

## Chapter 2

# Nonlinear Regression, Experimental Design, and Phase I Clinical Trials

In typical Phase I studies in the development of relatively benign drugs, the drug is initiated at low doses and subsequently escalated to show safety at a level where some positive response occurs, and healthy volunteers are often used as study subjects. In Sect. 2.2 we describe some basic pharmacologic principles and models underlying dose determination. These models are typically nonlinear in certain parameters and therefore nonlinear regression models are used. Section 2.1 gives an introduction to nonlinear regression and also describes in this connection nonlinear mixed effects models (NONMEMs), which play a central role in population pharmacokinetics and pharmacodynamics in Sect. 2.2. In connection with Phase I studies, Sect. 2.3 gives an overview of the theory of optimal experimental design. The design and analysis of Phase I studies are described in Sect. 2.4.

This paradigm in Sect. 2.4 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 after enrollment, but toxicities manifest themselves in a relatively short time period. Therefore, patients (rather than healthy volunteers) 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. The objective of Phase I cancer trials is to find a *maximum tolerated dose* (MTD) with the ethical constraint of protecting the study subjects from toxicities in excess of what they can tolerate. To address this constraint, 3 + 3 designs are often used and they are described in Sect. 2.5.1. However, simulation studies by O’Quigley et al. (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 the MTD, or more generally the dose level  $x$  such that the probability  $F(x)$  of a toxic event is

$p$  ( $1/3$  in the case of MTD). Section 2.5.2 describes the CRM and other model-based designs. However, because of the ethical demands for treating patients in the study at safe doses even though they may not be effective, 3+3 designs and their variants are still widely used despite their inadequacy in generating dose-toxicity information for the posttrial estimate of the MTD, for which the model-based designs are more efficient. Bartroff and Lai (2010) have provided a mathematical representation of this dilemma between safe treatment of current patients in the dose-finding cancer trial and efficient experimentation to gather information about the MTD for future patients. The next chapter will describe their formulation of a stochastic optimization problem that addresses this dilemma and summarize their solution of the problem, leading to a class of hybrid designs.

## 2.1 Nonlinear Regression Models

### 2.1.1 Nonlinear Least Squares

As in linear regression models, the method of least squares is commonly used to estimate the unknown parameter vector  $\boldsymbol{\theta}$  in the nonlinear regression model

$$y_j = f_{\boldsymbol{\theta}}(\mathbf{x}_j) + \varepsilon_j, \quad j = 1, \dots, n, \quad (2.1)$$

in which  $f_{\boldsymbol{\theta}}(\cdot)$  is a given nonlinear function of  $\boldsymbol{\theta}$  and  $\varepsilon_j$  are unobservable independent random errors with zero means and

- (a)  $\text{var}(\varepsilon_j) = \sigma^2$  (constant variance error models), or
- (b)  $\text{var}(\varepsilon_j) = f_{\boldsymbol{\theta}}^2(\mathbf{x}_j)\sigma^2$  (constant coefficient of variation error models), or
- (c)  $\text{var}(\varepsilon_j) = f_{\boldsymbol{\theta}}(\mathbf{x}_j)\sigma^2$  (Poisson-type error models).

We can estimate  $\boldsymbol{\theta}$  by generalized least squares (GLS), that is, by minimizing

$$S(\boldsymbol{\theta}) = \sum_{j=1}^n w_j [y_j - f_{\boldsymbol{\theta}}(\mathbf{x}_j)]^2, \quad (2.2)$$

where the weights are inversely proportional to  $\text{var}(\varepsilon_j)$ .

To compute the minimizer  $\hat{\boldsymbol{\theta}}$  of (2.2), we write  $f_{\boldsymbol{\theta}}(\mathbf{x}_j) = f(\boldsymbol{\theta}, \mathbf{x}_j)$ , initialize with  $\hat{\boldsymbol{\theta}}^{(0)}$  and approximate  $f(\boldsymbol{\theta}, \mathbf{x}_j)$  after the  $k$ th iteration, which yields  $\hat{\boldsymbol{\theta}}^{(k)}$ , by

$$f(\boldsymbol{\theta}, \mathbf{x}_j) \approx f(\hat{\boldsymbol{\theta}}^{(k)}, \mathbf{x}_j) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{(k)})^T \nabla f(\hat{\boldsymbol{\theta}}^{(k)}, \mathbf{x}_j)$$

so that (2.1) can be approximated by the linear regression model

$$y_j - f(\hat{\boldsymbol{\theta}}^{(k)}, x_j) = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{(k)})^T \nabla f(\hat{\boldsymbol{\theta}}^{(k)}, x_j) + \varepsilon_j. \quad (2.3)$$

The GLS estimate  $\hat{\boldsymbol{\theta}}^{(k+1)}$  of  $\boldsymbol{\theta}$  in (2.3) is given explicitly, and the iterative scheme is called the *Gauss–Newton algorithm*.

The Gauss increment  $\delta_{k+1} := \hat{\boldsymbol{\theta}}^{(k+1)} - \hat{\boldsymbol{\theta}}^{(k)}$  may produce an increase in  $S(\boldsymbol{\theta})$  when it is outside the region where the linear approximation holds. To ensure a decrease in  $S(\boldsymbol{\theta})$ , use a step factor  $0 < \lambda \leq 1$  so that  $S(\hat{\boldsymbol{\theta}}^{(k)} + \lambda \delta^{(k)}) < S(\boldsymbol{\theta}^{(k)})$ . A commonly used method is to start with  $\lambda = 1$  and halve it until we have  $S(\boldsymbol{\theta}^{(k+1)}) < S(\boldsymbol{\theta}^{(k)})$ . A commonly used criterion for numerical convergence is the size of the parameter increment relative to the parameter value. Another criterion is that the relative change in  $S(\boldsymbol{\theta})$  be small. A third criterion is that  $Y - \eta(\boldsymbol{\theta}^{(k)})$  be nearly orthogonal to the tangent space of  $\eta(\boldsymbol{\theta}) := (f(\boldsymbol{\theta}, \mathbf{x}_1), \dots, f(\boldsymbol{\theta}, \mathbf{x}_n))^T$  at  $\boldsymbol{\theta}^{(k)}$ . The Gauss–Newton algorithm is aborted at the  $k$ th step when one gets a singular (or nearly singular) coefficient matrix in the linear equation defining GLS. It may also stop after reaching a prescribed upper bound on the number of iterations without convergence. When one does not get an answer from the Gauss–Newton algorithm, one should choose another starting value and repeat the algorithm.

### 2.1.2 Nonlinear Mixed Effects Models

As will be explained in the next section, two important pharmacologic models are the poly-exponential model  $f_{\boldsymbol{\theta}}(t) = \sum_{k=1}^K \alpha_k e^{-\lambda_k t}$ , with  $\boldsymbol{\theta} = (\alpha_1, \dots, \alpha_K, \lambda_1, \dots, \lambda_K)^T$  and  $t$  denoting time, and the Michaelis–Menten model  $f_{\boldsymbol{\theta}}(u) = vu/(\alpha + u)$ , with  $\boldsymbol{\theta} = (v, \alpha)^T$  and  $u$  denoting drug concentration. A Phase I trial collects data from  $I$  subjects, yielding  $(y_{ij}, x_{ij})$ ,  $i = 1, \dots, I$ ,  $j = 1, \dots, n_i$ . In the analysis of these data, it is more flexible to allow subject-specific parameters  $\boldsymbol{\theta}_i$  in (2.1). This leads to a NONMEM of the form

$$y_{ij} = f_i(t_{ij}, \boldsymbol{\theta}_i) + \varepsilon_{ij}, \quad \boldsymbol{\theta}_i = \mathbf{g}(\mathbf{x}_i, \boldsymbol{\beta}) + \mathbf{b}_i \quad (1 \leq j \leq n_i, 1 \leq i \leq I), \quad (2.4)$$

in which  $\boldsymbol{\theta}_i$  is a  $1 \times r$  vector of the  $i$ th subject's parameters whose regression function on the subject's observed covariate  $\mathbf{x}_i$  is given by  $\mathbf{g}(\mathbf{x}_i, \boldsymbol{\beta})$  with  $1 \times s$  parameter vector  $\boldsymbol{\beta}$ , which is the “fixed effect” to be estimated. The “random effects”  $\mathbf{b}_i$  in (2.4) are assumed to be independent and identically distributed, having common distribution  $G$  with mean 0. The  $i$ th subject's response  $y_{ij}$  at  $t_{ij}$  has mean  $f_i(t_{ij}, \boldsymbol{\theta}_i)$ , in which  $f_i$  is a known function and  $t_{ij}$  may represent time or some covariate value (such as drug concentration) at that time. Given  $\boldsymbol{\theta}_i$ , the random errors  $\varepsilon_{ij}$  are assumed to be normal with mean 0 and standard deviation  $\sigma w(\boldsymbol{\theta}_i)$ , in which  $w$  is a given

function and  $\sigma$  is an unknown parameter. The regression function  $\mathbf{g}$  relates  $\boldsymbol{\theta}_i$  to the  $i$ th subject's physiologic characteristics that constitute the covariate vector  $\mathbf{x}_i$  in (2.4). The first equation of (2.4) is often called the *individual measurement model* and the second equation the *population structure model*. The population distribution  $G$  is usually assumed to be normal with mean 0 and covariance matrix  $\boldsymbol{\Sigma}$  so that  $\boldsymbol{\beta}$ ,  $\sigma$ ,  $\boldsymbol{\Sigma}$  can be estimated by maximum likelihood. However, unlike linear mixed effects (LME) models in which the normal assumption on  $G$  yields closed-form expressions of the likelihood, the normality of  $G$  in NONMEM leads to computationally intensive likelihoods that involve  $I$  integrals. A commonly used approach, as adopted in the software package NONMEM (Beal and Sheiner 1992) or the nlme procedure in R, is to develop iterative schemes based on first-order approximations of  $f_i(t_{ij}, \mathbf{g}(\mathbf{x}_i, \boldsymbol{\beta}) + \mathbf{b}_i)$  in (2.4), so that the normal assumption on  $G$  can be used to reduce the problem to that of a linear Gaussian mixed effects model at each iterative step.

