Biometrika (1999), **86**, 3, *pp.* 517–529
© 1999 Biometrika Trust
Printed in Great Britain

Interim analyses using repeated confidence bands

By X. JOAN HU and STEPHEN W. LAGAKOS
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.
joanhu@sdac.harvard.edu lagakos@biostat.harvard.edu

Summary

Clinical trials often include interim analyses that compare treatment groups with respect to the mean function of a response process. Sometimes it is unclear how the mean functions of the groups might differ, and thus one cannot confidently prespecify a simple metric upon which a stopping rule or repeated confidence interval can be based. This motivated us to extend the repeated confidence intervals approach for a finite-dimensional parameter (Jennison & Turnbull, 1989) to the use of repeated confidence bands for the mean function of a response process. Formal tests of hypotheses are easily constructed from the repeated confidence bands. We also describe how inferences for the mean function can be adaptively restricted to different subsets of its domain at different interim analyses. An example is given involving an AIDS clinical trial.

Some key words: Clinical trial; Group sequential analysis; Mean function; Nonparametric inference.

1. Introduction

Preplanned interim analyses of clinical trials are commonly used to monitor patient safety and the evolving efficacy of treatment groups (Jennison & Turnbull, 1990; Whitehead, 1997). Usually, the interim analyses are guided by prespecified termination criteria, or stopping rules, that specify both the metric on which the relative efficacy of the groups is measured and how the overall type I error rate will be spent during the multiple examinations of the data (Pocock, 1977; O'Brien & Fleming, 1979; Lan & DeMets, 1983). In some circumstances, however, relatively little may be known about the nature and magnitude of the expected treatment effect or the clinical significance of certain types of differences in the outcome measure. This is common in AIDS studies, for instance. In such settings, a repeated confidence intervals approach (Jennison & Turnbull, 1989) is attractive but this requires the metric of interest to be expressed as a finite-dimensional parameter. When the response is a stochastic process, such as repeated measures of a biological marker or a failure time process, there may be little confidence in the type of magnitude of treatment differences that will occur and hence no obvious metric for comparing groups.

Motivated by these considerations, we extend the repeated confidence intervals approach for a finite-dimensional parameter to the use of repeated confidence bands for the mean function of a general response process. Special cases include outcomes such as repeated measures of a laboratory marker, failure time endpoints and recurrent event data. Assuming that there is an estimator of the mean function that, when standardised, is asymptotically Gaussian, we develop a general procedure for the construction of repeated confidence bands. This allows us to learn about the mean function of a specific treatment group or the difference between two groups at each interim analysis. Also, should the trial be terminated, a valid confidence band is provided for use in the interpretation of the study results. The same techniques can also be used to carry out formal hypothesis tests. A unique feature of the approach is that inferences for the mean function can be adaptively restricted to different subsets of its domain at each interim analysis.

In § 2 we give a general definition of repeated confidence bands, describe three settings in which the techniques can be applied, and present the construction of a specific set of repeated confidence bands. In § 3 we discuss the use of different subsets of the domain of the mean function at different interim analyses. We consider the formation of hypothesis tests from repeated confidence bands in § 4. Section 5 gives an illustrative example based on a recent AIDS trial, and § 6 comments on some related issues.

2. SETTING AND CONSTRUCTION OF REPEATED CONFIDENCE BANDS

2·1. *Definitions and assumptions*

Consider a clinical trial that begins at calendar time 0 and that enrols and follows n subjects over time. Let $X_i(s)$ denote the value of the response process at time s for subject i, where s denotes time measured from entry into the trial. Denote the common mean function by $\mu(.)$; that is $\mu(s) = E\{X_i(s)\}$ for $s \in [0, S]$. Throughout we use t to denote calendar or external time and s to denote study or internal time. Suppose K analyses of the trial are conducted at prespecified calendar times $T_1, \ldots, T_K$, where $0 < T_1 < \ldots < T_K$. Let $\mathscr{H}_k$ denote the available information from the trial up to calendar time T_k. For example, if a_i is the calendar time of entry for subject i, $\mathscr{H}_k$ might include the values of $X_i(s)$ that are observed for $s \leqslant T_k - a_i$ for those subjects who are enrolled by time T_k.

Suppose that $B_k = B_k(\mathscr{H}_k)$ is a random subset of the space $\mathscr{S}[0, S]$ of all functions defined on $[0, S]$. Then $\{B_k : k = 1, \ldots, K\}$ is called a set of level $1 - \alpha$ repeated confidence bands for the mean function $\mu(.)$ on $[0, S]$ if

$$\text{pr}\{\mu(.) \in B_k \text{ for } 1 \leqslant k \leqslant K\} \geqslant 1 - \alpha. \tag{2·1}$$

This definition of repeated confidence bands is a natural generalisation of the notion of repeated confidence intervals for a finite-dimensional parameter (Jennison & Turnbull, 1989). We define the spending function for $\{B_k : k = 1, \ldots, K\}$ by $\{\pi_1, \ldots, \pi_K)$, where, for $k = 1, \ldots, K$,

$$\text{pr}\{\mu(.) \in B_l \text{ for } 1 \leqslant l \leqslant k-1;\ \mu(.) \notin B_k\} = \pi_k.$$

Note that $\sum_{k=1}^{K} \pi_k \leqslant \alpha$. There are some applications where the interest at the kth analysis is on values of $\mu(.)$ in a subset of $[0, S]$, $\mathscr{E}_k$ say. As we show later, these situations are accommodated in the above definition by defining B_k to be noninformative about $\mu(s)$ for $s \notin \mathscr{E}_k$.

We shall assume that there is an estimator of the mean function, $\hat{\mu}_n(s; t)$, say, based on the information available at external time t, such that, as $n \to \infty$, the K-dimensional process $n^{\frac{1}{2}}\{\hat{\mu}_n(s) - \mu(s)\mathbb{1}\}$ converges weakly to a K-dimensional zero-mean Gaussian process $\mathscr{G}(s)$ for $s \in \mathscr{E} \subseteq [0, S]$, where $\hat{\mu}_n(s) = (\hat{\mu}_n(s; T_1), \ldots, \hat{\mu}_n(s; T_K))'$ and $\mathbb{1}$ denotes the K-dimensional vector of ones. The set $\mathscr{E}$ denotes the values s for which $\hat{\mu}_n(.)$ weakly converges, and, as the following examples illustrate, will usually depend on the underlying process governing the values of s for which $X_i(s)$ is observable. Without loss of generality, we take $\mathscr{E} = [0, S]$

except for Example 1 below. We denote the covariance function for $\mathscr{G}(s)$ by the $K \times K$ matrix function $\operatorname{cov}(s, u)$ with (l, m) element being the function

$$\operatorname{cov}(s, u)_{lm} = \operatorname{cov}\{\mathscr{G}_l(s), \mathscr{G}_m(u)\},$$

where $\mathscr{G}_l(s)$ is the lth component of $\mathscr{G}(s)$.

