Methodol Comput Appl Probab (2018) 20:719–738
DOI 10.1007/s11009-017-9582-3

Monitoring Phase II Comparative Clinical Trials with Two Endpoints and Penalty for Adverse Events

Sotiris Bersimis[1] · **Athanasios Sachlas**[1] · **Takis Papaioannou**[1]

Received: 21 April 2015 / Revised: 29 February 2016 /
Accepted: 14 July 2017 / Published online: 3 August 2017

Abstract Adverse events in Phase II comparative clinical trials have received limited attention in the literature. Bersimis et al. (Stat Med 34:197–214, 2014) in proposed a class of comparative sequential designs with bivariate endpoints, where as a special case, the termination of the clinical trial due to the occurrence of a severe adverse event is treated. In this paper, using the Markov chain embedding technique, we extend this class of designs proposing two new designs, which treat cases where the development of an adverse event does not immediately stop the clinical trial, but penalizes appropriately the treatment that caused it. In both designs the penalty can be chosen either by assessing the severity of the adverse event or by optimizing the power. The numerical results show an excellent performance, achieving small expected sample sizes in conjunction with large values for power, satisfying in this way the ethical requirement for small sample sizes and fast decisions in clinical practice. The formulation of the procedure as a stochastic process is elegantly accomplished while it offers the necessary mathematical framework for further generalizing the designs covering more cases such as group sequential designs, etc.

Keywords Bivariate sequences of trials · Clinical trials involving two binary endpoints · Markov chain embeddable random variables · Phase II clinical trials · Decision rules · Waiting time distribution · Adverse events · Penalization

✉ Sotiris Bersimis
sbersim@unipi.gr

Athanasios Sachlas
asachlas@unipi.gr

Takis Papaioannou
takpap@unipi.gr

[1] Department of Statistics & Insurance Science, University of Piraeus, 80 Karaoli & Dimitriou str., 185 34, Piraeus, Greece

Mathematics Subject Classification (2010) 60J20 · 62P10

1 Introduction

Comparative Phase II clinical trials with two or more endpoints have received little attention in the literature. Recently, Bersimis et al. (2014) introduced a class of designs for randomized Phase II comparative clinical trials where pairs of patients enter the trial sequentially and are randomly assigned between an experimental (E) and a reference (R) treatment. Patients' responses are sequentially measured in terms of two characteristics, Y_1^g, Y_2^g, $g = R, E$, where each of them takes two values: 1 if the treatment is successful and 0 in the opposite case. Since the enrollment is sequential, the information is also accrued sequentially as usual and the two dependent dichotomous responses are available relatively soon after treatment is administered. If Y_1^g, Y_2^g are not readily available, surrogate endpoints, or complete and partial responses, etc may be used (see for example Armitage et al. (2002), Thall and Cheng (1999), Lu et al. (2005), Thall et al. (2006), O'Connor et al. (2006), and Pryseley et al. (2010)). The binary (dichotomous) endpoints may be two efficacy endpoints or one efficacy and one safety endpoint. To early terminate the clinical trial, Bersimis et al. (2014) defined a set of decision rules based on how early a specific number of cases showing improvement due to "Treatment E" or "Treatment R" in at least one of the characteristics is observed. This is similar in nature to curtailed sampling procedures (see among others Herrmann and Szatrowski (1982, 1985) and Kunz and Kieser (2012)).

The class of designs proposed by Bersimis et al. (2014) requires immediate response which is common in Phase II clinical trials. For example, Yamanaka et al. (2003) presented the results of a clinical trial where patient's blood cells were pulsed with a tumour lysate and patients were monitored for immediate toxicities. Similar trials are encountered in oncology, cardiology, surgery, etc. (see among others Bradnock et al. (1995); Solomon et al. (2002) and Suffoletto et al. (2006)). Several authors have proposed sequential designs for monitoring clinical trials with immediate response (see for example Pryseley et al. (2010); Pocock (1977); O'Brien and Fleming (1979); Lan and DeMets (1983, 1989); Kim and Tsiatis (1990); Salvan (1990) and Kim and DeMets (1992)).

The above mentioned designs cover, among others, the case of terminating the clinical trial after the occurrence of a severe adverse event (for example, metastasis, coma, death). All cases are treated under a unified framework using the Markov chain embedding technique (see e.g. Lou (1996), Glaz et al. (2001), and Balakrishnan and Koutras (2002)) with minor modifications in the sub-matrices which constitute the transition probability matrix.

However, the design of terminating the clinical trial after an adverse event has occurred, may not be suitable for non-severe adverse events. Thus, in this paper we propose two designs that efficiently treat the case where the development of an adverse event does not immediately terminate the clinical trial, but penalizes the treatment due to which it was developed.

The paper is organized as follows: In Section 2 two new designs are defined giving the general form as well as the mathematical background. In addition, two examples of penalization are described, and the use of the Markov chain embedding technique is illustrated. In Section 3 we present the results of an extensive numerical experimentation in order to assess the performance of the new designs. More specifically, we make a power exploration with respect to the parameters of the design and we numerically justify the importance of the new design. Finally, in Section 4 we give some concluding remarks.

2 Assessing the Presence of Adverse Events

In this section we propose two designs - the standard and the superiority one - that efficiently treat the case where the development of an adverse event penalizes the responsible treatment, instead of immediately terminating the clinical trial. The standard design takes into account cumulatively all successes of the two treatments while the superiority design applies a "ties correction" strategy, which is common in clinical trials making the comparison of the two treatments more straightforward.

2.1 Penalizing the Treatment - The Standard Design

Our aim is to compare two treatments (an experimental E and a reference R one) with respect to two dependent categorical characteristics (endpoints), $Y_1^g, Y_2^g, g = R, E$. Each of the characteristics takes three values, taking into account the presence of a severe adverse event; 1 if the treatment is successful, 0 if it is unsuccessful, and $*$ if the treatment causes an adverse event, i.e. for $i = 1, 2$

$$Y_i^g = \begin{cases} 1, & \text{if treatment g is successful} \\ 0, & \text{if treatment g is unsuccessful} \\ *, & \text{if treatment g causes an adverse event.} \end{cases}$$

The potential adverse events are assumed to be associated with the characteristics we measure. This is a key assumption that is often encountered in clinical research. For example, if characteristic Y_1 is blood pressure reduction, then $Y_1^g = 1$ if g reduces blood pressure, $Y_1^g = 0$ if g does not reduce blood pressure, and $Y_1^g = *$ if g causes syncope due to low pressure (Leitch et al. 1991; Figueroa et al. 2010). Similarly, if endpoint Y_2 is diabetes regulation, then $Y_2^g = 1$ if g regulates diabetes, $Y_2^g = 0$ if g does not regulate diabetes, and $Y_2^g = *$ if g leads to a coma due to hypoglycemia (Bending et al. 1985). Adverse events (AE) not related to the characteristics have been treated in Bersimis et al. (2014).

As it is natural, the random variables Y_1^g, Y_2^g are dependent for $g = R, E$ with joint probabilities for the two dimensional random variable $\left(Y_1^g, Y_2^g\right)$

$$\pi_{ij}^g = P\left(Y_1^g = i, Y_2^g = j\right) \text{ and } \pi_{**}^g = P_g(\text{AE}),$$

where $i, j = 0, 1$, $g = R, E$ and AE stands for $\left\{Y_1^g = *, Y_2^g = *\right\} \cup \left\{Y_1^g = *, Y_2^g = 1\right\} \cup \left\{Y_1^g = *, Y_2^g = 0\right\} \cup \left\{Y_1^g = 1, Y_2^g = *\right\} \cup \left\{Y_1^g = 0, Y_2^g = *\right\}$. Note that the probability of an AE depends on the treatment. More specifically, the AE rate may be different from the control (reference) arm to the treatment arm.

Whenever treatment g causes an adverse event in either characteristic, we penalize treatment g with a penalty of size m, $m = 1, 2, 3, \ldots$. Thus the outcome under penalization for each patient receiving treatment g is given by the random variables L_s^R, L_s^E defined as

$$L_s^g = \begin{cases} 2, & \text{if } \left(Y_1^g, Y_2^g\right) = (1, 1) \\ 1, & \text{if } \left(Y_1^g, Y_2^g\right) = (1, 0) \text{ or } (0, 1) \\ 0, & \text{if } \left(Y_1^g, Y_2^g\right) = (0, 0) \\ \text{-m}, & \text{if } \left(Y_1^g, Y_2^g\right) = (*, *) \text{ or } (*, 1) \text{ or } (*, 0) \text{ or } (1, *) \text{ or } (0, *). \end{cases}$$

$g = R, E$ while s is the number of pair and $*$ means that an adverse event appears. L_s^R and L_s^E score the cases showing improvement due to "Treatment R" and "Treatment E" in each pair while they take the value $-m$ whenever "Treatment R" or "Treatment E" cause an adverse event, respectively. In other words, the design parallels two study endpoints,

combines their outcomes using a success count, and equalizes the effect of adverse events to a fixed negative integer. The design can straightforwardly be extended to assign different weights to the two characteristics, by appropriately defining the Y_i's.

The value m of penalty may be chosen depending on the severity of the adverse event and remains constant from pair to pair and from treatment to treatment. The more serious the adverse event the higher the value of m. For example, if a drug causes fever (mild adverse event) we can use $m = 1$; if it causes hemorrhage (moderate adverse event) we can use $m = 2$ while for anaphylactic shock (severe adverse event) we can use $m = 3$. This strategy gives researchers the opportunity to actively participate in the study design, deciding the value of m, using clinical criteria. Later on, we will provide mathematical directions for choosing m in order to maximize power.