Unless otherwise stated, we shall assume throughout the sequel that the random errors  $\varepsilon_{ij}$  in model (2.4) have common variance  $\sigma^2$  (so  $w(\boldsymbol{\theta}) \equiv 1$ ). The likelihood function  $L(\boldsymbol{\beta}, \sigma, \boldsymbol{\Sigma})$  is proportional to

$$|\boldsymbol{\Sigma}|^{-I/2} \prod_{i=1}^I \int_{\mathbb{R}^r} \sigma^{-n_i} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n_i} [y_{ij} - f_i(t_{ij}, \mathbf{g}(\mathbf{x}_i, \boldsymbol{\beta}) + \mathbf{b}_i)]^2 - \frac{1}{2} \mathbf{b}_i \boldsymbol{\Sigma}^{-1} \mathbf{b}_i^T \right\} d\mathbf{b}_i, \quad (2.5)$$

where  $|\boldsymbol{\Sigma}|$  denotes the determinant of  $\boldsymbol{\Sigma}$ . For the case of more general  $w(\boldsymbol{\theta}_i)$ , simply replace  $\sigma$  in (2.5) by  $\sigma w(\mathbf{g}(\mathbf{x}_i, \boldsymbol{\beta}) + \mathbf{b}_i)$ . Computing the maximum likelihood estimate of  $(\boldsymbol{\beta}, \sigma, \boldsymbol{\Sigma})$  via numerical integration and nonlinear optimization becomes prohibitively difficult for large  $I$ . Letting  $\boldsymbol{\eta} = (\sigma, \boldsymbol{\Sigma})$ , Lindstrom and Bates (1990) proposed the following iterative procedure that involves successive linear approximations to  $f_i(t_{ij}, \mathbf{g}(\mathbf{x}_i, \boldsymbol{\beta}) + \mathbf{b}_i)$ . At the  $m$ th iteration, the Lindstrom–Bates procedure consists of a pseudo-data step and a LME step:

(a) *The pseudo-data step*

Given the current estimate  $\hat{\boldsymbol{\eta}}^{(m)}$  of  $\boldsymbol{\eta}$ , compute  $\hat{\boldsymbol{\beta}}^{(m)} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\eta}}^{(m)})$  and  $\hat{\mathbf{b}}_i^{(m)} = \hat{\mathbf{b}}_i(\hat{\boldsymbol{\eta}}^{(m)})$ ,  $1 \leq i \leq I$ , that jointly minimize

$$\sum_{i=1}^I \left\{ (\hat{\sigma}^{(m)})^{-2} S_i(\boldsymbol{\beta}, \mathbf{b}) + \mathbf{b}_i \left( \hat{\boldsymbol{\Sigma}}^{(m)} \right)^{-1} \mathbf{b}_i^T \right\}, \quad (2.6)$$

where

$$S_i(\boldsymbol{\beta}, \mathbf{b}) = \sum_{j=1}^{n_i} [y_{ij} - f_i(t_{ij}, \mathbf{g}(\mathbf{x}_i, \boldsymbol{\beta}) + \mathbf{b})]^2.$$

This can be carried out by modifying a standard nonlinear least squares routine; see Sect. 6.1 of Lindstrom and Bates (1990). Define the  $s \times n_i$ ,  $r \times n_i$ , and  $1 \times n_i$  matrices

$$\begin{aligned}
\mathbf{X}_i^{(m)} &= \left( \frac{\partial f_i}{\partial \boldsymbol{\beta}} \left( t_{ij}, \mathbf{g}(\mathbf{x}_i, \boldsymbol{\beta}) + \hat{\mathbf{b}}_i^{(m)} \right) \right) \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(m)}} \Big|_{1 \leq j \leq n_i}, \\
\mathbf{Z}_i^{(m)} &= \left( \frac{\partial f_i}{\partial \mathbf{b}_i} \left( t_{ij}, \mathbf{g}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}^{(m)}) + \mathbf{b}_i \right) \right) \Big|_{\mathbf{b}_i=\hat{\mathbf{b}}_i^{(m)}} \Big|_{1 \leq j \leq n_i}, \\
\mathbf{Y}_i^{(m)} &= \left( y_{ij} - f_i \left( t_{ij}, \mathbf{g}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}^{(m)}) + \hat{\mathbf{b}}_i^{(m)} \right) \right) \Big|_{1 \leq j \leq n_i} + \hat{\boldsymbol{\beta}}^{(m)} \mathbf{X}_i^{(m)} + \hat{\mathbf{b}}_i^{(m)} \mathbf{Z}_i^{(m)}.
\end{aligned}$$

(b) *The LME step*

Linear approximation to  $f_i(t_{ij}, \mathbf{g}(\mathbf{x}_i, \boldsymbol{\beta}) + \mathbf{b}_i)$  around  $(\hat{\boldsymbol{\beta}}^{(m)}, \hat{\mathbf{b}}_i^{(m)})$  leads to the LME model

$$\mathbf{Y}_i^{(m)} = \boldsymbol{\beta} \mathbf{X}_i^{(m)} + \mathbf{b}_i \mathbf{Z}_i^{(m)} + (\varepsilon_{i1}, \dots, \varepsilon_{in_i}). \quad (2.7)$$

The integrals in (2.5) for the likelihood function of the LME model (2.7) (instead of (2.4)) have closed-form expressions, yielding maximum likelihood estimates of the form

$$\hat{\boldsymbol{\beta}} = \left( \sum_{i=1}^I \mathbf{Y}_i^{(m)} \mathbf{V}_{i,m}^{-1} \mathbf{X}_i^{(m)T} \right) \left( \sum_{i=1}^I \mathbf{X}_i^{(m)} \mathbf{V}_{i,m}^{-1} \mathbf{X}_i^{(m)T} \right)^{-1}, \quad (2.8)$$

where  $\mathbf{V}_{i,m} = \mathbf{Z}_i^{(m)T} \hat{\boldsymbol{\Sigma}} \mathbf{Z}_i^{(m)} + \hat{\sigma}^2 \mathbf{I}_{n_i}$  and  $\hat{\boldsymbol{\eta}} = (\hat{\sigma}, \hat{\boldsymbol{\Sigma}})$  is computed via the Newton–Raphson algorithm to maximize the likelihood.

Wolfinger (1993) derives the above pseudo-data step by using Laplace’s approximation arguments. Vonesh (1996) directly approximates the integrals in (2.5) with  $\sigma, \boldsymbol{\beta}, \boldsymbol{\Sigma}$  fixed, by using Laplace’s asymptotic formula

$$\int_{\mathbb{R}^r} e^{\ell_i(\mathbf{b})} d\mathbf{b} \sim (2\pi)^{r/2} \left\{ \det \left( -\ddot{\ell}_i(\hat{\mathbf{b}}_i) \right) \right\}^{-1/2} e^{\ell_i(\hat{\mathbf{b}}_i)}, \quad (2.9)$$

where  $\hat{\mathbf{b}}_i$  is the maximizer of  $\ell_i(\mathbf{b})$  and  $\ddot{\ell}_i$  is the Hessian matrix of second partial derivatives of  $\ell_i$  with respect to the components of  $\mathbf{b}$ . Noting that Laplace’s approximation to an integral corresponds to adaptive Gaussian quadrature with one quadrature point, Pinheiro and Bates (1995) use adaptive Gaussian quadrature with  $q$  quadrature points to compute the integrals in (2.5). Lai and Shih (2003b) have developed a hybrid method that uses (2.9) if the minimum eigenvalue  $\lambda_{\min}(-\ddot{\ell}_i(\hat{\mathbf{b}}_i))$  exceeds a prescribed threshold and uses Monte Carlo simulations otherwise. Lai et al. (2006b) introduce importance sampling to refine the Monte Carlo component of the hybrid method. They also point out the importance of approximating the likelihood function adequately with relative ease for selecting good predictive models  $\mathbf{g}(\mathbf{x}_i, \boldsymbol{\beta})$ .

Since the normality assumption on  $G$  only provides numerically tractable maximum likelihood estimates after various approximations, a natural alternative is to try estimating  $G$  nonparametrically by a distribution with finite support, with the number of support points depending on the sample size. However, even for the simple case  $n_i \equiv n$  and  $f_i(t_{ij}, \boldsymbol{\theta}_i) = \boldsymbol{\theta}_i$  with known  $\boldsymbol{\beta}$  and  $\sigma$ , it is difficult to estimate  $G$  well since the optimal rate of convergence of the estimate to  $G$  is very slow when  $G$  has a smooth density function, as pointed out by Fan (1991). Lai and Shih (2003a) have developed a nonparametric maximum likelihood estimator (MLE) of  $G$  when there are  $I' \leq I$  subjects whose  $\boldsymbol{\theta}_i$  can be well estimated by the nonlinear least squares estimator  $\tilde{\boldsymbol{\theta}}_i$  based on  $\{(y_{ij}, t_{ij}) : 1 \leq j \leq n_i\}$ . Because of the low resolution in estimating  $G$  nonparametrically, however, the nonparametric approach does not yield a better estimate of  $f(\cdot, \cdot)$  in the simulation study reported by Lai and Shih (2003a) who consider the case of  $f_i$  being all equal (to  $f$ ).

## 2.2 Pharmacokinetics and Pharmacodynamics

The nonlinear regression and NONMEM in the preceding section are basic statistical methods in pharmacology, which is the science dealing with interactions between living systems and molecules, especially chemicals introduced from outside the system. This broad definition includes clinical pharmacology (whose objective is to prevent, diagnose, and treat diseases with drugs) and the pathogenesis of diseases due to chemicals in the environment; see Katzung (1995). A drug is defined as a small molecule that, when introduced into the body, alters the body's function. The component of a cell or organism that interacts with a drug and initiates the 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), which is the study of drug effects and their mechanisms of action. The relation between the dose of a drug and its clinically observed effects can be quite complex. In carefully controlled in vitro systems, however, the relation between the concentration of a drug at the site(s) of action and its effects can often be described by relatively simple mathematical models. How a drug dose produces its effects involves not only pharmacodynamics but also *pharmacokinetics* (PK). The latter is concerned with the concentration–time curve that is associated with the following “history” of a single administration of a drug:

- (a) *Absorption phase of the drug into the body*: Transfer of the drug from its site of administration (via oral, or inhalational, or intravenous, or other route) into the bloodstream.
- (b) *Distribution phase*: Distribution of the drug to different compartments of the body, including receptor-binding sites in the target tissue, and resulting in rapid decline in plasma concentration.
- (c) *Elimination phase*: Excretion of chemically unchanged drug or elimination via metabolism that converts the drug into one or more metabolites (e.g., at the liver).