2·2. *Examples*

Example 1: *Repeated measures responses.* Suppose that $X_1, \ldots, X_n$ are independent and identically distributed, and $X_i(s_j)$ is the value of a measured outcome for subject i at study time s_j, for $j = 1, \ldots, J$. For example, in an AIDS trial, $X_i(s_j)$ might denote a measure of viral burden observed at study month s_j. Here $\mu(s_j)$ denotes the mean viral load at s_j. A natural commonly used nonparametric estimator of $\mu(s_j)$ based on information available at external time t is given by

$$\hat{\mu}_n(s_j;\, t) = \frac{\sum_{i=1}^n \delta_i(s_j;\, t) X_i(s_j)}{\sum_{i=1}^n \delta_i(s_j;\, t)}, \tag{2·2}$$

where $\delta_i(s;\, t)$ is an indicator of whether or not $X_i(s)$ is observed at time t and is assumed to be independent of the response process. That is, the estimator of $\mu(s_j)$ at time t is simply the average of the observed values of $X_i(s_j)$ at time t. For completeness, $\hat{\mu}_n(s_j;\, t)$ can be defined arbitrarily when $\sum_{i=1}^n \delta_i(s_j;\, t) = 0$. The set $\mathscr{E}$ consists of those s_j for which $X_i(s)$ is observed with positive probability. When t and n are large enough, the estimator in (2·2) is well defined and unbiased for $s \in \mathscr{E}$, and under mild conditions the vector-valued process $n^{\frac{1}{2}}\{\hat{\mu}_n(s) - \mu(s)\mathbb{1}\}$ converges to a K-dimensional zero-mean Gaussian process for $s \in \mathscr{E}$. Proofs of this and the weak convergence results noted in the following two examples are available upon request from the authors.

For the special case where the observed values of $X_i(s)$ are caused entirely by staggered entry of subjects into the trial, $\delta_i(s;\, t) = I(s \leqslant t - a_i)$, where a_i is the entry time of subject i. Here the (l, m) element of the asymptotic covariance function of the vector process $n^{\frac{1}{2}}\{\hat{\mu}_n(s) - \mu(s)\mathbb{1}\}$ can be shown to equal

$$\operatorname{cov}(s, u)_{lm} = \frac{\operatorname{cov}\{X_1(s), X_1(u)\}}{H(\max\{(T_l - s), (T_m - u)\})}$$

for $s, u \in \mathscr{E}$ and $l, m = 1, \ldots, K$, where $H(.)$ is the cumulative distribution function of a_i.

In general, the covariance function $\operatorname{cov}(s, u)$ is estimated consistently by the moment estimator

$$\widehat{\operatorname{cov}}(s, u)_{lm} = \frac{n \sum_{i=1}^n \delta_i(s;\, T_l)\delta_i(u;\, T_m)}{\sum_{i=1}^n \delta_i(s;\, T_l) \sum_{j=1}^n \delta_j(u;\, T_m)} \operatorname{cov}\{X_1(s), X_1(u)\}.$$

When $\operatorname{cov}\{X_1(s), X_1(u)\}$ is unknown, it may be replaced at the kth analysis by a consistent estimator, such as

$$\hat{c}_n(s, u;\, T_k) = \frac{\sum_{i=1}^n \delta_i(s;\, T_k)\delta_i(u;\, T_k)\{X_i(s) - \hat{\mu}_n(s;\, T_k)\}\{X_i(u) - \hat{\mu}_n(u;\, T_k)\}}{\sum_{i=1}^n \delta_i(s;\, T_k)\delta_i(u;\, T_k)}$$

or

$$\tilde{c}_n(s, u;\, T_k) = \frac{\sum_{i=1}^n \delta_i(s;\, T_k)\delta_i(u;\, T_k) X_i(s) X_i(u)}{\sum_{i=1}^n \delta_i(s;\, T_k)\delta_i(u;\, T_k)} - \hat{\mu}_n(s;\, T_k)\hat{\mu}_n(u;\, T_k).$$

Example 2: *Recurrent events.* Suppose that $X_1, \ldots, X_n$ are independent and identically distributed, and let $X_i(s)$ denote the number of occurrences of some recurrent event by internal time s, so that $\mu(s)$ is the expected number of events by internal time s. For example, $X_i(s)$ may denote the number of asthma attacks in child i by time s. Suppose further that at calendar time t we observe $X_i(s)$ for $0 \leqslant s \leqslant s_i(t)$, where $s_i(t)$ is independent of $X_i(.)$. Then, with $\delta_i(s;\, t) = I\{s \leqslant s_i(t)\}$, a natural nonparametric estimator of $\mu(.)$ is given by

$$\hat{\mu}_n(s;\, t) = \sum_{i=1}^{n} \int_0^s \frac{\delta_i(u;\, t) I\{\sum_{j=1}^n \delta_j(u;\, t) > 0\}}{\sum_{j=1}^n \delta_j(u;\, t)}\, dX_i(u). \tag{2·3}$$

When the observation process $\delta_i(s;\, t)$ arises from staggered entry, (2·3) is used by Cook and Lawless (1996). This statistic also arises as the Nelson–Aalen estimator (Fleming & Harrington, 1991, p. 121) of the cumulative intensity function when the response is a Poisson process with mean function $\mu(.)$, for here the cumulative intensity at time s equals the expected number of events by s.

