CLINICAL TRIALS FOR DRUG DEVELOPMENT: SOME STATISTICAL PROBLEMS

by

Tze Leung Lai

Technical Report No. 2004-34
December 2004

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065
CLINICAL TRIALS FOR DRUG DEVELOPMENT: SOME STATISTICAL PROBLEMS

by

Tze Leung Lai
Department of Statistics
Stanford University

Technical Report No. 2004-34
December 2004

This research was supported in part by the National Institute of Health 5R01-CA 088890

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://www-stat.stanford.edu
Clinical Trials for Drug Development: Some Statistical Problems
By Tze Leung Lai

About two years ago, the editor of the ICSA Bulletin invited me to contribute an article giving an overview of my research. I thanked him for the invitation but told him that I did not know how to proceed since during the past thirty years I had been working in various areas of statistics and probability and different fields of application that might be difficult to provide a unified and not too lengthy overview. On the other hand, I asked him if he would agree to the possibility of narrowing the scope of the article to something related to my plenary talk (which was still vague in my mind at that time) in the 2003 ICSA Applied Statistics Symposium (later postponed to 2004 because of SARS), entitled “Current Statistical Issues in Clinical Trials for Drug Development.” His answer was enthusiastically positive, and after my talk in June, he reminded me of what I was supposed to send him. The “Statistical Problems” in the title of this article, therefore, are closely related to several “issues” I discussed in that talk. They also reflect some of my research interests during the past decade related to clinical trials for drug development. I have been attracted to them not only because of their practical relevance but also because they are challenging and fundamental statistical problems that have far-reaching implications beyond biopharmaceutics and clinical trials.

1. Efficient group sequential designs and interim analysis of clinical trials

In standard clinical trial designs, the sample size is determined by the power at a given alternative. In practice, however, it is often difficult for investigators to specify a realistic alternative at which sample size determination can be based, especially for new treatments about which there is little information concerning the magnitude of the treatment effect before actual data are collected. Moreover, the choice of the alternative to determine the sample size is usually guided not only by published results on the magnitude of the treatment difference to be expected but also by economic considerations related to funding and duration for the trial and by administrative considerations related to other trials that compete for patients and investigators. Although the protocol of a trial typically justifies its choice of sample size by stating some conventional level, such as 80% or 90%, of power at a plausible alternative, there are actually many other feasibility considerations that are difficult to quantify and much harder to explain than the precise but oversimplified statement of some prescribed power at the alternative.

Group sequential designs that can adapt to information about unknown parameters during interim analyses provide natural ways to address the above difficulty in clinical trial
designs due to lack of information at the design stage. A few years ago, together with Mei-Chiung Shih who had just finished her Ph.D. dissertation at Stanford under my supervision, I undertook the project of developing flexible and efficient group sequential designs that can ‘self-tune’ to the unknown parameters during the course of the trial, under pre-specified constraints on the maximum sample size and Type I error probability. Previous work on efficient group sequential designs uses the expected sample size at an alternative, or more generally a weighted average of expected sample sizes over a set of parameter values, as the optimization criterion while controlling the error probabilities under the null hypothesis and a specified alternative at prescribed levels; see Pocock (1982), Wang & Tsiatis (1987), Kim & DeMets (1987), Eales & Jennison (1992, 1995) and Barber & Jennison (2002). There are several practical difficulties with this approach to efficient group sequential design. First, even though the mean of the random sample size is minimized at some alternative, the maximum sample size can be substantially larger than the mean and also the fixed sample size. Secondly, the optimization problem requires precise specification of the relative sizes of all groups, e.g. equal group sizes, but it is often not feasible to do so prior to the trial because interim analyses are usually scheduled at calendar times for administrative reasons; see Chapter 7 of Jennison & Turnbull (2000). Thirdly, as pointed out in the preceding paragraph, it may be difficult to come up with a realistic alternative before data are collected from the trial, but the optimization problem depends on the chosen alternative.

Clearly efficiency of a group sequential test depends not only on the design of the stopping rule but also on the test statistics used. To fix the ideas, we began by focusing on the generic special case of a one-parameter exponential family $f_\theta(x) = e^{\theta x - \psi(\theta)}$ of densities with respect to some measure on the real line, and consider the problem of testing the one-sided hypothesis $H_0 : \theta \leq \theta_0$ at significance level $\alpha$ and taking no more than $M$ observations $X_1, X_2, \ldots$. Sufficient statistics are the sample means which are maximum likelihood estimators of $\psi'(\theta)$, and the Kullback-Leibler information number is given by

$$I(\theta, \lambda) = E_\theta[\log{f_\theta(X_i)}/f_\lambda(X_i)] = (\theta - \lambda)\psi'(\theta) - \{\psi(\theta) - \psi(\lambda)\}. \quad (1)$$

Letting $S_n = X_1 + \ldots + X_n$, the fixed sample size test that rejects $H_0$ if $S_M \geq c_\alpha$ has maximal power at any alternative $\theta > \theta_0$, in particular at the alternative $\theta(M)$ ‘implied’ by $M$ (in the sense that $M$ can be derived from the assumption that the above fixed sample test has some prescribed power $1 - \alpha$ at $\theta(M)$) when one does not have much information on which to base a realistic alternative. Under the constraint of $M$ on the maximum sample size, it is desirable to adapt to the information on the actual $\theta$ gathered during the course of the
trial, allowing early stopping at times of interim analysis, so that the test has nearly optimal expected sample size under a wide range of alternatives but with small loss in power from the fixed sample size test.

To achieve these goals in a group sequential test with \( k \) groups and group sizes \( n_1, n_2 - n_1, \ldots, n_k - n_{k-1} \) so that \( n_k = M \), our approach was to make use of the theory of optimal sequential tests, which was relatively complete in the fully sequential framework (Lai, 2001), and to modify it for the group sequential setting. This theory leads to a group sequential test with rejection region of the form \( S_{nk} \geq c \) at the \( k \)th analysis, where \( c > c_\alpha \) but \( c \) does not differ much from \( c_\alpha \). Let \( \bar{X}_n = S_n / n \). During the first \( k - 1 \) analyses, the test uses the maximum likelihood estimator \( \hat{\theta}_{ni} = (\psi')^{-1}(\bar{X}_{ni}) \) to estimate \( \theta \) and a stopping rule of the form

\[
\hat{\theta}_{ni} > \theta_0 \text{ and } n_i I(\hat{\theta}_{ni}, \theta_0) \geq b, \quad \text{or} \quad \hat{\theta}_{ni} < \theta(M) \text{ and } n_i I(\hat{\theta}_{ni}, \theta(M)) \geq \tilde{b},
\]

for \( 1 \leq i \leq k - 1 \). If (2a) holds, the test rejects \( H_0 \) upon stopping. If stopping occurs with (2b), it accepts \( H_0 \). The thresholds \( b, \tilde{b} \) and \( c \) are so chosen that \( P_{\theta_0}(\text{Test rejects } H_0) = \alpha \) and that the power of the test at \( \theta(M) \) does not differ much from its upper bound \( 1 - \alpha \). A simple way of choosing \( b, \tilde{b} \) and \( c \) satisfying these properties is as follows. Let \( 0 < \epsilon < \frac{\alpha}{2} \) and define \( \tilde{b} \) by the equation

\[
P_{\theta(M)}(\hat{\theta}_{ni} < \theta(M) \text{ and } n_i I(\hat{\theta}_{ni}, \theta(M)) \geq \tilde{b} \text{ for some } 1 \leq i \leq k - 1) = \epsilon \alpha.
\]

