistics.* 2003;2:87-102.

14. Senn S. *Statistical issues in drug development.* Wiley. 2007.

15. Ryu E, Agresti A. Modeling and inference for an ordinal effect size measure. *Statistics in Medicine.* 2007 (in press).

16. Grissom RJ. Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology.* 1994;79:2:314-316.

17. Vargha A, Delaney HD. The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational and Behavioural Statistics.* 1998;23:170-192.

18. Hauck WW, Hyslop T, Anderson S. Generalized treatment effects for clinical trials. *Statistics in Medicine.* 2000;19:887-899.

19. Zhou W. Statistical inference for P(X < Y). *Statistics in Medicine*. 2007 in press.

20. McGraw KO, Wong SP. A common language effect size statistic. *Psychological Bulletin.* 1992;111:2:361-365.

21. Acion L, Peterson JJ, Temple S, Arndt S. Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statistic in Medicine.* 2006;25:591-602.

22. Senn SJ. Letter to the editor on “Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects”. *Statistic in Medicine.* 2006;25:3944-3948.

23. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology.* 1975;12:387-415.

24. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982;143:29-36.

25. Munzel U, Hauschke D. A nonparametric test for proving noninferiority in clinical trials with ordered categorical data. *Pharmaceutical statistics.* 2003;2:31-37.

26. Ananth CV, Kleinbaum DG. Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology.* 26;6:1323-1333.

27. Stokes ME, Davis CS, Koch GG. *Categorical Data Analysis using the SAS System.* SAS Institute Inc., Cary, NC. 1995;Section 13.8:418-422.

28. O´Gorman TW, Woolson RF, Jones MP. A comparison of two methods of estimating a common risk difference in a stratified analysis of a multicenter clinical trial. *Controlled Clinical Trials.* 1994;15:135-153.

29. Koch GG, Tangen CM, Jung JW, Amara IA. Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine*. 1998;17:1863-1892.

30. Rao CR. *Linear statistical inference and its applications* (second edition). Wiley. 1973; Section 6a.2:389-391.

31. Lees KR, Zivin JA, Ashwood T, Davalos A, Davis SM, Diener HC, Grotta J, Lyden P, Shuaib A, Hårdemark HG, Wasiewski WW. NXY-059 for acute ischemic stroke. *New England Journal of Medicine.* 2006;354:588-600.

32. Saver JL. Clinical impact of NXY-059 demonstrated in the SAINT I trial: Derivation of number needed to treat for benefit over entire range of functional disability. *Stroke*. 2007;38:1515-1518.

33. Kraemer HC. Reconsidering the odds ratio as a measure of 2x2 association in a population. *Statistics in Medicine*. 2004;23:257–270.

Mathematical statistics
September 2009

www.math.su.se

Mathematical statistics
Department of Mathematics
Stockholm University
SE-106 91 Stockholm