Design and Monitoring of Multi-Arm Multi-Stage Clinical Trials

Pranab Ghosh, Lingyun Liu, P. Senchaudhuri, Ping Gao, and Cyrus Mehta

1Cytel Inc., Cambridge, Massachusetts, U.S.A.
2Boston University, Boston, Massachusetts, U.S.A.
3The Medicines Company, Parsippany, New Jersey, U.S.A.
4Harvard T.H. Chan School of Public Health, Boston, Massachusetts, U.S.A.

Summary. Two-arm group sequential designs have been widely used for over 40 years, especially for studies with mortality endpoints. The natural generalization of such designs to trials with multiple treatment arms and a common control (MAMS designs) has, however, been implemented rarely. While the statistical methodology for this extension is clear, the main limitation has been an efficient way to perform the computations. Past efforts were hampered by algorithms that were computationally explosive. With the increasing interest in adaptive designs, platform designs, and other innovative designs that involve multiple comparisons over multiple stages, the importance of MAMS designs is growing rapidly. This article provides break-through algorithms that can compute MAMS boundaries rapidly thereby making such designs practical. For designs with efficacy-only boundaries the computational effort increases linearly with number of arms and number of stages. For designs with both efficacy and futility boundaries the computational effort doubles with successive increases in number of stages.

Key words: Adaptive designs; Computation of stopping boundaries; Early efficacy stopping; Futility stopping; Generalized Dunnett tests; Group sequential; Multiple comparisons; Platform designs; Spending functions; Strong control of type-I error.

1. Introduction

Group sequential design and monitoring of randomized clinical trials has played an important role in clinical drug development for over 40 years. The literature on group sequential methods is vast and has been presented in a unified manner in outstanding textbooks by Jennison and Turnbull (2000) and Proschan, Lan, and Wittes (2006). Throughout its 40-year history, the development and use of group sequential methodology has been largely confined to two-arm trials, in which a proposed new medical intervention is compared to a standard-of-care control arm with respect to a single primary endpoint. There is increasing interest, however, in extending the methodology to multiple pair-wise comparisons to a common control – the so-called multi-arm multi-stage (MAMS) case. Parmar, Carpenter, and Sydes (2014) have argued persuasively that multiarm trials allow more treatments to be assessed in less time than could ever be done in a series of two-arm trials. Equally important, for patients and policy makers, the multiarm trial produces contemporaneous results for all research treatments. Jaki (2016) compared the sample size requirements of separate two-arm trials to those of multi-arm trials for Alzheimer’s disease treatments and showed savings of up to 50%.

A major hurdle to adopting MAMS designs routinely is the computational effort of obtaining stopping boundaries that guarantee strong control of type-I error. This problem has not, so far, been handled satisfactorily. Attempts were made by, among others, Hughes (1993), Follman, Proschan, and Geller (1994), Stallard and Todd (2003), Chen, DeMets, and Lan (2010) and Magirr, Jaki, and Whitehead (2012) with varying degrees of success, as we discuss in Section 5. MAMS designs face two major computational hurdles – unless their special structure is exploited, the computational complexity grows exponentially with number of arms as well as with number of stages. In all previous work on MAMS designs, the fact that the independent increments property, so successfully exploited for 2-arm trials (see, for e.g., Armitage et al., 1969), could be generalized to the multi-arm setting was never recognized. Additionally, multi-arm designs involve multivariate normal integration for which special techniques were developed only in the 1990s (see, for e.g., Genz and Bretz, 2009), and hence were not available to the early proponents of MAMS designs.

In this article, we present an algorithm that exploits independent increments and also utilizes the quasi-Monte Carlo methods of Genz and Bretz (2009, page 91) to perform multivariate normal integration efficiently. As a result, the computational effort required to create designs with efficacy-only boundaries is linear in both number of arms and number of stages, while the computational effort required to create designs with both efficacy and futility, boundaries remains linear in number of arms but doubles with each successive increase in number of stages. Furthermore, because we utilize quasi-Monte Carlo methods, the error in the numerical integration can be quantified and the accuracy of the
resulting stopping boundaries can be bounded to any desired level of confidence. This is not possible with the other methods discussed above since they utilize numerical quadrature in their algorithms.

We formulate the problem in Section 2 and standardize it in Section 3 so as to eliminate dependence on the variance–covariance matrix of the test statistic. The numerical algorithm is presented in Section 4. Sections 5 and 6 discuss how the method compares with alternative methods that have been developed for MAMS designs. The method is applied, for illustrative purposes, to a multi-arm clinical trial of chronic obstructive pulmonary disease in Section 7. We end in Section 8 with some concluding remarks.

2. Problem Formulation

2.1. Stopping Boundaries for Normally Distributed Data

Consider a group sequential clinical trial in which $D$ treatment arms, indexed by $i = 1, 2, \ldots, D$, are compared in a pairwise manner to a common control arm, indexed by $i = 0$. The trial comprises up to $J$ looks at accumulating data, indexed by $j = 1, 2, \ldots, J$. As subjects enter the trial they are randomized to either the $i$th treatment arm or the control arm in accordance with a pre-specified allocation ratio of $\lambda_i$. Let $X_{ij} \sim N(\mu_i, \sigma^2_i)$ denote the response of the $i$th subject who is enrolled between look $j - 1$ and look $j$. Let $\delta_i = \mu_i - \mu_0$. We shall be interested in testing the global null hypothesis $H_0: \delta_i = 0$ for all $i = 1, 2, \ldots, D$ against the one side alternative hypothesis $H_1: \delta_i > 0$ for at least one $i$, through a group sequential hypothesis test strategy that ensures strong control of type-1 error. To that end, let $n_{i1} < n_{i2} \cdots < n_{ij}$ be the cumulative sample sizes on arm $i$, $\bar{X}_{ij}$ be the sample mean for the $n_{ij}$ subjects on arm $i$ and $\delta_{ij} = \bar{X}_{ij} - \bar{X}_{0j}$ be the maximum likelihood estimate of $\delta_i$ at look $j$. The Fisher information for $\delta_i$ at look $j$, denoted by $\mathcal{I}_{ij}$, is estimated by the square inverse of the standard error of $\delta_{ij}$. Thus it is easy to show that $\mathcal{I}_{ij} = n_{0j}/\Lambda_i$ where