After determining \( \tilde{b} \), define \( b \) and then \( c \) by the equations

\[
\sum_{j=1}^{k-1} P_{\theta_0}(\hat{\theta}_{nj} > \theta_0 \text{ and } n_j I(\hat{\theta}_{nj}, \theta_0) \geq b, n_i I(\hat{\theta}_{ni}, \theta_0) 1_{\{\hat{\theta}_{ni} > \theta_0\}} < b \text{ and } n_i I(\hat{\theta}_{ni}, \theta(M)) 1_{\{\hat{\theta}_{ni} < \theta(M)\}} < \tilde{b} \text{ for } i < j) = \epsilon \alpha,
\]

\[
P_{\theta_0}(S_{nk} \geq c, n_i I(\hat{\theta}_{ni}, \theta_0) 1_{\{\hat{\theta}_{ni} > \theta_0\}} < b \text{ and } n_i I(\hat{\theta}_{ni}, \theta(M)) 1_{\{\hat{\theta}_{ni} < \theta(M)\}} < \tilde{b} \text{ for } i < k)
\]

\[
= (1 - \epsilon) \alpha.
\]

Note that although the rejection region in favor of the treatment at each of the \( k \) times is one-sided, involving the thresholds \( b \) and \( c \), there is also a "futility" boundary in (2b) with threshold \( \tilde{b} \) that stops the trial when it becomes unlikely to demonstrate efficacy of the treatment within the resources allocated to the trial. We call these tests the modified Haybittle-Peto tests as they use a more flexible choice of \( b \) than \( |S_{ni}|/\sqrt{n_i} \geq 3\sigma \) for \( 1 \leq i \leq k - 1 \) in the group sequential test proposed by Haybittle (1971) and Peto et al. for normally distributed \( X_i \).
Noting that the \( n_j I(\hat{\theta}_{n_j}, \theta_*) \) used in the stopping rule (2) are generalized likelihood rates (GLR) statistics for testing \( \theta = \theta_* \), we then extended these group sequential tests to the multiparameter exponential family and to multi-armed clinical trials by using appropriate GLR statistics in these problems. The basic idea behind the proposed class of tests, therefore, is to estimate the unknown parameter \( \theta \) by maximum likelihood during the course of the trial and to use it to replace \( \theta \) in an approximately optimal sequential test that assumes known \( \theta \). These group sequential tests, which are shown in our recent paper (Lai & Shih, 2004) to have nearly optimal power and expected sample size properties over a wide range of alternatives, do not require pre-specification of the group sizes, nor do they require estimation of the “maximum information” of the trial in the error-spending approach (see Chapter 7 of Jennison & Turnbull (2000)). They are, therefore, very flexible and can be easily extended to more complex settings. An important extension is related to the design and interim analysis of clinical trials for comparing the failure times between two treatment groups; see Gu & Lai (1991, 1998). A general weak convergence theory for certain time-sequential censored rank statistics under the null hypothesis of no treatment difference and under local alternatives has been developed, showing that these time-sequential statistics behave asymptotically like cumulative sums of independent normal random variables, with the number of summands up to time \( t_j \) known only at the calendar time \( t_j \) of the \( j \)th interim analysis. This is a consequence of the independent increments property of the limiting Gaussian process, whose increments have variances that are not specified in advance but have to be estimated from the data. Despite such complexity, the modified Haybittle-Peto tests can be easily modified to develop efficient time-sequential tests subject to prescribed constraints on the type I error probability and maximum study duration, in contrast with the error-spending approach that uses simulations at each interim analysis to estimate the maximum information under the null hypothesis; see Scharfstein & Tsiatis (1998).

Although group sequential designs are attractive because they allow for early termination while preserving the overall significance level of the test and can adapt to information gathered during the course of the trial, they may also introduce substantial bias when one applies standard point and interval estimates for the parameters of interest following the test, and this potential bias has inhibited the use of group sequential methodology. After reviewing previous work in this area, I began a systematic investigation of the problem of constructing valid confidence intervals following group sequential tests in the mid-nineties. Siegmund’s (1978) seminal paper introduced an exact method, based on ordering the sample space in a certain way, to construct exact confidence intervals for the mean of a normal
population with known variance following a repeated significance test. Tsiatis, Rosner & Mehta (1984) extended Siegmund’s method to the group sequential tests of Pocock (1977) and O’Brien & Fleming (1979). Alternative orderings of the sample space were subsequently introduced by Chang & O’Brien (1986), Rosner & Tsiatis (1988), Chang (1989) and Emerson & Fleming (1990). For samples of fixed size, an important methodology for constructing confidence intervals without distributional assumptions is Efron’s (1987) bootstrap method. This prompted me to try resampling methods to adjust for the bias due to the possibility of early stopping. In Chuang & Lai (1998), we studied bootstrap confidence intervals for a population mean in a group sequential setting as an alternative to the exact methods. We found that, since the stopping rule makes the approximate pivots in nonsequential bootstrap methods highly “non-pivotal”, the bootstrap method does not yield reliable confidence intervals in a group sequential setting. However, by integrating the main ideas behind the exact and bootstrap methods, we were able to develop a resampling method for the construction, after a group sequential test, of confidence intervals whose coverage probabilities are nearly equal to the nominal ones.