Drug administration can be divided into two phases, a PK phase in which the kinetics of drug absorption, distribution, and elimination translate into drug concentration–time relationships in the body, and a PD phase in which the drug concentration at the site(s) of action leads to the response/effects produced. Knowledge of both phases is important for the design of a dosage regimen to achieve the therapeutic objective. Since both the desired response and toxicity of the drug are functions of the drug concentration at the site(s) of action, the therapeutic objective can be achieved only when the drug concentration lies within a “therapeutic window,” outside which the therapy is either ineffective or has unacceptable toxicity. Drug concentrations, however, can rarely be measured directly at the sites of action and are typically measured at the plasma, which is a more accessible site. An optimal dosage regimen can therefore be defined as one that maintains the plasma concentration of a drug within the therapeutic window. This can be achieved for many drugs by giving an initial dose to yield a plasma concentration within the therapeutic window and then maintaining the concentration within this window by periodic doses to replace the drug lost over time.

A basic goal of PD models is to describe and quantify the steady-state relationship of drug concentration ( $C$ ) at an effector site to the drug effect ( $E$ ). The simplest PD model for one drug is the so-called Emax model defined by  $E = e_{\max}C/(C + c_{50})$ , where  $e_{\max}$  is the maximum effect that the drug can produce and  $c_{50}$  is the concentration that yields 50% of  $e_{\max}$ . This equation is the same as the Michaelis–Menten model in enzyme kinetics. A generalization to incorporate the baseline effect  $e_0$  leads to

$$E = e_0 + e_{\max}C/(C + c_{50}). \quad (2.10)$$

A convenient surrogate for the drug concentration at an effector site, which is difficult to measure directly, is dose ( $D$ ). In empirical studies,  $C$  and  $c_{50}$  in (2.10) are replaced by  $D$  and  $ED_{50}$ .

There is a large literature on PK models, which can roughly be classified as “mechanistic” and “empirical”; see [Rowland and Tozer \(1989\)](#). In mechanistic models, the body is viewed in terms of kinetic compartments between which the drug distributes and from which elimination occurs. The kinetics is often described by a linear system of ordinary differential equations, which have explicit solutions involving exponential functions. On the other hand, the rate constants of a compartmental model may be functions of the concentration of the drug itself or another metabolite/interacting drug, leading to a system of nonlinear differential equations that have to be solved numerically. Empirical PK models are typically poly-exponential models of the form  $\sum \alpha_i e^{-\lambda_i t}$ . One such model that is commonly used is the one-compartment model

$$y_j = \frac{Dk_a}{V(k_a - k_e)}(e^{-k_e t_j} - e^{-k_a t_j}) + \varepsilon_j, \quad 1 \leq j \leq n, \quad (2.11)$$

in which  $y_j$  is the concentration at time  $t_j$  after the administration of a single oral dose  $D$ . Here  $V$ ,  $k_a$ ,  $k_e$  are the volume of distribution, absorption rate constant,

and elimination rate constant, respectively. Note that (2.11) has the form of a bi-exponential model  $\alpha_1 e^{-\lambda_1 t} + \alpha_2 e^{-\lambda_2 t}$  with  $\alpha_1 = -\alpha_2$ .

So far we have considered estimation of the PK/PD parameters of a subject from the data in a study on the subject. In many PK/PD studies, however, 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 embedding the individual parametric PK/PD models in a population model. For example, the  $y_j$  in (2.11) are now replaced by  $y_{ij}$ , where  $i$  denotes the subject number. Since the dose, volume of distribution, absorption, and elimination rate constants may vary from subject to subject, we also have to replace  $D, V, k_a, k_e, n$  by  $D_i, V_i, k_{ai}, k_{ei}$ , and  $n_i$  in (2.11). Let  $\theta_i$  be the vector consisting of the logarithms of the PK parameters  $V_i, k_{ai}, k_{ei}$ . The unknown  $\theta_i$  may vary with certain covariates, such as the subject's age and body weight. How can the individual subjects' data be used to analyze such relationships for the target population, of which the subjects can be regarded as a sample? The NONMEM provides a valuable tool to address this problem. The subject's data are often too sparse to provide an adequate estimate  $\hat{\theta}_i$  of  $\theta_i$  so that  $h(\hat{\theta}_i)$  can be used to estimate  $h(\theta_i)$ . If  $\beta$ ,  $\sigma$ , and  $G$  are known, then a natural estimate of  $h(\theta_i)$  in the mixed effects model is the posterior mean  $E_{\beta, \sigma^2, G}[h(\theta_i) | \text{subject } i\text{'s data}]$ . Without assuming  $\beta$ ,  $\sigma^2$ , and  $G$  to be known, the empirical Bayes approach replaces them by their estimates  $\hat{\beta}$ ,  $\hat{\sigma}^2$ , and  $\hat{G}$  from the  $I$  studies so that  $h(\theta_i)$  is estimated by  $\widehat{h(\theta_i)} = E_{\hat{\beta}, \hat{\sigma}^2, \hat{G}}[h(\theta_i) | \text{subject } i\text{'s data}]$ .

Returning to the PD model (2.10), the variable  $C$  refers to concentration at an effector site. It is usually impossible to measure  $C$  directly, so some surrogate for  $C$  has to be used. On the other hand, if one has a kinetic model for  $C$ , then it can be used to impute the value of  $C$  from the blood/urine measurements. Chapter 9 of [Davidian and Giltinan \(1995\)](#) illustrates how population PK/PD models can be synthesized for such tasks.

## 2.3 Theory of Optimal Design

The conditions under which an experiment is performed affect the quality of information arising from the experiment. *Optimal design of experiments* (or simply *optimal design*) concerns how to choose these conditions, or “settings,” in order to maximize the amount of information coming from an experiment and thus optimize the quality of statistical inference that is possible. In the context of clinical trials, these settings may be the treatment dose or dosing regimen, the treatment type, or the characteristics of the patient who may be randomized into one of multiple treatment groups. In what follows we give a brief introduction to the theory emphasizing general concepts over technicalities; for a more complete mathematical treatment, see [Fedorov \(1972\)](#) or [Silvey \(1980\)](#).



### 2.3.1 Optimal Design Theory in Linear Regression Models

Consider a random variable  $Y \sim p(y|\mathbf{x}, \boldsymbol{\theta}, \sigma)$  such that  $\text{Var}(Y|\mathbf{x}, \boldsymbol{\theta}, \sigma) = \sigma^2$  and

$$E(Y|\mathbf{x}, \boldsymbol{\theta}, \sigma) = \boldsymbol{\theta}^T \mathbf{x}, \quad (2.12)$$

where  $\mathbf{x} = (x_1, \dots, x_k)^T \in \mathcal{X}$ , the *design space*, is a vector of control variables which may be chosen by the experimenter and  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^T$  and  $\sigma$  are unknown parameters. The linear regression model  $Y \sim N(\boldsymbol{\theta}^T \mathbf{x}, \sigma^2)$  will be referred to as the normal case for which the linearity of (2.12) in  $\boldsymbol{\theta}$  greatly simplifies the problem of choosing  $\mathbf{x}$  in order to get the maximal information about  $\boldsymbol{\theta}$  out of  $Y$ . Before proceeding to the problem, we make two remarks about the assumptions. First, (2.12) can be extended to  $E(Y|\mathbf{x}, \boldsymbol{\theta}, \sigma) = \theta_1 f_1(\mathbf{x}) + \dots + \theta_k f_k(\mathbf{x})$ , where  $\mathbf{f} = (f_1, \dots, f_k)$  and the  $f_i$  are known functions. Replacing  $\mathbf{x}$  by  $\mathbf{f}$  and the design space  $\mathcal{X}$  by  $\mathbf{f}(\mathcal{X})$  reduces to the original problem. Second, the variance  $\sigma^2$  could be replaced by  $\sigma^2 v(\mathbf{x})$  for any known function  $v$  because this case can again be reduced to the original one with  $\tilde{Y} = Y/\sqrt{v(\mathbf{x})}$  replacing  $Y$ .

Suppose we are planning to perform  $n$  independent experiments with input variables  $\mathbf{x}_1, \dots, \mathbf{x}_n$  which will result in the independent observations  $Y_1, \dots, Y_n$ . The least squares estimator  $\hat{\boldsymbol{\theta}}$  of  $\boldsymbol{\theta}$ , or equivalently the MLE in the normal case, has covariance matrix

$$\sigma^2 \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \quad (2.13)$$

when the  $\mathbf{x}_1, \dots, \mathbf{x}_n$  are such that  $\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$  is invertible. Two key properties of (2.13) are that it does not depend on  $\boldsymbol{\theta}$ , which is a direct result of the linear structure of (2.12), and that it depends on  $\sigma$  but in a special way such that the minimizer of any function of (2.13) does not depend on  $\sigma$ . If the desire is to make (2.13) “small” in some sense, then this is equivalent to making the *information matrix*

$$\mathbf{M} = \mathbf{M}(\mathbf{x}_1, \dots, \mathbf{x}_n) = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \quad (2.14)$$

“large.” Since  $\mathbf{M}$  is a matrix, there are various criteria for judging  $\mathbf{M}$  to be “large” so that the optimal design problem is to find the  $\mathbf{x}_1, \dots, \mathbf{x}_n$  that maximize  $\Psi(\mathbf{M})$ , for some real-valued function  $\Psi$ . Some popular choices for  $\Psi$  include the following:

*D-optimality:* Under the normality assumption, the volume of the confidence ellipsoid for  $\boldsymbol{\theta}$  is proportional to  $(\det \mathbf{M})^{-1/2}$ , and minimizing this is equivalent to maximizing  $\Psi(\mathbf{M}) = \log \det(\mathbf{M})$ .