Under some mild conditions, the vector process $n^{\frac{1}{2}}\{\hat{\mu}_n(s) - \mu(s)\mathbb{1}\}$ converges weakly to a K-dimensional, zero-mean Gaussian process with covariance function $\mathrm{cov}(s, u)$ that is consistently estimated by

$$\widehat{\mathrm{cov}}(s, u)_{lm} = \int_0^s \int_0^u \frac{n \sum_{i=1}^n \delta_i(v;\, T_l)\delta_i(w;\, T_m)}{\sum_{i=1}^n \delta_i(v;\, T_l) \sum_{j=1}^n \delta_j(w;\, T_m)}\, \mathrm{cov}\{dX_1(v), dX_1(w)\}.$$

The covariance function $\mathrm{cov}\{dX_1(s), dX_1(u)\}$ in the formula above may be replaced with a consistent estimator.

Example 3: *Survival data.* A special case of Example 2 is where the response is the time until an event, such as death or clinical progression of a disease. Let $X_i(s)$ indicate whether or not subject i has reached the event by internal time s, so that $1 - \mu(s)$ is the probability of survival to time s. Given the right-censored observations as in Example 2, the Kaplan–Meier estimator (Kaplan & Meier, 1958) of the survival function using data up to external time t can be expressed as

$$\hat{\mu}_n(s;\, t) = 1 - \prod_{u<s} \left[1 - \frac{\sum_{i=1}^n \delta_i(u;\, t)\, dX_i(u)}{\sum_{j=1}^n \delta_j(u;\, t)\{1 - X_j(u^-)\}} \right], \tag{2·4}$$

and the vector process $n^{\frac{1}{2}}\{\hat{\mu}_n(s) - \mu(s)\mathbb{1}\}$ converges weakly to a zero-mean Gaussian process. When the interval of observation arises from staggered entry with no other censoring, $\delta_i(s;\, t) = I(s \leqslant t - a_i)$ and the covariance function of the Gaussian process has (l, m) element

$$\mathrm{cov}(s, u)_{lm} = \{1 - \mu(s)\}\{1 - \mu(u)\} \int_0^{\min\{s,u\}} \frac{\mathrm{var}\{x_1(v)\}}{H(\max\{T_l, T_m\} - v)\{1 - \mu(v)\}}\, dv,$$

where $x_1(s)\, ds = dX_1(s)$ and $H(.)$ is the cumulative distribution of the a_i. Note that $\mu(s) = 1 - \exp\{-\Lambda(s)\}$, where $\Lambda(s)$ is the cumulative hazard function of the survival time. An alternative estimator for $\mu(.)$ at the external time t is thus

$$\tilde{\mu}_n(s;\, t) = 1 - \exp\{-\hat{\Lambda}_n(s;\, t)\}, \quad s \in [0, S], \tag{2·5}$$

where $\hat{\Lambda}_n(.;\, t)$ is the usual Nelson–Aalen estimator of Λ based on information collected up to time t:

$$\hat{\Lambda}_n(s;\, t) = \sum_{i=1}^{n} \int_0^s \frac{\delta_i(u;\, t)}{\sum_{j=1}^{n} \delta_j(u;\, t)\{1 - X_j(u^-)\}}\, dX_i(u).$$

Estimators (2·4) and (2·5), when standardised, have the same limiting distribution.

2·3. *Construction of repeated confidence bands*

In this section we present the construction of a specific set of repeated confidence bands for the setting where each B_k is intended to be informative about $\mu(s)$ for all $s \in [0, S]$. For $k = 1, \dots, K$, define

$$S_n(s;\, T_k) = \{n/\text{var}(s;\, T_k)\}^{\frac{1}{2}}\{\hat{\mu}_n(s;\, T_k) - \mu(s)\}, \tag{2·6}$$

where $\text{var}(s;\, T_k) = \text{cov}(s, s)_{kk}$ denotes the asymptotic variance of $n^{\frac{1}{2}}\{\hat{\mu}_n(s;\, T_k) - \mu(s)\}$. Under the assumptions in § 2·1 about the asymptotic behaviour of $\hat{\mu}_n(s;\, t)$, the vector process with kth component $S_n(.;\, T_k)$ converges weakly to a Gaussian process. A set of level $1 - \alpha$ repeated confidence bands for $\mu(.)$ can be constructed as

$$B_k = \Bigg\{ v(.) : \text{for all } s \in [0, S],$$

$$v(s) \in \left[\hat{\mu}_n(s;\, T_k) - c_k \left\{\frac{\text{var}(s;\, T_k)}{n}\right\}^{\frac{1}{2}},\ \hat{\mu}_n(s;\, T_k) + c_k \left\{\frac{\text{var}(s;\, T_k)}{n}\right\}^{\frac{1}{2}} \right] \Bigg\}, \tag{2·7}$$

for $k = 1, \dots, K$, where the constants c_k satisfy

$$\text{pr}\left\{ \sup_{s \in [0,S]} |S_n(s;\, T_l)| \leqslant c_l,\ 1 \leqslant l \leqslant k-1;\ \sup_{s \in [0,S]} |S_n(s;\, T_k)| > c_k \right\} = \pi_k. \tag{2·8}$$

Bands with different shapes can be formed by allowing c_k to depend on s. In the following we proceed as if the covariance function were known. In practice, it will usually be unknown and replaced with a consistent estimator.

We see that $S_n^{(k)}(s) = (S_n(s;\, T_1), \dots, S_n(s;\, T_k))'$ has a limiting zero-mean multivariate Gaussian process, denoted by $\xi_k(s)$. The (l, m) element of the covariance matrix $\text{cov}\{\xi_k(s), \xi_k(u)\}$ is

$$\text{cov}(s, u)_{lm}/\{\text{var}(s;\, T_l)\, \text{var}(u;\, T_m)\}^{\frac{1}{2}} \quad (l, m = 1, \dots, k).$$

In theory, we can thus choose the c_k to approximate (2·8) based on the asymptotic Gaussian distribution of the $S_n^{(k)}(.)$. However, the computations are generally difficult except in special cases, such as Example 1, where $\xi_k(.)$ is multivariate normal.