This hybrid resampling approach was subsequently developed further in Chuang & Lai (2000), where we showed that it also works well in other situations where the bootstrap method fails. For group sequential tests, the hybrid resampling approach in Chuang & Lai (1998, 2000) assumes the group sizes to be pre-determined constants. To extend the approach to random group sizes, Wenshi Li and I recently found that the ordering method introduced by Siegmund (1978) still works in the case of sample means. To begin with, suppose $Z_1, Z_2, \ldots$ are i.i.d. normal random variables with known variance 1 and unknown mean $\mu$ and $T$ is a two-sided stopping rule of the form $T = \inf\{n \in J : S_n \geq b_n \text{ or } S_n \leq a_n\}$, where $S_n = Z_1 + \ldots + Z_n$ and $J$ is a finite set of positive integers. Siegmund (1978) orders the sample space of $(T, S_T)$ as follows: $(t, s) > (t', s')$ whenever (i) $t = t'$ and $s > s'$, or (ii) $t < t'$ and $s \geq b_t$, or (iii) $t > t'$ and $s' \leq a_{t'}$. Let $\mu_c$ denote the value of $\mu$ for which $P_{\mu_c}(T, S_T) \geq (t, s)_{\text{obs}} = c$, where $(t, s)_{\text{obs}}$ denotes the observed value of $(T, S_T)$. Siegmund’s confidence interval is $\mu_{c} \leq \mu \leq \mu_{1-c}$, which has coverage probability $1 - 2\alpha$. Note that this ordering only involves considering possible sample paths that stop (or do not stop if downcrossing of the lower boundary is observed) prior to the observed stopping time. Hence a hybrid resampling version that removes the assumption of normality in Siegmund’s method does not require one to generate data beyond the observed stopping time $t$; see Chuang & Lai (1998) for details and the second-order accuracy of the method. Although the set $J$ considered by Siegmund (1978) is nonrandom, the argument of Chuang & Lai (1998) is still
applicable to the case of random \( J = \{n_1, \ldots, n_k \} \) by conditioning on \((n_1, \ldots, n_k)\), thereby establishing the second-order accuracy of the hybrid resampling method when the random vector \((n_1, \ldots, n_k)\) is independent of \(\{Z_i, i \geq 1\}\). The main motivation behind my work with Li was the construction of confidence intervals for survival parameters in time-sequential clinical trials with survival endpoints, which will be described below.

Suppose that a clinical trial to compare times to failure between two treatment groups \(X\) and \(Y\) involves \(n\) patients who enter the trial serially, are randomly assigned to treatment \(X\) or \(Y\) and are then followed until they fail or withdraw from the study or until the study is terminated. Let \(T'_i \geq 0\) denote the entry time and \(X_i > 0\) the survival time (or time to failure) after entry of the \(i\)th subject in treatment group \(X\), and let \(T''_j\) and \(Y_j\) denote the entry time and survival time after entry of the \(j\)th subject in treatment group \(Y\). Thus the data at calendar time \(t\) consist of \((X_i(t), \delta'_i(t))), i = 1, \ldots, n',\) and \((Y_j(t), \delta''_j(t))), j = 1, \ldots, n''\), where \(X_i(t) = X_i \wedge \xi'_i \wedge (t - T'_i)^+, Y_j(t) = Y_j \wedge \xi''_j \wedge (t - T''_j)^+, \delta'_i(t) = I_{\{X_i(t) = X_i\}}, \delta''_j(t) = I_{\{Y_j(t) = Y_j\}}, m'_{n,t}(s) = \sum_{i=1}^{n'} I_{\{X_i(t) \geq s\}}, m''_{n,t}(s) = \sum_{j=1}^{n''} I_{\{Y_j(t) \geq s\}},\) and \(\xi'_i(\xi''_j)\) denotes the withdrawal time, possibly infinite, of the \(i\)th \((j)\)th subject in treatment group \(X\) \((Y)\). At a given calendar time \(t\), one can compute, on the basis of the observed data from the two treatment groups, a rank statistic of the general form considered by Tsiatis (1982):

\[
S_n(t) = \sum_{i=1}^{n'} \delta'_i(t)Q_n(t, X_i(t)) \left\{ 1 - \frac{m'_{n,t}(X_i(t))}{m'_{n,t}(X_i(t)) + m''_{n,t}(X_i(t))} \right\} - \sum_{j=1}^{n''} \delta''_j(t)Q_n(t, Y_j(t)) \frac{m''_{n,t}(Y_j(t))}{m'_{n,t}(Y_j(t)) + m''_{n,t}(Y_j(t))},
\]

(3)

where \(Q_n(t, s)\) is some weight function satisfying certain measurability assumptions. The case \(Q_n \equiv 1\) corresponds to the logrank statistic. Let \(H_{n,t}\) denote a product-limit-type estimator of the common distribution function of the two treatment groups under the null hypothesis, based on \(\{(X_i(t), \delta_i(t), Y_j(t), \delta_j(t)) : i \leq n', j \leq n''\}\). For a general weight function of the form \(Q_n(t, s) = \psi(H_{n,t}(s))\) in (3), Minggao Gu and I have shown that \(\{S_n(t)/\sqrt{n}, t \geq 0\}\) converges weakly to a Gaussian process with independent increments and variance function \(V(t)\) under the null hypothesis, and contiguous alternatives. Note that \(V(t)\) is called the “information time” by Lan & DeMets (1989). The mean function of the limiting Gaussian process is 0 under the null hypothesis and is of the form \(\mu_0(t)\) under contiguous alternatives that satisfy

\[
\int_0^t \left| \frac{d\Delta_G}{d\Delta_F} - 1 \right| d\Lambda_F = O\left( \frac{1}{\sqrt{n}} \right), \quad \sqrt{n}\left\{ \frac{d\Delta_G}{d\Delta_F}(s) - 1 \right\} \to g(s)
\]
as \( n \to \infty \), uniformly over closed subintervals of \( \{ s \in [0, t^*] : F(s) < 1 \} \), where \( \Lambda_F \) and \( \Lambda_G \) are the cumulative hazard functions of \( F \) and \( G \); see Gu & Lai (1991), in which we have also provided consistent estimates \( V_n(t) \) of \( V(t) \) and shown that \( \mu_g(t) = V(t) \) in the case of asymptotically optimal score functions \( \psi(\cdot) = g(F^{-1}(\cdot)) \). In practice, the actual alternatives are unknown and \( \mu_g \) need not even be monotone when \( \psi \) is not optimal for the actual alternatives, such as using logrank statistics for non-proportional hazards alternatives, for which we have shown in Gu & Lai (1998) that time-sequential tests based on \( S_n(t) \) can achieve both savings in study duration and increase in power over the fixed-duration test based on \( S_n(t^*) \). It is widely recognized that tests of treatment effects based on the rank statistics (3) may lose substantial power when the effects of other covariates are strong. In nonsequential trials, a commonly used method to remedy this when logrank statistics are used is to assume the proportional hazards regression model and to use Cox's partial likelihood approach to adjust for other covariates. Tsiatis, Rosner & Tritchler (1985), Gu & Ying (1995) and Bilias, Gu & Ying (1997) have developed group sequential tests using this approach. Instead of relying on the proportional hazards model to adjust for concomitant variables, it is useful to have other methods for covariate adjustment, especially in situations where other score functions than the logrank are used in (3) to allow for the possibility of non-proportional hazards alternatives. Lin (1992) and Gu & Lai (1998) have proposed alternative covariate adjustment methods based on rank estimators and \( M \)-estimators in accelerated failure time models.

