Biometrical Journal **42** (2000) 7, 877–886

# A Minimum-Cost Strategy for Cluster Recruitment

WENYAW CHAN

Program in Biometry
School of Public Health
University of Texas – Houston
U.S.A.

NAN FU PENG

Institute of Statistics
National Chiao Tung University
Hsinchu
Taiwan

*Summary*

It is becoming increasingly common for the design of a clinical study to involve cluster samples. Very few researches investigated the appropriate number of clusters. None of them treat cluster size and the number of clusters as random variables. In reality, the recruitment of clusters can not be reached at one time and the cluster sizes are usually random. The longer the recruitment takes the more expensive the total study costs will be. This paper provides a strategy for sequential recruitment of clusters, which can minimize the total study cost. By treating the number of additional observational subjects required at each time point as a Markov Chain, we derive an iterative procedure for optimal strategy and study the property of this strategy, especially the duration of the cluster recruitment. This strategy is also extended to search for an optimal number of centers in a multi-center clinical trial.

*Key words:* Cluster sample; Markov chain; Principle of optimality; Sequential method.

## 1. Introduction

The number of clusters has been an important planning issue in a clinical study involving cluster samples, especially when study budget is limited. Some researches have proposed the calculation of the sample size for the observational subjects in cluster randomization trials [see, for example, DONNER (1992) and DONNER and KLAR (1994)]. Others have discussed the appropriate number of clusters based on power analysis [see, for example HSIEH (1988)]. Very few investigated this issue from the viewpoint of study costs [see, for example, MCKINLAY (1994) and FELDMAN, MCKINLAY and NIKNIAN (1996)]. None of them treat cluster size and the number of clusters as random variables. In reality, the recruitment of the study subjects (and hence the clusters) can not be reached at one time and the cluster sizes can rarely be determined in advance. The longer the recruitment takes the more expensive the total study costs will be. This paper is intended to study the optimal strategy for sequential recruitment of clusters in order to minimize the total cost of a clinical study. A typical example is to study the individuals (observational subjects) from different families (clusters) or to study the students (observational subjects) from different schools (clusters). An example can be related to a hypothetical study, which is to investigate the elevation of systolic blood pressure (SBP) of female adolescents after menarche. Eligible participants are female adolescents who just reached menarche within three months. This hypothetical study will measure subjects' SBP two times: one at participation and the other at the sixth month. The difference of these two measurements will be the major outcome for studying elevation of SBP and the sample size for the paired $t$ test has been determined. The subjects will be enlisted from seven graders of the participated schools. In this hypothetical study, the number of eligible participants in each school is random and the number of schools can be recruited per week is also random. To find an optimal strategy for sequential recruitment of schools, which minimizes the total study cost, will be the goal of this paper.

This paper will focus on the mathematical model and the theoretical property of the optimal strategy for cluster recruitment that will be applicable for a clinical study. The logistic matters of the study, unrelated to the cost, will not be investigated in this paper.

Section 2 presents the mathematical model as well as the criterion for optimization. Section 3 describes the strategy for minimizing the total cost and investigates its properties. The moments of the duration of cluster recruitment under this strategy are calculated in section 4. In section 5, the model is extended to find the optimal number of centers in a multi-center clinical trial when the sample size is pre-determined. Section 6 discusses the computing efficiency and offers an alternative strategy when sample size is very large and the optimal strategy presented in this paper may be computationally expensive.

## 2. The Model

Let $n$ be the number of observational subjects needed for a clinical study which can be determined by a statistical power analysis. Assume that the number of observational subjects obtained from each recruited cluster is independently and identically distributed, denoting its distribution as $p(k)$, $k = 0, 1, 2, \ldots$ Let $N_t$, $t = 1, 2, \ldots$ denote the number of clusters recruited in the $t$th week. Here we assume all cluster recruitments will occur at the beginning of the week. Our optimal strategy is to find $N_1, N_2, \ldots, N_m$ such that $\sum_{i=1}^{m} \sum_{j=1}^{N_i} X_{ij} \geq n$ and the expected total cost for the study

$$E(\Omega_n) = E\left[C_1 \sum_{i=1}^{m} \sum_{j=1}^{N_i} X_{ij} + C_2 \sum_{i=1}^{m} N_i + mC_3\right] \tag{1}$$