The probabilities associated with L_s^R, L_s^E, i.e. $p_{uv} = P(L_s^R = u, L_s^E = v)$, $u, v = 0, 1, 2, *$, are

$$
\begin{array}{lll}
p_{00} = \pi_{00}^R \cdot \pi_{00}^E, & p_{11} = \pi_{10}^R \cdot \pi_{01}^E + \pi_{01}^R \cdot \pi_{10}^E + \pi_{01}^R \cdot \pi_{01}^E + \pi_{10}^R \cdot \pi_{10}^E, & p_{22} = \pi_{11}^R \cdot \pi_{11}^E, \\
p_{01} = \pi_{00}^R \cdot \pi_{10}^E + \pi_{00}^R \cdot \pi_{01}^E, & p_{02} = \pi_{00}^R \cdot \pi_{11}^E, & p_{12} = \pi_{10}^R \cdot \pi_{11}^E + \pi_{01}^R \cdot \pi_{11}^E, \\
p_{10} = \pi_{10}^R \cdot \pi_{00}^E + \pi_{01}^R \cdot \pi_{00}^E, & p_{20} = \pi_{11}^R \cdot \pi_{00}^E, & p_{21} = \pi_{11}^R \cdot \pi_{01}^E + \pi_{11}^R \cdot \pi_{10}^E, \\
p_{0*} = \pi_{00}^R \cdot \pi_{**}^E, & p_{1*} = \pi_{10}^R \cdot \pi_{**}^E + \pi_{01}^R \cdot \pi_{**}^E, & p_{2*} = \pi_{11}^R \cdot \pi_{**}^E, \\
p_{*0} = \pi_{**}^R \cdot \pi_{00}^E, & p_{*1} = \pi_{**}^R \cdot \pi_{01}^E + \pi_{**}^R \cdot \pi_{10}^E, & p_{*2} = \pi_{**}^R \cdot \pi_{11}^E, \\
p_{**} = \pi_{**}^R \cdot \pi_{**}^E, & &
\end{array}
\tag{1}
$$

In order to terminate the clinical trial, we use the same rule as in Bersimis et al. (2014):

- $(sr)_1$: The study is terminated in favor of "Treatment E" when a total of k cases (patients) showing improvement due to "Treatment E" in at least one of the characteristics are observed **early enough, say before the c-th pair of patients** or
- $(sr)_2$: The study is terminated in favor of "Treatment R" when a total of k cases (patients) showing improvement due to "Treatment R" in at least one of the characteristics are observed **early enough, say before the c-th pair of patients** or
- $(sr)_3$: The study is terminated with a decision that the two drugs are equivalent when a total of k cases (patients) showing improvement due to "Treatment R" or due to "Treatment E" are not observed **until the c-th pair of patients**.

To determine k and c that appear in $(sr)_1$ - $(sr)_3$ we shall use the cumulative sum of L_s^g's under the restriction that it cannot take negative values. More specifically, we define a random variable S, which counts the number of pairs of patients until one of the rules is realized. It is evident that S equals n if and only if upon the completion of the n-th comparison, we have

$$Z_s^R \geq k \text{ or } Z_s^E \geq k,$$

where

$$Z_s^g = \max\{Z_{s-1}^g + L_s^g, 0\}, s = 1, 2, \ldots$$

and none of the events stated above have occurred before the n-th pair. We take $Z_0^g = 0$. Z_s^g is the cumulative sum of L_s^g under the restriction that it cannot take negative values. As we have already mentioned, L_s^g, $g = R, E$ counts the cases showing improvement due to "Treatment g" in each pair, so it is meaningless to let Z_s^g to take negative values. In other words, S counts the number of pairs enrolled until the first appearance of either k cases showing improvement due to "Treatment R" in at least one of the characteristics or k cases showing improvement due to "Treatment E" in at least one of the two characteristics. S is a discrete variable taking values $1, 2, \ldots$.

It is obvious that the Z_s^R, Z_s^E, $s = 1, 2, \ldots$ constitute a random walk on the discrete plane, which is directly related to the maximum number of patients that will be needed

in order to make a decision, and the distribution of the "waiting time" S can be studied using the Markov chain embedding technique, where an absorbing barrier depending on k is introduced in the random walk (see e.g. Bersimis et al. (2014)).

Terminating the clinical trial in favor of Treatment R or E is associated with the probabilities of Eq. 1. The procedure for the early termination of the clinical trial can be represented as a test of hypothesis based on the distribution of S, and will be presented below. The threshold value k, the penalty m and the probabilities of Eq. 1 are parameters of the distribution of S while c is an appropriate percentile of the distribution of S, a critical value that denotes the maximum number of patients to be enrolled.

A visualization of the procedure with $m = 1$ and $m = 2$ penalizations on the first twelve pairs of patients of a clinical trial is given in Table 1. Let us first see what happens for the $m = 1$ case. At the first stage "Treatment R" is unsuccessful with respect to both characteristics and this leads to $L_1^R = 0$. Contrary, at the same stage "Treatment E" is successful with respect to both characteristics and thus we have $L_1^E = 2$. At the second stage "Treatment R" is successful with respect to the first characteristic but unsuccessful with respect to the second one. Thus we have $L_2^R = 1$. At the same stage "Treatment E" is again successful with respect to both characteristics, so we have $L_2^E = 2$. An adverse event due to "Treatment R" on the second characteristic appears for the first time at the fourth pair. This leads to $L_4^R = -1$. As far as Z_s^g, $g = R, E$ is concerned at the first stage we have $Z_1^R = L_1^R = 0$ and $Z_1^E = L_1^E = 2$. At the second stage we have $Z_2^R = Z_1^R + L_2^R = 0+1 = 1$ and $Z_2^E = Z_1^E + L_2^E = 2 + 2 = 4$. Every time that an adverse event appears, we subtract a unit from the Z_s^g, $g = R, E$. Thus, at the fourth stage that we have an adverse event due to "Treatment R", we have $Z_4^R = Z_3^R + L_4^R = 3 - 1 = 2$. At the sixth pair, "Treatment E" causes an adverse event on the first characteristic and this leads to the subtraction of a unit from the Z_s^E. At pairs 7, 8, and 9 we also have subtraction of a unit from the Z_s^R. In the case of the $m = 2$, the procedure differs only in that at stages 4, 7, 8, and 9 we subtract two units from the Z_s^R, while in the sixth stage we subtract two units from the Z_s^E. Due to the restriction that Z_s^R and Z_s^E cannot take negative values we do not make the subtraction from Z_s^R at stages 8 and 9. This restriction is used for the first time at the eighth stage of the $m = 2$ case where instead of $Z_8^R = Z_7^R + L_8^R = 1 - 2 = -1$ we set $Z_8^R = 0$. Here we have

Table 1 The clinical trial with the outcomes of the first twelve pairs of patients for the $m = 1$ and $m = 2$ penalizations under the standard design

Pair of patients (s)		1	2	3	4	5	6	7	8	9	10	11	12	…
Treatment R	Y_1^R	0	1	1	1	1	0	*	*	*	0	0	0	…
	Y_2^R	0	0	1	*	1	0	1	1	0	1	0	1	…
Treatment E	Y_1^E	1	1	0	1	1	*	0	1	0	1	1	1	…
	Y_2^E	1	1	1	0	1	1	1	0	0	1	0	0	…
$m = 1$	L_s^R	0	1	2	−1	2	0	−1	−1	−1	1	0	1	…
	L_s^E	2	2	1	1	2	−1	1	1	0	2	1	1	…
	Z_s^R	0	1	3	2	4	4	3	2	1	2	2	3	…
	Z_s^E	2	4	5	6	8	7	8	9	9	11	12	13	…
$m = 2$	L_s^R	0	1	2	−2	2	0	−2	−2	−1	1	0	1	…
	L_s^E	2	2	1	1	2	−2	1	1	0	2	1	1	…
	Z_s^R	0	1	3	1	3	3	1	0	0	1	1	2	…
	Z_s^E	2	4	5	6	8	6	7	8	8	10	11	12	…

to note that the adverse event prevails over the success or failure. As already mentioned, the proposed method can be easily modified in order to accommodate groups of patients (leading to a group sequential design). However, this while being more practical it would lead to larger sample sizes.

2.2 The Exact Distribution of S

To find the exact cumulative distribution of S, we use the Markov Chain embedding technique as in Bersimis et al. (2014). We have

$$P(S \leq n) = P(V_n = \alpha_d) = \boldsymbol{\pi}_0' \boldsymbol{\Lambda}_0^n \mathbf{e}_d,$$

and

$$P(S = n) = P(S \leq n) - P(S \leq n-1) = \boldsymbol{\pi}_1' (\boldsymbol{\Lambda}_1)^{n-1} \mathbf{h},$$

where α_d is an appropriate absorbing state, $\boldsymbol{\pi}_0'$ is the (row) vector of initial probabilities of the Markov chain, $\boldsymbol{\Lambda}_0$ is the transition probability matrix, $\mathbf{e}_d'$ is an $1 \times d$ vector of the form $(0, 0, \ldots, 1)$, $\boldsymbol{\pi}_1$ is a $(d-1) \times 1$ column vector that contains all the entries of the initial probability vector $\boldsymbol{\pi}_0$ except the last one (which, for non-degenerate cases, vanishes) and $\boldsymbol{\Lambda}_1$ is matrix $\boldsymbol{\Lambda}_0$ without the last column and the last row. The vector $\mathbf{h}$ is calculated by $\mathbf{h} = \mathbf{1} - \boldsymbol{\Lambda}_1 \mathbf{1}$, where $\mathbf{1}$ is a $(d-1) \times 1$ column vector of ones; using the fact that $\boldsymbol{\Lambda}_0$ is a stochastic matrix. The $\mathbf{h}$ contains the transition probabilities from the corresponding state to the absorbing state (for more details see Balakrishnan et al. (2009) and references therein).