The recent work of Lai & Shih (2004) showing the efficiency of modified Haybittle-Peto tests relative to the error spending approach suggests that one can focus on the modified Haybittle-Peto-type boundary in constructing a time-sequential test based on (3), with \( t \) restricted to the set \( \{ t_1, \ldots, t_k (= t^*) \} \) of calendar times at which interim analyses are conducted, since it yields a statistically efficient stopping rule that can also circumvent the difficulty of “calendar time” versus “information time”. The Monte Carlo simulation method provides a flexible and practical way to compute power and expected duration of these complex trials, and also to check the adequacy of the normal approximation to the type I error probability under various scenarios of baseline survival, censoring pattern, noncompliance, and accrual rate. To provide the clinical trial designer with a tool to perform these Monte Carlo simulations, Gu & Lai (1999) developed a simulation program which gives the user some options for choosing the stopping boundary, including the modified Haybittle-Peto-type boundary. The program also allows the user to choose the score function \( \psi \) in \( Q_n(t, s) = \psi(H_{n,t}(s)) \) from a general family proposed by Self (1991). It enables the clinical
trialist, who can download it from a website listed in the paper, to select the test statistic most sensitive to the anticipated kind of departures from the null hypothesis. Gu & Lai (1999) also incorporated this power calculation program into another program that computes the sample size of a group sequential trial having a prescribed power at given baseline and alternative distributions.

The above software only has the design module, but its basic programs can be modified for the development of an analysis module that can be used for interim analysis of clinical trials data. I am currently working with my Ph.D. student Zheng Su from the Department of Computer Science at Stanford to develop an analysis module that will also include software for terminal analysis of the trial. Concerning such terminal analysis, some progress towards constructing valid confidence intervals following time-sequential tests was recently made in Lai & Li (2004). Consider the logrank statistic (3) with $Q_n \equiv 1$ and a group sequential test of $H_0 : F = G$ with stopping rule of the form $\tau = \min \{t_j : |S_n(t_j)| \geq b_j V_{n}^{1/2}(t_j)\}$, where $t_1 < \ldots < t_k$ denote the calendar times of interim analysis and $V_n(t)$ is either Mantel’s (1966) estimate of the null variance of $S_n(t)$ or (total number of deaths up to time $t$)/4. Assuming the proportional hazards model $\Lambda_F = \theta \Lambda_G$, the null hypothesis can be rephrased as $H_0 : \theta = 1$, where $\theta$ is the hazard ratio. Let $P_\theta$ denote the probability measure under which $\theta = e^\beta$. Thus, $\beta$ is the regression parameter in Cox’s hazard regression model with covariate that takes the value 1 (under $F$, representing treatment) or 0 (under $G$, representing control), for which $S_n(t)$ is the efficient score statistic. For small $\beta$ (such that $\sqrt{n} \beta \to \mu$), $\{(S_n(t_j)/\sqrt{n}, V_n(t_j)/n) : 1 \leq j \leq K\}$ converges in distribution to $\{(W(V(t_j)), V(t_j)) : 1 \leq j \leq K\}$, where $W(\cdot)$ is a Wiener process with drift coefficient $\mu$. Analogy with the case of normal means and random group sizes described above in connection with hybrid resampling led us to the following ordering of the sample space, which reduces to Siegmund’s (1978) ordering in the case of confidence intervals for means. Let $\Psi_t = S_n(t)/V_n(t)$. Order the sample space of $(\tau, \Psi_\tau)$ by:

$$
(\tau_1, \Psi_{\tau_1}) \leq (\tau_2, \Psi_{\tau_2}) \quad \text{if and only if} \quad \Psi_{\tau_1} \leq \Psi_{\tau_2}.
$$

(4)

Similar to the normal mean case, let $p(\beta) = P_\beta\{(\tau, \Psi_\tau) > (\tau, \Psi_{\tau_{\text{obs}}})\}$. Then $\{\beta : \alpha < p(\beta) < 1 - \alpha\}$ is a confidence set for $\beta$ with coverage probability $1 - 2\alpha$. Even if the baseline distribution $G$ should the known, the probability $p(\beta)$ has to be evaluated by simulation. In practice $G$ is unknown and we can replace it by Breslow’s estimate $\hat{G}$ from all the data at the end of the trial. This suggests replacing $p(\beta)$ by

$$
\hat{p}(\beta) = P\{(\tau^{(\beta)}, \Psi_{\tau^{(\beta)}}(\beta)) > (\tau, \Psi_{\text{obs}})\},
$$

(5)
where the superscript \((\beta)\) means that the observations are generated by hybrid resampling from the baseline distribution \(\hat{G}\), with \(\beta\) as the hazard ratio. Since \(\hat{p}(\beta)\) is an increasing function of \(\beta\), the confidence set \(\{\beta : \alpha < \hat{p}(\beta) < 1 - \alpha\}\) with coverage probability \(1 - 2\alpha + O(n^{-1})\) becomes an interval whose endpoints \(\hat{\beta} < \tilde{\beta}\) are defined by \(\hat{p}(\beta) = \alpha, \hat{p}(\tilde{\beta}) = 1 - \alpha\). Details are given in Lai & Li (2004), where simulation studies and analytic results show that the confidence intervals thus constructed have coverage probabilities close to nominal values. Zheng Su and I are currently working on extensions of this approach to other (non-proportional hazards) survival models and other test statistics. We are also developing more efficient simulation procedures based on importance sampling and resampling to speed up the computation of (5) by Monte Carlo.

2. Population PK/PD and nonlinear mixed effects models

Pharmacology is the science dealing with interactions between living systems and molecules, especially chemicals introduced from outside the system. The component of a cell or organism that interacts with a drug and initiates a chain of biochemical events leading to the drug’s therapeutic and toxic effects is called a receptor. The receptor concept has become the central focus of investigation of pharmacodynamics (PD) – the study of drug effects and their mechanisms of action. How a drug dose produces its effects involves not only PD but also pharmacokinetics (PK), which is concerned with the concentration-time curve associated with the absorption, distribution and elimination phases of a single administration of the drug. In Lai, Shih & Zhu (2003), we have given an overview of the basic principles, models and statistical methods in PK/PD and of their roles in drug development. In many PK/PD studies, data are collected from a number of subjects, some of whom may have intensive blood sampling while others only have sparse data. A primary objective of these studies is to study the PK/PD characteristics of the entire population, such as how they vary with certain covariates. This requires a population model, and nonlinear mixed effects modeling provides a valuable tool to address this problem.