A simple approach for computing the c_k is obtained by extending an idea used in Lin, Wei & Ying (1993). Consider the setting of Example 1, and define

$$\tilde{S}_n(s;\, T_k) = \{n/\text{var}(s;\, T_k)\}^{\frac{1}{2}} \sum_{i=1}^{n} \left[\frac{\delta_i(s;\, T_k)\{X_i(s) - \hat{\mu}_n(s;\, T_k)\}}{\sum_{j=1}^{n} \delta_j(s;\, T_k)} \right] Z_i, \quad s \in \mathscr{E}, \tag{2·9}$$

where $Z_1, \dots, Z_n$ are independent random variables from the standard normal distribution, and are independent of $X_i(.)$ and $\delta_i(.)$, for $i = 1, \dots, n$. It is shown in the Appendix that $\tilde{S}_n^{(k)}(s) = (\tilde{S}_n(s;\, T_1), \dots, \tilde{S}_n(s;\, T_k))'$, conditional on $\{X_i, \delta_i\}$, has the same limiting distribution as $S_n^{(k)}(s)$. Since realisations from the conditional distribution of $\tilde{S}_n^{(k)}(s)$ can be computer generated, these can be used to approximate the critical values c_k. Specifically, given $c_1, \ldots, c_{k-1}$, we choose c_k as follows.

Algorithm

Step (a). *Generate M_k sets of independent realisations of* $(Z_1, \ldots, Z_n)$ *and, for each set, compute $\tilde{S}_n^{(k)}(.)$, denoted by $(\tilde{S}_n(.; T_1)^{(m)}, \ldots, \tilde{S}_n(.; T_k)^{(m)})'$, for $m = 1, \ldots, M_k$.*

Step (b). *Define $L_n(T_l)^{(m)} = \sup_{s \in [0,S]} |\tilde{S}_n(s; T_l)^{(m)}|$ for $l = 1, \ldots, k$, and select c_k so that a proportion π_k of the M_k realisations satisfy $L_n(T_l)^{(m)} \leqslant c_l$ for $1 \leqslant l \leqslant k-1$ and $L_n(T_k)^{(m)} > c_k$.*

The same type of algorithm can be used for other estimators. For example, in Example 2, we consider the vector process with kth component

$$\tilde{S}_n(s; T_k) = \{n/\mathrm{var}(s; T_k)\}^{\frac{1}{2}} \sum_{i=1}^{n} \left[\int_0^s \frac{\delta_i(u; T_k)}{\sum_{j=1}^{n} \delta_j(u; T_k)} d\{X_i(u) - \hat{\mu}_n(u; T_k)\} \right] Z_i, \qquad (2{\cdot}10)$$

where var$(s; t)$ is the asymptotic variance function of the Gaussian process associated with (2·3). Then the Algorithm may be employed to compute the c_k. Similarly, in the setting of Example 3, we employ the Algorithm by using

$$\tilde{S}_n(s; T_K) = \{n/\mathrm{var}^*(s; T_k)\}^{\frac{1}{2}} \sum_{i=1}^{n} \left[\int_0^s \frac{\delta_i(u; T_K)\{1 - X_i(u^-)\}}{\sum_{j=1}^{n} \delta_j(u; T_K)\{1 - X_j(u^-)\}} dX_i(u) \right] Z_i, \qquad (2{\cdot}11)$$

where var$^*(s; t)$ is the asymptotic variance in $n^{\frac{1}{2}}\{\hat{\Lambda}_n(s; t) - \Lambda(s)\}$ with $\hat{\Lambda}_n(s; t)$ defined following equation (2·5). Here we have used the fact that the process with kth component given by (2·6) has the same limiting distribution as the process with kth component $\{n/\mathrm{var}^*(s; T_k)\}^{\frac{1}{2}}\{\hat{\Lambda}_n(s; T_k) - \Lambda(s)\}$.

More generally, suppose that for $k = 1, \ldots, K$ we have an estimator of $\mu(.)$ such that the function in (2·6) can be expressed as, or is asymptotically equivalent to,

$$\{n/\mathrm{var}(s; T_k)\}^{\frac{1}{2}} \sum_{i=1}^{n} Y_{ni}(s; T_k \mid \mu),$$

where the Y_{ni} are independent, given the observation processes. Then it can be shown that this process has the same limiting distribution as

$$\tilde{S}_n(s; T_k) = \{n/\mathrm{var}(s; T_k)\}^{\frac{1}{2}} \sum_{i=1}^{n} Y_{ni}(s; T_k \mid \hat{\mu}_n) Z_i \qquad (2{\cdot}12)$$

conditional on $\{X_i, \delta_i\}$, where the Z_i are independent standard normal random variables, and thus the latter process can be used in an algorithm similar to the above algorithm to approximate the c_k.

The choice of $M_1, \ldots, M_K$ will depend in part on the number of interim analyses and the spending function. For a specific k, a simple guideline is to monitor the resulting value of c_k periodically as the number of simulations increases and stop when these values stabilise.

3. Varying the domain of inference

A unique feature of repeated inferences about a function, as opposed to a scalar parameter, is that inferences for $\mu(.)$ can be restricted to different subsets of its domain at each interim analysis. In general, $\mathscr{E}_k$ can be adaptively selected based on $\mathscr{H}_{k-1}$ and $\mathscr{Q}_k$, where $\mathscr{Q}_k$ denotes any information up to time T_k that is independent of the response process, such as the set of the enrolment times of subjects who enter the study by time T_k. As we shall see, this adds to the flexibility of the repeated confidence bands approach and has practical implications for efficiency.

One reason for varying the domain of interest at different interim analyses is that there may be insufficient information for a precise estimate of $\mu(s)$ at certain study times. For example, when the observed data are wholly a result of the staggered entry of subjects into the trial very few of the subjects may have completed the period of observation at the time of the first several interim analyses.

Another reason is a result of lack of knowledge about the outcome. For example, in an AIDS trial that partly motivated our interest in this problem, treatment efficacy is based on monthly observations of the patients' viral loads. The types of difference between treatment groups that might justify early termination of a trial were not known in advance, yet it was clear to the investigators that, if the trial were not terminated early, the primary basis for comparing the treatment groups at the time of the final analysis should be $\mu(S)$, the mean response at the largest observation time. In such a situation we might want to construct repeated confidence bands of $\mu(.)$ over its entire domain during the first $K-1$ interim analyses but then only for $\mu(S)$ at the final analysis. That is, if $\mathscr{E}_k$ denotes the domain of $\mu(.)$ of interest in the kth of the K analyses, we might choose $\mathscr{E}_k = [0, S]$ for $k < K$ and $\mathscr{E}_K = \{S\}$. More generally, the domain $\mathscr{E}_k$ could be chosen adaptively based on some rule and the information $(\mathscr{H}_{k-1}, \mathscr{Q}_k)$.