The steps and notation are the same as in Bersimis et al. (2014). The main difference is that instead of using the pivotal variables $\sum_{s=1}^{n} L_s^R$ and $\sum_{s=1}^{n} L_s^E$ we use Z_s^R and Z_s^E due to the penalization and the restriction that their cumulative sum cannot take negative values. The general form of the transition probability matrix $\boldsymbol{\Lambda}_1$ is

$$\boldsymbol{\Lambda}_1 = \begin{bmatrix}
\mathbf{A}_0 + \boldsymbol{\Delta}_0 & \mathbf{B}_0 & \boldsymbol{\Gamma}_0 & & & & & & & & & & & & \\
\boldsymbol{\Delta}_1 & \mathbf{A}_1 & \mathbf{B}_1 & \boldsymbol{\Gamma}_1 & & & & & & & & & & & \\
\vdots & & \ddots & \ddots & \ddots & & & & & & & & & & \\
\vdots & & & \ddots & \ddots & \ddots & & & & & & & & & \\
\vdots & & & & \ddots & \ddots & \ddots & & & & & & & & \\
\boldsymbol{\Delta}_m & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{A}_m & \mathbf{B}_m & \boldsymbol{\Gamma}_m & & & & & & & \\
 & \boldsymbol{\Delta}_{m+1} & \mathbf{0} & \mathbf{0} & & \mathbf{0} & \mathbf{A}_{m+1} & \mathbf{B}_{m+1} & \boldsymbol{\Gamma}_{m+1} & & & & & & \\
 & & \ddots & \ddots & & & \ddots & \ddots & \ddots & \ddots & & & & & \\
 & & & \ddots & & & \ddots & \ddots & \ddots & \ddots & \ddots & & & & \\
 & & & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & & & \\
 & & & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & & \\
 & & & & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\
 & & & & & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
 & & & & & & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
 & & & & & & & & & & \ddots & \ddots & \ddots & \mathbf{A}_{k-3} & \mathbf{B}_{k-3} & \boldsymbol{\Gamma}_{k-3} \\
 & & & & & & & & & & & \ddots & \ddots & \mathbf{0} & \mathbf{A}_{k-2} & \mathbf{B}_{k-2} & \boldsymbol{\Gamma}_{k-2} \\
 & & & & & & & & & & & \boldsymbol{\Delta}_{k-1} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{k-1} & \mathbf{B}_{k-1} \\
 & & & & & & & & & & & & \boldsymbol{\Delta}_k & \mathbf{0} & \mathbf{0} & \mathbf{A}_k
\end{bmatrix}$$

where $\mathbf{A}_i = \mathbf{A}$, $i = 0, 1, \ldots, k$, $\mathbf{B}_i = \mathbf{B}$, $i = 0, 1, \ldots, k-1$, $\boldsymbol{\Gamma}_i = \boldsymbol{\Gamma}$, $i = 0, 1, \ldots, k-2$, and $\boldsymbol{\Delta}_i = \boldsymbol{\Delta}$, $i = 0, 1, \ldots, k$ are $k \times k$ matrices. Matrices $\mathbf{A}$, $\mathbf{B}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Delta}$ are described below.

Matrix $\mathbf{A}$ includes the probabilities of the transitions of the Markov chain from states of the form (j_1, j_2) to states of the form (j_1, j_2+x), where $x = 0, 1, 2$; it describes transitions of the Markov chain due to different numbers of successes of "Treatment E" while "Treatment R" is totally unsuccessful. The transition probabilities from states of the form (j_1, j_2) to states of the form $(j_1 + 1, j_2 + x)$, where $x = 0, 1, 2$ are contained in matrix $\mathbf{B}$. In other words this matrix describes transitions of the Markov chain due to different numbers of successes of "Treatment E" while "Treatment R" is successful to one of the characteristics. The transition probabilities from states of the form (j_1, j_2) to states of the form $(j_1 + 2, j_2 + x)$, where $x = 0, 1, 2$ are contained in matrix $\mathbf{\Gamma}$. Alternatively, matrix $\mathbf{\Gamma}$ describes transitions of the Markov chain due to different numbers of successes of "Treatment E" while "Treatment R" is successful in both characteristics. The transition probabilities from states of the form (j_1, j_2) to states of the form $(j_1 + y, j_2 + x)$, where $x = -m, -(m-1), \ldots, 0, 1, 2$ and $y = m$ if $j_i \geq m$ or $y = j_i$ if $j_i < m$ are contained in matrix $\mathbf{\Delta}$. In other words $\mathbf{\Delta}$ describes transitions of the Markov chain due to the penalization of "Treatment R". The empty cells of the above matrices are equal to zero. The Markov Chain describes the possible values taken by Z_s^R and Z_s^E under the restriction that Z_s^R and Z_s^E do not get negative values.

The first three elements of the first line of $\mathbf{\Lambda}_1$ are $\mathbf{A}_0 + \mathbf{\Delta}_1$, $\mathbf{B}_0$, $\mathbf{\Gamma}_0$, respectively. The sequence of $\mathbf{A}_i$, $\mathbf{B}_i$, $\mathbf{\Gamma}_i$, $i = 1, 2, \ldots, k$ begins from the second element of the second line and moves one position to the right - the full sequence $(\mathbf{A}_i, \mathbf{B}_i, \mathbf{\Gamma}_i)$ appears until the $k-1$ line of $\mathbf{\Lambda}_0$. From the second to the $m+1$ line, matrix $\mathbf{\Delta}_i$, $i = 1, \ldots, m$ appears in the first column of $\mathbf{\Lambda}_1$. From the $m+2$ line, $\mathbf{\Delta}_i$, $i = m+1, \ldots, k$ moves one position to the right. Between $\mathbf{\Delta}_i$ and $\mathbf{A}_i$, $i = m, \ldots, k$ there are $m-1$ zero matrices $\mathbf{0}$. The form of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{\Gamma}$, and $\mathbf{\Delta}$ for general m is given in Appendix. The empty cells of $\mathbf{\Lambda}_1$ are filled by $k \times k$ zero matrices $\mathbf{0}$.

It is obvious that the exact distribution of S depends on the π_{ij}^R's, the π_{ij}^E's and m and k. Moreover, m specifies matrix $\mathbf{\Delta}$ and the position of matrix $\mathbf{\Delta}$ in the transition probability matrix $\mathbf{\Lambda}_1$.

2.3 Penalizing the Treatment - The Superiority Design

A modification of the standard design, following the rationale of the Design 4.2 proposed in Bersimis et al. (2014), is to take into account only superiority of one treatment applying in that way a "tie correction" strategy. The main characteristic of this design is that none of the treatments scores in the decision process if the two treatments are both successful in terms of the same characteristic. This means that whenever, for example we have $(Y_1^R = 1, Y_2^R = 1)$ and $(Y_1^E = 1, Y_2^E = 1)$, L_s^R and L_s^E do not take the value 2 but the value 0.

The methodology remains the same as in the standard design. The stopping rules and S are as defined in Section 2.1. This means that we can find the exact cumulative distribution of S, exploiting the Markov Chain already defined without changing the general form of $\mathbf{\Lambda}_1$. The only change is associated to some of the probabilities $p_{uv} = P(L_s^R = u, L_s^E = v)$, $u, v = 0, 1, 2$, given in Eq. 1. Because of the fact that only superiority of one treatment is taken into account, it holds that $p_{22} = p_{12} = p_{21} = 0$. In this case p_{21} of Eq. 1 has been added to p_{10}, p_{12} of Eq. 1 has been added to p_{01}, and p_{22} of Eq. 1 has been added to p_{00}. Furthermore, the last two components of p_{11} of Eq. 1 has been added to p_{00}. Thus the probabilities that change are

$$
\begin{aligned}
p_{00} &= \pi_{00}^R \cdot \pi_{00}^E + \pi_{01}^R \cdot \pi_{01}^E + \pi_{10}^R \cdot \pi_{10}^E + \pi_{11}^R \cdot \pi_{11}^E, \quad p_{22} = 0, \quad p_{11} = \pi_{10}^R \cdot \pi_{01}^E + \pi_{01}^R \cdot \pi_{10}^E, \\
p_{01} &= \pi_{00}^R \cdot \pi_{10}^E + \pi_{00}^R \cdot \pi_{01}^E + \pi_{10}^R \cdot \pi_{11}^E + \pi_{01}^R \cdot \pi_{11}^E, \quad p_{12} = 0, \\
p_{10} &= \pi_{10}^R \cdot \pi_{00}^E + \pi_{01}^R \cdot \pi_{00}^E + \pi_{11}^R \cdot \pi_{01}^E + \pi_{11}^R \cdot \pi_{10}^E, \quad p_{21} = 0.
\end{aligned}
$$

Table 2 visualizes the procedure on the first twelve pairs of patients of the clinical trial described in Table 1, under the superiority case. Note that for pair 2 we have one tie and thus $(L_2^R, L_2^E) = (0, 1)$ instead of $(1, 2)$. As in the standard case, we subtract one or two units from Z_s^R at stages 4, 7, 8, and 9, for the $m = 1$ and $m = 2$ cases, respectively, while we subtract one or two units from Z_s^E at the sixth stage, for the $m = 1$ and $m = 2$ cases, respectively. The difference of the superiority design is that whenever we have a tie (i.e. both treatments are successful or unsuccessful on a characteristic) none of the treatments scores. Again, due to the restriction that Z_s^R and Z_s^E cannot take negative values we force them to be zero whenever the case arise. The tie correction leads to smaller Z_s^R and Z_s^E than in the standard case.