$$\Lambda_i = \left(\sigma^2_0 + \sigma^2_i/\lambda_i\right)^{-1}.$$

Denote the efficient score statistic by $W_{ij} = \delta_{ij}/\mathcal{I}_{ij}$ and let $W_j = (W_{1j}, W_{2j}, \ldots, W_{Dj})$. Then $W_j, j = 1, 2, \ldots, J$, is a multivariate Brownian process with the following properties:

$$E(W_{ij}) = \delta_{ij},$$
$$\text{var}(W_{ij}) = \mathcal{I}_{ij},$$
$$\text{cov}(W_{ij}, W_{i'j}) = \mathcal{I}_{ij} \text{ if } j \geq j',$$
$$\text{cov}(W_{i1j}, W_{i2j}) = \Lambda_{i1} \Lambda_{i2} \sigma^2_0 n_{0j} \text{ if } i_1 \neq i_2.$$

Property (2.1) implies that the $W_{ij}$ process has independent increments, that is, $W_{i1} = W_{i2} - W_{i1}$ are independent.

Let $\delta = (\delta_1, \delta_2, \ldots, \delta_D)$ and $\text{max}(W_j) = \text{max}(W_{ij}, i = 1, 2, \ldots, D)$. A $J$-look level-$$\alpha$$ group sequential design with early stopping only for efficacy involves evaluation of efficacy boundaries $b_1, b_2, \ldots, b_J$ that satisfy

$$\sum_{j=1}^{J} P_{\delta_j}(\bigcap_{l=1}^{j-1} \text{max}(W_l) < b_l \land \text{max}(W_j) \geq b_j) = \alpha$$

(2.2)

where $P_{\delta_j}(\cdot)$ denotes probability under $\delta = b$. These efficacy boundaries are typically computed recursively. The overall type-1 error $\alpha$ is first split into $J$ positive components $(a_1, a_2, \ldots, a_J)$ in accordance with some pre-specified error spending function (Lan and DeMets, 1983) such that $\sum a_i = \alpha$. Now suppose that $b_1, \ldots, b_{j-1}$ have already been evaluated. Then $b_j$ is obtained from the $j$th term in equation (2.2) by solving

$$P_{\delta_j}(\bigcap_{l=1}^{j-1} \text{max}(W_l) < b_l \land \text{max}(W_j) \geq b_j) = a_j \quad (2.3)$$

Notice that these efficacy boundaries have been evaluated under the complete null hypothesis $\delta = 0$. It can be shown, however, that because they utilize $\text{max}(W_j)$ at each look, they do provide strong control of type-1 error (see, Magirr et al., 2012).

Once the efficacy boundaries have been evaluated, one can compute $\beta$, the type-2 error of the design at any alternative hypothesis, say $\delta = \delta_1$, as

$$\beta = P_{\delta_1}(\bigcap_{j=1}^{J} \text{max}(W_j) < b_j).$$

(2.4)

One can also incorporate futility boundaries, $(a_1, a_2, \ldots, a_{j-1}, a_j = b_j)$, into the design such that the trial may be stopped for futility if $\text{max}(W_j) \leq a_j$ for any $j$, in which case the type-2 error is

$$\beta = \sum_{j=1}^{J} P_{\delta_1}(\bigcap_{l=1}^{j-1} a_l < \text{max}(W_l) < b_l \land \text{max}(W_j) \leq a_j).$$

(2.5)

In practice, the futility boundaries are made nonbinding by first evaluating the efficacy boundaries $b_1, b_2, \ldots, b_J$. Next, an arbitrary value is selected for $\beta$, the type-2 error for any pre-specified sample size, and is split into $J$ components $(\beta_1, \beta_2, \ldots, \beta_j), \sum_{j=1}^{J} \beta_j = \beta$, in accordance with some desired error spending function. Now suppose that $a_1, \ldots, a_{j-1}$ have already been evaluated. Then at look $j$, $a_j$ can be evaluated by equating the $j$th term in equation (2.4) with $\beta_j$. Thus

$$P_{\delta_1}(\bigcap_{l=1}^{j-1} a_l < \text{max}(W_l) < b_l \land \text{max}(W_j) \leq a_j) = \beta_j \quad (2.6)$$

This entire procedure is iterated with different values of $\beta$ until the $a_j = b_j$, that is, the boundaries must meet at the last look. The final value of $\beta$ thus evaluated will be the actual type-2 error of the study.

We show in Appendix S1 of the Supplementary Materials that the probability equations (2.3) and (2.6) can be solved
by repeated evaluation of “component probabilities” of the form
\[ P_{\bar{H}} \left( \bigcap_{j=1}^{J} \max \{ W_{ij} \} < c_j \right) \]  
(2.7)
where \( c_j \) is either \( a_i \) or \( b_i \), depending on the context. We also show in Appendix 1 that at any look \( j \), it is necessary to evaluate \( 2^J \), such individual components each being of order \( j \times D \).