Apart from more naturally focusing attention on subsets of $[0, S]$ of interest, repeated confidence bands based on varying domains can have good efficiency properties. For example, if a conservative spending function is used, the resulting confidence interval for $\mu(S)$ at the time of the final analysis in the example above will be similar in width to the nominal confidence interval that disregards the multiple previous examinations of the data.

To construct repeated confidence bands with varying domains, suppose that $\mathscr{E}_k$ is chosen based on some function of $\mathscr{H}_{k-1}$ and $\mathscr{Q}_k$, that is $\mathscr{E}_k = \mathscr{E}_k(\mathscr{H}_{k-1}, \mathscr{Q}_k)$, for $k = 1, \ldots, K$. By '$\mu(.) \in B_l$' in the following we mean that $\mu(s) \in B_l(s)$ for $s \in \mathscr{E}_l$, where

$$B_l(s) = \{v(s) : v(.) \in B_l \subseteq \mathscr{S}[0, S]\}.$$

Note that B_k is informative for $\mu(s)$ only when $s \in \mathscr{E}_k$. That is, we can view $B_k(s)$ as $(-\infty, \infty)$ for $s \notin \mathscr{E}_k$. Then, if we choose B_k to be a random subset of $\mathscr{S}[0, S]$ such that

$$\operatorname{pr}\{\mu(.) \notin B_k \mid \mu(.) \in B_l,\ 1 \leqslant l \leqslant k-1;\ \mathscr{E}_k\} = \frac{\pi_k}{1 - \sum_{1 \leqslant l \leqslant k-1} \pi_l}, \tag{3·1}$$

with $\sum_{1 \leqslant l \leqslant 0} \pi_l = 0$, $\{B_k : k = 1, \ldots, K\}$ is a set of level $1-\alpha$ repeated confidence bands. This follows by mathematical induction since, first,

$$\operatorname{pr}\{\mu(.) \in B_1\} = 1 - E[\operatorname{pr}\{\mu(.) \notin B_1 \mid \mathscr{E}_1\}] = 1 - \pi_1$$

and, secondly, provided $\operatorname{pr}\{\mu(.) \in B_l,\ 1 \leqslant l \leqslant k-1\} = 1 - \sum_{1 \leqslant l \leqslant k-1} \pi_l$ for $k \geqslant 2$,

$$\begin{aligned}
&\operatorname{pr}\{\mu(.) \in B_l,\ 1 \leqslant l \leqslant k-1;\ \mu(.) \notin B_k\} \\
&\quad = \operatorname{pr}\{\mu(.) \in B_l,\ 1 \leqslant l \leqslant k-1\} \\
&\qquad \times E[\operatorname{pr}\{\mu(.) \notin B_k \mid \mu(.) \in B_l,\ 1 \leqslant l \leqslant k-1;\ \mathscr{E}_k\} \mid \mu(.) \in B_l,\ 1 \leqslant l \leqslant k-1] \\
&\quad = \left(1 - \sum_{1 \leqslant l \leqslant k-1} \pi_l\right) \times \left(\frac{\pi_k}{1 - \sum_{1 \leqslant l \leqslant k-1} \pi_l}\right) \times E[1 \mid \mu(.) \in B_l,\ 1 \leqslant l \leqslant k-1] = \pi_k,
\end{aligned}$$

from which (2·1) follows.

A specific confidence band B_k with adaptive domain can be computed similarly to (2·7) as

$$B_k = \left\{ v(.) : v(s) \in \left[\hat{\mu}_n(s;\, T_k) - c_k \left\{ \frac{\text{var}(s;\, T_k)}{n} \right\}^{\frac{1}{2}},\ \hat{\mu}_n(s;\, T_k) + c_k \left\{ \frac{\text{var}(s;\, T_k)}{n} \right\}^{\frac{1}{2}} \right],\ \text{for } s \in \mathscr{E}_k \right\},$$

where c_k is chosen to satisfy (3·1). As in § 2·3, bands with different shapes can be obtained by allowing c_k to depend on s. As we will illustrate later, the same type of algorithm described in § 2·3 can be used for constructing repeated confidence bands with varying domains, provided that c_k is chosen to reflect the rule used for selecting $\mathscr{E}_k$ and that (3·1) is satisfied.

4. Tests derived from repeated confidence bands

As in other inference problems, there is a duality between confidence regions and tests of hypotheses in the current setting that can be used to construct formal hypothesis tests from a set of repeated confidence bands. To illustrate, let $\mu(.)$ denote the mean function for the difference in the response processes of two treatment groups, so that the hypothesis that the mean functions for the two groups are identical is given by

$$H_0 : \mu(s) = 0, \quad s \in [0, S].$$

Suppose that $B_1, \ldots, B_K$ is a set of level $1 - \alpha$ repeated confidence bands with domains $\mathscr{E}_1 = \ldots = \mathscr{E}_K = [0, S]$. Then, if we define a rejection region at the kth interim analysis by any outcome in which B_k does not wholly contain the zero function, it follows that, under H_0, the probability of rejecting H_0 for the first time at the kth analysis is π_k, and hence that the overall probability of rejecting H_0 at some analysis is at most α. Furthermore, the results of a test can be summarised by a significance level, provided an ordering of the sample space is specified in advance (Whitehead, 1997, Ch. 5).

The approach with varying domains introduced in § 3 can also be used in the test setting. To illustrate, consider the example in § 3 where $\mathscr{E}_k = [0, S]$ for $k < K$ and $\mathscr{E}_K = \{S\}$. Here the resulting test rejects H_0 if B_k does not wholly contain the zero function from 0 to S at the kth interim analysis or if the final confidence interval for $\mu(S)$ does not contain zero. Note that the test at the kth interim analysis considers evidence for alternatives to $\mu(s)$ for all $s \in [0, S]$, whereas at the final analysis the resulting test is sensitive only to alternatives to H_0 in which $\mu(S) \neq 0$. Thus, while this repeated testing procedure has the desired level, it may be insensitive to many types of departure from H_0.

The $\mathscr{E}_k$ can also be chosen adaptively based on a rule or procedure when constructing a hypothesis test. At first glance this may seem inappropriate because $\mathscr{E}_k$ is selected in the light of the data available at calendar time T_{k-1}. However, the choice of the boundary values c_k in the construction of B_k depends on the rule of selecting $\mathscr{E}_k$, and the argument in § 3 shows that the resulting testing procedure will have level α.