is minimized, where $\Omega_n$ denotes the total cost for this study if the required sample size is $n$, $m$ is the number of weeks required for the cluster recruitment, $C_1$ is the cost for each study subject, $C_2$ is the cost for each cluster, $C_3$ is the cost per week for the study unrelated to clusters or observational subjects, and $X_{ij}$ represents the number of subjects enrolled at $j$th cluster in the $i$th week. Note that in (1), the cost is specified to be a linear function of costs for subjects, clusters and other expenses, such as overhead, unrelated to either subjects or clusters. The unit costs $C_1, C_2$ and $C_3$ are assumed to be known constants, $X_{ij}s$ are independent random variables with a common distribution specified as $p(k)$ and $m$ is an random variable depending on $X_{ij}$ and the choice of $N_i' s$.

Let $n_t$ denote the number of additional subjects needed to achieve the required sample size at the beginning of $t$th week and $N(\ )$ be a strategy function mapping from $n_t$ to the number of selected clusters for that week, i.e. $N(n_t) = N_t$. Mathematically, we will treat $\{n_t\}$ as a discrete-time Markov chain with a lower triangular transition probability matrix defined as

$$Q = (q_{ij}) = \Pr\left(n_{t+1} = j \mid n_t = i\right) = \begin{cases} 1 & \text{if} \quad i = j = 0 \\ 0 & \text{if} \quad j > i \\ p^*(i-j) & \text{if} \quad 0 < j \leq i \\ \sum_{l=i}^{\infty} p^*(i) & \text{if} \quad j = 0,\ i > 0 \end{cases}, \tag{2}$$

where $p^*(\ )$ represents the distribution for $\sum_{k=1}^{N_t} X_{t,k}$. Note that $p^*(\ )$ is the $N_t$-fold convolution of $p(\ )$ and depends on $N_t$ only through $n_t = i$. To clarify the notation, we may denote $p^*(\ )$ as $p_i{}^*(\ )$ whenever necessary. The above assumption that $\{n_t\}$ is a Markov chain implies that the study has a time-independent ability to recruit observational units. This ability depends only on the current study status specifying number of observational units needed to add to the study.

In this paper, we will search for an optimal strategy function $N(\ )$ among the collection of all functions mapping from the nonnegative integers (number of additional subjects needed) to the nonnegative integers (number of clusters selected). This strategy will minimize the expected total cost defined in (1).

## 3. The Optimal Strategy

From the principle of optimality [see Puterman (1994)] which states that, whatever the initial state and the initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision, we can rewrite (1) as the sum of the cost of the first week plus the cost of the remaining weeks when the optimal strategies for both periods were applied, i.e.

$$\begin{aligned} E(\Omega_n) &= \sum_{i=0}^{n-1} p_1^*(i)\,[iC_1 + N(n_1)\,C_2 + C_3 + E(\Omega_{n-i})] + \sum_{i=n}^{\infty} p_1^*(i)\,[iC_1 + N(n)\,C_2 + C_3] \\ &= p_1^*(0)\,E(\Omega_n) + \sum_{i=1}^{n-1} p_1^*(i)\,E(\Omega_{n-i}) + C_1 \sum_{i=0}^{\infty} ip_1^*(i) + [N(n)\,C_2 + C_3] \sum_{i=0}^{\infty} p_1^*(i) \end{aligned} \tag{3}$$

where $p_1^*(\ )$ is the $N_1$-fold convolution of $p(\ )$. Note that in (3), $i$ represents the number of study subjects recruited in the first week when $N(n_1)$ clusters were selected at the beginning of that week. Since $N(n) = N(n_1) = N_1$ and $\sum_{i=0}^{\infty} ip_1^*(i) = E\left(\sum_{k=1}^{N_1} X_{1,k}\right)$, a routine algebraic manipulation of (3) leads to

$$E(\Omega_n) = \left[C_1 E\left(\sum_{k=1}^{N_1} X_{1,k}\right) + [N(n)C_2 + C_3] + \sum_{i=1}^{n-1} p_1^*(i)\,E(\Omega_{n-i})\right] \Big/ [1 - p_1^*(0)]\,. \tag{4}$$