2.2. Generalization to Discrete and Time-to-Event Models
Although the above formulation for designing a MAMS group sequential design assumes that the underlying data are normally distributed, it is applicable more generally to clinical trials with continuous endpoints, binary endpoints, time-to-event endpoints, and even to regression models in which the test statistics are derived as efficient scores from the likelihood function, by a straightforward application of methods by Jennison and Turnbull (1997) (JT) for two-arm group sequential designs. The basic result is Theorem 2 of JT which by Jennison and Turnbull (1997) (JT) for two-arm group sequential design assumes that the underlying data are normally distributed, it is applicable more generally to clinical models in which \( \bar{H} \) denotes the coefficients of the covariates and \( c \), by a straightforward application of methods

3. Standardized Score Statistics
In order to evaluate component probabilities of the form (2.7), it is desirable to transform \( W_{ij} \) such that the sample size affects the transformed statistic only through its mean and not through its covariance matrix, for then the efficacy boundaries will not depend on sample size. Accordingly, we define a new stochastic process
\[ U_{ij} = \frac{W_{ij}}{\sqrt{I_{\text{max}}}}, \quad j = 1, 2, \ldots, J, \]  
(3.8)
where \( I_{\text{max}} = \max \{ I_{ij}, i = 1, 2, \ldots, D \} \) is the maximum information at the final look, over all the \( \delta_i \) parameters. Let \( t_j = n_{0j}/n_{ij} \), \( I_{\text{max}} = \max \{ I_{ij}, i = 1, 2, \ldots, D \} \), and define the “drift parameter” for treatment \( i \) by
\[ \eta_i = \delta_i \left( \frac{\Lambda_i}{I_{\text{max}}} \right) \sqrt{I_{\text{max}}} = \delta_i \left( \frac{\Lambda_i}{I_{\text{max}}} \right) \sqrt{n_{ij}\Lambda_{ij}}. \]

Then \( U_{ij} \) is a multivariate brownian process of independent increments with:
\[ E(U_{ij}) = t_{ij}, \]  
\[ \var(U_{ij}) = t_{ij} \left( \frac{\Lambda_i}{I_{\text{max}}} \right), \]  
\[ \text{cov}(U_{ij}, U_{ij'}) = t_{ij} \left( \frac{\Lambda_i}{I_{\text{max}}} \right) \text{ if } f_{ij} > f_{ij'}, \]  
\[ \text{cov}(U_{ij}, U_{ij''}) = t_{ij} \rho_{ij'}^{\frac{1}{2}} \left( \frac{\Lambda_i}{I_{\text{max}}} \right) \text{ if } i_{ij} \neq i_{ij''}. \]

Define the \( D \times D \) matrix \( \rho \) by
\[ \rho_{lm} = \begin{cases} \sigma_l^2 \left( \frac{\Lambda_l}{I_{\text{max}}} \right) & \text{if } l \neq m \\ \frac{\Lambda_l}{I_{\text{max}}} & \text{if } l = m. \end{cases} \]

Let \( \eta = (\eta_1, \eta_2, \ldots, \eta_D) \) and \( t = (t_1, t_2, \ldots, t_D) \). Then, \( U_{ij} \sim N(t_i\eta, t_i\rho) \) with independent increments. Equation (2.7) is thus transformed into
\[ P \left( \bigcap_{j=1}^{J} \max \{ U_{ij} \} < d_i | t, \eta, \rho \right) \]  
(3.9)
where \( d_i = c_i/\sqrt{I_{\text{max}}} \). We will develop an efficient numerical algorithm for evaluating (3.9).

4. Numerical Algorithm
4.1. Integration Over a Unit Hypercube
The independent increment structure of \( U_{ij} \) enables us to factor (3.9) into a product of nested integrals that can be solved
recursively. To see this, observe that

$$P(t_{\text{rec}} \max(U_l) < d_l | \eta, \rho) = \int_{\mathbb{R}^D} \cdots \int_{\mathbb{R}^D} f(u_1, u_2, \ldots, u_D | \eta, \rho) du_1 du_2 \cdots du_D$$

where $f(u_1, u_2, \ldots, u_D | \eta, \rho)$ is the joint density of $(U_1, U_2, \ldots, U_D)$ and $u_{il} < d_l$ means that $u_{il} < d_l$ for all $i = 1, 2, \ldots, D$. Factoring $f(.)$ as a product of conditional probabilities, we can write (4.10) as

$$f(u_1; t_{1\eta}, t_{1\eta}) f(u_2; t_{2\eta}, t_{2\eta}) \cdots \times f(u_D; t_{D\eta}, t_{D\eta}) du_1 du_2 \cdots du_D,$$

or more conveniently as

$$f(u_{il}; t_{l\eta}, t_{l\eta}) = \frac{1}{(2\pi)^{D/2} \sqrt{\det(t_{l\rho})}}$$

The next step is to replace (4.12) with a product of univariate integrals. To this end, let $\rho = CC'$ be the Cholesky decomposition of $\rho$ where $C$ is a lower triangular matrix. Using the linear transformation $(t_{il} - t_{il})/\sqrt{t_{l\rho}} = C Y_l$, $l = 1, 2, \ldots, j$, as discussed in Genz and Bretz (2009, page 29), we have

$$\left( \frac{t_{il} - t_{il}}{\sqrt{t_{l\rho}}} \right) \rho^{-1} = Y_l$$

and the Jacobian of the transformation is given by

$$J = \left[ \frac{\partial u_1}{\partial Y_1}, \frac{\partial u_2}{\partial Y_2}, \ldots, \frac{\partial u_D}{\partial Y_D} \right] = \sqrt{t_{l\rho}} C.$$
to show explicit dependence of \( e_{il} \) only on \( e_{mi} \), \( m = 1, 2, \ldots, i - 1 \), and on \( Q_{m,l-1} = \sum_{i=1}^{l-1} \Phi^{-1}(e_{mi} Q_{m,i}) \sqrt{t(l)} \), \( m = 1, 2, \ldots, i \). This relationship shows that although (4.17) appears to be a product of independent integrals on the unit hypercube, the integrands, \( e_{il} \), are linked recursively. Note, for future reference, the recursion

\[
Q_{ml} = Q_{m,l-1} + \Phi^{-1}(e_{ml} Q_{ml}) \sqrt{t(l)}, \quad \text{for } l = 2, \ldots, j.
\]

We shall evaluate (4.17) by quasi-Monte Carlo integration.

