Title
Designing Studies for Dose Response

Authors
Weng-Kee Wong
Peter A. Lachenbruch

Publication Date
2011-10-24

STATISTICS IN MEDICINE, VOL. 15, 343–359 (1996)

TUTORIAL IN BIOSTATISTICS
DESIGNING STUDIES FOR DOSE RESPONSE

WENG KEE WONG

UCLA Department of Biostatistics, Los Angeles, CA 90024-1772, U.S.A.

AND

PETER A. LACHENBRUCH

FDA/CBER/OELPS HFM-215, 1401 Rockville Pike, Rockville, MD 20852, U.S.A.

SUMMARY

'Dose response' refers to the regression of a response on a stimulus. We review a number of options for dose–response designs, and compare various designs which may be used in practice. We start with two group designs. Next, we introduce basic optimal approximate design theory for simple linear and quadratic regression illustrating different criteria of optimality and their effect on the allocation of the levels of the dose. Then we obtain the efficiencies of these optimal approximate designs and some simple designs which have intuitive appeal (symmetry, equal spacing of treatments, reduced numbers of observations at the highest and lowest doses).

1. INTRODUCTION

The regression of response on stimulus may be represented graphically as a curve [as in Figure 1]. When the stimulus is in the form of a 'dose' (e.g., of a drug, or possibly of an applied force or some other source), this may be called a 'dose response curve'. (Kotz and Johnson[1]). In its simplest form, a dose–response curve is a simple linear or polynomial regression. More complex 'dose–response' curves may involve, for example, a transcendental function. Others may involve transformations of the dose in the regressions. For example, dose–response models often use the logarithm of dose. This function of the dose is called the dose metameter. In some cases, the response is quantal (yes/no) and the dose–response technique is a probit analysis or logit analysis. Figure 1 shows dose–response curves for linear and quadratic models.

The determination of a threshold dose is also a dose–response problem (see Figure 1). Here the response is A below a dose x_0 and B above that point. That is, the model is $E(Y|x) = A$ if $x < x_0$ and $E(Y|x) = B$ if $x \geqslant x_0$ where $E(Y|x)$ is the expected value of the response Y given $X = x$. The model requires that we estimate the value of x_0, A and B. In this tutorial paper we will consider the dose–response cases in which response and the dose are continuous and the regression functions are either simple linear or quadratic models, that is

$$E(Y|x) = A + Bx$$

or

$$E(Y|x) = A + Bx + Cx^2.$$

We also assume throughout that x is coded so that $0 \leqslant x \leqslant 1$.

CCC 0277–6715/96/040343–17

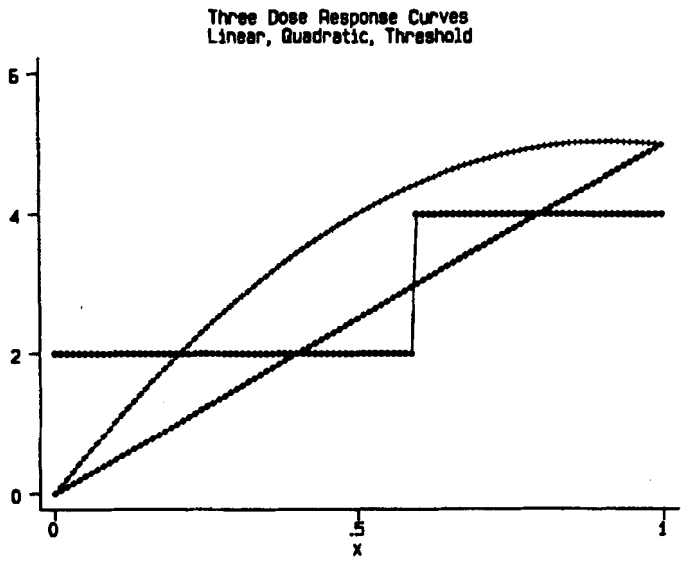


Figure 1. Three dose response curves: linear, quadratic and threshold

The steps in conducting a dose–response study (in idealized form) consist of the following:

(a) Assume a form or model for the curve (for example, linear, quadratic or threshold).
(b) Select the dose metameter (dose or log (dose)).
(c) Design the study so 'good' information is obtained; this includes obtaining estimates of the model coefficients with small standard errors and having the ability to test for model failures (such as testing for a quadratic model when a linear one has been assumed or testing for non-normal errors).
(d) Collect the data.
(e) Perform the analyses. For the simplest models, these include a linear regression, followed by model diagnostics such as testing for common variance (homoscedasticity), normality, and whether the model is linear or quadratic, and examining for outliers.
(f) Prepare a report describing the steps in the study, including the limitations of the study.