Since the seminal work of Sheiner & Beal (1980), nonlinear mixed effects models of the form

\[
y_{ij} = f_i(t_{ij}, \theta_i) + \epsilon_{ij}, \quad \theta_i = g(x_i, \beta) + b_i \quad (1 \leq j \leq n_i, \ 1 \leq i \leq I),
\]

have become widely used in population PK/PD. In (6), \(\theta_i\) is a \(1 \times \tau\) vector of the \(i\)th subject’s parameters whose regression function on the subject’s observed covariate \(x_i\) is \(g(x_i, \beta)\) with a \(1 \times s\) parameter vector \(\beta\), which is the “fixed effect” to be estimated. The “random effects” \(b_i\) in (6) are assumed to be i.i.d. and their nonzero components have a common distribution \(G\) with mean 0. The \(i\)th subject’s response \(y_{ij}\) at \(t_{ij}\) has mean \(f_i(t_{ij}, \theta_i)\), in which \(f_i\) is a known
function. Given $\theta_i$, the random errors $\varepsilon_{ij}$ are assumed to be independent normal random variables with mean 0 and standard deviation $\sigma \omega_{ij}(\theta_i)$, in which $\omega_{ij}$ is a given function and $\sigma$ is an unknown parameter. In PK, the concentrations at times $t_{ij}$ after the administration of a single dose $D_i$ are often modeled by the one-compartment model

$$y_{ij} = \frac{D_i k_i \tilde{k}_i}{Cl_i(k_i - \tilde{k}_i)} (e^{-k_i t_{ij}} - e^{-k_i t_{ij}}) + \varepsilon_{ij}, \quad 1 \leq j \leq n_i. \quad (7)$$

Here $Cl_i, k_i, \tilde{k}_i$ are the $i$th subject’s total body clearance, absorption rate, and elimination rate, respectively, and their logarithms constitute the vector $\theta_i$ in (6). The regression function $g$ relates $\theta_i$ to the $i$th subject’s physiologic characteristics that constitute the covariate vector $x_i$. The population distribution $G$ is usually assumed to be normal with unknown parameters which, together with $\beta$ and $\sigma$, can be estimated by maximum likelihood. Unlike linear mixed effects models in which the normality assumption on $G$ yields closed-form expressions of the likelihood, the normality of $G$ in nonlinear mixed effects models leads to computationally intensive likelihoods that involve $I$ multiple integrals. A commonly used approach, as adopted in the software package NONMEM (Beal & Sheiner, 1992), and the nlmme procedure in S-Plus due to Lindstrom & Bates (1990), is to develop iterative schemes based on first-order approximations of $f_i(t_{ij}, g(x_i, \beta) + b_i)$ in (6), so that the normality assumption on $G$ can be used to reduce the problem to that of a linear Gaussian mixed effects model at each iterative step.

A basic issue with this approximation is that when some of the subjects have sparse data there are considerable errors in approximating the likelihood function via these first-order approximations, as noted by Yafune et al. (1998) who propose to use Monte Carlo integration to evaluate the $I$ multiple integrals in the likelihood function for Phase 1 studies but point out that the computational time (already taking 22 hours in their particular Phase 1 trial) may be too long for Phase 2 (or later) trials to be of practical interest. Another issue is that the actual population distribution may be highly nonnormal. Since there is no computational advantage in using normal $G$ when first-order approximations to reduce to linear Gaussian mixed effects models are not used, it may be more appropriate to try more flexible parametric families for $G$. Davidian & Gallant (1992), Fattinger, Sheiner & Verotta (1995) and Magder & Zeger (1996) have proposed certain parametric families that incorporate skewness and multimodality, but they are too computationally intensive for routine use.

In Lai & Shih (2003b) and Lai, Shih & Wong (2004a), we address these issues by developing a “hybrid” approach that uses first-order approximations based on Laplace’s
method to evaluate the likelihood when the subject has sufficient data, in combination with Monte Carlo approximations of the likelihood involving relatively few simulation runs when the subject has sparse data. To begin with, suppose the distribution $G$ is normal with mean 0 and covariance matrix $\Sigma$. For given values of $\beta$, $\sigma$ and $\Sigma$, the integral for the $i$th subject in the likelihood function can be written as an expectation $E\psi_i(b)$, which can be computed by Monte Carlo simulations of the random vector $b$ with the normal density function $\phi_{\Sigma}$ having mean 0 and covariance matrix $\Sigma$. Alternatively, letting $e^i(b) = \psi_i(b)\phi_{\Sigma}(b)$, we can use Laplace’s method to approximate the integral

$$\int \ldots \int e^{\ell_i(b)}db^{(1)} \ldots db^{(r)} \div \sqrt{(2\pi)^{r/2} |\tilde{\ell}_i(\hat{b}_i)|^{-1/2}} e^{\ell_i(\hat{b}_i)},$$

where $\tilde{\ell}_i$ is the Hessian matrix of second partial derivatives of $\ell_i$ with respect to the components $b^{(j)}$, and $\hat{b}_i$ is the maximizer of $\ell_i(b)$. Laplace’s approximation basically approximates $\ell_i(b)$ by a quadratic function in a neighborhood of the maximizer $\hat{b}_i$ as $\lambda_{\text{min}}(\cdot) \rightarrow \infty$, where $\lambda_{\text{min}}(\cdot)$ denotes the minimum eigenvalue of a symmetric matrix. If the observations $(y_{ij}, t_{ij}), 1 \leq j \leq n_i$, are sufficiently informative about the $i$th subject’s parameter vector $\theta_i = g(x_i, \beta_0) + b_i$, then for $(\beta, \sigma)$ near the true value $(\beta_0, \sigma_0)$, $\ell_i(b)$ becomes peaked around $\hat{b}_i$ and can be well approximated by the quadratic function $\ell_i(\hat{b}_i) + (b - \hat{b}_i)^T \tilde{\ell}_i(\hat{b}_i)(b - \hat{b}_i)/2$.

Laplace’s approximation is also applicable when $\lambda_{\text{min}}(\Sigma^{-1})$ is large, which occurs when the distribution of $b$ is concentrated around 0.