Note that $p_1^*(0) = [p(0)]^{N_1}$. For $n = 1$, we have $E(\Omega_1) = [C_1 N(1)\,E(X_{1,1}) + N(1)\,C_2 + C_3]/[1 - p^{N_1}(0)]$ and hence an optimal strategy for $N(1)$ can be found if $p(0)$ and $E(X_{1,1})$ are known. Iteratively, the optimal strategy for $N(2)$, $N(3)$, $N(n)$ can be found, since, in (4) $E(\Omega_n)$ is a function of $E(\Omega_k)$'s, $k \le n-1$ and other presumably known quantities.

**Remark 1:** In the case that the number of clusters available at a particular week is limited (deterministic or random), the minimum-cost strategy will be $N' = \min(N, N^*)$, where $N$ is the minimum-cost strategy described above and $N^*$ is the available number of clusters. Denoting $\Omega_{n,N}$ the total costs for selecting the strategy $N$, then $E(\Omega_{n,N'}) = E\left[E(\Omega_{n,N'} \mid N^*)\right]$ is minimized since $E(\Omega_{n,N'} \mid N^*)$ is minimized among all possible realizations of random strategy $N^*$.

**Proposition 1:** Under the minimum-cost strategy proposed above, the expected total cost is a monotone increasing function of the sample size, i.e. $E(\Omega_{n-1}) \le E(\Omega_n)$, for any $n$.

**Proof:** We will prove the contrapositive. Suppose that $E(\Omega_{n-1}) > E(\Omega_n)$. Let $E(\Omega'_{n-1})$ denote the total cost resulted from choosing $N'(n-1) = N(n)$, $N'(n-2) = N(n-1)$, $\ldots N'(1) = N(2)$ and stopping when $\sum\sum X_{t,k} \ge n-1$. Then $E(\Omega'_{n-1}) \le E(\Omega_n)$, since both strategies allocate the same numbers of clusters but the stopping rule for $E(\Omega'_{n-1})$ stops before the strategy used for $E(\Omega_n)$. This implies that $E(\Omega'_{n-1}) \le E(\Omega_n) < E(\Omega_{n-1})$ which contradicts the fact that $E(\Omega_{n-1})$ is the expected total cost for the optimal strategy.

## 4. Moments of the Duration of the Recruitment

Let $I_i$ be an indicator random variable assigning value 1 if the Markov chain $\{n_t\}$ ever reaches $i$, and 0 otherwise, i.e. $I_i = \begin{cases} 1 & \text{if} \quad n_t = i \quad \text{for some } t \\ 0 & \text{otherwise} \end{cases}$. For $i = 0, 1, 2, \dots n$, let $\pi_i = \Pr(I_i = 1)$. Since $n_1 = n$ and hence $\pi_n = 1$, we have

$$\begin{aligned}\pi_{n-1} &= \Pr(I_{n-1} = 1 \mid I_n = 1)\,\pi_n + \Pr(I_{n-1} = 1 \mid I_n = 0)\,(1 - \pi_n) \\ &= \sum_{i=1}^{\infty} [p_n^*(0)]^i\, q_{n,n-1} = q_{n,n-1}/(1 - p_n^*(0))\,,\end{aligned}$$

where the summand in the summation term indicates number of times the process stay at $n$. Similarly, for $k = 2, 3, \dots, n-1$,

$$\begin{aligned}\pi_{n-k} &= \sum_{j=0}^{k-1} \Pr(I_{n-k} = 1, I_{n-l} = 0\,,\ \text{for } k > l > j \mid I_{n-j} = 1)\Pr(I_{n-j} = 1) \\ &= \sum_{j=0}^{k-1} \frac{q_{n-j,n-k}}{1 - p_{n-j}^*(0)} \Pr(I_{n-j} = 1) = \sum_{j=0}^{k-1} \frac{q_{n-j,n-k}}{1 - p_{n-j}^*(0)}\, \pi_{n-j}\,. \qquad (5)\end{aligned}$$

In (5), the first equality is held since, if the Markov chain $\{n_t\}$ ever reaches $i$, it has to come from a previous state, and the second equality is the calculation of $\Pr(I_{n-k} = 1,\ I_{n-l} \neq 1,\ \text{for } k > l > j \mid I_{n-j} = 1)$ which can be obtained similarly to the calculation of $\pi_{n-1}$ above.