4.2. Quasi-Monte Carlo Integration

Denote the multiple integral (4.17) by

\[
I = \int_0^1 \cdots \int_0^1 f(x) dx_{11} \cdots dx_{Dj}
\]

where \( x = (x_{11}, x_{21}, \ldots, x_{D1}, x_{12}, x_{22}, \ldots, x_{D2}) \) and \( f(x) = \prod_{m=1}^{D} \prod_{j=1}^{N} e_{ji} \). The integrand \( f(x) \) is computed by evaluating each \( e_{il} \) recursively as described below.

**Look 1**: At look \( l = 1 \), each \( e_{i1}, i = 1, 2, \ldots, D \), is evaluated as a function of the preceding \( e_{m1}, m = 1, 2, \ldots, i - 1 \), according to the formula

\[
e_{i1} = \Phi \left\{ \frac{1}{C_{il}} \left[ \frac{1}{\sqrt{t(l)}} (d_1 - t_1 \eta_l) - \sum_{m=1}^{i-1} C_{im} \Phi^{-1}(e_{m1} Q_{m1}) \right] \right\}.
\]

Finally the terms \( Q_{i1} = \Phi^{-1}(e_{i1} x_{i1}) \sqrt{t(l)} \), \( i = 1, 2, \ldots, D \), are saved for use at look \( l = 2 \).

**Look 2**: At look \( l = 2 \), each \( e_{i2}, i = 1, 2, \ldots, D \), is evaluated as a function of the preceding \( e_{m2}, m = 1, 2, \ldots, i - 1 \), and by the \( Q_{i1}, i = 1, 2, \ldots, D \) that were saved at look \( l = 1 \), according to the formula

\[
e_{i2} = \Phi \left\{ \frac{1}{C_{il}} \left[ \frac{1}{\sqrt{t(2)}} (d_2 - t_2 \eta_l) - \sum_{m=1}^{i-1} C_{im} \Phi^{-1}(e_{m2} Q_{m1}) \right] \right\}.
\]

Finally the terms \( Q_{i2} = \Phi^{-1}(e_{i2} x_{i2}) \sqrt{t(2)} + Q_{i1}, i = 1, 2, \ldots, D \), are saved for use at look \( l = 3 \).

**Look 1**: At any general look \( l \) each \( e_{il}, i = 1, 2, \ldots, D \), is evaluated as a function of the preceding \( e_{m1}, m = 1, 2, \ldots, i - 1 \), and by the \( Q_{i,l-1}, i = 1, 2, \ldots, D \) that were saved at look \( l - 1 \), according to the formula

\[
e_{il} = \Phi \left\{ \frac{1}{C_{il}} \left[ \frac{1}{\sqrt{t(l)}} (d_l - t_l \eta_l) - \sum_{m=1}^{i-1} C_{im} \Phi^{-1}(e_{m1} Q_{m1}) \right] \right\}.
\]

Finally the terms \( Q_{il} = \Phi^{-1}(e_{il} x_{il}) \sqrt{t(l)} + Q_{i,l-1}, i = 1, 2, \ldots, D \), are saved for use at look \( l + 1 \).

These computations proceed look by look for \( l = 1, 2, \ldots, j \) and \( f(x) \) is finally evaluated as the product of the individual \( e_{il} \) terms. Let \( x = (x_{11}, x_{21}, \ldots, x_{D1}, x_{12}, x_{22}, \ldots, x_{D2}) \), \( Q = (Q_{11}, Q_{21}, \ldots, Q_{D1}) \), and \( e_i = e_{i1} e_{i2} \cdots e_{il} \). Then the above computations imply that

\[
f(x) = e_{i1}(x_{i1}) e_{i2}(x_{i2}) \cdots e_{i,l-1}(x_{i,l-1}).
\]

One can obtain a crude Monte Carlo estimate for \( I \) by sampling \( x \) uniformly from a \( D \times j \) dimensional unit hypercube. Let \( x^{(1)}, x^{(2)}, \ldots, x^{(N)} \) be the \( N \) sampled values. Then

\[
I_N = \frac{1}{N} \sum_{n=1}^{N} f(x^{(n)})
\]

is an unbiased estimate of \( I \) with standard error of order \( O(N^{-1/2}) \). It is possible to improve the accuracy of the crude Monte Carlo estimate by a quasi-Monte Carlo (QMC) procedure that combines the above randomly generated points with a fixed sequence of points created by lattice rules. We shall use a QMC procedure proposed by Genz and Bretz (2009, page 47-48). A rank-1 lattice is a point set of the form

\[
L_P = \left\{ \frac{i v_P}{P} \mod 1, i = 1, 2, \ldots, P \right\}
\]

where \( P \) is a prime number and (in our setting) \( v_P \) is a \( D \times j \) dimensional vector that depends on \( P \). Many choices have been proposed for \( (P, v_P) \). In order to achieve \( O(N^{-1+\epsilon}) \) integration errors, good lattice parameters must be determined. Genz (2000, http://www.math.wsu.edu/~genz/software/fort77/mvnpack.f+) has created a table of \( (P, v_P) \) combinations based on the component by component algorithms of Dick and Kuo (2004), and Nuyens and Cools (2006) that are well suited to multivariate normal integration. Our computations utilize \( (P, v_P) \) values from this table, a subset of which is displayed in Figure S1 of the Supplementary Materials. The QMC procedure is outlined below.