2.4 The Exact Distribution of S with $m = 1$ and $m = 2$ Penalizations

In this subsection we present the special cases that we penalize the responsible for an adverse event treatment with a penalty of size 1 and 2.

The transition probability matrix of the Markov Chain, for $m = 1$ and $m = 2$, is

$$\mathbf{\Lambda}_{1,m=1} = \begin{bmatrix} \mathbf{A}_0+\mathbf{\Delta}_1 & \mathbf{B}_0 & \mathbf{\Gamma}_0 & & & & & \\ \mathbf{\Delta}_1 & \mathbf{A}_1 & \mathbf{B}_1 & \mathbf{\Gamma}_1 & & & & \\ & \mathbf{\Delta}_2 & \mathbf{A}_2 & \mathbf{B}_2 & \mathbf{\Gamma}_2 & & & \\ & & \mathbf{\Delta}_3 & \mathbf{A}_3 & \mathbf{B}_3 & \ddots & & \\ & & & \ddots & \ddots & \ddots & \ddots & \\ & & & & \ddots & \ddots & \ddots & \ddots \\ & & & & & \mathbf{A}_{k-3} & \mathbf{B}_{k-3} & \mathbf{\Gamma}_{k-3} \\ & & & & & \mathbf{\Delta}_{k-2} & \mathbf{A}_{k-2} & \mathbf{B}_{k-2} & \mathbf{\Gamma}_{k-2} \\ & & & & & & \mathbf{\Delta}_{k-1} & \mathbf{A}_{k-1} & \mathbf{B}_{k-1} \\ & & & & & & & \mathbf{\Delta}_k & \mathbf{A}_k \end{bmatrix} \quad \text{and} \quad \mathbf{\Lambda}_{1,m=2} = \begin{bmatrix} \mathbf{A}_0+\mathbf{\Delta}_1 & \mathbf{B}_0 & \mathbf{\Gamma}_0 & & & & & \\ \mathbf{\Delta}_1 & \mathbf{A}_1 & \mathbf{B}_1 & \mathbf{\Gamma}_1 & & & & \\ \mathbf{\Delta}_2 & \mathbf{0} & \mathbf{A}_2 & \mathbf{B}_2 & \mathbf{\Gamma}_2 & & & \\ & \mathbf{\Delta}_3 & \mathbf{0} & \mathbf{A}_3 & \mathbf{B}_3 & \ddots & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & & & \ddots & \mathbf{A}_{k-3} & \mathbf{B}_{k-3} & \mathbf{\Gamma}_{k-3} \\ & & & & & \mathbf{0} & \mathbf{A}_{k-2} & \mathbf{B}_{k-2} & \mathbf{\Gamma}_{k-2} \\ & & & & & \mathbf{\Delta}_{k-1} & \mathbf{0} & \mathbf{A}_{k-1} & \mathbf{B}_{k-1} \\ & & & & & & \mathbf{\Delta}_k & \mathbf{0} & \mathbf{A}_k \end{bmatrix},$$

where $\mathbf{A}_i = \mathbf{A}$, $i = 0, 1, \ldots, k$, $\mathbf{B}_i = \mathbf{B}$, $i = 0, 1, \ldots, k-1$, $\mathbf{\Gamma}_i = \mathbf{\Gamma}$, $i = 0, 1, \ldots, k-2$, and $\mathbf{\Delta}_i = \mathbf{\Delta}$, $i = 0, 1, \ldots, k-1$ are $k \times k$ matrices and $\mathbf{A}$, $\mathbf{B}$, $\mathbf{\Gamma}$, $\mathbf{\Delta}$ are as given in Section 2.2. The pair $(0, 0)$ is the starting point of the Markov Chain, thus, $\boldsymbol{\pi}_0' = (1, 0, 0, \ldots, 0)_{1\times d}$.

$\mathbf{\Lambda}_{1,m=2}$ only differs from $\mathbf{\Lambda}_{1,m=2}$ in the position of the $\mathbf{\Delta}_i$, $i = 2, \ldots, k$. These are one position to the left and thus a $k \times k$ zero matrix is between $\mathbf{\Delta}_i$ and $\mathbf{A}_i$, $i = 2, \ldots, k$.

Table 2 The clinical trial with the outcomes of the first twelve pairs of patients for the $m = 1$ and $m = 2$ penalizations under the superiority design

Pair of patients (s)		1	2	3	4	5	6	7	8	9	10	11	12	…
Treatment R	Y_1^R	0	1	1	1	1	0	$*$	$*$	$*$	0	0	0	…
	Y_2^R	0	0	1	$*$	1	0	1	1	0	1	0	1	…
Treatment E	Y_1^E	1	1	0	1	1	$*$	0	1	0	1	1	1	…
	Y_2^E	1	1	1	0	1	1	1	0	0	1	0	0	…
$m = 1$	L_s^R	0	0	1	-1	0	0	-1	-1	-1	0	0	1	…
	L_s^E	2	1	0	0	0	-1	0	0	0	1	1	1	…
	Z_s^R	0	0	1	0	0	0	0	0	0	0	0	1	…
	Z_s^E	2	3	3	3	3	2	2	2	2	3	4	5	…
$m = 2$	L_s^R	0	0	1	-2	0	0	-2	-2	-2	0	0	1	…
	L_s^E	2	1	0	0	0	-2	0	0	0	1	1	1	…
	Z_s^R	0	0	1	0	0	0	0	0	0	0	0	1	…
	Z_s^E	2	3	3	3	3	2	2	2	2	3	4	5	…

Analogously we can create the matrix $\mathbf{\Lambda}_1$ for $m > 2$. We recall that the transition probability matrix is the same for both the standard and the superiority designs.

The method can be easily modified in order to accommodate different penalty values (m value) depending on which characteristic the adverse event is associated to, by appropriately manipulating the transition probabilities and the transition probability matrix. This is a key advantage of the proposed mathematical framework which is based on the formulation of the method using an appropriate stochastic process (Markov chain).

2.5 Setting up a Hypothesis Test Using the Distribution of S

Let us assume that for the cure of an illness we use a reference treatment following a distribution $F_{Y_1^R Y_2^R}$. A new treatment for the same illness is proposed which we assume that follows a distribution $F_{Y_1^E Y_2^E}$. In order to compare the two treatments using the standard or the superiority design we should set up a hypothesis test.

The hypothesis test, based on the distribution of S, is of the form

$$H_0 : F_{Y_1^R Y_2^R} = F_{Y_1^E Y_2^E},$$

i.e. the probabilities of the treatment being successful with respect to the two characteristics and the probabilities of the adverse event for both treatments are equal for both characteristics. The alternative hypothesis is

$$H_1 : F_{Y_1^R Y_2^R} \neq F_{Y_1^E Y_2^E},$$

Design parameters of the procedure are k and c.

The procedure for determining k and c and then applying the decision rule may now be described in the form of Algorithm 1. The algorithm holds for both designs. At each stage we examine all three sub-rules $(sr)_1$, $(sr)_2$, $(sr)_3$ to see if they are satisfied or not. The above procedure has the following features: (i) terminates the study as soon as the outcome is known, (ii) does not involve repeated significance testing on accumulating data (thus it protects from an increased overall significance level), (iii) allows the exact power to be easily calculated under different alternatives.

Algorithm 1 Hypothesis testing procedure

1: Define $F_{Y_1^R Y_2^R}$ (H_0) and $F_{Y_1^E Y_2^E}$ (H_1).
2: Define the desired probability of type I error, α and power γ.
3: Set the appropriate penalty m.
4: For various k's,
(i) calculate $c_{k,m}$ using $P(S \leq c_{k,m} | k, m, H_0) = \alpha$.
(ii) calculate $\gamma_{k,m}$ using $P(S \leq c_{k,m} | k, m, H_1)$.
5: Choose k and $c_{k,m}$ that give $\gamma_{k,m} \geq \gamma$, preserving α to the desired level.
6: Run the clinical trial, tracking Z_s^R, Z_s^E. If for the first time at the n-th pair,
$n \leq c$ and $Z_s^R \geq k$ interrupt the study in favor of "Treatment R".
$n \leq c$ and $Z_s^E \geq k$ interrupt the study in favor of "Treatment E".
$n > c$ and both Z_s^R, $Z_s^E \geq k$ interrupt the study with a decision that the two treatments are equivalent.

3 Power Study and Numerical Comparisons

In this section we provide a power investigation in order to evaluate the two designs. We first present numerical results for the case that the probability of an adverse event is the same (constant for the two treatments) while in the sequel we give results for the case that the probability of an adverse event is different. Finally, we numerically justify the necessity of the new designs.

3.1 Constant Adverse Event Probabilities

Assume that we have a placebo (reference treatment) following the distribution $F^{H_0}_{Y_1Y_2} = F_{Y_1^R,Y_2^R}$ given in Table 3 and that we want to identify the case where an experimental treatment is better than the reference one with level of significance α at most 0.05, power $\gamma \geq 0.90$ and keeping the sample size as small as possible. Table 4 gives the power, the expected number of pairs (in brackets) and c (i.e the maximum number of pairs of patients required to be enrolled) for $k = 5$ to 10 and the $m = 1$ and the $m = 2$ penalizations for both the standard and the superiority designs.