Note that, if $p(0) < 1$, then $p^*(0) < 1$ and hence $\Pr(n_t \text{ reaches } 0) = 1$. That is $\pi_0 = 1$. From (5), we have $\pi - \boldsymbol{e} = (Q' - P_\Delta)\,R_\Delta \pi$, where $\pi = (\pi_0, \pi_1, \dots \pi_n)'$, $\boldsymbol{e} = (0, \dots, 0, 1)'$ is a $n+1$ dimensional vector, $P_\Delta = \text{diag}\,(1, p_1^*(0),\ p_2^*(0), \dots,\ p_n^*(0))$ is a diagonal matrix of dimension $(n+1) \times (n+1)$, $R_\Delta = \text{diag}\,(0, 1/1 - p_1^*(0),\ 1/1 - p_2^*(0), \dots, 1/1 - p_n^*(0))$ is a diagonal matrix of dimension $(n+1) \times (n+1)$ and $Q'$ is the transpose of the transition matrix defined in (2). Note that $(Q' - P_\Delta)\,R_\Delta$ is the transpose of the transition probability matrix of $\{n_t\}$ conditioned on the event that $\{n_t\}$ does not stay in the same state. The explicit form for $\pi$ can be stated in the following lemma.

**Lemma 1:** Under the condition $p(0) < 1$, $\pi = [I - (Q' - P_\Delta)\,R_\Delta]^{-1}\,\boldsymbol{e}$, if the inverse exists.

**Remark 2:** By the nature of $(Q' - P_\Delta)\,R$, $[(Q' - P_\Delta)\,R_\Delta]^{n+1} = 0$, the zero matrix. Therefore $[I - (Q' - P_\Delta)\,R_\Delta]^{-1}$ is equal to a finite sum $\sum_{i=0}^{n} [(Q' - P_\Delta)\,R_\Delta]^i$, instead of an infinite sum.

Let $m_i$ denote the number of times that $\{n_t\}$ stays at $i$. From the Markovian property, we can derive that the conditional distribution of $m_i$ given $I_i = 1$ follows a geometric distribution with mean $1/1 - p_i^*(0)$ (consequently, variance equal to $p_i^*(0)/1 - p_i^*(0))$ and the conditional distribution of $m_i$ given $I_i = 0$ is identically 0. Therefore, by repeatedly using the above property and the property of the conditional expectation we have

$$E(m_i) = (1/1 - p_i^*(0))\, P(I_i = 1) \tag{6}$$

and

$$\begin{aligned}\mathrm{Var}\,(m_i) &= \mathrm{Var}\,(E(m_i \mid I_i)) \;+\; E(\mathrm{Var}\,(m_i \mid I_i)).\\ &= E[(E(m_i \mid I_i))^2] - [E(E(m_i \mid I_i))]^2 + E(\mathrm{Var}\,(m_i \mid I_i))\\ &= (E(m_i \mid I_i = 1))^2\, \pi_i - [E(m_i)]^2 + \mathrm{Var}\,(m_i \mid I_i = 1)\, \pi_i\\ &= \left[\frac{1}{1 - p_i^*(0)}\right]^2 \pi_i - \left[\frac{1}{1 - p_i^*(0)}\right]^2 (\pi_i)^2 + \frac{p_i^*(0)}{[1 - p_i^*(0)]^2}\, \pi_i\,.\end{aligned} \tag{7}$$