1. Select a \( (P, v_P) \) combination and construct the lattice \( L_P \) consisting of \( P \) points each of dimension \( D \times j \).
2. For each lattice point, \( z_p \in L_P \) generate \( M \) random vectors \( \Delta_{p,m} \sim U[0, 1]^{D \times j} \), \( m = 1, 2, \ldots, M \).
3. For a given \( m \), let

\[
I_{P,m} = \frac{1}{2P} \sum_{p=1}^{P} \left( f(2[z_p + \Delta_{p,m}] - 1) + f(1 - 2[z_p + \Delta_{p,m}] - 1) \right)
\]

where \( [a] = a \mod 1 \). Then a QMC estimate of \( I \) is given by

\[
I_{P,M} = \frac{1}{M} \sum_{m=1}^{M} I_{P,m}
\]
with variance

$$\sigma_{P,M}^2 = \frac{1}{M(M-1)} \sum_{m=1}^{M} (I_{P,m} - I_{P,M})^2.$$ 

4. Further, variance reduction is achieved by repeating steps 1 to 3 with additional \((P, v_P)\) combinations. Suppose, we utilize \(K\) rank-1 lattices, generated by combinations \((P_k, v_{P_k})\) and resulting in the estimates \(I_{P_k,M}\). Then the final estimate of \(I\), due to LePage (1978), is

$$I = \bar{\sigma}^2 \sum_{k=1}^{K} \frac{I_{P_k,M}}{\sigma_{P_k,M}^2}$$

with standard error

$$\bar{\sigma} = \left( \sum_{k=1}^{K} \frac{1}{\sigma_{P_k,M}^2} \right)^{-1/2}.$$ 

4.3. Accuracy

Standard error estimates of the crude Monte Carlo (MC) and quasi-Monte Carlo (QMC) procedures are displayed in Table 4.2 for various choices of \(D\) and \(J\). The MC estimates are displayed for 50,000 independent samples. The QMC estimates are displayed for \(M = 6\) and a suitable value of \(K\) such that \(\sum_{k=1}^{K} P_k \approx 50,000\). Thus, for the comparison of the two procedures with respect to accuracy, the number of sampled points is approximately the same.

### Table 1

| Looks | Comparisons | Standard error of integral estimate |
|-------|-------------|-------------------------------------|
| \(J\) | \(D\)       | QMC Integration | MC Integration |
| 2     | 4           | 0.000075         | 0.029537       |
| 3     | 4           | 0.000156         | 0.016844       |
| 5     | 4           | 0.000302         | 0.027771       |
| 2     | 5           | 0.000421         | 0.029855       |
| 3     | 5           | 0.000359         | 0.041026       |
| 4     | 5           | 0.000052         | 0.031194       |
| 5     | 4           | 0.000637         | 0.035716       |
| 2     | 6           | 0.000566         | 0.040911       |
| 3     | 6           | 0.000591         | 0.043898       |
| 4     | 6           | 0.000595         | 0.047661       |
| 5     | 6           | 0.000662         | 0.048621       |
| 2     | 5           | 0.000739         | 0.050674       |
| 3     | 5           | 0.001083         | 0.047483       |
| 4     | 5           | 0.001324         | 0.049185       |
| 5     | 5           | 0.000995         | 0.041286       |

Based on 50,000 MC and approximately 50,000 QMC simulations

These results ensure that the 99.9% confidence interval for \(I\) based on QMC will be at least as accurate as \(I \pm 0.001\) even for a design with 5 comparisons to a common control and 5 looks at accumulating data. Larger problems are unlikely to be encountered in the clinical trials setting.

4.4. A Non-Technical Explanation of Quasi-Monte Carlo

Multivariate normal probability computations typically involve semi-definite integration. For one-dimensional problems semi-definite integration is performed by quadrature methods. A multi-dimensional integral can be expressed as repeated one-dimensional integrals by applying Fubini’s theorem. But the computational effort of using numerical quadrature grows exponentially with the number of dimensions. To avoid that, one usually considers Monte Carlo methods for multi-dimensional integration.

The Monte Carlo method requires the multivariate integral to be a definite integral. The transformations in Section 4.1 convert the semi-definite multivariate normal integral (4.10) into a definite integral over a hypercube. The regular Monte Carlo method averages the repeated evaluations of the integrand function over a random sequence of grid points. However, as we demonstrated in Table 4.2, using random sequences in this way is not sufficiently efficient for MAMS boundary computations. This led us to the quasi-Monte Carlo integration described in Section 4.2. In quasi-Monte Carlo integration some points are chosen deterministically using low-discrepancy measures to make the computations converge faster. These low-discrepancy points are combined in a specific way with random sequences of points so as to make the computation robust. The creation of these sequences of low-discrepancy points utilize number theory methods that are outside the scope of this article. In particular, we have used the Korobov (2002) sequence of low-discrepancy points and the Merseenne-Twister method (Matsumoto and Nishimura, 1998) for combining them with random sequences.

5. Comparison with Alternative Algorithms

The most recent published method for generating stopping boundaries and computing sample size for MAMS designs is by Magirr, Jacki, and Whitehead (2012). An R program implementing this method and maintained by Jaki is available at https://cran.r-project.org/web/packages/+MAMS/index.html

Tables 2 and 3 compare the execution times of the new algorithm described in Section 4, hereafter referred to as NEW and the R implementation of the Magirr, Jackie, and Whitehead (2012) method, hereafter referred to as MJW, for a range of treatment comparisons and looks at the accruing data. All computations were executed on a Lenovo Think Pad, Model T440P with Intel i7 processor and 8 core CPUs. Table 2 displays results for designs with efficacy-only boundaries while Table 3 displays results for both efficacy and futility boundaries.

It is evident from these tables that, for designs with efficacy-only boundaries, the computing times of NEW are linear in both \(D\) and \(J\), while they increase as \(2^J\) for designs with both efficacy and futility boundaries. In contrast the
Execution Time Comparison of QMC and MJW Methods for MAMS Designs with Efficacy Boundary Only

| Looks | Comparisons | Execution times (secs) |   |   |
|-------|-------------|------------------------|---|---|
| (J)   | (D)         | NEW         | MJW       |   |   |
| 2     | 1           | 2           | 2          |   |   |
| 3     | 1           | 2           | 2          |   |   |
| 2     | 4           | 2           | 2          |   |   |
| 5     | 2           | 2           | 2          |   |   |
| 2     | 1           | 138         | 2          |   |   |
| 3     | 1           | 148         | 2          |   |   |
| 3     | 4           | 148         | 2          |   |   |
| 5     | 2           | 158         | 2          |   |   |
| 2     | 1           | >28,800     | 2          |   |   |
| 3     | 1           | >28,800     | 2          |   |   |
| 4     | 2           | >28,800     | 2          |   |   |
| 5     | 2           | >28,800     | 2          |   |   |
| 2     | 1           | >28,800     | 2          |   |   |
| 3     | 2           | >28,800     | 2          |   |   |
| 5     | 4           | >28,800     | 2          |   |   |
| 5     | 2           | >28,800     | 2          |   |   |

Total Sample Size is 600 for all Designs

Execution times of the MJW algorithm increase exponentially with J for designs with or without futility boundaries and break down entirely for J > 3.