5. Example

To illustrate the approach discussed in this paper, we reanalyse the results of a recent clinical trial comparing the drugs AZT and ddI for people infected with HIV. For purposes of illustration, we only consider the AZT and low dose ddI arms used in the trial. For a detailed discussion of the study, see Kahn et al. (1992). The primary efficacy endpoint was AIDS-free survival; that is, time until the development of AIDS or death, whichever occurred first. A total of 602 patients were randomised to the AZT and low-dose ddI arms between October 1989 and April 1991, and follow-up ended in November 1991. By the end of the study, 125 of the 304 subjects assigned to AZT and 94 of the 298 assigned to ddI were observed to develop AIDS or die. With Y_i denoting the AIDS-free survival time of subject i, we let $X_i(s) = I(Y_i \geqslant s)$, so that $\mu(.)$ is the survival function of Y_i, which we denote by $\mu_A(.)$ for the AZT group and by $\mu_d(.)$ for the ddI group.

We reconstruct the results of this trial based on conducting two interim analyses and a final analysis at $T_1 = 50$, $T_2 = 70$ and $T_3 = 110$ weeks after the trial began. These times correspond approximately to the times of the actual interim and final analyses conducted during the study. We take $\alpha = 0{\cdot}05$ and a conservative O'Brien–Fleming type of spending function with $\pi_1 = 0{\cdot}001$, $\pi_2 = 0{\cdot}016$ and $\pi_3 = 0{\cdot}034$.

Repeated confidence bands for the survival function of each treatment group, together with the estimated values of $\mu_A(.)$ and $\mu_d(.)$ obtained from the Kaplan–Meier estimator (Kaplan & Meier, 1958), are shown in Fig. 1 at the three analyses, where we have taken $\mathscr{E}_1 = [0, 50]$, $\mathscr{E}_2 = [0, 70]$ and $\mathscr{E}_3 = [0, 110]$. The bands were obtained by evaluating (2·4) using (2·7), with c_k determined by the algorithm described in § 2·3 with $M_1, M_2, M_3 = 4000$. By their construction, the three bands for a treatment group simultaneously include the true mean function with probability at least 95%, and thus each individual band does also.

Next consider the difference between the survival functions of the ddI and AZT groups, which we denote by $\mu_\Delta(.) = \mu_d(.) - \mu_A(.)$. Figure 2 gives the estimates of $\mu_\Delta(.)$ and corresponding repeated confidence bands based on domains $\mathscr{E}_1 = [0, 40]$, $\mathscr{E}_2 = [0, 60]$ and $\mathscr{E}_3 = \{60\}$. We choose the first two domains because there was not much information collected beyond study times 40 and 60 at the two interim reviews. At the time of the first interim analysis, the point estimate of $\mu_\Delta(.)$ is generally positive, indicating a larger survival probability in the ddI group. However, the confidence band is quite wide, indicating that it is plausible that either group may be superior. At the time of the second interim analysis, the estimate of $\mu_\Delta(s)$ continues to increase with s, and the lower bound of the confidence band approaches 0. This suggests that it is unlikely that AIDS-free survival is superior in the AZT group. At the time of the final analysis, Fig. 2(c) shows that the confidence interval for the difference between the probabilities of an AIDS-free survival beyond 60 weeks is [0·033, 0·311] and thus only includes values for which ddI has superior AIDS-free survival probability. Also, each of the previous repeated confidence bands in this example will be level 95%. Thus, in addition to providing a confidence interval for $\mu_\Delta(60)$, the confidence band for $\mu_\Delta(s)$, for $0 \leqslant s \leqslant 60$, at the time of the second interim analysis is still a level 95% band.

One consequence of using a conservative spending function is that the confidence band at the final analysis will be similar to that of a nominal confidence band that does not account for the interim analyses. In this example, the nominal confidence interval for $\mu_\Delta(60)$ is [0·056, 0·288] and thus is only slightly narrower than that obtained from the repeated confidence bands approach.

The same procedures used to construct the confidence band for $\mu_\Delta(.)$ can also be used to test the hypothesis $H_0 : \mu_\Delta(s) = 0$, for $0 \leqslant s \leqslant 60$. Since the repeated confidence bands at the time of the first and second interim analyses wholly contain the zero function, H_0 would not have been rejected at these analyses. However, H_0 would be rejected at final analysis since the confidence interval excludes zero. Although the final analysis provided evidence favouring ddI over AZT for values of s prior to week 60, it is clear that the test might have had little power to detect early and transient differences between the treatment