The objective of this paper is to discuss design issues in a dose–response experiment. Specifically, we consider the problem of allocating the dose, x, in $[0, 1]$ to estimate $E(Y|x)$. We motivate these issues by simulating data from three designs and two dose–response functions. Each data set has 20 observations with a standard deviation of 2. In all cases, the data are normally distributed. The goal is to estimate the dose–response function, which may be linear or quadratic. The first design allocates half of the data at $x = 0$ and half at $x = 1$. The second design has one-quarter of the observations at four points, $x = 0$, $x = 1/3$, $x = 2/3$ and $x = 1$. These are examples of *uniform* designs which have equal spacing of the x values, and equal number of observations at each x. The third design has half of the data at $x = 0$ and half at $x = 0{\cdot}75$. This design might be used if the investigator were concerned over possible toxic effects of the highest dose (at $x = 1$). The allocation of data is shown as histograms in Figure 2. The first response function is $E(Y|x) = 5x$. The second response function is $E(Y|x) = 11x - 6x^2$. This function was chosen to reach a maximum inside the interval $[0, 1]$. It also agrees with the linear function at 0 and 1. The data are given

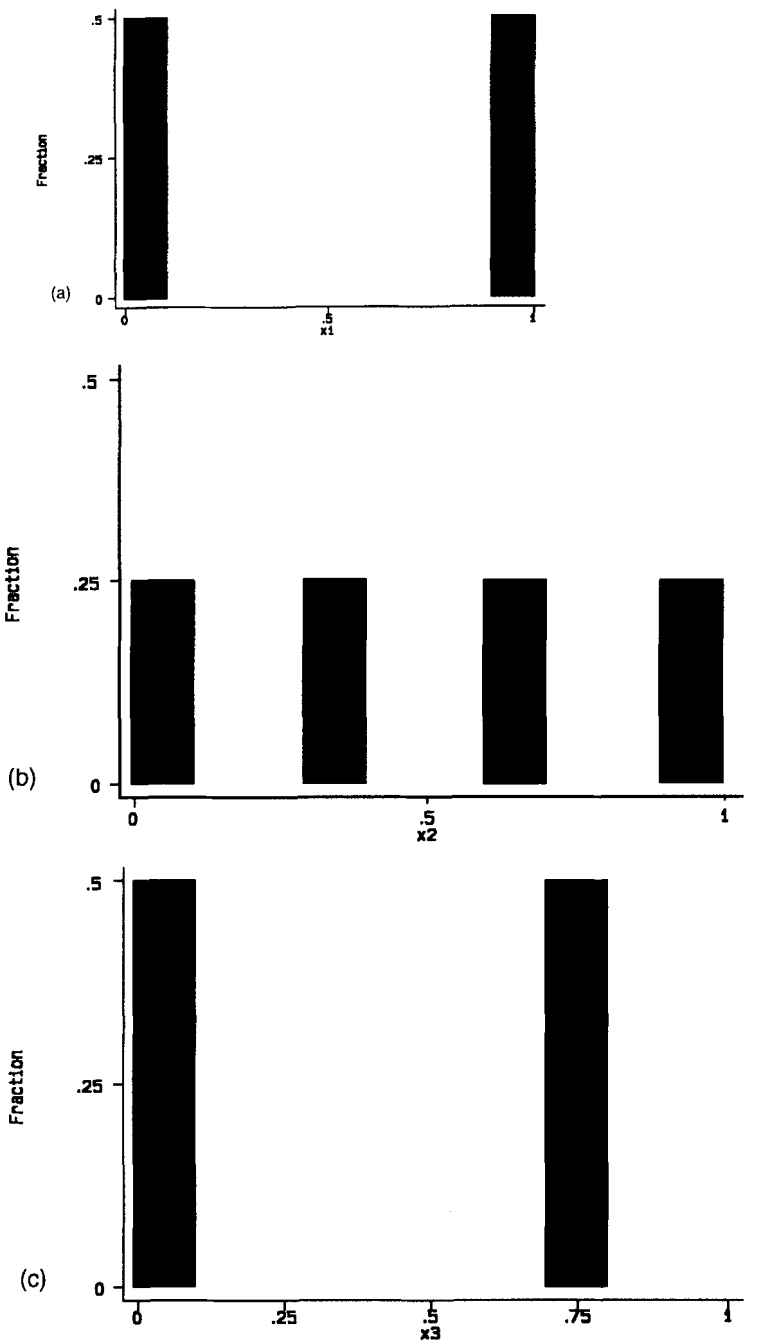


Figure 2. Histogram of X values for three designs

in Table I(a) and I(b) rounded by 5 digits (your analyses with these data should be similar, but not identical to ours). To distinguish the data of Table I(a) from that of Table Ib) we label the simulated dependent variable data in Table Ib) as W.

Table I. Data and analysis from three simulated dose–response studies

(*a*) Data generated by $Y_i = 5x_i + \varepsilon$, where ε is normal with $\mu = 0$ and $\sigma = 2$

	Design 1		Design 2		Design 3	
	x_1	y_1	x_2	y_2	x_3	y_3
1	0	− 3·69894	0	− 1·77029	0	− 3·95725
2	0	0·88581	0	1·04965	0	1·42738