It is insightful to analyze why MJW, unlike NEW, is computationally explosive as the number of stages increase. Both algorithms must compute the probability of the very same event

\[ R(\theta) = \bigcap_{i=1}^{D} \left( \bigcup_{j=1}^{J} \left( \bigcap_{l=1}^{i} I_{ij} < T_{ij} < u_{j} \right) \right) \]  

for a normally distributed statistic \( T_{ij} \) and suitably standardized efficacy and futility boundaries \( l_{ij} \) and \( u_{j} \). The difference is that MJW uses the Wald statistic \( \delta \sqrt{T_{ij}} \) for \( T_{ij} \), whereas NEW uses the score statistic \( \delta_{ij} \) for \( T_{ij} \), for computing the event (5.21). This initial choice of test statistic dooms the MJW method for it can no longer utilize the underlying stage-wise independent increments structure of the problem. Instead, the problem gets transformed into the form

\[ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{j=1}^{D} \left[ \sum_{i=1}^{j} \Phi_{i}[L_{ij}, U_{ij}] \right] d\Phi(t_{1}) \cdots d\Phi(t_{j}) \]  

where, for the \( j \)th treatment comparison, \( \Phi_{i}[L_{ij}, U_{ij}] \) denotes the result of integrating the \( j \)-dimensional multivariate normal density over a region defined by a vector of lower limits \( L_{ij} \) and a vector of upper limits \( U_{ij} \). Decomposition of (5.22) into a product of univariate normal integrals such as we have obtained in (4.17) is clearly impossible. Evaluation of (5.22) is by numerical quadrature. Suppose, each of the \( J \) dimensions of the outer integral \( \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{j=1}^{D} [\sum_{i=1}^{j} \Phi_{i}[L_{ij}, U_{ij}]] \) is divided into \( G \) grid points. For each grid point, the inner product-sum \( \prod_{j=1}^{D} [\sum_{i=1}^{j} \Phi_{i}[L_{ij}, U_{ij}]] \) is evaluated by repeated calls to a function such as mvtnorm (Genz et al., 2016). For each \( i = 1, 2, \ldots, D \), there are \( J \), repeated calls to mvtnorm, in which each call evaluates a region \( [L_{ij}, U_{ij}] \) of a \( j \)-dimensional multivariate normal density. This computation must be repeated for \( j = 1, 2, \ldots, J \). It follows that the MJW algorithm must make \( \sum_{j=1}^{J} G^{D-x/j} \) calls to mvtnorm to evaluate (5.22). Assuming that \( G = 20 \), the current implementation of MJW in R, one can see why the problem breaks down entirely for \( J > 3 \), even for the balanced case where only \( \sum_{j=1}^{J} G^{J} \) calls to mvtnorm are needed.

In contrast the NEW algorithm, by exploiting the independent increments structure of \( W_{ij} \), is able to transform the problem into a simple product of univariate integrals of the form (4.17) from which the stopping boundaries, type-1 error or power can be obtained by a recursive computation that is linear in \( D \) or \( J \) when only efficacy boundaries are evaluated, and linear in \( D \) but increasing like \( 2^{D} \) when both efficacy and futility boundaries are evaluated.

Prior to MJW, methods to obtain group sequential boundaries for MAMS designs were proposed by Hughes (1993), Follman, Proschan, and Geller (1994), Stallard and Todd (2003) and Chen, DeMets and Lan (2010). All these methods utilized the Wald statistic for the computations and so suffer from the same limitation as MJW. Hughes (1993) simply utilizes the boundaries of a two-arm clinical trial, relying on binding futility rules established via simulation, for dropping non-performing arms in mid-course, and thereby preserving the type-1 error conservatively. There is no guarantee that this approach will provide strong control of type-1 error. Follman et. al. (1994) start out by computing Bonferroni based stopping boundaries and then adjusting them by...
6. Comparison with Combination P-value Methods

An alternative method to test multiple treatment arms against a common control arm with possible treatment selection at one or more interim analysis time points is by combining p-values from the different stages with pre-specified weights and using a closed test for the final analysis. Since the p-values from the separate stages are independent and uniformly distributed under the null hypothesis (or stochastically larger than uniform in the discrete case), their combination with pre-specified weights is also uniformly distributed so that valid level-α tests may be constructed. In this approach, p-values are required both to test elementary hypotheses of the form $H_i; \delta_i = 0$ as well as intersection hypotheses of the form $H_i \cap H_j \cap H_k; \delta_i = \delta_j = \delta_k = 0$. The latter p-values are adjusted for multiplicity by methods proposed by, among others, Dunnett, Bonferroni, and Simes. An excellent reference to the combination p-value method is the paper by Posch et al., (2005).

Although the MAMS and combination p-value approaches tackle essentially the same problem, the two approaches are fundamentally different. MAMS, having its roots in Markov processes, exploits the known correlation structure of the sequentially computed score statistic when computing the early stopping boundaries. The p-value combination approach on the other hand, having its roots in multiple comparisons methodology, exploits closed testing (Markus et al., 1976) to control the type-1 error. Moreover since, in this approach the independent incremental data from each stage are combined, correlations between sequentially computed cumulative statistics across stages are not exploited. It would thus be of interest to make power comparisons between the two methods.

There are actually three definitions of power in a multi-arm setting; global power, disjunctive power, and conjunctive power. Global power is the probability that at least one treatment arm will attain statistical significance. Disjunctive power is the probability that at least one non-null treatment arm will attain statistical significance. Conjunctive power is the probability that all non-null treatment arms will attain statistical significance.