*c-optimality:* For a given  $k$ -vector  $\mathbf{c}$ , the least squares estimate (or MLE in the normal case) of the linear combination  $\mathbf{c}^T \boldsymbol{\theta}$  is  $\mathbf{c}^T \hat{\boldsymbol{\theta}}$ , which has variance proportional to

$$\mathbf{c}^T \mathbf{M}^{-1} \mathbf{c}, \quad (2.15)$$

hence,  $\Psi(\mathbf{M}) = -\mathbf{c}^T \mathbf{M}^{-1} \mathbf{c}$  is the function to be maximized for this criterion.

*E-optimality*: Closely related to  $\mathbf{c}$ -optimality is the criterion which seeks to minimize the maximum of (2.15) over all  $\mathbf{c}$  in the  $k$ -dimensional unit sphere, that is, to minimize

$$\max_{\mathbf{c}: \|\mathbf{c}\|=1} \mathbf{c}^T \mathbf{M}^{-1} \mathbf{c}.$$

Kiefer (1974) showed that this is equivalent to maximizing the minimum eigenvalue of  $\mathbf{M}$ , which  $\Psi$  is taken to be for this criterion.

Because of the discreteness of the problem of maximizing  $\Psi(\mathbf{M})$  over all choices for  $\mathbf{x}_1, \dots, \mathbf{x}_n$ , standard numerical optimization techniques often have difficulty, especially when  $n$  is large. Moreover, the value of  $n$  itself may not be well motivated in the experimenter's mind prior to the experiment. An elegant solution comes with the identification of  $\mathbf{x}_1, \dots, \mathbf{x}_n$  to a certain probability measure over the design space  $\mathcal{X}$ , that is, the discrete measure placing mass  $1/n$  on each point  $\mathbf{x}_1, \dots, \mathbf{x}_n$ , and enlarging the search to include all such probability measures has led to the *approximate theory* of linear optimal design (Kiefer 1974). Letting  $\mu$  denote a probability measure on  $\mathcal{X}$  and

$$\mathbf{M}(\mu) = E_\mu(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T), \quad (2.16)$$

where  $\tilde{\mathbf{x}}$  denotes the random variable with distribution  $\mu$ , the optimization problem is equivalent to finding the measure  $\mu$  that maximizes  $\Psi(\mathbf{M}(\mu))$ . Closed-form analytic solutions are available in some cases, but in general, iterative algorithms are necessary to find optimal designs; see Fedorov (1972, Sect. 2.10).

### 2.3.2 Elfving's Method for $\mathbf{c}$ -Optimal Design

In order to give concrete examples we next focus on  $\mathbf{c}$ -optimal designs because, in low dimensions, optimal designs can often be found exactly by using an elegant geometric method of Elfving (1952). Assume that a linear model (2.12) is given and that the design space  $\mathcal{X} \subseteq \mathbb{R}^k$  is compact, that is, closed and bounded. For a given vector  $\mathbf{c} \in \mathbb{R}^k$ , the problem is to find the measure  $\mu$  on  $\mathcal{X}$  maximizing

$$\Psi(\mathbf{M}(\mu)) = -\mathbf{c}^T \mathbf{M}(\mu)^{-1} \mathbf{c}$$

(or equivalently, minimizing  $\mathbf{c}^T \mathbf{M}(\mu)^{-1} \mathbf{c}$ ), where  $\mathbf{M}(\mu)$  is given by (2.16). It follows from the facts that  $\Psi$  is a concave function (of matrices), the space of all matrices  $\mathbf{M}(\mu)$  is convex, and Carathéodory's theorem (see Silvey 1980, p. 72) that a maximizer of  $\Psi(\mathbf{M}(\mu))$  can be found among the measures  $\mu$  with at most  $k$  support points, that is,  $\mu$  of the form

$$\mu = \sum_{i=1}^k p_i \delta_{\mathbf{x}_i}, \quad \text{where } \sum_{i=1}^k p_i = 1, \mathbf{x}_i \in \mathcal{X} \text{ and } p_i \geq 0 \text{ for all } i = 1, \dots, k, \quad (2.17)$$

in which  $\delta_{\mathbf{x}}$  is the degenerate measure putting mass 1 at  $\mathbf{x}$ . Therefore, we can restrict our search for a maximizer of  $\Psi(\mathbf{M}(\mu))$  to probability measures of the form (2.17).

Elfving's (1952) method for finding this discrete measure is the following. Let  $\mathcal{X}^- = \{-\mathbf{x} : \mathbf{x} \in \mathcal{X}\}$  denote the reflection of  $\mathcal{X}$  through the origin, and let  $\mathcal{S}$  denote the *convex hull* of  $X \cup X^-$ , that is,  $\mathcal{S}$  is the collection of all points of the form  $\sum_{i=1}^k p_i \mathbf{z}_i$ , where  $\sum_{i=1}^k p_i = 1$ ,  $p_i \geq 0$  and  $\mathbf{z}_i \in X \cup X^-$  for all  $i = 1, \dots, k$ . Extend a ray from the origin through the point  $\mathbf{c}$  and let  $\mathbf{s}^* \in \mathcal{S}$  be the point where this ray pierces the boundary of  $\mathcal{S}$ . By the definition of  $\mathcal{S}$ ,  $\mathbf{s}^*$  can be written as

$$\mathbf{s}^* = \sum_{i=1}^k \pm p_i \mathbf{x}_i$$

for some choice of signs, where the  $p_i$  and  $\mathbf{x}_i$  satisfy the conditions in (2.17). Then the design measure  $\sum_{i=1}^k p_i \delta_{\mathbf{x}_i}$  is  $\mathbf{c}$ -optimal, that is, the design that places weight  $p_i$  at point  $\mathbf{x}_i$ ,  $i = 1, \dots, k$ ; see Chernoff (1972) for a sketch of the proof.

*Example 2.1.* Suppose that independent responses  $Y_i$  to a drug with dose  $x_i \in [0, a]$  ( $a > 0$  the known “maximum dose”) are given by

$$Y_i = \alpha x_i + \beta x_i^2 + \varepsilon_i, \quad i = 1, \dots, n,$$

where the  $\varepsilon_i$  are i.i.d.  $N(0, \sigma^2)$  random variables. This model fits into the form (2.12) by taking  $\mathbf{x} = (x, x^2)^T$ ,  $\boldsymbol{\theta} = (\alpha, \beta)^T$ , and  $k = 2$ . Not worrying for the moment about what value of  $n$  to use, suppose that the ultimate objective of the  $n$  measurements to be taken is to estimate optimally the mean response at some critical dose  $x_0$ ,  $0 < x_0 \leq a$ . Thus, it is appropriate to consider the  $\mathbf{c}$ -optimal design with  $\mathbf{c} = (x_0, x_0^2)^T$  for optimal estimation of the mean response  $\alpha x_0 + \beta x_0^2 = \mathbf{c}^T \boldsymbol{\theta}$  at dose  $x_0$ . The design space

$$\mathcal{X} = \{(x, x^2) \in \mathbb{R}^2 : 0 \leq x \leq a\},$$

as well as  $\mathcal{X}^-$ , are truncated parabolas. Let  $\gamma = \sqrt{2} - 1 = .4142136\dots$

*Case 1.* If  $0 < x_0 \leq \gamma a$ , then the ray in direction  $(x_0, x_0^2)$  pierces  $\mathcal{S}$  at the point

$$\left( \frac{a^2 \gamma (1 - \gamma)}{a(\gamma^2 + 1) - x_0(\gamma + 1)}, x_0 \cdot \frac{a^2 \gamma (1 - \gamma)}{a(\gamma^2 + 1) - x_0(\gamma + 1)} \right). \quad (2.18)$$

Setting (2.18) equal to

$$p(\gamma a, \gamma^2 a^2) + (1 - p)(-a, -a^2)$$

and solving for  $p$  gives

$$p = \frac{1 - x_0(\gamma + 1) + a\gamma}{(\gamma + 1)[a(\gamma^2 + 1) - x_0(\gamma + 1)]}. \quad (2.19)$$

Thus, the  $\mathbf{c}$ -optimal design is  $\mu = p\delta_{(\gamma a, \gamma^2 a^2)} + (1 - p)\delta_{-(a, a^2)}$  which, in other words, puts the fraction  $p$  of observations at dose  $x = \gamma a$  and the remaining fraction  $1 - p$  at dose  $x = a$ . Note that the design may not be implementable in practice, since  $np$ , with  $p$  given by (2.19), may not be an integer. This is a consequence of using the optimal design formulation that uses a probability measure rather than a discrete collection of  $n$  design points to represent a design. In practice, if  $np$  is not an integer, then choose the closest integer.

*Case 2.* If  $\gamma a \leq x_0 \leq a$ , then the ray in direction  $(x_0, x_0^2)$  pierces  $\mathcal{S}$  precisely at  $(x_0, x_0^2)$ ; hence, the  $\mathbf{c}$ -optimal design is simply  $\delta_{(x_0, x_0^2)}$ , that is, the design that puts all measurements at dose  $x = x_0$ .