When $\lambda_{\text{min}}(\Sigma^{-1})$ is not sufficiently large and the $i$th subject has sparse data, Laplace’s method may give a poor approximation to the left hand side of (8), which will be denoted by $L_i(\beta, \sigma, \Sigma)$. These considerations led to the following hybrid method in Lai & Shih (2003b) for evaluating $L_i(\beta, \sigma, \Sigma)$, which combines Laplace’s with the Monte Carlo approximation. Choose a threshold $c$ and let $V_i = -\tilde{\ell}_i(\hat{b}_i)$.

(i) If $\lambda_{\text{min}}(V_i) < c$, evaluate $L_i(\beta, \sigma, \Sigma)$ by the Monte Carlo approximation $B^{-1} \sum_{j=1}^B \psi_i(\Sigma^{1/2}z_j)$, where $z_j, j = 1, \ldots, B$, are independent random vectors from the standard normal distribution. Note that $\Sigma^{1/2}z_j$ is normal with mean 0 and covariance matrix $\Sigma$.

(ii) If $\lambda_{\text{min}}(V_i) \geq c$, evaluate $L_i(\beta, \sigma, \Sigma)$ by its Laplace approximation $(2\pi)^{r/2} |V_i|^{-1/2} e^{\ell_i(\hat{b}_i)}$.

By performing simple diagnostics on the appropriateness of using Laplace’s approximation to evaluate the integral in (8) for the $i$th subject, the hybrid approach preserves the computational simplicity of Laplace’s method when it can be used and switches to the Monte Carlo method when Laplace’s method fails. If the $i$th subject has enough data so that $\ell_i(b)$ is peaked around $\hat{b}_i$ for $(\beta, \sigma)$ near $(\beta_0, \sigma_0)$, the Monte Carlo approach becomes unreliable.
unless B is very large or importance sampling is used to generate the B samples from a distribution that is peaked around \( \hat{b}_i \), so Laplace's method gives a better approximation to \( L_i(\beta, \sigma, \Sigma) \) in this case. On the other hand, if the ith subject has sparse data and \( \ell_i(b) \) is relatively flat in b, then applying the Monte Carlo approach is tantamount to choosing a random distribution \( G_i \), which is the empirical distribution of a sample of B random vectors \( \Sigma^{1/2}z_j \) with standard normal \( z_j \), to approximate \( G \). As there is no need for “high resolution” in the random distribution used to approximate the actual \( G \) (which may not even be normal), using \( 50 \leq B \leq 200 \) samples in the Monte Carlo method should be able to provide enough statistical detail while maintaining a low computational cost comparable to that of the first-order method that can be derived from Laplace's approximation.

In Lai, Shih & Wong (2004a), we improve the Monte Carlo method in (i) above by using importance sampling instead of sampling directly from \( \phi_{\Sigma} \). Specifically, we evaluate \( L_i(\beta, \sigma, \Sigma) \) by the importance sampling estimate

\[
\sum_{j=1}^{B} \psi_i(\zeta_j)w_j / \sum_{j=1}^{B} w_j,
\]

where \( P\{ \zeta_j = \Sigma^{1/2}z_j \} = p = 1 - P\{ \zeta_j = \hat{b}_i + (V_i + \epsilon I)^{-1/2}z_j \} \) with standard normal \( z_j \), which corresponds to sampling \( \zeta_j \) from a mixture of the prior normal distribution with density \( \phi_{\Sigma} \) and the posterior normal distribution with mean \( \hat{b}_i \) and covariance matrix \( (V_i + \epsilon I)^{-1} \), choosing some small \( \epsilon > 0 \) to ensure that the covariance matrix is invertible. Denoting the density function of this mixture distribution by \( \lambda \), note that \( \lambda(x) = p \phi_{\Sigma}(x) + (1-p)\phi_{(V_i+\epsilon I)^{-1}}(x - \hat{b}_i) \). The \( w_j \) in (9) are the importance weights given by \( w_j = \phi_{\Sigma}(\zeta_j) / \lambda(\zeta_j) \). Note that the special case \( p = 1 \) reduces to direct Monte Carlo in (i) above, whereas the case \( p = 0 \) corresponds to a Monte Carlo implementation of Laplace’s method, and recommend choosing \( p \) in the range \( 0.2 \leq p \leq 0.5 \).

Following Lindstrom & Bates (1990), the iterative procedure used to maximize the logarithm of \( \prod_{i=1}^{I} L_i(\beta, \sigma, \Sigma) \) first maximizes over \( \beta \) for fixed \( \eta = (\sigma, \Sigma) \) and then maximizes over \( \eta \) for fixed \( \beta \), repeating until convergence or until a prespecified maximum number of iterative steps is reached. To avoid numerical instability in differentiating \( \log L_i \) with respect to \( \beta \), care should be taken when \( L_i \) computed by (9) is small, in which case we can circumvent the difficulty by simply replacing \( L_i \) by its Laplace approximation whose logarithm is convenient for differentiation. Details on the choice of the threshold \( c \) and starting values for \( \beta, \sigma, \Sigma \) can be found in Section 3.2 of Lai & Shih (2003b). In particular, for typical population PK studies that involve both healthy volunteers from whom intensive
blood sampling is conducted and clinical patients who only have sparse blood samples, one can first single out "potentially good" studies and check their $\lambda_{\min}(V_i)$ values. It is usually adequate to choose a threshold $c$ as low as 10 for $\lambda_{\min}(V_i)$ to determine if these potentially good studies indeed qualify for using Laplace's approximation to $L_i(\beta, \sigma, \Sigma)$. Moreover, for such experimental designs, good starting values can be obtained by using only those studies that have sufficient data so that their $\theta_i$ can be well estimated by the nonlinear least squares estimate based on $(y_{ij}, t_{ij})$, $1 \leq j \leq n_i$.