Table 4 displays disjunctive and conjunctive power comparisons between the MAMS method and three commonly used combination p-value methods – Bonferroni, Simes, and Dunnett. These power comparisons are for 3, 4, and 5-arm designs with three equally spaced looks, 50 subjects per arm, the Lan and DeMets (1987), O’Brien-Fleming (1979) type boundary for early efficacy stopping, $\delta/\sigma = 0.5$.

### Table 4

**Power Comparisons of P-value Combination and MAMS Procedures** (all entries based on 100,000 simulations)

| Arms | Bonferroni | Simes | Dunnett | MAMS |
|------|------------|-------|---------|------|
| 3    | 0.717      | 0.732 | 0.732   | 0.766 |
| 4    | 0.722      | 0.735 | 0.746   | 0.805 |
| 5    | 0.726      | 0.736 | 0.750   | 0.835 |

7. **The INHANCE Clinical Trial of COPD**

Indacaterol to Help Achieve New COPD Treatment Excellence (INHANCE) was a randomized clinical trial for the treatment of chronic obstructive pulmonary disease in which four doses (75 mg, 150 mg, 300 mg, 500 mg) of inhaled indacaterol, a once-daily long-acting β2-agonist bronchodilator were compared to placebo (Donohue et al., 2010). The primary efficacy objective was to show the superiority of at least one dose over placebo at week 12 with respect to 24-hour post-dose (trough) forced expiratory volume in 1 second (FEV1). The improvement in FEV1 for indacaterol versus placebo, denoted by $\delta$, was expected to be between 0.14 and 0.18 liters and the between-subject variability was assumed to be $\sigma = 0.5$. Although the actual trial had only two stages and utilized closed testing for preserving the type-1 error (Lawrence et al., 2014), we will use this setting to illustrate a MAMS design comprising three pair-wise comparisons (150 mg, 300 mg, 500 mg) to placebo over up to four equally-spaced looks at the accumulating data. The design will utilize the Lan and DeMets (1983) error spending function to generate O’Brien-Fleming (1979) type boundaries (LOD boundary) to generate early stopping efficacy boundaries. Table 5 displays these boundaries on the Wald scale and contrasts them with corresponding boundaries for a conventional four-look group sequential design (GSD) with only one treatment arm versus placebo.

The 4-arm boundaries are stricter than the 2-arm ones because, with three comparisons to placebo, there are three times as many chances for declaring efficacy under the global
null hypothesis in the MAMS design compared to the 2-arm GSD.

Suppose, we wish the COPD trial to have 90% global power under the alternative hypothesis \( H_1: \delta_1 = 0.14, \delta_2 = 0.16, \delta_3 = 0.18 \). We will adopt the LD(OF) boundaries for early efficacy stopping with type-1 error = 0.025 to reject the global null hypothesis \( H_0: \delta_1 = \delta_2 = \delta_3 = 0 \). These efficacy boundaries will preserve the type-1 error conservatively if treatment arms are dropped in midstream for any reason, including excess toxicity. It is nevertheless desirable to incorporate formal futility boundaries into the design so as to have objective criteria for dropping non-performing arms at one or more interim analysis time points. Table 6 displays the maximum sample size, expected sample size under \( H_1 \), and expected sample size under \( H_0 \) for designs with between 1 and 4 looks, LD(OF) efficacy boundaries and non-binding LD(OF) futility boundaries.

There are large savings in expected sample size, with diminishing returns as the number of looks increase. The efficacy and futility boundaries for the 4-look design are displayed on the Wald scale in Figure S2 of the Supplementary Material.

The results in Table 6 have not taken into consideration the possibility of recomputing the efficacy boundary whenever a treatment arm is dropped. We now show that substantial additional efficiency gains can be obtained by boundary recomputation. Boundary re-computation if treatment arms are dropped is not a trivial problem, however. If done incorrectly, the type-1 error will be inflated. We utilize a generalization of the conditional error rate principle, proposed by Müller and Schäfer (2001) for adaptive designs, to recompute the efficacy boundary without inflating the type-1 error. For expository purposes, we will explain the principle in the context of a three-look, four-arm design from which a single arm is dropped at look 1. Let \((W_{11}, W_{21}, W_{31})\) be the score statistics for the three treatment comparisons and let \(b_j\) be the corresponding efficacy boundary at look \(j, j = 1, 2, 3\). Suppose that at look 1, with \((W_{11}, W_{21}, W_{31}) = (w_{11}, w_{21}, w_{31})\), treatment 1 is dropped. Denote the recomputed boundaries for looks 2 and 3 by \(b_2^*\) and \(b_3^*\), respectively. Define the conditional error rate of the original design as

\[
\epsilon(w_{11}, w_{21}, w_{31}) = P_0[\max(W_{12}, W_{22}, W_{32}) \geq b_2 \text{ or } \max(W_{13}, W_{23}, W_{33}) \geq b_3 | (w_{11}, w_{21}, w_{31})]
\]

and the conditional error rate of the design with the modified boundaries as

\[
\epsilon^*(w_{21}, w_{31}) = P_0[\max(W_{22}, W_{32}) \geq b_2^* \text{ or } \max(W_{23}, W_{33}) \geq b_3^* | (w_{21}, w_{31})].
\]

The Müller and Schäfer error rate principle states that the type-1 error will be preserved if the procedure used to compute the modified boundaries \((b_2^*, b_3^*)\) always satisfies the conditional error rate requirement

\[
\epsilon^*(w_{21}, w_{31}) \leq \epsilon(w_{11}, w_{21}, w_{31}). \tag{7.23}
\]

Additional constraints similar to (7.23) are imposed for other intersection hypotheses.

To study the impact of recomputing the efficacy boundaries, we ran 100,000 simulations with and without boundary recomputation for two and three-look COPD designs with \( \delta_1 = 0.14, \delta_2 = 0.16, \delta_3 = 0.18 \), and LD(OF) efficacy boundaries. At the first interim analysis the treatment arm with the greatest standardized treatment effect was selected and the other two treatment arms were dropped. The gains in power due to recomputing the boundaries are displayed in Table 7.