### 2.3.3 Extension to Nonlinear Models

A key feature of the linear design theory in the previous section is that the information matrix (2.14) does not depend on  $\boldsymbol{\theta}$ . In this section we consider the more general case where

$$Y \sim p(y|\mathbf{x}, \boldsymbol{\theta}) \quad \text{and} \quad E(Y|\mathbf{x}, \boldsymbol{\theta}) = \eta(\boldsymbol{\theta}, \mathbf{x}) \quad (2.20)$$

for some function  $\eta$ , where we have absorbed the parameter  $\sigma$  of the previous section into  $\boldsymbol{\theta}$  for notational simplicity since the distinction between parameters of interest and nuisance parameters does not matter in the nonlinear case. To generalize the notion of information matrix used above, we note that (2.13) is the inverse of the Fisher information matrix of independent observations  $Y_1, \dots, Y_n$ , and therefore it is natural to define  $\mathbf{M}(\mu) = \mathbf{M}(\mu, \boldsymbol{\theta})$  in the nonlinear case as the Fisher information of the design  $\mu$

$$\mathbf{M}(\mu, \boldsymbol{\theta}) = \int_{\mathcal{X}} E \left[ -\frac{\partial^2 p(Y|\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right] d\mu(\mathbf{x}), \quad (2.21)$$

where the expectation in (2.21) is taken over  $Y$ . As the notation suggests, the information matrix  $\mathbf{M}(\mu, \boldsymbol{\theta})$  depends on  $\boldsymbol{\theta}$  in general. The problem of optimal design now becomes more difficult as the optimal design for inference about  $\boldsymbol{\theta}$  now depends on  $\boldsymbol{\theta}$  itself. Application of linear optimal design theory leads to *locally optimal designs*, that is, designs that are optimal for a given value of  $\boldsymbol{\theta}$ . A globally optimal design for nonlinear models has to proceed in a sequential fashion, computing a locally optimal design at the current estimate of  $\boldsymbol{\theta}$  to obtain a new measurement or measurements, and then updating the estimate and repeating the process until the criterion function or sequence of estimates is judged to converge; see Fedorov (1972, Sect. 4.4). Here, the role of prior information about  $\boldsymbol{\theta}$  is important, particularly for beginning the sequential process. If there is prior

information about the true value of  $\theta$ , then the sequential process can begin at that value. Such information may be present if the current experiment is a continuation of a previous experiment or if theoretical knowledge about the current or similar settings is available, and in either of these situations, the prior information may be encoded in a prior distribution on  $\theta$  in the Bayesian sense, from which an estimate of the true value of  $\theta$  can be obtained. In the absence of such prior information, “preliminary” observations should be performed using some nondegenerate design so that an estimate of  $\theta$  can be obtained from them, and then the sequential procedure described above can begin.

Closely related to this sequential approach is the Bayesian approach which puts a prior distribution  $\Pi$  on  $\theta$  and maximizes

$$\int \Psi(\mathbf{M}(\mu, \theta)) d\Pi(\theta) \quad (2.22)$$

rather than simply  $\Psi(\mathbf{M}(\mu, \hat{\theta}))$ , where  $\hat{\theta}$  is the current estimate of  $\theta$ . In order to produce Bayesian designs for clinical trials that control the chance of overdosing, [Haines et al. \(2003\)](#) propose to modify the Bayesian criterion (2.22) by including a penalty for high doses. That is, for scalar doses  $x$  and an unknown target dose  $x^*$  with prior distribution  $\rho$  induced by  $\Pi$ , the problem becomes to find the design measure  $\mu$  maximizing (2.22) subject to the constraint

$$P_{\mu, \rho}(x \geq x^*) = \int_{\mathcal{X}} \rho(\{x^* : x \geq x^*\}) d\mu(x) \leq \varepsilon,$$

for some small chosen value of  $\varepsilon > 0$ . For clinical trials in which patients are assigned doses sequentially, [Haines et al. \(2003\)](#) further extend their method by adding a sequential aspect by replacing the Bayesian information (2.22) by the sequential analog at the  $(k+1)$ st stage, given by finding the  $(k+1)$ st dose  $x_{k+1}$  maximizing

$$\int \Psi(\{kM(\mu_k, \theta) + M(\delta_{x_{k+1}}, \theta)\}/(k+1)) d\Pi_k(\theta) \quad \text{subject to}$$

$$P_{\rho_k}(x_{k+1} \geq x^*) = \rho_k(\{x^* : x_{k+1} \geq x^*\}) \leq \varepsilon,$$

where  $\mu_k$  is the empirical measure of the first  $k$  doses,  $\delta_x$  is the degenerate measure at  $x$ , and  $\Pi_k$  and  $\rho_k$  are the posterior distributions based on the first  $k$  doses and responses.

## 2.4 Phase I Clinical Trials for Relatively Benign Drugs

The primary objective of a Phase I clinical trial is to determine the dose and dosing regimen of a new drug and to collect information about drug-related side effects.

The secondary objective is to use the data collected to evaluate the effectiveness of the treatment. Before the Phase I trial, preclinical in vitro and animal studies are conducted to evaluate toxicity and the pharmacologic actions of the drug, thereby coming up with estimates of a good starting dose for Phase I trials with human subjects. Because of safety considerations for subjects in the trial, the drug is usually initiated at a low, safe dose and sequentially escalated to show safety at a level where some therapeutic response occurs. As noted in Sect. 2.2, the PK/PD models are nonlinear, and nonlinear design theory described in Sect. 2.3.3 is particularly suited for efficient estimation of the model parameters. On the other hand, the ultimate goal is not just estimation of these parameters per se, but to find a dose within the therapeutic window. For relatively benign drugs, Phase I trials involve healthy volunteers from whom intensive blood sampling is conducted over time. The next section describes a different paradigm for Phase I trials of cytotoxic treatments in cancer.

Although intersubject variability is seldom considered at the design stage of Phase I trials, such variability should be examined in the analysis of the data. Thus, while a nonlinear regression model of the type (2.1) with the same  $\theta$  for all subjects is assumed at the design stage, nonlinear mixed models of the type (2.4) with subject-specific  $\theta_i$  can be used to analyze the data. An example is given by Lai et al. (2006b), in which an orally administered cancer drug, temozolomide, was given to 65 adult patients with advanced cancer in four Phase I trials sponsored by the Schering–Plough Research Institute. Once such trial for treating patients who had advanced cancer that was refractory to standard forms of therapy was reported by Newlands et al. (1992). Each of these 65 patients had 10–15 drug concentration measurements from 10 min to 16 h after a single dose, and a total of 756 concentration measurements were collected. These concentrations were modeled by the one-compartment open model (2.11) to identify the influence of patient characteristics on the PK; the patient covariates forming the vector  $\mathbf{x}_i$  in the analysis were body surface area, gender, age, and creatinine clearance.

## 2.5 Early Phase Clinical Trials for Cytotoxic Cancer Treatments

### 2.5.1 Up-and-Down and Related Designs

Up-and-down designs are sequential (or cohort-by-cohort) designs for a discrete dose set in which the “next” dose is always equal or adjacent (the next higher or lower) to the current dose, hence the name “up-and-down.” The original idea is often credited to Dixon and Mood (1948), but an earlier paper by Wilson and Worcester (1943) proposed the idea for clinical uses. These designs have a wide range of applications such as for bioassays, explosives testing, metallurgy, and educational testing. In the dose-finding setting, they have the intuitive appeal of not making

large jumps within the dose space. Most up-and-down designs are *random walk rules*, sometimes called *first-order Markov procedures*, which choose the next dose based only on the most recent dose and observation. Because of this simplicity, the properties of random walk rules such as the limiting stationary distribution of the dose allocation and its speed of convergence can be obtained exactly using random walk theory.

*Example 2.2.* The biased coin design of [Durham and Flournoy \(1994\)](#) for estimating the  $p$ th quantile,  $0 < p \leq 1/2$ , of a response curve using available dose set

$$d_1 < d_2 < \cdots < d_L \quad (2.23)$$

utilizes a biased coin that lands heads with probability  $p/(1-p)$  and chooses the  $(k+1)$ st dose  $x_{k+1}$  as follows: If the  $k$ th dose and observed toxicity are  $x_k = d_\ell$  and  $y_k \in \{0, 1\}$ , respectively, then

$$x_{k+1} = \begin{cases} d_{(\ell-1) \vee 1} & \text{if } y_k = 1, \\ d_{(\ell+1) \wedge L} & \text{if } y_k = 0 \text{ and the coin lands heads,} \\ d_\ell & \text{if } y_k = 0 \text{ and the coin lands tails.} \end{cases}$$

[Durham and Flournoy \(1994\)](#) show that, if  $F(d) := P(y_k = 1 | x_k = d)$  is non-increasing in  $d$ , the limiting distribution of the dose allocation of this up-and-down rule is unimodal with mode essentially equal to the  $p$ th quantile of  $F(x)$ .

To derive the limiting distribution and to understand up-and-down designs more generally, describe an up-and-down design by its *transition probabilities*

$$p_{\ell,m} = P(x_{k+1} = d_m | x_k = d_\ell), \quad \ell, m \in \{1, \dots, L\},$$

which is the probability of stepping to the  $m$ th dose  $d_m$ , given that the current dose is  $d_\ell$ . For random walk rules in which dose levels are never skipped, we will have  $p_{\ell,m} = 0$  whenever  $|\ell - m| > 1$  and hence

$$p_{\ell,\ell-1} \mathbf{1}\{\ell > 1\} + p_{\ell,\ell} + p_{\ell,\ell+1} \mathbf{1}\{\ell < L\} = 1.$$

As a Markov chain, the random walk  $\{x_k\}$  has a tri-diagonal transition probability matrix  $\mathbf{P} = \{p_{\ell,m}\}_{\ell,m=1}^L$ . Given any initial treatment distribution

$$(P(x_1 = d_1), P(x_1 = d_2), \dots, P(x_1 = d_L))$$

and a  $\mathbf{P}$  such that any dose level in (2.23) can be eventually reached from any other, the limiting treatment distribution  $\pi_\ell = \lim_{k \rightarrow \infty} P(x_k = d_\ell)$ ,  $\ell = 1, \dots, L$ , can be found by solving  $L$  linear *balance equations*

$$\pi_\ell = \pi_{\ell-1} p_{\ell-1,\ell} \mathbf{1}\{\ell > 1\} + \pi_\ell p_{\ell,\ell} + \pi_{\ell+1} p_{\ell+1,\ell} \mathbf{1}\{\ell < L\}, \quad \ell = 1, \dots, L, \quad (2.24)$$

or equivalently,  $\mathbf{P}^T \boldsymbol{\pi} = \boldsymbol{\pi}$ , where  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_L)^T$ . The unique solution is

$$\pi_\ell \propto \prod_{j=\ell}^{L-1} \frac{p_{j+1,j}}{p_{j,j+1}}, \quad \ell = 1, \dots, L, \quad (2.25)$$

with the convention  $\prod_{j=L}^{L-1} = 1$ , and the proportionality constant in (2.25) is

$$\pi_L = \left( 1 + \sum_{\ell=1}^{L-1} \prod_{j=\ell}^{L-1} \frac{p_{j+1,j}}{p_{j,j+1}} \right)^{-1}. \quad (2.26)$$

The form (2.25) of the solution can be used to find the mode of the limiting distribution  $\boldsymbol{\pi}$  since it implies that  $\pi_\ell \geq \pi_{\ell-1}$  if and only if  $p_{\ell-1,\ell} \geq p_{\ell,\ell-1}$ . In particular, for [Durham and Flournoy's \(1994\)](#) biased coin design,

$$p_{\ell-1,\ell} = [1 - F(d_{\ell-1})]p / (1 - p) \quad \text{and} \quad p_{\ell,\ell-1} = F(d_\ell),$$

hence

$$\pi_\ell \geq \pi_{\ell-1} \iff \frac{F(d_\ell)}{1 - F(d_{\ell-1})} \leq \frac{p}{1 - p},$$

which shows that this design's limiting distribution has its mode at the discrete  $p$ th quantile of  $F(x)$ .