Let the experimental treatment follow the distribution $F^{H_1^5}_{Y_1Y_2}$. For the standard design with $m = 1$, the obvious choice is $k = 6$, as this is the smallest k which gives $\gamma \geq 0.90$. The value of the critical point c equals 9, which corresponds to an $\alpha = 0.046$. The expected number of pairs is approximately 6, so the expected number of enrolled patients is 12. For the $m = 2$ penalization, we choose $k = 7$, as this is the smallest k which gives $\gamma \geq 0.90$. The value of the critical point c equals 12, which corresponds to an $\alpha = 0.043$. The expected number of pairs is approximately 8, so the expected number of patients to be enrolled is 16 - 4 more than the previous case. The same table presents results for the other four alternative hypotheses, given above, keeping the probability of development of an adverse event constant.

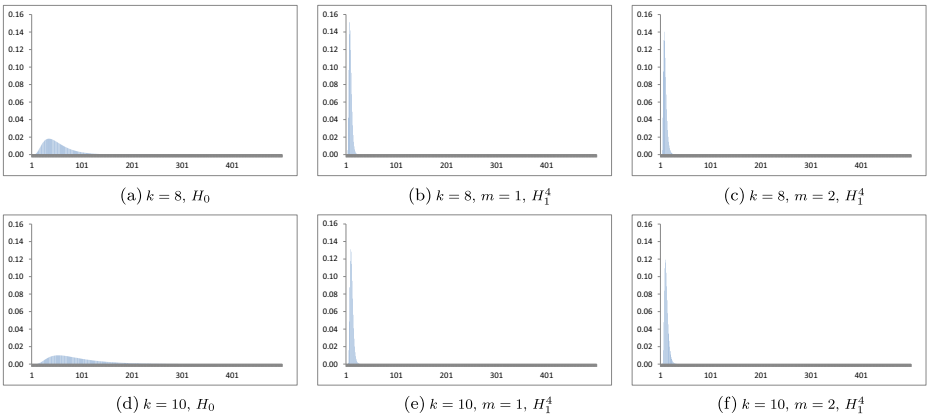
For the superiority design with $m = 1$ the obvious choice is $k = 6$. The value of the critical point c equals 10, which corresponds to an $\alpha = 0.039$. The expected number of pairs is approximately 7, so the expected number of patients to be enrolled is 14. For the $m = 2$ penalization, we choose $k = 7$, as this is the smallest k which gives $\gamma \geq 0.90$. The value of the critical point c equals 15, which corresponds to an $\alpha = 0.052$. The expected number of pairs is approximately 9, so the expected number of patients to be enrolled is 18 - 4 more than the previous case. We observe that we choose the same k ($k = 6$ for $m = 1$ and $k = 7$ for $m = 2$) regardless of the design. However, for the standard design we have to recruit two patients less, regardless m. In most cases the two designs give similar results. However, because we use a discrete test statistic we cannot have exactly the same α, and thus we cannot make straightforward comparisons.

Table 3 The null and alternative hypotheses with constant adverse event probability

		H_0		H_1^1		H_1^2		H_1^3		H_1^4		H_1^5	
		Y_2		Y_2		Y_2		Y_2		Y_2		Y_2	
		0	1	0	1	0	1	0	1	0	1	0	1
Y_1	0	0.645	0.080	0.590	0.085	0.485	0.090	0.375	0.100	0.235	0.140	0.070	0.160
	1	0.080	0.045	0.085	0.090	0.090	0.185	0.100	0.275	0.140	0.335	0.160	0.460
	AE	0.150		0.150		0.150		0.150		0.150		0.150	

Table 4 Power, expected number of pairs and c for $k = 5$ to 10 for the $m = 1$ and the $m = 2$ penalizations for several alternative hypothesis with constant probability of developing an adverse event

		Standard design							Superiority design						
		H_0	H_1^1	H_1^2	H_1^3	H_1^4	H_1^5	c	H_0	H_1^1	H_1^2	H_1^3	H_1^4	H_1^5	c
$k = 5$	$m = 1$	0.038	0.081	0.232	0.429	0.631	0.855	6	0.039	0.080	0.228	0.424	0.622	0.842	7
		(22.506)	(17.024)	(11.048)	(8.040)	(6.247)	(4.714)		(26.747)	(20.362)	(13.104)	(9.430)	(7.273)	(5.439)	
	$m = 2$	0.037	0.078	0.216	0.396	0.582	0.796	6	0.038	0.076	0.210	0.386	0.567	0.779	7
		(24.845)	(19.100)	(12.366)	(8.889)	(6.808)	(5.061)		(30.048)	(23.456)	(15.057)	(10.643)	(8.059)	(5.905)	
$k = 6$	$m = 1$	0.046	0.106	0.319	0.570	0.781	0.945	9	0.039	0.088	0.280	0.522	0.736	0.922	10
		(28.998)	(21.532)	(13.583)	(9.715)	(7.497)	(5.610)		(34.989)	(26.071)	(16.209)	(11.443)	(8.750)	(6.498)	
	$m = 2$	0.044	0.097	0.283	0.506	0.704	0.890	9	0.050	0.102	0.291	0.515	0.712	0.892	11
		(32.951)	(25.047)	(15.653)	(10.957)	(8.301)	(6.082)		(40.748)	(31.503)	(19.384)	(13.280)	(9.897)	(7.148)	
$k = 7$	$m = 1$	0.047	0.117	0.377	0.659	0.861	0.977	12	0.046	0.110	0.364	0.646	0.850	0.973	14
		(35.823)	(26.204)	(16.169)	(11.443)	(8.775)	(6.547)		(43.781)	(32.011)	(19.344)	(13.473)	(10.236)	(7.575)	
	$m = 2$	0.043	0.101	0.315	0.565	0.771	0.932	12	0.050	0.110	0.332	0.586	0.786	0.934	15
		(41.739)	(31.519)	(19.185)	(13.193)	(9.868)	(7.176)		(52.664)	(40.509)	(24.032)	(16.046)	(11.783)	(8.430)	
$k = 8$	$m = 1$	0.044	0.121	0.419	0.722	0.908	0.990	15	0.047	0.122	0.424	0.729	0.910	0.990	18
		(42.911)	(30.972)	(18.740)	(13.134)	(10.032)	(7.459)		(53.025)	(38.115)	(22.467)	(15.478)	(11.709)	(8.642)	
	$m = 2$	0.050	0.118	0.375	0.651	0.846	0.968	16	0.049	0.109	0.354	0.631	0.832	0.962	19
		(51.063)	(38.411)	(22.790)	(15.385)	(11.406)	(8.230)		(65.532)	(50.368)	(28.861)	(18.814)	(13.659)	(9.697)	
$k = 9$	$m = 1$	0.050	0.144	0.498	0.807	0.954	0.997	19	0.045	0.128	0.471	0.787	0.945	0.996	22
		(50.215)	(35.820)	(21.317)	(14.834)	(11.293)	(8.382)		(62.647)	(44.343)	(25.584)	(17.479)	(13.182)	(9.714)	
	$m = 2$	0.050	0.125	0.413	0.707	0.890	0.983	20	0.050	0.114	0.392	0.689	0.880	0.980	24
		(60.932)	(45.746)	(26.529)	(17.637)	(12.964)	(9.305)		(79.450)	(61.147)	(33.896)	(21.628)	(15.547)	(10.935)	
$k = 10$	$m = 1$	0.050	0.161	0.560	0.863	0.976	0.999	23	0.048	0.147	0.540	0.852	0.972	0.999	27
		(57.699)	(40.727)	(23.886)	(16.520)	(12.548)	(9.298)		(72.587)	(50.664)	(28.688)	(19.471)	(14.652)	(10.783)	
	$m = 2$	0.050	0.127	0.440	0.747	0.919	0.991	24	0.050	0.115	0.417	0.731	0.911	0.989	29
		(70.861)	(53.265)	(30.295)	(19.866)	(14.510)	(10.367)		(93.564)	(72.398)	(39.002)	(24.433)	(17.430)	(12.245)	



(a) $k = 8$, H_0 (b) $k = 8$, $m = 1$, H_1^4 (c) $k = 8$, $m = 2$, H_1^4

(d) $k = 10$, H_0 (e) $k = 10$, $m = 1$, H_1^4 (f) $k = 10$, $m = 2$, H_1^4

Fig. 1 The pmf of S for $k = 8, 10$ and $m = 1, 2$ for H_0 and H_1^4 under the superiority design

In both cases, as k increases the power γ increases. Furthermore, as k increases the difference between the power of the $m = 1$ and the power of the $m = 2$ case increases. The power also increases as the alternative hypothesis departs from the null hypothesis.

Figures 1a–f present the probability mass function (pmf) of S for $k = 8$ and $k = 10$ and probabilities from H_0 for the superiority design. In each row three graphs are presented - the left graph corresponds to H_0 while the second one to the alternative hypothesis H_1^4 for $m = 1$ and the third one to the same alternative hypothesis for $m = 2$. It is evident that the distribution of S has excellent discrimination properties. Similar graphs hold for the standard design as well.

3.2 Varying the Adverse Event Probability

Table 6 shows the results for $m = 1$ and $m = 2$ for both designs, allowing the probability of occurrence of an adverse event to vary. Assume that the reference treatment follows the distribution $F_{Y_1Y_2}^{H_0}$ given in Table 5 and that we want to identify the case where an experimental treatment is better than the reference one with α at most 0.05, $\gamma \geq 0.90$ and keeping the sample size as small as possible. Let the experimental treatment follow the distribution $F_{Y_1Y_2}^{H_1^4}$ in Table 5.