In the case where the $I$ studies contain many good ones (in the preceding sense), our companion paper Lai & Shih (2003a) has developed nonparametric maximum likelihood estimates of $G, \beta$ and $\sigma$. Previous work in this direction by Mallet (1986, 1992) and Mentré & Mallet (1994) assumes that the $x_i$ are i.i.d. so that $\beta$ can be estimated via the joint distribution of $(x_i, b_i)$. By using the good studies to initialize the nonparametric maximum likelihood estimate of $(G, \beta, \sigma)$, we remove in Lai & Shih (2003a) the restrictive assumption that $x_i$ be i.i.d. and to estimate the finite-dimensional parameter $\beta$ directly without going through the much more difficult infinite-dimensional problem of estimating the joint distribution of $(x_i, b_i)$. A major finding of Lai & Shih (2003a), however, is that even when $G$ is highly non-normal (e.g., has a bimodal distribution), the parametric estimates of $\beta$ and $\sigma$ that assume normal $G$ compare favorably with the nonparametric estimates. An asymptotic theory explaining this is given in Lai & Shih (2003b). Since the nonparametric maximum likelihood estimate $\hat{G}$ has relatively low resolution (with very slow rate of convergence to $G$ as the total sample size $n_1 + \ldots + n_I$ becomes infinite), approximating the population distribution $G$ by a normal distribution (with covariance matrix to be estimated from the data), or by the random distribution $G_i$ when $\ell_i(b)$ is relatively flat in the hybrid method, is usually an innocuous assumption in population PK/PD models.

Laplace's asymptotic formula (8) was also used by Breslow & Clayton (1993) and Lee & Nelder (1996) to derive their estimators for generalized linear models, and by Lin & Zhang (1999) in their extension of generalized linear to generalized additive mixed models. Since Laplace's approximation may be inappropriate for individuals with sparse longitudinal observations, we have recently developed in Lai, Shih & Wong (2004b) a hybrid estimation scheme that combines Laplace's approximation with Monte Carlo computations. Moreover, instead of generalized linear or additive models, our approach uses univariate regression splines and their tensor products as basis functions. Not only can these basis functions model the covariate effects and their interactions effectively, but they also involve linear parameters that can be estimated by the same procedure for generalized linear mixed models.
once the knots are specified. Using this hybrid method to compute the likelihood function, we have developed likelihood-based inference and model selection schemes that can be used to determine the smoothing parameter (e.g., the number of knots for regression splines) and variables to be included in the regression model.

3. Other topics and concluding remarks

A topic that has attracted much recent interest in pharmaceutical biostatistics is mid-course adaptive designs of clinical trials; see e.g. the Controversial Statistical Issue of the *ICSA Bulletin* (Jan. 2003, pp. 37-51). In estimating the sample size of a controlled clinical trial testing a new drug, one often faces such difficulties as that the published information about the control drug may be unreliable because the dosing methods may have changed, or that new technology has been introduced in the measurements, or that new criteria are used to assess efficacy. Therefore, the problem of sample size re-estimation based on an observed treatment difference at some time before the prescheduled end of the trial has attracted considerable attention during the past decade; see e.g. Gould & Shih (1992), Herson & Wittes (1993), Bauer & Köhne (1994), Proschan & Hunsberger (1995), Wassmer (1998) and Jennison & Turnbull (2000, §14.2). Moreover, there are concerns from the regulatory perspective regarding possible inflation of the type I error probability when such sample size adjustments are used in pharmaceutical trials; see O’Neill (1995). For normally distributed outcome variables with common known variance, Fisher (1998), Cui, Hung & Wang (1999), Lehmacher & Wassmer (1999), Posch & Bauer (1999) and Shen & Fisher (1999) have proposed ways to adjust the test statistics after mid-course sample size modification so that the type I error probability is maintained at the prescribed level. Jennison & Turnbull (2003) recently gave a general form of these methods and showed that they performed considerably worse than group sequential tests. Tsiatis & Mehta (2003) independently came to the same conclusion, pointing out their inefficiency because the adjusted test statistics are not functions of the sufficient statistic \((T, S_T)\).

It is possible to adhere to efficient generalized likelihood ratio statistics in a mid-course adaptive design if one uses the non-normal (due to the mid-course adaptation) sampling distribution of the test statistic instead of ignoring the nonnormality and thereby resulting in type I error inflation. I am currently working on a resampling method for testing and interval estimation in mid-course adaptive designs. Besides using efficient test statistics, other statistical issues need to be addressed to make the current generation of adaptive designs more efficient. In particular, there is uncertainty in the mid-course parameter estimates that
determine the final sample size of the adaptive design, and the uncertainty depends on the choice of the initial sample size. This problem was already recognized in Simon’s (1989) seminal paper on optimal two-stage designs for testing $H_0 : p \leq p_0$, in which $p$ is the success probability of a treatment and the optimization problem is to choose the initial sample size so that the expected sample size under $p_0$ is minimized subject to prescribed type I and type II error probabilities at the simple null hypothesis $p_0$ and a given alternative $p_1$ ($> p_0$). Therefore, besides using efficient test statistics, the efficiency of mid-course adaptive designs depends also on when and how to carry out mid-course adaptation, and sequential testing theory, which has provided us with important clues in the development of efficient group sequential designs, should provide useful insights into this problem. The increasing complexity of medical treatments and the lack of a priori knowledge about the treatment effects on different outcome variables have often made it difficult to come up with which endpoints, or which functions thereof, should be included, and which of two or more competing treatment regimens should be used for confirmatory testing of a new drug. Since the maximum resource allocation to a Phase III trial being designed typically depends on what can be claimed in labeling the drug, it is sometimes difficult for a pharmaceutical company to commit a large budget to the trial being planned in the midst of too many uncertainties, so the maximum sample size for a group sequential design may be difficult to pinpoint at the design stage. In this case, which is clearly much more complicated than the oversimplified prototype of a single normally distributed outcome variable with known variance in the literature on the subject, an adaptive design, which finalizes certain design features like maximum sample size, endpoints to be included and treatment regimen to be used after observing an initial sample, is particularly appealing.

Clinical trials are sometimes conducted to compare the efficacy and toxicity of two drugs. Moreover, efficacy is often measured by more than one response variable and so is toxicity. Although univariate methods for assessing each response variable individually have been widely used in this setting, there is often additional need for a single, overall comparison. Combining the univariate comparison by Bonferroni’s inequality ignores the correlations between the response variables and may lack power for alternatives at which the response variables are strongly correlated. Beginning with O’Brien’s (1984) seminal paper, the problem of comparing multivariate treatment effects has received much attention in the literature, and various fixed sample size and group sequential tests have been developed; see e.g. Pocock, Geller & Tsiatis (1987), Tang, Gnecco & Geller (1989), Laska, Tang & Meisner (1992), Jennison & Turnbull (1993), Tang, Geller & Pocock (1993), Follman (1993) and
Thall & Cheng (1999). These methods are mostly based on the multivariate normal distribution and involve linear combinations or maxima of the response variables. In Bloch, Lai & Tubert-Bitter (2001), we have introduced a new formulation of the multiple endpoint problem in clinical trials to compare two treatments based on their sample means and covariance matrices, incorporating the essential univariate and multivariate features of the treatment effects to be compared. An important ingredient of our approach is to demonstrate that the new treatment is non-inferior to the active control for all endpoints and is superior for some endpoint. We are currently working on extensions of this approach to incorporate different outcome variables, including discrete or continuous immediate response variables and censored survival times, and to accommodate nonlinear and nonparametric tests statistics. The increasing complexity of medical therapies and the technological advances in obtaining a wide variety of measurements from study subjects have made multiple endpoints and adjustments for multiple testing increasingly important issues in the design and analysis of clinical trials. Whereas it is now relatively easy to observe many outcome variables from each subject, it is still difficult to recruit a large enough number of subjects so that the trial has reasonable power for demonstrating that the new treatment is better than the active control even for a single outcome variable. Controlling the overall type I error for multiplicity of testing all outcome variables often leads to an unaffordable sample size for a clinical trial to achieve reasonable power. Moreover, there are also technical difficulties in achieving tight control of the overall type I error probability because the correlations of the p-values of the individual tests are difficult to model realistically and to incorporate into the overall significance level, so conservative Bonferroni bounds are often used instead. This is the motivation behind our approach introduced in Lai, Bloch & Tubert-Bitter (2001).