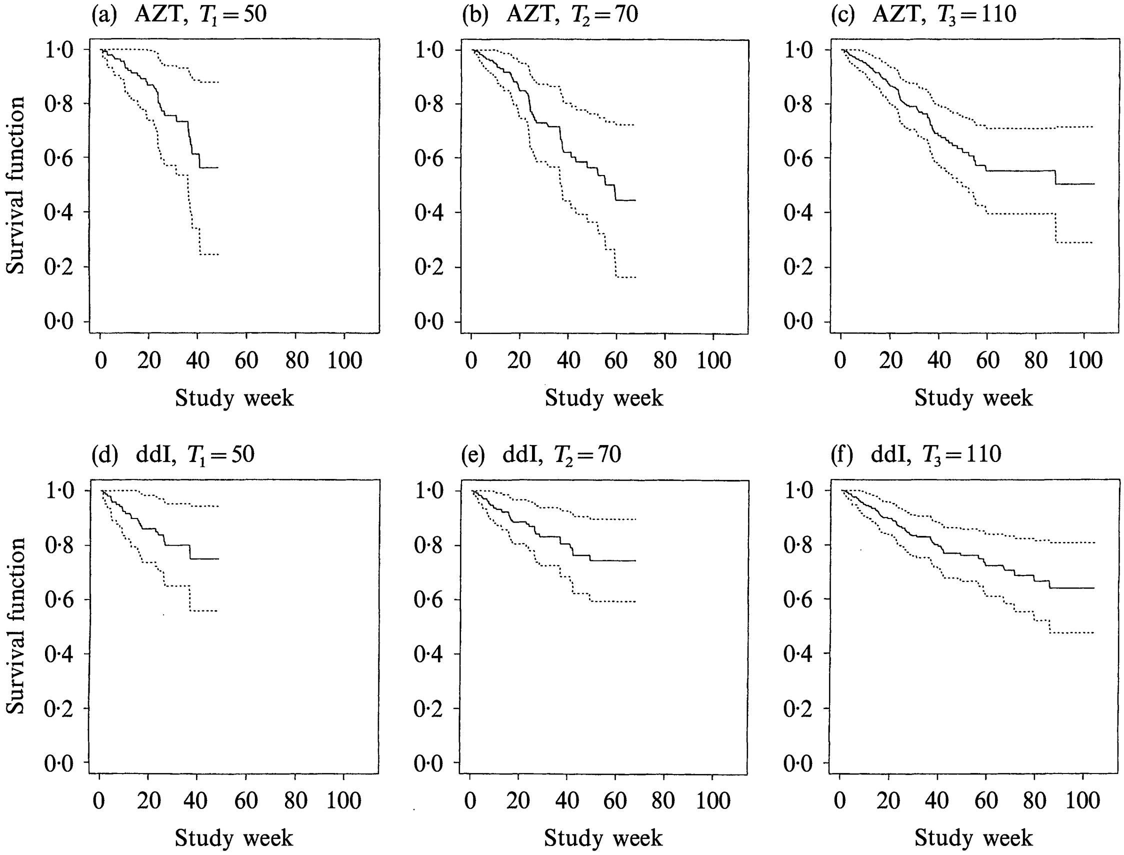

Fig. 1: Example. Repeated confidence bands at calendar times $T_1 = 50$, $T_2 = 70$ and $T_3 = 110$ for AZT group, (a)–(c), and for ddI group, (d)–(f). Solid line represents Kaplan–Meier estimate and dotted lines represent upper and lower limits of bands.

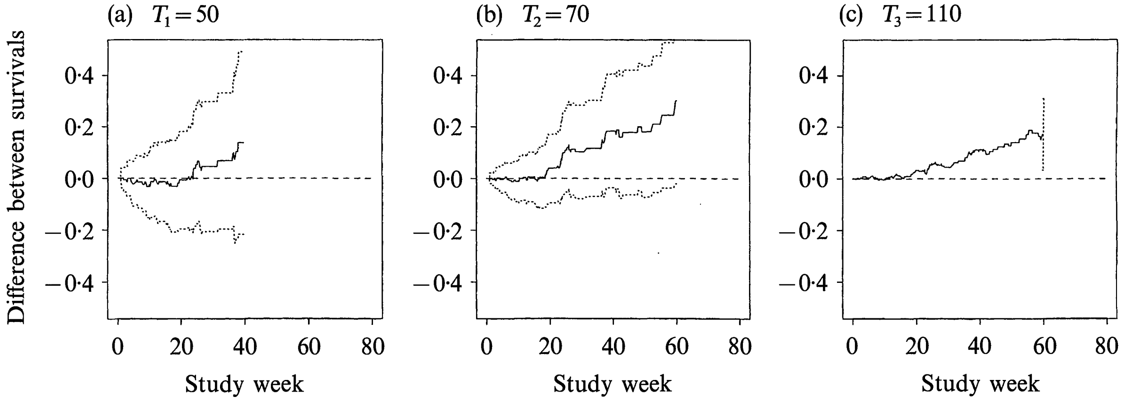

Fig. 2: Example. Repeated confidence bands for difference between ddI and AZT groups at calendar times $T_1 = 50$ in (a), $T_2 = 70$ in (b) and $T_3 = 110$ in (c), using domains $[0, 40]$, $[0, 60]$ and $\{60\}$, respectively. Solid line represents Kaplan–Meier estimate, dotted lines at T_1 and T_2 represent upper and lower limits of bands, and vertical dotted line at T_3 represents confidence interval for difference at $s = 60$. Horizontal dashed line represents the hypothesis of treatment group equality.

groups because of the domain used in this example. Indeed, for small trials or trials with very conservative spending functions, the widths of the confidence bands for early interim analyses may be so wide as to be of little value.

6. Discussion

A useful extension of the approach is to consider situations where the response is vector valued. This would be applicable, for example, where there are multiple types of response and a joint confidence region is desired, or where there are more than two treatment groups and some may be dropped during the course of a trial. The proposed methods can be extended in a natural way for vector responses; further details are available from the authors upon request.

In many applications, the error probabilities π_k are pre-determined. However, since the critical values c_k used to construct the bands B_k are defined sequentially with pre-determined domains, the π_k can be chosen adaptively based on $(\mathscr{H}_{k-1}, \mathscr{Q}_k)$ with no real modification to the Algorithm. A formal proof and related results are available upon request. Another way of adding flexibility to the proposed approach is to allow the interim analysis times to be determined adaptively, such as is done by Slud & Wei (1982) and Lan & DeMets (1983) for inferences about a scalar parameter. The same procedures proposed in this paper could be used when $T_k = T_k(\mathscr{H}_{k-1}, \mathscr{Q}_k)$, provided that $n^{\frac{1}{2}}\{\hat{\mu}_n(s) - \mu(s)\}$ could be shown to converge to a Gaussian process.

The methods introduced in this paper can also be applied to settings other than those discussed in § 2, and need not to be restricted to inferences about a mean function. For example, consider a comparison of two groups with respect to a failure time endpoint while adjusting for other covariates. Suppose a Cox regression model (Cox, 1972) is used with a regression coefficient, say $\beta(s)$, for the treatment group indicator that is time dependent to allow for nonproportional treatment hazard functions. Then, if one can develop an estimator that, when standardised, has a known asymptotic distribution, the ideas in § 2·2 can be extended to form repeated confidence bands for $\beta(.)$. This would allow the treatment groups to be monitored in a more flexible way than if a proportional treatment hazard ratio were assumed.