As far as the standard design with $m = 1$ is concerned, we choose $k = 6$ that gives $\gamma = 0.925$. The c equals 9 and the expected number of pairs is approximately 7 (i.e. 14 patients). For $m = 2$, we choose $k = 7$ that gives $\gamma = 0.949$, $c = 12$, and the expected

Table 5 The null and alternative hypotheses with varying adverse event probability

		H_0		H_1^1		H_1^2		H_1^3		H_1^4		H_1^5	
		Y_2		Y_2		Y_2		Y_2		Y_2		Y_2	
		0	1	0	1	0	1	0	1	0	1	0	1
Y_1	0	0.645	0.080	0.590	0.105	0.485	0.125	0.375	0.145	0.235	0.190	0.070	0.215
	1	0.080	0.045	0.105	0.090	0.125	0.185	0.145	0.275	0.190	0.335	0.215	0.460
	AE	0.150		0.110		0.080		0.060		0.050		0.040	

Table 6 Power, expected number of pairs and c for $k = 5$ to 10 for the $m = 1$ and the $m = 2$ penalizations for several alternative hypotheses with varying probability of developing an adverse event

		Standard design							Superiority design						
		H_0	H_1^1	H_1^2	H_1^3	H_1^4	H_1^5	c	H_0	H_1^1	H_1^2	H_1^3	H_1^4	H_1^5	c
$k = 5$	$m = 1$	0.038	0.101	0.313	0.569	0.792	0.966	6	0.039	0.101	0.314	0.569	0.786	0.957	7
		(22.506)	(14.605)	(9.063)	(6.599)	(5.223)	(4.043)		(26.747)	(17.248)	(10.535)	(7.593)	(5.982)	(4.608)	
	$m = 2$	0.037	0.098	0.302	0.550	0.769	0.947	6	0.038	0.097	0.300	0.547	0.761	0.934	7
		(24.845)	(16.070)	(9.638)	(6.858)	(5.365)	(4.114)		(30.048)	(19.414)	(11.358)	(7.950)	(6.174)	(4.702)	
$k = 6$	$m = 1$	0.046	0.141	0.448	0.747	0.925	0.996	9	0.039	0.120	0.407	0.707	0.899	0.991	10
		(28.998)	(18.177)	(10.972)	(7.892)	(6.223)	(4.795)		(34.989)	(21.616)	(12.776)	(9.092)	(7.134)	(5.478)	
	$m = 2$	0.044	0.131	0.418	0.710	0.897	0.988	9	0.050	0.143	0.444	0.737	0.909	0.989	11
		(32.951)	(20.574)	(11.828)	(8.256)	(6.418)	(4.889)		(40.748)	(25.249)	(14.009)	(9.594)	(7.397)	(5.603)	
$k = 7$	$m = 1$	0.047	0.166	0.544	0.846	0.972	0.999	12	0.046	0.161	0.541	0.844	0.969	0.999	14
		(35.823)	(21.801)	(12.884)	(9.200)	(7.933)	(5.566)		(43.781)	(26.050)	(15.006)	(10.592)	(8.290)	(6.358)	
	$m = 2$	0.043	0.147	0.493	0.801	0.949	0.997	12	0.050	0.166	0.536	0.835	0.961	0.998	15
		(41.739)	(25.319)	(14.051)	(9.676)	(7.483)	(5.686)		(52.664)	(31.489)	(16.682)	(11.243)	(8.624)	(6.514)	
$k = 8$	$m = 1$	0.044	0.182	0.616	0.904	0.989	1.000	15	0.047	0.190	0.639	0.916	0.990	1.000	18
		(42.911)	(25.454)	(14.781)	(10.491)	(8.234)	(6.326)		(53.025)	(30.518)	(17.216)	(12.081)	(9.442)	(7.233)	
	$m = 2$	0.050	0.184	0.601	0.890	0.983	1.000	16	0.049	0.176	0.598	0.890	0.982	0.999	19
		(51.063)	(30.247)	(16.270)	(11.078)	(8.538)	(6.470)		(65.532)	(38.055)	(19.352)	(12.880)	(9.847)	(7.421)	
$k = 9$	$m = 1$	0.050	0.227	0.722	0.957	0.998	1.000	19	0.045	0.211	0.711	0.953	0.997	1.000	22
		(50.215)	(29.124)	(16.671)	(11.783)	(9.237)	(7.092)		(62.647)	(35.003)	(19.415)	(13.568)	(10.594)	(8.109)	
	$m = 2$	0.050	0.206	0.675	0.936	0.994	1.000	20	0.050	0.199	0.677	0.939	0.994	1.000	24
		(60.932)	(35.351)	(18.493)	(12.482)	(9.595)	(7.260)		(79.450)	(44.936)	(22.021)	(14.515)	(11.070)	(8.330)	
$k = 10$	$m = 1$	0.050	0.263	0.797	0.980	0.999	1.000	23	0.048	0.254	0.796	0.980	0.999	1.000	27
		(57.699)	(32.803)	(18.552)	(13.071)	(10.238)	(7.854)		(72.587)	(39.493)	(21.604)	(15.052)	(11.745)	(8.985)	
	$m = 2$	0.050	0.221	0.729	0.961	0.998	1.000	24	0.050	0.213	0.733	0.964	0.998	1.000	29
		(70.861)	(40.529)	(20.707)	(13.879)	(10.649)	(8.047)		(93.564)	(51.963)	(24.677)	(16.146)	(12.291)	(9.238)	

number of pairs is approximately 8 (i.e. 16 patients). Thus for the $m = 2$ penalization we have to recruit 2 more patients obtaining, however, higher power (Table 6).

For the superiority design with $m = 1$ we choose $k = 7$, which gives power equal 0.969. The c equals 14 and the expected number of pairs is approximately 9 (i.e. 18 patients). For the $m = 2$ penalization we choose $k = 6$ that gives $\gamma = 0.909$, $c = 11$, and the expected number of pairs is approximately 8 (i.e. 16 patients). This means that we want 2 patients less on the $m = 2$ case but with lower power.

These results demonstrate an important property of the proposed method. Our method is a flexible one, in the sense that it enables researchers to choose the values of the parameters k and m with respect to the power γ and the sample size. For example, let a researcher use the superiority design and fixes k to 6. He will then choose $m = 2$ as this gives greater power ($\gamma = 0.909$ vs $\gamma = 0.899$). If the researcher fixes m to 1 (he may assume that the adverse event is non-severe) he will choose $k = 7$ as this gives power $\gamma \geq 0.90$. However, recruiting only one more pair of patients he can choose $k = 8$ that gives higher power ($\gamma = 0.990$).

3.3 Choosing *m*

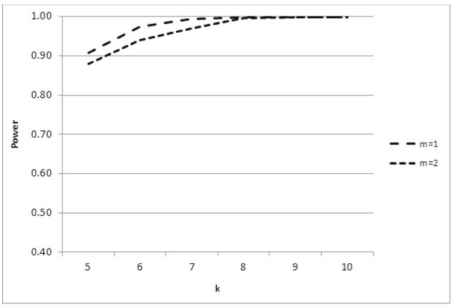
A reasonable question is how to choose the value of m. In Section 2.1 we mentioned that the doctor may choose m according to the severity of the adverse event. The more serious the adverse event the higher the value of m. However, m can be optimized in such a way to obtain the maximum power as the value of m is influenced by the probabilities of Eq. 1.

Figure 2a–d present the power for the $m = 1$ and $m = 2$ penalization for the superiority design allowing one of the parameters to change. Similar graphs hold for the standard design as well. More specifically, Fig. 2a shows the power for $m = 1$ and $m = 2$ when k changes from 5 to 10 with constant probability of adverse event. The larger the k the larger the power for both $m = 1$ and $m = 2$, with $m = 1$ giving higher power. Figure 2b shows the power for $m = 1$ and $m = 2$ with k from 5 to 10 when the reduction of the probability of adverse event affects only the probability of treatment being successful. The larger the k the larger the power for both $m = 1$ and $m = 2$, with $m = 2$ giving higher power. Figure 2c shows the power for $m = 1$ and $m = 2$ when the probability of adverse event $\pi^E_{\star\star}$ increases. The smaller the $\pi^E_{\star\star}$ the larger the power for both $m = 1$ and $m = 2$, an excellent feature of the method. Figure 2d presents the power for $m = 1$ and $m = 2$ when the difference $\delta = \pi^E_{11} - \pi^R_{11}$ increases from 10% to 50%. The larger the difference δ the larger the power for both $m = 1$ and $m = 2$. Comparing the $m = 1$ and $m = 2$ penalizations, the power curves converge as k and δ increase while the curves diverge as $\pi^E_{\star\star}$ increases.

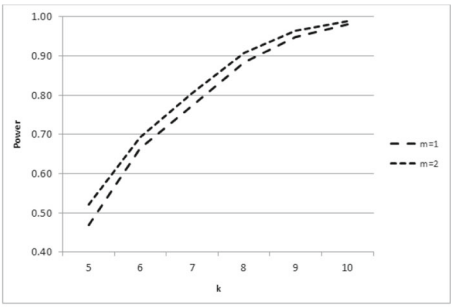
Summarizing, we can select $m = 1$ when we have constant probability of adverse event. Contrary, we can select $m = 2$ when we can assume that the reduction of the probability of adverse event affects only the probability of treatment being successful. Moreover, the interested reader can experiment numerically himself using specific practical scenarios and decide himself which m is suitable, using the Mathematica program which is available under request, or by programming Algorithm 1 by himself.