Closely related to the problem of toxicity and efficacy endpoints in Phase III studies is the design and analysis of Phase I studies to determine the dose and dosing regimen. In typical Phase I studies in the development of relatively benign drugs, the drug is initiated at low doses and slowly escalated to show safety at a level where some positive response occurs, and healthy volunteers are used as study subjects. This paradigm does not work for diseases like cancer, for which a non-negligible probability of severe toxic reaction has to be accepted to give the patient some chance of a favorable response to the treatment. Moreover, in many such situations, the benefits of a new therapy may not be known for a long time (perhaps years) after enrollment but toxicities manifest themselves in a relatively short time period (days or weeks). Therefore patients are used as study subjects, and given the hoped-for (rather than observed) benefit for them, one aims at an acceptable level of toxic response in
determining the dose. Due to the absence of a comprehensive methodology, a number of ad hoc protocols for Phase I trials involving new cancer treatments are commonly used. Storer (1989) gives a review of such designs for estimating the maximum tolerated dose (MTD) that he defines to be the dose at which there is 1/3 probability of experiencing a toxic event in connection with a commonly used dose escalation design. The design, which is sequential in nature, treats groups of 3 patients sequentially, starting with the smallest of an ordered set of doses. Escalation occurs if no toxicity is observed in all three patients; otherwise an additional 3 patients are treated at the same dose level. If only one of the 6 patients has toxicity, escalation again continues; otherwise the trial stops with the current dose declared as the MTD. As pointed out by Storer (1989), these designs are difficult to analyze since even a strict quantitative definition of MTD is lacking, “although it should be taken to mean some percentile of a tolerance distribution with respect to some objective definition of clinical toxicity,” and the “implicitly intended” percentile seems to be the 33rd percentile (related to 2/6). Storer (1989) also considered three other “up and down” sequential designs for quantile estimation in the bioassay literature and performed simulation studies of their performance in estimating the 33rd percentile. Subsequent simulation studies by O’Quigley, Pepe & Fisher (1990) showed the performance of these designs to be “dismal”, for which they provided the following explanation: “Not only do (these designs) not make efficient use of accumulated data, they make use of no such data at all, beyond say the previous three, or sometimes six, responses.” They proposed an alternative design, called the “continual reassessment method” (CRM), which uses parametric modeling of the dose-response relationship and a Bayesian approach to estimate parameters and to sequentially determine the dose level x such that the probability p(x) of a toxic event is p_0 (e.g. 1/3). It is natural to try incorporating pharmacokinetic and other covariate information, such as age, gender, performance status (e.g. Karnofsky rating) and disease duration, into the parametric model for dose-response, and to choose a prior distribution that incorporates the current state of knowledge from the medical literature and preclinical studies. Piantadosi & Liu (1996) have demonstrated in simulation studies that incorporating an additional PK parameter (AUC) into the logistic dose-response model of CRM can improve the design of a dose escalation study.

I am currently working on other enhancements of CRM, while also considering other approaches to improved Phase I and II designs. For cancer treatments, the primary objective of a typical Phase I trial is to determine the MTD and dose limiting toxicities of the treatment so that the MTD can be used for the ensuing Phase II trial to evaluate antitumor response. Since MTD seldom has a strict quantitative definition in the protocols
of these trials, defining it formally as the dose that yields a target probability of toxicity response, and using this definition of MTD for the dose to be estimated from a Phase I trial and to be used in the ensuing Phase II trial as in Storer (1989), O’Quigley, Pepe & Fisher (1990) and Piantadosi & Liu (1996), may not best reflect the clinical considerations and constraints on the trial. A traditional dose escalation design, which does not provide adequate information for estimating some quantile of the tolerance distribution, may still be sufficiently informative for estimating something else that can be used to determine the dose of a subsequent Phase II study. In particular, I believe that a Phase I study should also generate useful PK/PD information. By combining this with the dose-limiting toxicities observed, a PK/PD modeling approach can be used to determine the dose of the Phase II study, instead of using the traditional MTD as the Phase II dose. This approach involves developing a PK/PD model from Phase I/II data and using it to evaluate via computer simulations the drug’s pharmacologic actions or therapeutic/toxic responses under dosing schemes not yet studied. Thus, instead of having to perform actual clinical experiments at these dosing schemes to come up with a dose recommendation for a Phase II or III trial, one can use computer experiments to search for the dose. Although one should then conduct a pilot study at this dose to check the actual performance, this simulation approach has eliminated many clinical experiments at intermediate “trial-and-error” doses. Monte Carlo methods to carry out these computer simulations efficiently are therefore of great interest in the application of PK/PD modeling to drug development. In their position paper to the American College of Clinical Pharmacology, Derendorf et al. (2000) point out that “the area of simulation science (i.e., using drug and disease models in a simulation mode to address relevant questions) is still at a very early stage in the discipline of clinical pharmacology” and that “pharmaceutical science educational programs should actively integrate with statistical educational programs at universities.”

In conclusion, drug development is moving towards increasing reliance on mathematical/statistical modeling and increasing size of data sets (measurements collected per subject, though perhaps not more subjects). Computing intensive statistical methodologies, which together with statistics in biotech research constitute the theme of the 2004 ICSA Applied Statistics Symposium, are therefore becoming increasingly important tools for tackling the complexity of new treatments and their underlying biopharmaceutical models.

References

Biometrics 50, 1029-1041.


Lai, TL & Shih, MC (2004). Power, sample size and adaptation considerations in the design
of group sequential trials. *Biometrika* 91, 507-528.