The numbers in column 6 indicate the additional patients, one would have to enroll to get the same gain in power without boundary recomputation. For example, to boost the power from 90 to 94.4% in a 3-look design without boundary recomputation would require 160 additional patients, a 20% increase in sample size.

8. Discussion

We have shown how one may create MAMS designs efficiently, illustrated their application to trials involving multiple doses and found them to have greater power than competing designs based on combining p-values from independent stages. On the face of it the computational problem appears intractable. If one attempts to solve it by numerical quadrature as was attempted by Magirr, Jaki, and Whitehead (2012) (MJW), the complexity increases exponentially with number of looks and breaks down entirely for designs with more than 3 looks at the accumulating data. Our major contribution was to show that if the problem is viewed in the proper framework it can be transformed such that the computational complexity is linear in both number of looks and number of arms for designs with efficacy-only boundaries and doubles with successive stages for designs with both efficacy and futility boundaries. Thus, for example, designs with six arms (including the control arm) and five stages can be created in

| Look | Information fraction | 2-arm | 4-arm |
|------|----------------------|-------|-------|
| 1    | 0.25                 | 4.3326| 4.5654|
| 2    | 0.5                  | 2.9631| 3.2655|
| 3    | 0.75                 | 2.359 | 2.7225|
| 4    | 1.00                 | 2.0141| 2.4142|

| Look | Information fraction | 2-arm | 4-arm |
|------|----------------------|-------|-------|
| 1    | 0.25                 | 4.3326| 4.5654|
| 2    | 0.5                  | 2.9631| 3.2655|
| 3    | 0.75                 | 2.359 | 2.7225|
| 4    | 1.00                 | 2.0141| 2.4142|

### Table 5
Comparison of 2-arm and 4-arm LD(OF) Boundaries
(1-sided \( \alpha = 0.025 \))

| Look | Information fraction | Wald scale boundaries |
|------|----------------------|-----------------------|
| 1  | 0.25 | 4.3326 | 4.5654 |
| 2  | 0.5  | 2.9631 | 3.2655 |
| 3  | 0.75 | 2.359  | 2.7225 |
| 4  | 1.00 | 2.0141 | 2.4142 |

### Table 6
Sample Size Requirements for COPD Trial with 1 to 4 Looks

| Number of looks | Sample sizes for 90% power |
|-----------------|-----------------------------|
|                 | Maximum | Under \( H_1 \) | Under \( H_0 \) |
| 1               | 624     | 624 | 624 |
| 2               | 644     | 558 | 446 |
| 3               | 668     | 522 | 420 |
| 4               | 684     | 502 | 401 |
under two minutes by the NEW algorithm. This type of performance frees the trial designer to experiment with different design options including number of looks, types of stopping boundaries and sample size, under alternative scenarios for the treatment effects. This is a crucially important consideration for optimizing trial design. If one had to wait several hours or days in between scenarios, it is unlikely that one would consider more than one or two design options and might miss out on the best possible design for the situation under consideration.

The full potential of our algorithm has not yet been realized. We have seen from Table 2, that once the design has more than one interim analysis the computational effort of the MJW algorithm can increase by two orders of magnitude. For example, a 3-look design with efficacy-only boundaries involving three comparisons to a common control takes 1 second with the NEW algorithm but 148 seconds with the MJW algorithm. While 148 seconds might not be a limitation for one-time computation of early-stopping boundaries, it can be an impassable barrier for evaluating the operating characteristics of more sophisticated MAMS designs. Increasingly, there is interest in efficient adaptive designs where doses are dropped, the sample size is reestimated, and the stopping boundaries are recomputed at interim looks (see, Magirr, Stallard and Jaki, 2014; Gao, Liu and Mehta, 2014). In such designs the type-1 error is preserved at 0.025 or less. This suggests that at the design stage of a confirmatory clinical trial one must verify through extensive simulation what nominal type-1 error will guarantee that the 1-sided actual type-1 error remains below 0.025. A systematic study of the rate of convergence to normality of the various test statistics commonly used for normal, binomial and time-to-event data would be desirable, and is made possible by the ready availability of the relevant stopping boundaries through the present work.

9. Supplementary Materials

Appendices 1, 2 and Figures S1, S2, referenced in Sections 2.1, 4.1, 4.2, and 7, respectively, are available with this article at the Biometrics website on Wiley Online Library.

Software for designing MAMS trials is implemented in the East® software package and is freely available to instructors and students for course work by contacting support@cytel.com.

Acknowledgements

The authors thank the associate editor and a referee for helpful suggestions that strengthened this article greatly. The authors would like to thank Professor Ralph D’Agostino, Sr. for his valuable contributions.

References

Armitage, P., McPherson, C. K., and Rowe, B. C. (1969). Repeated significance test on accumulating data. Journal of the Royal Statistical Society, Series A, 132, 235–244.
Chen, Y. H. J., DeMets, D. L., and Lan, K. K. G. (2010). Some drop-the-loser designs for monitoring multiple doses. Statistics in Medicine, 29, 1793–1807.
Dick, J. and Kuo, F. Y. (2004). Reducing the construction cost of the component-by-component construction of good lattice rules. Mathematics of Computation, 73, 1967–1988.
Donohue, J. F., Fogarty, C., Lotvall, J., et al. (2010). Once-daily bronchodilators for chronic obstructive pulmonary
Design and Monitoring of Multi-Arm Multi-Stage Clinical Trials

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63, 655–660.

Matsumoto, M. and Nishimura, T. (1998). Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, 8, 3–30.

Müller, H.-H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical Trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, 57, 886–891.

Nuyens, D. and Cools, R. (2006). Fast algorithms for component-by-component construction of rank-1 lattice rules in shift-invariant reproducing kernel Hilbert spaces. *Mathematics of Computation*, 75, 903–920.

Received August 2016. Revised January 2017. Accepted February 2017.