3.4 Comparison with the Design of Stopping Immediately the Clinical Trial for an Adverse Event

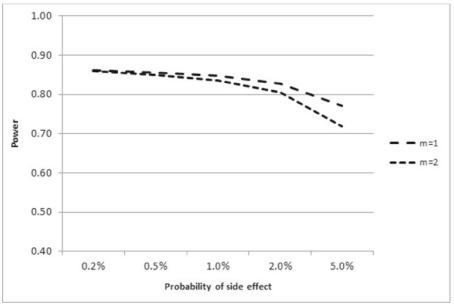
In this subsection we compare the new design using the penalization of the treatment in case of an adverse event with Design 4.3 of Bersimis et al. (2014) where the development of an adverse event immediately terminates the clinical trial. Assume that a reference treatment



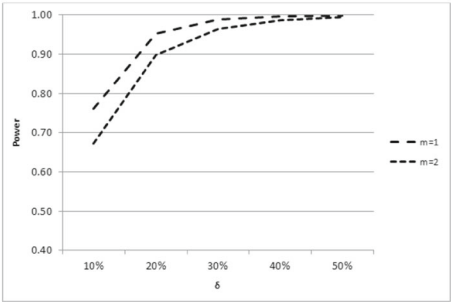
(a) $\pi_{00}^R = 0.735$, $\pi_{11}^R = 0.035$, $\pi_{00}^E = 0.235$, $\pi_{11}^E = 0.335$, $\pi_{\star\star}^R = \pi_{\star\star}^E = 0.150$



(b) $\pi_{00}^R = 0.735$, $\pi_{11}^R = 0.035$, $\pi_{\star\star}^R = 0.150$, $\pi_{00}^E = 0.235$, $\pi_{11}^E = 0.200$, $\pi_{\star\star}^E = 0.015$



(c) $k = 6$, $\pi_{1,1}^R = 0.948$, $\pi_{1,2}^R = 0.050$, $\pi_{1,3}^R = 0.002$, $\pi_{1,2}^E = 0.050$



(d) $k = 6$, $\pi_{00}^R = 0.72$, $\pi_{11}^R = 0.010$, $\pi_{\star\star}^R = \pi_{\star\star}^E = 0.010$

Fig. 2 The power against to (**a**) k with constant probability of adverse event, (**b**) k with decreasing probability of adverse event, (**c**) probability of adverse event $\pi_{1,3}^E$, and (**d**) difference $\delta = \pi_{11}^E - \pi_{11}^R$ for $m = 1$ and $m = 2$ and the superiority design

is described by the distribution $H_0 : \pi_{00}^R = 0.899, \pi_{11}^R = 0.003, \pi_{\star\star}^R = 0.004$ and that we want to identify the case where "Treatment E" with distribution $H_1 : \pi_{00}^E = 0.487, \pi_{11}^E = 0.090, \pi_{\star\star}^E = 0.004$ is better than "Treatment R" with α at most 0.05, $\gamma \geq 0.95$ and keeping the sample size as small as possible. We assume that we have a rare adverse event (e.g. the patient experiences a coma), so its probability is too small ($\pi_{\star\star}^R = 0.004$). Table 7 shows the results for the Design 4.3 and the $m = 1$ and $m = 2$ penalization under the superiority design. Design 4.3 terminates the clinical trial early enough but with very low power. In fact, as k increases, the power decreases. Using the case of penalization we select $k = 5$ which gives c equal to 18 for both $m = 1$ and $m = 2$. In both cases the expected number of pairs is approximately 10, so the expected number of patients to be enrolled is 20.

Assume now that a reference treatment causes a common non-severe adverse event such as dizziness or sleepiness. Usually such adverse events occur with high probability. For example, let a reference treatment is described by H_0 shown in Table 3 and that we want to identify the case that "Treatment E" described by H_1^5 shown in Table 3 is better than "Treatment R" with α at most 0.05, $\gamma \geq 0.90$ and keeping the sample size as small as possible. From Table 7 we see that Design 4.3 gives $c = 1$ for all k. This means that this

Table 7 Comparison of the power, expected number of pairs and c for $k = 5$ to 10 for two sets of hypotheses with low and high probability of adverse event for the Design 4.3 and the $m = 1$ and $m = 2$ penalizations under the superiority design

		Design 4.3			$m = 1$			$m = 2$		
		H_0	H_1	c	H_0	H_1	c	H_0	H_1	c
$\pi_{00}^R = 0.899, \pi_{11}^R = 0.003, \pi_{\star\star}^R = 0.004,$	$k = 5$	0.047	0.312	6	0.050	0.987	18	0.051	0.985	18
$\pi_{00}^E = 0.487, \pi_{11}^E = 0.090, \pi_{\star\star}^E = 0.004$		(32.411)	(8.271)		(40.969)	(9.084)		(41.230)	(9.122)	
	$k = 6$	0.050	0.263	7	0.048	0.997	24	0.048	0.996	24
		(38.582)	(9.816)		(50.538)	(10.857)		(50.921)	(10.907)	
	$k = 7$	0.050	0.143	7	0.050	0.999	31	0.050	0.999	31
		(44.405)	(11.340)		(60.229)	(12.628)		(60.745)	(12.691)	
	$k = 8$	0.050	0.084	7	0.050	1.000	38	0.050	0.999	38
		(49.883)	(12.842)		(70.016)	(14.399)		(70.674)	(14.474)	
$\pi_{00}^R = 0.645, \pi_{11}^R = 0.045, \pi_{\star\star}^R = 0.150,$	$k = 5$	0.278	0.278	1	0.039	0.842	7	0.038	0.779	7
$\pi_{00}^E = 0.070, \pi_{11}^E = 0.460, \pi_{\star\star}^E = 0.150$		(3.510)	(2.650)		(20.362)	(5.439)		(30.048)	(5.905)	
	$k = 6$	0.278	0.278	1	0.039	0.922	10	0.050	0.892	11
		(3.561)	(2.855)		(34.989)	(6.498)		(40.748)	(7.148)	
	$k = 7$	0.278	0.278	1	0.046	0.973	14	0.050	0.934	15
		(3.585)	(3.022)		(43.781)	(7.575)		(52.664)	(8.430)	
	$k = 8$	0.278	0.278	1	0.047	0.990	18	0.049	0.962	19
		(3.595)	(3.149)		(53.025)	(8.642)		(65.532)	(9.697)	

design tends to terminate the clinical trial at the first pair of patients and this leads to a low power. However, this is reasonable since it is likely an adverse event to immediately occur due to its high probability. Conversely, with the $m = 1$ penalization we select $k = 7$ which gives $c = 14$, $\gamma = 0.973$ and 8 pairs of patients. With the $m = 2$ penalization we select $k = 8$ which gives $c = 19$, $\gamma = 0.962$ and 10 pairs of patients.

4 Discussion

In this paper, we proposed two designs for Phase II comparative clinical trials: the standard and the superiority one. Here we consider that the development of an adverse event penalizes the corresponding treatment by m points. We presented the general m penalization of the corresponding treatment and the special cases of $m = 1$ and $m = 2$ penalization.

The new designs fall within the family of cases that can be handled with the unifying and flexible framework based on the Markov chain embedding technique introduced by Bersimis et al. (2014). The formulation of the procedure as a stochastic process can be easily accomplished and it offers the indisputable advantage of further generalizing the designs to cover more cases such as group sequential designs, more than two binary responses, more complex scoring systems implying unequal weights to the outcomes, etc., through slight modifications in the transition probability matrix.

The numerical illustration showed a very good performance of the new designs. The test terminates the clinical trial early enough (involving a small number of patients) with high power. The penalization case should be preferred over Design 4.3 proposed by Bersimis et al. (2014).

The method also offers the necessary tools in order to select penalty m based on either power values or medical practice. A rational choice for researchers is to select m with respect to the severity of the adverse event. This means that researchers can use a low penalty with a mild adverse event, a moderate penalty with a moderate adverse event and a larger penalty with a severe adverse event. An alternative solution is to select k with respect to desirable power γ and then to select m which gives the maximum power or the minimum number of pairs of patients (i.e. minimum number of patients). The numerical illustration showed that m is also influenced by the probabilities of success and adverse event (i.e. the probabilities defined in Eq. 1). For example, when the probability of adverse event is constant, the difference between H_0 and H_1 is small, and k is small, larger m gives higher power. As H_1 departs more from H_0, the $m = 1$ penalization gives higher power regardless of k. The $m = 2$ penalization also gives higher power when the reduction of the probability of adverse event is accounted for only by the probability of the experimental treatment being successful. Because several parameters are involved the researcher may use ready Mathematica programs.

The practitioner having in hand a specific problem (i.e. knowing the joint distribution for the reference treatment and assuming the joint distribution for the experimental one) can decide which of the proposed designs is preferable to his problem, guided by the power and the sample size required. The best choice is the design that gives high power with small sample size. A Mathematica program for computing the power for any choice of H_0, H_1, k, and m is available upon request.

The method evaluates individuals randomized in pairs, and involves two responses, binary or continuous converted to binary, a scoring system implying equal weights for the two outcomes and an integer penalty for the occurrence of adverse events associated with each of the outcomes. However, as already mentioned, the proposed method can

be appropriately modified to handle cases where greater weight is given in one of the characteristics or two or more adverse events are encountered.

The proposed method is essentially a pair randomization design. The pair randomization may introduce imbalance in baseline covariables as other types of randomization, such as simple randomization or block randomization, may do as well. To overcome this, the researcher should match patients according to their characteristics and then evaluate similar pairs. Another choice is to adopt a risk adjusted procedure. For the last case, the transition probabilities are computed dynamically through a logit model and thus, they depend on the patients' characteristics.