### 3+3 Designs

The widely used 3+3 design (see [Korn et al. 1994](#)) can be viewed as a truncated mixture of two up-and-down designs. There are many variations on the 3+3 design, but in its simplest form, the design begins at the lowest dose  $d_1$  and, treating patients in cohorts of 3, escalates to the next highest dose level if 0 of 3 experiences toxicity, stays at the same level if 1 of 3 experiences toxicity, and de-escalates or stops the trial if at least 2 of 3 experience toxicity. As pointed out earlier 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). In particular, the 3+3 design tends to not have the reliable convergence properties of random walk designs and has been widely criticized in dose-finding clinical trials, such as [Reiner et al. \(1999\)](#) who conclude that its “risk of choosing the incorrect level is large.”



## Stochastic Approximation

Another class of designs related to up-and-down designs consists of stochastic approximation procedures (Lai and Robbins 1979; Robbins and Monro 1951), one distinguishing feature being that dose selection under a stochastic approximation procedure will typically converge to a point, whereas random walk up-and-down design points converge to a distribution, as mentioned above. If  $F(x) = E(y|x)$  is the mean of the outcome  $y = y(x)$  at level (e.g., dose)  $x$ , then the goal of stochastic approximation is to produce a sequence  $\{x_n\}$  of estimates converging to the unique root  $x^*$  of the equation  $F(x) = y^*$ , for given  $y^*$ . Robbins and Monro (1951) introduced stochastic approximation procedures of the form

$$x_{n+1} = x_n - \frac{(y_n - y^*)}{nb}$$

for some constant  $b > 0$  and established that  $x_n \rightarrow x^*$  in probability under the assumption  $\sup_x E[y(x)^2] < \infty$ . Moreover, if  $b < 2F'(x^*)$ , then  $\sqrt{n}(x_n - x^*)$  converges to the  $N(0, \sigma^2/[b(2F'(x^*) - b)])$  distribution, where  $\sigma^2 = \lim_{x \rightarrow x^*} \text{Var}[y(x)]$ , and the choice of  $b$  is thus crucial to the performance of this stochastic approximation procedure (Sacks 1958). Since the optimal choice of  $b$  depends on the unknown slope  $F'(x^*)$ , Lai and Robbins (1979) proposed an adaptive stochastic approximation scheme in which  $b$  is replaced by an adaptively chosen sequence  $b_n$  that is strongly consistent for  $F'(x^*)$ . They also study the global cost

$$\sum_{n=1}^N (x_n - x^*)^2 \quad (2.27)$$

of the stochastic approximation sequence  $\{x_n\}_1^N$  and show that it is of order  $\sigma^2 \log N$  as long as  $b < 2F'(x^*)$ . Although this suggests that adaptive stochastic approximation may be a good choice to use in Phase I dose finding, its “out of the box” application to finite dose spaces and logistic regression models has been less than successful than model-based methods, since it is essentially nonparametric and the sample sizes of Phase I studies are typically small. For example, Bartroff and Lai (2010, 2011) have shown that myopic model-based methods perform considerably better than stochastic approximation in terms of “global” cost functions like (2.27) for  $N$  patients and that the performance can be further improved by utilizing approximate dynamic programming techniques, as will be discussed further in Sect. 3.8.

### 2.5.2 Model-Based Designs

Even though 3+3 designs and their variants are widely used in Phase I cancer trials, it has also been widely recognized as unsatisfactory on both ethical and efficiency grounds because it results in mostly subtherapeutic doses and inadequate information to estimate the MTD for a subsequent Phase II trial. To address this difficulty, [Eisenhauer et al. \(2000\)](#) suggest to use (a) methods to determine more informative starting doses, (b) pharmacokinetics-guided dose-escalation methods, and (c) model-based methods for dose determination, which are discussed next.

In model-based methods, a patient's response  $y$  to treatment at dose level  $x$  is usually modeled by a binary random variable taking values 0 or 1, such that  $y = 1$  indicates a DLT and whose distribution depends on  $x$  and an unknown vector  $\theta$  of parameters through the function

$$F_{\theta}(x) = P(y = 1 | \text{dose} = x).$$

We assume that  $F_{\theta}(x)$  is an increasing function of  $x$ , approaching 0 as  $x \rightarrow -\infty$  and 1 as  $x \rightarrow \infty$ . In a sequential trial with  $n$  patients, we assume that  $y_1, \dots, y_n$  are independent, except possibly through the choice of the dose levels  $x_1, \dots, x_n$ , since  $x_{k+1}$  will typically be chosen as a function of the previous doses and responses  $(x_1, y_1), \dots, (x_k, y_k)$ . As defined above, the MTD is then the  $p$ th quantile of  $F_{\theta}$ , that is,  $\text{MTD} = F_{\theta}^{-1}(p)$ . Because of its prevalence in the literature and for simplicity, here we take as our working model the two-parameter logistic regression model

$$F_{\theta}(x) = 1 / \left( 1 + e^{-(\alpha + \beta x)} \right) \quad (2.28)$$

where  $\theta = (\alpha, \beta)$ . For the two-parameter logistic model,  $\text{MTD} = [\log(p/(1-p)) - \alpha]/\beta$ . The methods that follow are not restricted to the model (2.28) and can be applied to other models such as the probit, gamma, and hyperbolic tangent models (see e.g., [O'Quigley et al. 1990](#)).

Noting that the nonparametric approach in stochastic approximation seems too ambitious for moderate sample sizes, [Wu \(1985\)](#) proposed to use a parametric modification of the stochastic approximation scheme in Sect. 2.5.1, taking  $x_{k+1}$  to be the  $p$ th quantile of  $F_{\hat{\theta}_k}$ , where  $\hat{\theta}_k$  is the MLE of  $\theta$  based on the doses and responses of the first  $k$  patients. [O'Quigley et al. \(1990\)](#) proposed a similar design but from a Bayesian point of view, called the CRM, that estimates the MTD at each stage by the posterior mean of  $\theta$  with respect to a chosen prior distribution. [O'Quigley \(2002\)](#) extends CRM to allow early stopping through the use of a sequential stopping rule.

[Babb et al. \(1998\)](#) pointed out that the CRM dose, being the mean of the MTD's posterior distribution, can be viewed as the Bayesian design with respect to squared error loss. That is, letting  $\mathcal{F}_k$  denote the information set generated by the first  $k$  doses and responses, that is, by  $(x_1, y_1), \dots, (x_k, y_k)$ , CRM chooses the  $(k+1)$ st dose  $x_{k+1}$  to be that minimizing  $E[h(x_{k+1}) | \mathcal{F}_k]$ , for

$$h(x) = (x - \text{MTD})^2. \quad (2.29)$$

Babb et al. (1998) suggested that the symmetric nature of the squared error loss or its close relative, the absolute error loss, may not be appropriate for modeling the toxic response to a cancer treatment and proposed the “escalation with overdose control” (EWOC) method, which is a Bayesian design with respect to the asymmetric loss function

$$h(x) = \begin{cases} \omega(\text{MTD} - x) & \text{if } x \leq \text{MTD} \\ (1 - \omega)(x - \text{MTD}) & \text{if } x \geq \text{MTD} \end{cases} \quad (2.30)$$

where the chosen constant  $0 < \omega < 1/2$  is the so-called *feasibility bound*. Note that this loss function penalizes an overdose  $x = \text{MTD} + \delta$  more than an underdose  $x = \text{MTD} - \delta$  of the same magnitude  $\delta > 0$ . EWOC can be shown to be equivalent to estimating the MTD at each stage by the  $\omega$ th quantile of the posterior distribution of the MTD. In the examples in Babb et al. (1998),  $\omega$  is chosen to be slightly less than  $p$ .

Whereas the step-up/down design in traditional Phase I cancer trials focuses on the safety of patients in the study at the expense of being inefficient for the posttrial estimate of the MTD, there has also been much work on  $c$ - and  $D$ -optimal experimental designs for such estimation from binary responses. Haines et al. (2003) proposed sequential Bayesian  $c$ - and  $D$ -optimal designs, subject to a prescribed upper-bound  $\varepsilon$  on the probability of doses exceeding the MTD, as described in the last paragraph of Sect. 2.3.3.

Despite their shortcomings and the development of alternative Bayesian approaches since 1990, conventional dose-escalation designs are still widely used in Phase I cancer trials because of the ethical issue of safe treatment of patients currently in the trial. However, a Phase I design also has the goal of determining the MTD for a future Phase I cancer trial, and needs an informative experimental design to meet this goal. Von Hoff and Turner (1991) have documented that the overall response rates in Phase I trials are low and that substantial numbers of patients are treated at doses that are retrospectively found to be nontherapeutic. Eisenhauer et al. (2000, p. 685) have pointed out that “with a plethora of molecularly defined antitumor targets and an increasingly clear description of tumor biology, there are now more antitumor candidate therapies requiring Phase I study than ever” and that “unless more efficient approaches are undertaken, Phase I trials may be a rate-limiting step in the process of evaluation of novel anticancer agents.” The hybrid designs of Bartroff and Lai (2010) that will be described in Sect. 3.8 were motivated by developing one such “more efficient” approach.