Furthermore, for $i < j$,

$$\begin{aligned}E(m_i m_j) &= \sum_{k,\,l=0 \text{ or } 1} E(m_i m_j \mid I_i = k,\, I_j = l)\, \Pr\,(I_i = k,\, I_j = l)\\ &= E(m_i m_j \mid I_i = 1,\, I_j = 1)\, \Pr\,(I_i = 1,\, I_j = 1)\\ &= E(m_i \mid I_i = 1)\, E(m_j \mid I_j = 1)\, \Pr\,(I_i = 1 \mid I_j = 1)\, \Pr\,(I_j = 1)\\ &= \left[\frac{1}{1 - p_i^*(0)}\right] \left[\frac{1}{1 - p_j^*(0)}\right] \Pr\,(I_i = 1 \mid I_j = 1)\, \pi_j\,,\end{aligned} \tag{8}$$

where the second equality is the result of $(m_i \mid I_i = 0) \equiv 0$, the third equality is due to the conditional independence of $m_i$ and $m_j$ given $I_i = 1$ and $I_j = 1$, and the last equality is the consequence of the geometric distribution of $m_i$ given $I_i = 1$. Note that $\Pr\,(I_i = 1 \mid I_j = 1) \neq \Pr\,(I_{n-(j-i))} = 1 \mid I_n = 1)$, since $N(n)$, $N(n-1), \ldots$ and $N(n-(j-i))$ may not be equal to $N(j)$, $N(j-1)$, $\ldots$, and $N(j-i)$ correspondingly. When $N(\;)$ is a constant, $\Pr\,(I_i = 1 \mid I_j = 1) = \pi_{n-(j-i).}$. Nevertheless $\Pr\,(I_i = 1 \mid I_j = 1)$ can be calculated using the method similar to lemma 1 by letting $n = j$.

From (6) the expected duration of recruitment is

$$E(m) = \sum_{i=1}^{n} E(m_i) = \sum_{i=1}^{n} (1/1 - p_i^*(0)\, \pi_i\,. \tag{9}$$

The explicit form for the expected duration of recruitment can be derived in proposition 2.

**Proposition 2:** Under the condition stated in lemma 1, $E(m) = \mathbf{1}'(I - Q + A)^{-1}\, \boldsymbol{e} - 1$, where $A$ is a $(n+1) \times (n+1)$ matrix with the (1,1) element equal to 1 and all other elements equal to 0, $\boldsymbol{1}$ is a $n+1$ dimensional vector with all elements equal to 1 and $\boldsymbol{e} = (0, \ldots, 0, 1)'$ as defined before.

**Proof:** By lemma 1 and (9), we have

$$\begin{aligned} E(m) &= \mathbf{1}'(I - P_\Delta + A)^{-1}\,\pi - 1 \\ &= \mathbf{1}'(I - P_\Delta + A)^{-1}\,[I - (P' - P_\Delta)\,R_\Delta]^{-1}\,\mathbf{e} - 1 \\ &= \mathbf{1}'[(I - (Q' - P_\Delta)\,R_\Delta)\,(I - P_\Delta + A)]^{-1}\,\mathbf{e} - 1 \\ &= \mathbf{1}'(I - Q' + A)^{-1}\,\mathbf{e} - 1\,, \end{aligned}$$

where the first equality results from the fact that $\pi_0 = 1$ and the last equality is the consequence of $R_\Delta A = 0_{(n+1)\times(n+1)}$, $R_\Delta(I - P_\Delta) = (I - A)$ and $Q'A = P_\Delta A$.

**Proposition 3:** The second moment of the duration of recruitment can be expressed as

$$E(m^2) = \sum_{i=1}^{n}\left[\frac{1 + p_i^*(0)}{\left(1 - p_i^*(0)\right)^2}\right]\pi_i + 2\sum_{i<j}\left[\frac{1}{1 - p_i^*(0)}\right]\left[\frac{1}{1 - p_j^*(0)}\right]\Pr\,(I_i = 1 \mid I_j = 1)\,\pi_j\,.$$

**Proof:** Since $m = \sum_{i=1}^{n} m_i$,

$$E(m^2) = \sum_{i=1}^{n} E(m_i^2) + 2\sum_{i<j} E(m_i m_j) = \sum_{i=1}^{n}\,[(E(m_i))^2 + \mathrm{Var}\,(m_i)] + 2\sum_{i<j} E(m_i m_j)\,.$$

Equations (6), (7) and (8) and a routine algebra complete the proof.