Combined with the proposed methodology, paired randomization leads to a significant reduction in sample size, which from the clinical researcher's point of view is very important in terms of time, resources, etc. The proposed method can also be modified to cover other randomization schemes or group sequential designs with group sizes larger than 2 with a slight modification in the transition probability matrix. The proposed design can also be used for Phase III trials provided that the patients' outcome is known relatively quickly.

Acknowledgements S. Bersimis and A. Sachlas are supported by the Greek General Secretariat for Research and Technology research funding action "ARISTEIA II". The authors would like to thank the Associate Editor and the two anonymous referees for their constructive comments that helped them to improve the manuscript.

Appendix: The Components of the Transition Probability Matrix

The matrices that compose the transition probability matrix $\mathbf{\Lambda}_1$ are

$$\mathbf{A} = \begin{array}{c|cccccccc} & (j_i,0) & (j_i,1) & (j_i,2) & \dots & (j_i,k-3) & (j_i,k-2) & (j_i,k-1) \\ \hline (j_i,0) & p_{00}+p_{0*} & p_{01} & p_{02} & & & & \\ (j_i,1) & p_{0*} & p_{00} & p_{01} & & & & \\ (j_i,2) & p_{0*} & & p_{00} & & & & \\ (j_i,3) & & p_{0*} & & \ddots & & & \\ \vdots & & & \ddots & & & & \\ (j_i,k-4) & & & & & p_{01} & p_{02} & \\ (j_i,k-3) & & & & & p_{00} & p_{01} & p_{02} \\ (j_i,k-2) & & & & & & p_{00} & p_{01} \\ (j_i,k-1) & & & & & p_{0*} & & p_{00} \end{array},$$

$$\mathbf{B} = \begin{array}{c|cccccccc} & (j_i+1,0) & (j_i+1,1) & (j_i+1,2) & \dots & (j_i+1,k-3) & (j_i+1,k-2) & (j_i+1,k-1) \\ \hline (j_i,0) & p_{10}+p_{1*} & p_{11} & p_{12} & & & & \\ (j_i,1) & p_{1*} & p_{10} & p_{11} & & & & \\ (j_i,2) & p_{1*} & & p_{10} & & & & \\ (j_i,3) & & p_{1*} & & \ddots & & & \\ \vdots & & & \ddots & & & & \\ (j_i,k-4) & & & & & p_{11} & p_{12} & \\ (j_i,k-3) & & & & & p_{10} & p_{11} & p_{12} \\ (j_i,k-2) & & & & & & p_{10} & p_{11} \\ (j_i,k-1) & & & & & p_{1*} & & p_{10} \end{array},$$

$$\mathbf{\Gamma} = \begin{array}{c|cccccccc} & (j_i+2,0) & (j_i+2,1) & (j_i+2,2) & \dots & (j_i+2,k-3) & (j_i+2,k-2) & (j_i+2,k-1) \\ \hline (j_i,0) & p_{20}+p_{2*} & p_{21} & p_{22} & & & & \\ (j_i,1) & p_{2*} & p_{20} & p_{21} & & & & \\ (j_i,2) & p_{2*} & & p_{20} & & & & \\ (j_i,3) & & p_{2*} & & \ddots & & & \\ \vdots & & & \ddots & & & & \\ (j_i,k-3) & & & & & p_{20} & p_{21} & p_{22} \\ (j_i,k-2) & & & & & & p_{20} & p_{21} \\ (j_i,k-1) & & & & & p_{2*} & & p_{20} \end{array},$$

$$\boldsymbol{\Delta} = \begin{array}{c|cccccccccccccccccc} & (j_i+y_i,0) & (j_i+y_i,1) & (j_i+y_i,2) & \cdots & (j_i+y_i,k-m+1) & (j_i+y_i,k-m) & (j_i+y_i,k-m-1) & \cdots & (j_i+y_i,k-4) & (j_i+y_i,k-3) & (j_i+y_i,k-2) & (j_i+y_i,k-1) \\ \hline (j_i,0) & p_{*0}+p_{**} & p_{*1} & p_{*2} & & & & & & & & & \\ (j_i,1) & p_{**} & p_{*0} & p_{*1} & & & & & & & & & \\ (j_i,2) & p_{**} & & p_{*0} & & & & & & & & & \\ \vdots & \vdots & & \ddots & \ddots & & & & & & & & \\ (j_i,k-m) & p_{**} & & & & p_{*0} & p_{*1} & p_{*2} & & & & & \\ (j_i,k-m-1) & & p_{**} & & & p_{*0} & p_{*1} & & p_{*2} & & & & \\ \vdots & \vdots & & \ddots & & \ddots & \ddots & & \ddots & & & & \\ (j_i,k-4) & & & & & & & & & p_{*0} & p_{*1} & p_{*2} & \\ (j_i,k-3) & & & & & & & & & & p_{*0} & p_{*1} & p_{*2} \\ (j_i,k-2) & & & & & & & p_{**} & & & & p_{*0} & p_{*1} \\ (j_i,k-1) & & & & & & p_{**} & & & & & & p_{*0} \end{array}.$$

$j_i = 0, 1, \ldots, k-1$ for $i = 1, 2$.

References

Armitage P, Berry G, Matthews JNS (2002) Statistical methods in medical research. 4nd Ed., Blackell Science Ltd

Balakrishnan N, Koutras MV (2002) Runs and scans with applications. Wiley, New York

Balakrishnan N, Bersimis S, Koutras MV (2009) Run and frequency quota rules in process monitoring and acceptance sampling. J Qual Technol 41(1):66–81

Bending JJ, Pickup JC, Keen H (1985) Frequency of diabetic ketoacidosis and hypoglycemic coma during treatment with continuous subcutaneous insulin infusion. Audit of medical care. Am J Med 79(6):685–691

Bersimis S, Sachlas A, Papaioannou T (2014) Flexible designs for phase II comparative clinical trials involving two response variables. Stat Med 34:197–214

Bradnock B, Law HT, Roscoe K (1995) A quantitative comparative assessment of the immediate response to high frequency ultrasound and low frequency ultrasound ('Longwave therapy') in the treatment of acute ankle sprains. Physiotherapy 82(2):78–84

Figueroa JJ, Basford JR, Low PA (2010) Preventing and treating orthostatic hypotension as easy as A, B, C. Cleveland Clin J Med 77:298–306

Glaz J, Naus J, Wallenstein S (2001) Scan statistics. Springer, New York

Herrmann N, Szatrowski TH (1982) Small sample properties of asymptotic formulas for expected sample size savings in curtailed binomial tests. Commun Stat C 1:221–245

Herrmann N, Szatrowski TH (1985) Curtailed binomial sampling procedures for clinical trials with paired data. Control Clin Trials 6:25–37

Kim K, Tsiatis AA (1990) Study duration for clinical trials with survival response and early stopping rule. Biometrics 46(1):81–92

Kim K, DeMets DL (1992) Sample size determination for group sequential clinical trials with immediate response. Stat Med 11:1391–1399

Kunz CU, Kieser M (2012) Curtailment in single-arm two-stage phase II oncology trials. Biom J 54(4):445–456

Lan KKG, DeMets DL (1983) Discrete sequential boundaries for clinical trials. Biometrika 70:659–663

Lan KKG, DeMets DL (1989) Changing frequency of interim analysis in sequential monitoring. Biometrics 45:1017–1020

Leitch JW, Klein GJ, Yee R, Leather RA, Kim YH (1991) Syncope associated with supraventricular tachycardia. an expression of tachycardia rate or vasomotor response? Circulation 85:1064–1071

Lou WYW (1996) On runs and longest run tests: a method of finite Markov chain embedding. J Am Stat Assoc 91(436):1595–1601

Lu Y, Jin H, Lamborn KR (2005) A design of phase II cancer trials using total and complete response endpoints. Stat Med 24:3155–3170

O'Brien PC, Fleming TR (1979) A multiple testing procedure for clinical trials. Biometrics 35:549–556

O'Connor PW, Li D, Freedman MS, Bar-Or A, Rice GPA, Confavreux C, Paty DW, Stewart JA, Scheyer R (2006) A Phase II study of the safety and efficacy of teriflunomide in multiple sclerosis with relapses. Neurology 66:894–900

Pocock SJ (1977) Group sequential methods in the design and analysis of clinical trials. Biometrika 64:191–199

Pryseley A, Tilahun A, Alonso A, Molenberghs G (2010) Using earlier measures in a longitudinal sequence as potential surrogate for a later one. Comput Stat Data Anal 54:1342–1354

Salvan A (1990) Planning sequential clinical trials: a review. Comput Stat Data Anal 9:47–56

Solomon MJ, Young CJ, Eyers AA, Roberts RA (2002) Randomized clinical trial of laparoscopic versus open abdominal rectopexy for rectal prolapse. Br J Surg 89:35–39

Suffoletto MS, Dohi K, Cannesson M, Saba S, Gorcsan J (2006) Novel Speckle-tracking radial strain from routine black-and-white echocardiographic images to quantify dyssynchrony and predict response to cardiac resynchronization therapy. Circulation 113:960–968

Thall PF, Cheng SC (1999) Treatment comparisons based on two-dimensional safety and efficacy alternatives in oncology trials. Biometrics 55:746–753

Thall PF, Wooten LH, Shpall EJ (2006) A geometric approach to comparing treatments for rapidly fatal diseases. Biometrics 62:193–201

Yamanaka R, Abe T, Yajima N, Tsuchiya N, Homma J, Kobayashi T, Narita M, Takahashi M, Tanaka R (2003) Vaccination of recurrent glioma patients with tumour lysate-pulsed dendritic cells elicits immune responses: results of a clinical phase I/II trial. Br J Cancer 89:1172–1179