## 2.6 Supplements and Problems

1. *Asymptotic theory of nonlinear least squares and Levenberg–Marquardt shrinkage.*

Let  $\hat{\theta}$  be the least squares estimate of  $\theta$  in the nonlinear regression model (2.1).

Let  $\theta_0$  denote the true value of  $\theta$ . Assuming  $w_j \equiv 1$  in (2.2), we have

$$E[S(\boldsymbol{\theta})] = \sum_{t=1}^n [f(\boldsymbol{\theta}, \mathbf{x}_t) - f(\boldsymbol{\theta}_0, \mathbf{x}_t)]^2 + n\sigma^2, \quad (2.31)$$

recalling that  $E(\varepsilon_t) = 0$  and  $\text{Var}(\varepsilon_t) = \sigma^2$ . Therefore,

$$E[S(\boldsymbol{\theta})] - n\sigma^2 \begin{cases} = 0 & \text{if } \boldsymbol{\theta} = \boldsymbol{\theta}_0 \\ \rightarrow \infty & \text{if } \boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \end{cases} \quad (2.32)$$

under the assumption

$$\sum_{t=1}^{\infty} [f_t(\boldsymbol{\theta}) - f_t(\boldsymbol{\theta}_0)]^2 = \infty \text{ for } \boldsymbol{\theta} \neq \boldsymbol{\theta}_0, \quad (2.33)$$

where  $f_t(\boldsymbol{\theta}) = f(\boldsymbol{\theta}, \mathbf{x}_t)$ . In the linear case  $f_t(\boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{x}_t$ , (2.33) is equivalent to the convergence of  $(\sum_{t=1}^n \mathbf{x}_t \mathbf{x}_t^T)^{-1}$  to  $\mathbf{0}$ . Since  $\hat{\boldsymbol{\theta}}$  is the minimizer of  $S(\boldsymbol{\theta})$ , (2.32) suggests that  $\hat{\boldsymbol{\theta}}$  is consistent. A rigorous proof involves considering  $S(\boldsymbol{\theta})$  as a random function of  $\boldsymbol{\theta}$  and requires additional assumptions.

Consistency of  $\hat{\boldsymbol{\theta}}$  leads easily to its asymptotic normality since we can approximate  $f_t(\hat{\boldsymbol{\theta}})$  by  $f_t(\boldsymbol{\theta}_0) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \nabla f_t(\boldsymbol{\theta}_0)$  when  $\hat{\boldsymbol{\theta}}$  is near  $\boldsymbol{\theta}_0$ , assuming that  $\nabla_t f(\boldsymbol{\theta})$  is uniformly continuous in  $t$  and  $\boldsymbol{\theta}$  belonging to some neighborhood of  $\boldsymbol{\theta}_0$ . The asymptotic properties of  $\hat{\boldsymbol{\theta}}$  are therefore the same as those of ordinary least squares (OLS):

$$\hat{\boldsymbol{\theta}} \approx N \left( \boldsymbol{\theta}_0, \sigma^2 \left( \sum_{t=1}^n \hat{\mathbf{x}}_t \hat{\mathbf{x}}_t^T \right)^{-1} \right), \quad (2.34)$$

where  $\hat{\mathbf{x}}_t = \nabla f_t(\hat{\boldsymbol{\theta}})$ . Moreover,  $\sigma^2$  can be consistently estimated by

$$\hat{\sigma}^2 = \sum_{t=1}^n \left( y_t - f_t(\hat{\boldsymbol{\theta}}) \right)^2 / n. \quad (2.35)$$

For smooth real-valued functions  $g(\boldsymbol{\theta}_0)$ , we apply the Taylor expansion  $g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta}_0) \doteq (\nabla g(\boldsymbol{\theta}_0))^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$  to approximate  $g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta}_0)$  by a linear function, providing the asymptotic normality of  $g(\hat{\boldsymbol{\theta}})$  with mean  $g(\boldsymbol{\theta}_0)$  and covariance matrix

$$\sigma^2 (\nabla g(\boldsymbol{\theta}_0))^T \left( \sum_{t=1}^n \hat{\mathbf{x}}_t \hat{\mathbf{x}}_t^T \right)^{-1} (\nabla g(\boldsymbol{\theta}_0)). \quad (2.36)$$

The square root of (2.36) also gives the estimated standard error for  $g(\hat{\boldsymbol{\theta}})$  if we replace the unknown  $\sigma$  and  $\boldsymbol{\theta}_0$  in (2.36) by  $\hat{\sigma}$  and  $\hat{\boldsymbol{\theta}}$ . The adequacy of this normal approximation to construct confidence intervals for  $g(\boldsymbol{\theta}_0)$  is questionable for highly nonlinear  $g$ , as the one-term Taylor expansion can be quite poor. An alternative to the asymptotic approximations is the *bootstrap method*, which uses Monte Carlo simulations to obtain standard errors and confidence intervals.

The nonlinear least squares procedure is implemented by many numerical software packages. The following are functions in R: `nls.lm`, `nls`. Since the Gauss–Newton scheme is aborted whenever  $\sum_{t=1}^n \hat{\mathbf{x}}_t^{(k)} \hat{\mathbf{x}}_t^{(k)T}$  is singular or nearly singular, where  $\hat{\mathbf{x}}_t^{(k)} = \nabla f_t(\hat{\boldsymbol{\theta}}^{(k)})$  and  $\hat{\boldsymbol{\theta}}^{(k)}$  is defined in Sect. 2.1.1. It is desirable to avoid such difficulties in matrix inversion. This has led to the modification that replaces  $\sum_{t=1}^n \hat{\mathbf{x}}_t^{(k)} \hat{\mathbf{x}}_t^{(k)T}$  by  $\sum_{t=1}^n \hat{\mathbf{x}}_t^{(k)} \hat{\mathbf{x}}_t^{(k)T} + \kappa \mathbf{D}$  for the OLS estimate in the  $k$ th iteration of the Gauss–Newton algorithm, corresponding to using shrinkage as in ridge regression. Here  $\mathbf{D}$  is a diagonal matrix whose diagonal elements are the same as those of  $\sum_{t=1}^n \hat{\mathbf{x}}_t^{(k)} \hat{\mathbf{x}}_t^{(k)T}$ , proposed by Marquardt as a refinement of an earlier proposal  $\mathbf{D} = \mathbf{I}$  by Levenberg.

## 2. Generalized linear mixed models.

The NONMEM in Sect. 2.1.2 have their counterparts for generalized linear models. These are called *generalized linear mixed models* (GLMM) and were introduced by Breslow and Clayton (1993) for longitudinal data  $Y_{it}$  to enhance generalized linear models by allowing subject-specific regression parameters  $\mathbf{b}_i$ , called “random effects,” thereby extending mixed effects models in linear regression to GLMM. The GLMM assumes the  $y_{it}$  to be conditionally independent given the observed covariates  $\mathbf{x}_{it}$  and  $\mathbf{z}_{it}$  and such that  $y_{it}$  has a conditional density of the form

$$f(y|\mathbf{b}_i, \mathbf{z}_{it}, \mathbf{x}_{it}) = \exp \{ [y\theta_{it} - \psi(\theta_{it})] / \sigma + c(y, \sigma) \}, \quad (2.37)$$

in which  $\sigma$  is a dispersion parameter and  $\mu_{it} = d\psi/d\theta|_{\theta=\theta_{it}}$  satisfies

$$\mu_{it} = g^{-1} \left( \boldsymbol{\beta}^T \mathbf{x}_{it} + \mathbf{b}_i^T \mathbf{z}_{it} \right), \quad (2.38)$$

where  $g^{-1}$  is the inverse of a monotone link function  $g$ , as in the standard generalized linear models for which  $\mu_{it} = g^{-1}(\boldsymbol{\beta}^T \mathbf{x}_{it})$ . The case  $g = d\psi/d\theta$  is called the “canonical link.” The random effects  $\mathbf{b}_i$  can contain an intercept term  $a_i$  by augmenting the covariate vector to  $(1, \mathbf{z}_{it})$  in case  $a_i$  is not included in  $\mathbf{b}_i$ ;  $\boldsymbol{\beta}$  is a vector of fixed effects and can likewise contain an intercept term. The density function (2.37) with  $\sigma = 1$  is that of an exponential family, which includes the Bernoulli and normal distributions as special cases. Breslow and Clayton assume the  $\mathbf{b}_i$  in (2.38) to have a common normal distribution with mean 0 and covariance matrix  $\boldsymbol{\Sigma}$  that depends on an unknown parameter vector  $\boldsymbol{\alpha}$ .

The likelihood function of the GLMM defined by (2.37) and (2.38) is of the form  $\prod_{i=1}^n L_i(\sigma, \alpha, \beta)$ , where

$$L_i(\sigma, \alpha, \beta) = \int \left\{ \prod_{t=1}^T f(y_{it}; \theta_{it}, \sigma) \right\} \phi_{\alpha}(\mathbf{b}) d\mathbf{b}, \quad (2.39)$$

in which  $\phi_{\alpha}$  denotes the normal density function with mean 0 and covariance matrix depending on an unknown parameter  $\alpha$ . Analogous to NONMEM described in Sect. 2.1.2, there are three methods to compute the likelihood function, the maximizer of which gives the MLE of  $\sigma$ ,  $\alpha$ , and  $\beta$ :

- (a) *Laplace's approximation.* Letting  $e^{l_i(\mathbf{b}|\sigma, \alpha, \beta)}$  be the integrand in the right-hand side of (2.39), Laplace's asymptotic formula for integrals yields the approximation

$$\int e^{l_i(\mathbf{b}|\sigma, \alpha, \beta)} d\mathbf{b} \approx (2\pi)^{q/2} \left\{ \det \left[ -\ddot{l}_i(\hat{\mathbf{b}}_i|\sigma, \alpha, \beta) \right] \right\}^{-1/2} \exp \left\{ l_i(\hat{\mathbf{b}}_i|\sigma, \alpha, \beta) \right\}, \quad (2.40)$$

where  $q$  is the dimension of  $\mathbf{b}_i$ ,  $\hat{\mathbf{b}}_i = \hat{\mathbf{b}}_i(\sigma, \alpha, \beta)$  is the maximizer of  $l_i(\mathbf{b}|\sigma, \alpha, \beta)$  and  $\ddot{l}_i$  denotes the Hessian matrix consisting of second partial derivatives of  $l_i$  with respect to the components of  $\mathbf{b}$ . The R package `lme4` computes the MLE by using the Laplace approximation (2.40) or the restricted pseudo-likelihood approach proposed by [Wolfinger and O'Connell \(1993\)](#), as the user-specified option.