The simulation technique introduced at the end of § 2 provides a simple way of approximating the boundaries c_k necessary to compute the repeated confidence bands, and can easily be modified to incorporate varying domains. For example, the values of c_k used to construct the confidence bands for $\mu(.)$ in Fig. 2 were obtained by modifying Step (b) in the Algorithm so that the supremum in the definition of $L_n(T_l)^{(m)}$ is taken over $s \in \mathscr{E}_l$. A different modification of the Algorithm is needed when the domains vary and are selected in an adaptive way. For example, suppose the domain $\mathscr{E}_3 = \{60\}$ in Fig. 2(c) had been selected based on the rule that the domain for the final analysis would be the singleton time-point that maximised the standardised difference $|S_n(.;\, T_2)|$ based on information available at the second interim analysis. Then, in addition to modifying the domain for which the supremum of $L_n(T_l)^{(m)}$ is taken in Step (b) of the Algorithm, the value of c_3 would be selected so that the proportion π_3 of the M_3 realisations satisfy $L_n(T_l)^{(m)} \leqslant c_l$ for $l = 1, 2$ and also $L_n(T_2)^{(m)} = |\tilde{S}_n(60;\, T_2)^{(m)}|$. This leads to $c_3 = 2{\cdot}40$ as compared to $c_3 = 2{\cdot}35$ when $\mathscr{E}_3 = \{60\}$ was predetermined.

Finally, more research is needed for the design of trials in which inferences are based on different subsets of the domain of the mean function at each analysis, especially when these domains are selected in an adaptive way.

Acknowledgement

This work was supported by a grant from the National Institute of Allergy and Infectious Diseases. The authors thank R. Betensky, L. J. Wei, a referee, an associate editor and the editor for their helpful comments.

Appendix

Limiting distribution of $\tilde{S}_n^{(k)}(s)$ conditional on $\{X_i, \delta_i\}$

Note that, in the setting of Example 1, the conditional covariance function

$$\mathrm{cov}[\tilde{S}_n(s;\, T_k), \tilde{S}_n(u;\, T_k) \,|\, \{X_i, \delta_i\}] = \frac{n\sum_{i=1}^n \delta_i(s;\, T_k)\delta_i(u;\, T_k)\{X_i(s) - \hat{\mu}_n(s;\, T_k)\}\{X_i(u) - \hat{\mu}_n(u;\, T_k)\}}{\{\mathrm{var}(s;\, T_k)\,\mathrm{var}(u;\, T_k)\}^{\frac{1}{2}}\sum_{i=1}^n \delta_i(s;\, T_k)\sum_{i=1}^n \delta_i(u;\, T_k)}$$

is asymptotically equivalent to

$$\frac{n\sum_{i=1}^n \delta_i(s;\, T_k)\delta_i(u;\, T_k)\{X_i(s) - \mu(s;\, T_k)\}\{X_i(u) - \mu(u;\, T_k)\}}{\{\mathrm{var}(s;\, T_k)\,\mathrm{var}(u;\, T_k)\}^{\frac{1}{2}}\sum_{i=1}^n \delta_i(s;\, T_k)\sum_{i=1}^n \delta_i(u;\, T_k)},$$

and thus almost surely converges to

$$\frac{\mathrm{cov}\{X(s), X(u)\}E\{\delta(s;\, T_k)\delta(u;\, T_k)\}}{\{\mathrm{var}(s;\, T_k)\,\mathrm{var}(u;\, T_k)\}^{\frac{1}{2}}G(s;\, T_k)G(u;\, T_k)}, \tag{A·1}$$

the asymptotic covariance function of $S_n(s;\, T_k)$. Furthermore, conditional on $\{X_i, \delta_i\}$, $\tilde{S}_n(s;\, T_k)$ is a summation of independent zero-mean normal random variables. It then follows from the Functional Central Limit Theorem (Pollard, 1990, Ch. 10) that $\tilde{S}_n(s;\, T_k)$ converges to a zero-mean Gaussian process with the covariance function (A·1). That is, conditional on $\{X_i, \delta_i(.;\, T_k)\}$, $\tilde{S}_n(s;\, T_k)$ has the same limiting Gaussian process as $S_n(s;\, T_k)$ for each k. With a minor modification of this argument, we can show that $\sum_{l=1}^k \alpha_l \tilde{S}_n(s;\, T_l)$ has the same limiting Gaussian process as $\sum_{l=1}^k \alpha_l S_n(s;\, T_l)$. Therefore $\tilde{S}_n^{(k)}$ and $S_n^{(k)}$ have the same limiting Gaussian process in the setting of Example 1.

Similar arguments can be applied to prove that, conditional on $\{X_i, \delta_i\}$, $\tilde{S}_n(s;\, T_k)$ defined by (2·10) in the setting of Example 2, or (2·11) in the setting of Example 3, or (2·12) for a more general setting has the same limiting Gaussian process as the corresponding $S_n(s;\, T_k)$.

References

Cook, R. J. & Lawless, J. F. (1996). Interim monitoring of longitudinal comparative studies with recurrent event responses. *Biometrics* **52**, 1311–23.

Cox, D. R. (1972). Regression models and life-tables (with Discussion). *J. R. Statist. Soc.* B **34**, 187–220.

Fleming, T. R. & Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* New York: John Wiley.

Jennison, C. & Turnbull, B. W. (1989). Interim analyses: the repeated confidence interval approach (with Discussion). *J. R. Statist. Soc.* B **51**, 305–61.

Jennison, C. & Turnbull, B. W. (1990). Statistical approaches to interim monitoring of medical trials: a review and commentary. *Statist. Sci.* **5**, 299–317.

Kahn, J. O., Lagakos, S. W., Richman D. D., Cross, A., Pettinelli, C., Liou, S., Brown, M., Volberding, P. A., Crumpacker, C. S., Beall, G., Sacks, H. S., Merigan, T. C., Beltangady, M., Smaldone, L. & Dolin, R. (1992). A controlled trial comparing continued zidovudine with didanosine in human immunodeficiency virus infection. *New Engl. J. Med.* **327**, 581–7.

Kaplan, E. L. & Meier, P. (1958). Non-parametric estimation from incomplete observations. *J. Am. Statist. Assoc.* **53**, 457–81.

Lan, K. K. G. & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659–63.

Lin, D. Y., Wei, L. J. & Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–72.

O'Brien, P. C. & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549–56.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191–9.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*, Regional Conference Series in Probability and Statistics 2. Hayward, CA: Institute of Mathematical Statistics.

Slud, E. V. & Wei, L. J. (1982). Two sample repeated significance tests based on the modified Wilcoxon statistic. *J. Am. Statist. Assoc.* **77**, 862–8.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*, 2nd ed. Chichester: John Wiley.

[*Received March* 1998. *Revised January* 1999]