## 5. Extension of the Model to Multi-Center Clinical Trials

In a multi-center clinical trial, the recruitment of the center is similar to the recruitment of the cluster as described in this paper, except for the calculation of the total cost. Ordinarily, the centers in the clinical trials are not dropped before its conclusion; therefore the costs of the centers will be cumulated over time. Detailed discussions on the issue of the number of centers in a multi-center clinical trial can be found in GOLDBERG and KOURY (1990). In this section, we will extend our model and use it to find the optimal number of centers for a clinical trial that will minimize the expected total cost. All centers will be recruited only at the beginning of the trials and they will be responsible for enrolling patients until the sample size is reached. In this section, the sample size for statistical analysis is supposed to have been calculated from a fixed-effects model and therefore, is independent of the number of centers. We will further assume that the duration of follow-up as the last patient is enrolled is a constant $T$ and no dropout will change the calculation of the cost. Therefore the expected total cost for this trial can be expressed as

$$E(\Omega_n) = E\left[C_1\sum_{i=1}^{m}\sum_{j=1}^{N} X_{ij}(m - i + 1) + C_1 T\sum_{i=1}^{m}\sum_{j=1}^{N} X_{ij} + C_2 N(m + T) + (m + T)\,C_3\right], \tag{10}$$

where $\Omega_n$, as in (1) denotes the total cost for this trial if the required sample size is $n$, $C_1$ denotes the cost per week for each patient, $C_2$ denotes the cost per week for each center, $C_3$ denotes the cost per week for the coordinating center, and $X_{ij}$ represents the number of subjects enrolled at $j$th center in the $i$th week, and patients will therefore be followed for $(m-i+1)$ weeks before the last patient recruitment. Note that, in (10) the first term on the right represents the cost of patients before the recruitment of last patient and the second term represents the cost of patients after the last patient recruitment. Similar to what described in section 2, the patient recruitment stopped if $\sum_{i=1}^{m}\sum_{j=1}^{N} X_{ij} \geq n$. Therefore $m$ is a random variable and has the following properties:

**Lemma 2:** With respect to the Markov process $Y_i = \sum_{j=1}^{N} X_{ij}$, $m$, $m+1$ and $\dfrac{m(m+1)}{2}$ are stopping times.

**Proof:** Since $m$ is defined to be such that $\sum_{i=1}^{m} Y_i \geq n$, $\{m=k\}$ is independent of $\{Y_{k+1}, Y_{k+2}, \ldots\}$ for all integer $k$ and hence $m$ is a stopping time with respect to $\{Y_i\}$. Now observe that $\{m+1=k\} = \{m=k-1\}$ is independent of $\{Y_k, Y_{k+1}, \ldots\}$, and hence is independent of $\{Y_{k+1}, Y_{k+2}, \ldots.\}$. Therefore $m+1$ is a stopping time. Similarly $\dfrac{m(m+1)}{2}$ is a stopping time, since for any possible value $k$ of $\dfrac{m(m+1)}{2}$, there is one and only one integer $k' \leq k$ such that $\left\{\dfrac{m(m+1)}{2} = k\right\} = \{m = k'\}$.

From lemma 2 and Wald's equation [see, for example ROSS (1983)] we can calculate (10) as

$$
\begin{aligned}
&E(\Omega_n)\\
&= C_1\left(E(Y_1)\,E[m(m+1)] - E\left[\sum_{i=1}^{m} iY_i\right]\right) + C_1TE(Y_1)\,E(m) + C_2N(E(m)+T)\\
&\quad + C_3(E(m)+T)\\
&= C_1E(Y_1)\left(E[m(m+1)] - E\left[\frac{m(m+1)}{2}\right]\right) + C_1TE(Y_1)\,E(m) + C_2N(E(m)+T)\\
&\quad + C_3(E(m)+T)\\
&= C_1E(Y_1)\left(E\left[\frac{m^2}{2}\right] + E\left[\frac{m}{2}\right] + TE(m)\right) + C_2N(E(m)+T) + C_3(E(m)+T)\,.
\end{aligned}
$$

From the above equation, if $E(m)$ and $E\left(\dfrac{m^2}{2}\right)$ are calculated, the optimal number of centers can be obtained from a routine integral programming procedure. The expected duration of recruitment $E(m)$ calculated in section 5 can now be simplified as $E(m) = (1/1 - p^*(0)) \sum_{i=1}^{n} P(I_i = 1)$, where $p^*(0)$ is a $N$-fold convolution of $p(0)$. Note that $\sum_{i=1}^{n} P(I_i = 1) \neq 1$. Similarly, $E(m^2)$ calculated in proposition 3 can be simplified as $E(m^2) = \sum_{i=1}^{n} \frac{1 + p^*(0)}{(1 - p^*(0))^2} \pi_i + 2 \sum_{i<j} \frac{1}{(1 - p^*(0))^2} \pi_{n-(j-i)} \pi_j$, when $N$ is a constant.