- (b) *Gauss–Hermite quadrature.* Laplace's asymptotic formula (2.40) is derived from the asymptotic approximation of  $l_i$  by a quadratic function of  $\mathbf{b}$  in a small neighborhood of  $\hat{\mathbf{b}}_i$  as  $\lambda_{\min}(-\ddot{l}_i(\hat{\mathbf{b}}_i|\sigma, \alpha, \beta)) \rightarrow \infty$ , where  $\lambda_{\min}(\cdot)$  denotes the minimum eigenvalue of a symmetric matrix. Therefore, such formula may produce significant approximation error for  $L_i$  if the corresponding  $\lambda_{\min}(-\ddot{l}_i(\hat{\mathbf{b}}_i|\sigma, \alpha, \beta))$  is not sufficiently large. One way to reduce the possible approximation error is to compute  $L_i$  by using an adaptive Gauss–Hermite quadrature rule, as in [Liu and Pierce \(1994\)](#). The software package SAS uses adaptive Gauss–Hermite quadrature in the *NLMIXED* procedure to compute (2.39); the R package `lmer()` also uses Gaussian quadrature to compute (2.39) but only for certain special cases of the exponential family (2.37). The numerical integration procedures demand a much higher computational effort and become computationally infeasible when  $n$  or  $q$  is large. To circumvent the issue of high-dimensional numerical integration, some authors propose putting prior distributions on the unknown parameters and estimate them by the Markov chain Monte Carlo (MCMC) method in a Bayesian way; see [Berry et al. \(2011\)](#) for logistic mixed models. The performance of the MCMC method, however, depends on how the prior

parameters are set as well as the convergence rate of the Markov chain to its stationary distribution, which may not even exist. Yafune et al. (1998) use direct Monte Carlo integration but point out that the computational time may be too long to be of practical interest.

- (c) *Hybrid method.* This is basically the same as the hybrid method for NONMEM, as pointed out by Lai and Shih (2003b, Sect. 5).

### 3. *Dose individualization and population PK/PD.*

Several physiologic (e.g., maturation of organs in infants) and pathologic (e.g., kidney failure, heart failure) processes require dosage adjustments in individual patients to modify specific PK parameters. Two basic parameters in this connection are *clearance* (a measure of the ability of the body to eliminate the drug) and *volume of distribution* (a measure of the apparent space in the body available to contain the drug). Drug clearance principles are similar to clearance concepts in renal physiology, in which creatinine or urea clearance is defined as the rate of elimination of the compound in the urine relative to the plasma concentration. Thus, clearance CL of a drug is the rate of elimination by all routes relative to the concentration  $C$  of the drug in a biologic fluid; it is perhaps the most important PK parameter to be considered in defining a rational drug dosage regimen. In most cases, the clinician would like to maintain steady-state drug concentrations  $C_{ss}$  within a known therapeutic window. Steady state will be achieved when the dosing rate (rate of active drug entering the systemic circulation) equals the rate of drug elimination. Therefore,

$$\text{Dosing rate} = \text{CL} \times C_{ss}.$$

The two major sites of drug elimination are the kidneys and the liver. Clearance of unchanged drug in the urine represents renal clearance. Within the liver, drug elimination occurs via biotransformation of the drug to one or more metabolites, or excretion of unchanged drug into the bile, or both. When no other organs are involved in elimination of the drug,  $\text{CL} = \text{CL}_{\text{renal}} + \text{CL}_{\text{liver}}$  since the liver and kidneys work in parallel. The volume of distribution ( $V$ ) is defined as

$$V = \text{Amount of drug in body} / C,$$

where  $C$  is the concentration of the drug in blood or plasma, depending on the fluid measured. It reflects the apparent space available in both the general circulation and the tissue of distribution. It does not represent a real volume but should be regarded as the size of the pool of blood fluids that would be required if the drug were distributed equally throughout all parts of the body. From mass balance and steady-state considerations,  $V$  is related to clearance via  $\text{CL} = k_e V$ , where  $k_e$  is the elimination rate constant. Note that both  $V$  and the elimination rate  $k_e$  appear in the one-compartment open model (2.11). Allowing these parameters and the absorption rate  $k_a$  in (2.11) to be subject-specific leads to a NONMEM that is used in the second paragraph of Sect. 2.4.

Dose individualization is a major practical goal of population PK. Since the efficacy and toxicity of a drug are directly related to drug concentrations at a target site, which are generally not available but for which blood concentrations are often good surrogates, criteria for determining the dose and dosing regimen for a specific subject often involve functions of the subject's concentrations or functions of the subject's parameter vector  $\boldsymbol{\theta} = g(\mathbf{x}, \boldsymbol{\beta}) + \mathbf{b}$  in (2.4). The subject's blood samples are often too sparse to provide an adequate estimate of  $\boldsymbol{\theta}$ . The empirical Bayes approach considered by Lai et al. (2006b) borrows information from healthy volunteers in Phase I studies who have undergone intensive blood sampling and also from clinical patients for whom intensive blood sampling is not feasible. Combining an individual patient's characteristics (as measured by  $\mathbf{x}$ ) and sparse concentration data with a large database from other subjects is one of the main motivations for building population PK models. Making use of the hybrid method, the last two paragraphs of Sect. 2.2 discuss how empirical Bayes estimates of  $h(\boldsymbol{\theta})$  can be computed from (a) the patient's data and (b) the population model fitted from other subjects' data.

An emerging trend in cancer therapeutics is to use biomarkers to personalize the treatment and treatment strategy for cancer patients; see Lai et al. (2012b). Personalization (or individualization) of treatments again falls in the domain of nonlinear/generalized LME models. Biomarker-guided personalized therapies for cancer require innovations in design and analysis of early phase and Phase III confirmatory clinical trials and may eventually lead to major breakthroughs in the methodology.

4. In the model  $Y = \alpha + \beta x + \varepsilon$ , where  $\varepsilon \sim N(0, \sigma^2)$  and  $-1 \leq x \leq 1$ , sketch the convex hull  $\mathcal{S}$  and use Elfving's method to find the optimal design for estimating (a) the slope  $\beta$  and (b) the mean response  $\alpha + \beta x_0$  at  $x = x_0$ , for arbitrary  $-1 \leq x_0 \leq 1$ .
5. In the setting of the example in Sect. 2.3.2, fix a value of  $a > 0$  and compute the value of

$$\frac{\mathbf{c}^T \mathbf{M}(\tilde{\boldsymbol{\mu}})^{-1} \mathbf{c}}{\mathbf{c}^T \mathbf{M}(\boldsymbol{\mu}^*)^{-1} \mathbf{c}} \quad \text{for } x_0 = i \cdot a/5, i = 1, \dots, 5,$$

where  $\tilde{\boldsymbol{\mu}}$  is the design putting weight 1/3 at each of the points  $x = 0, a/2$ , and  $a$ , and  $\boldsymbol{\mu}^*$  is the  $\mathbf{c}$ -optimal design found in the example.

6. Verify (2.18) and (2.19), and show that  $p$  given by (2.19) is in  $[0, 1]$  for arbitrary  $a > 0$  and all  $0 < x_0 \leq \gamma a$ .
7. In the example in Sect. 2.3.2, find the  $\mathbf{c}$ -optimal design if  $x_0$  is allowed to exceed  $a$ .
8. Find the Fisher information matrix

$$E \left[ -\frac{\partial^2 p(Y|\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right]$$

for the logistic regression model



$$P(Y = 1|\mathbf{x}, \boldsymbol{\theta}) = 1/(1 + e^{-(\alpha + \beta x)}), \quad P(Y = 0|\mathbf{x}, \boldsymbol{\theta}) = 1 - P(Y = 1|\mathbf{x}, \boldsymbol{\theta}),$$

where  $\mathbf{x} = (1, x)^T$  and  $\boldsymbol{\theta} = (\alpha, \beta)^T$ .

9. Making use of the asymptotic theory of nonlinear least squares described in (2.34) and (2.36), explain how optimal linear design theory can be used to construct locally optimal designs in nonlinear regression models.
10. *Discrete dose levels in Phase I cancer trials.*

As pointed out in Sect. 2.5, because of the traditional practice of using up-and-down designs in Phase I cancer trials, the dose levels in dose-finding studies of cancer drugs are usually chosen before the trial as a finite set  $\Lambda = \{\lambda_1, \dots, \lambda_d\}$  of possible doses, where  $\lambda_1 < \lambda_2 < \dots < \lambda_d$ , unlike the continuous doses we have assumed in Sect. 2.5.2. In this case the MTD has to be redefined as

$$\eta = \begin{cases} \max\{\lambda \in \Lambda : F_{\boldsymbol{\theta}}(\lambda) \leq p\} & \text{if } F_{\boldsymbol{\theta}}(\lambda_i) \leq p \text{ for some } i, \\ \lambda_1 & \text{otherwise.} \end{cases} \quad (2.41)$$

In many dose-finding trials, the number of discrete dose levels is relatively small, so one can use more robust order-restricted models of toxicity versus dose than the logistic regression model (2.28). Yin and Yuan (2009) have proposed a Bayesian model averaging design based on the monotone dose-toxicity relationship.

Sequential Experimentation in Clinical Trials  
Design and Analysis

Bartroff, J.; Lai, T.L.; Shih, M.-C.

2013, XVI, 240 p.,

ISBN: 978-1-4614-6114-2