## 6. Discussion

The technique proposed in this paper is different from the optimal stopping rule method applied in most decision questions. The usual method is to decide if the cluster recruitment should stop (even the patient recruitment still continuous until the planned sample size is reached) in order to minimize the expected total costs after observing the number of patients enrolled and the number of clusters recruited. The common method usually stops before the target sample size is reached and can not guarantee it will reach. In contrast, our method stops cluster recruitment after the target sample size is reached.

The proposed method is also different from the sample size re-estimation method appeared in the clinical trials literature. The later is a technique used to adjust sample size by an interim analysis without un-blinding the trial results in progress [see, for example, Shih (1992)]. The proposed methods is to present a strategy for cluster recruitment by monitoring the cumulative number of observations in each week, when keeping the study sample size as an unchanged target.

If the sample size $n$ is large, the minimum-cost strategy proposed in section 3 may be computationally expensive. A revision of our strategy, which is computationally efficient but not mathematically optimal, is to find a constant $N$ that depends on the sample size $n_t$ at each week $t$ only through $E(m)$. Observe that, from (1),

$$\begin{aligned} E(\Omega_n) &= C_1 E(m)\, N E(X_{11}) + C_2 N E(m) + C_3 E(m) \\ &= E(m)\, (C_1 N E(X_{11}) + C_2 N + C_3)\,. \end{aligned}$$

Since $E(m)$ can be calculated from Proposition 2, an optimal $N$ called $N^{(1)}$, for week 1 can be obtained from a routine integral programming procedure. Iteratively, replacing $n$ by $n - n_1$ we can calculate $E(m)$ and obtain $N^{(2)}$. This iterative procedure will be proceeded until the target sample size is reached. This revised strategy is better than the strategy using constant $N$ at each week and is computationally more efficient than the one proposed in section 3.

## Acknowledgements

Part of this research was done when Dr. Peng was visiting School of Public Health, University of Texas – Houston. He wishes thank National Science Council, Taiwan for financial support and UT-Houston for providing physical facilities.

*References*

DONNER, A., 1992: Sample size requirements for stratified cluster randomization designs. *Statistics in Medicine*. **11**, 743–750.

DONNER, A. and KLAR, N., 1994: Cluster randomization trials in epidemiology: Theory and application. *J. of Statistical Planning and Inference* **42**, 37–56.

GOLDBERG, J. D. and KOURY, K. J., 1990: *Design and Analysis of Multicenter Trials in Statistical Methods in the Pharmaceutical science* edited by Berry, D. Marcel Dekker, New York.

FELDMAN, H., MCKINLAY, S. and NIKNIAN, M., 1996: Batch sampling to improve power in a community trial: Experience from the Pawtucket Heart Health Program. *Evaluation Review* **20**, 244–274.

HSIEH, F., 1988: Sample size formulae for intervention studies with the cluster as unit of randomization. *Statistics in Medicine* **7**, 1195–1201.

MCKINLAY, S., 1994: Cost-efficient design of cluster unit trials. *Preventive Medicine* **23**, 606–611.

PUTERMAN, M. L., 1994: *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, New York.

ROSS, S. M., 1983: *Stochastic Processes.* Wiley, New York.

SHIH, W. J., 1992: Sample size re-estimation in clinical trials. Chapter 18. In: Karl Peace (ed).*: Biopharmaceutical Sequential Statistical Applications*. Marcel Dekker, New York, 285–301.

WENYAW CHAN<br>
Program in Biometry
School of Public Health
University of Texas – Houston
Houston, Texas 77030
U.S.A.

NAN FU PENG
Institute of Statistics
National Chiao Tung University
Hsinchu 30050
Taiwan

Received, August 1999
Revised, August 2000
Accepted, August 2000