3	0	4·24483	0	− 0·94450	0	0·45690
4	0	0·66185	0	0·99879	0	4·77256
5	0	− 1·37427	0	2·35421	0	− 1·48124
6	0	2·54575	0·33	4·84129	0	1·40287
7	0	− 0·84730	0·33	2·03856	0	1·28237
8	0	− 3·48385	0·33	0·47213	0	0·86857
9	0	− 0·97389	0·33	4·40601	0	− 0·98921
10	0	1·55282	0·33	0·29701	0	1·68386
11	1	8·21102	0·67	2·39918	0·75	1·52494
12	1	2·27225	0·67	5·94551	0·75	0·65396
13	1	7·51096	0·67	3·13616	0·75	8·59038
14	1	6·38050	0·67	7·11803	0·75	2·62081
15	1	6·58215	0·67	− 0·64908	0·75	5·79675
16	1	2·86575	1	1·60837	0·75	7·02660
17	1	7·94210	1	4·43520	0·75	4·30115
18	1	1·05613	1	3·62980	0·75	5·18883
19	1	8·61141	1	3·16162	0·75	4·42332
20	1	3·90077	1	3·44190	0·75	5·01722

(*b*) Data generated by $W_i = 11x - 6x^2 + \varepsilon$, where ε is normal with $\mu = 0$ and $\sigma = 2$

	Design 1		Design 2		Design 3	
	x_1	w_1	x_2	w_2	x_3	w_3
1	0	− 1·77029	0	− 3·69894	0	− 0·92930
2	0	1·04965	0	0·88581	0	− 0·36553
3	0	− 0·09445	0	4·24483	0	− 0·27144
4	0	0·99879	0	0·66185	0	− 1·70517
5	0	2·35421	0	− 1·37427	0	1·43117
6	0	3·19129	0·33	5·52235	0	− 2·25067
7	0	0·38856	0·33	2·12930	0	− 0·53015
8	0	− 1·17787	0·33	− 0·50725	0	− 1·74972
9	0	2·75601	0·33	2·00270	0	− 2·18116
10	0	− 1·35299	0·33	4·52942	0	3·14768
11	1	4·04918	0·67	7·88762	0·75	5·37605
12	1	7·59551	0·67	1·94885	0·75	2·13208
13	1	4·78616	0·67	7·18757	0·75	7·11342
14	1	8·76803	0·67	6·05710	0·75	5·91360
15	1	1·00092	0·67	6·25875	0·75	5·88210
16	1	1·60837	1	2·86575	0·75	7·22348
17	1	4·35200	1	7·94210	0·75	4·20166
18	1	3·62979	1	1·05613	0·75	1·18261
19	1	3·16162	1	8·61142	0·75	1·47634
20	1	3·44190	1	3·90078	0·75	4·61303

Table I. (Continued)

Correct model: $E(Y|x) = 5x$

(*c*) Regression analysis of Y_1 on X_1 (correct model)

Source	SS	d.f.	MS
Model	155·794907	1	155·794907
Residual	126·120725	18	7·00670695
Total	281·915632	19	14·8376649

Number of obs = 20
$F(1, 18) = 22·24$
Prob > $F = 0·0002$
R-squared = 0·5526
Adj R-squared = 0·5278
Root MSE = 2·647

y_1	Coefficient	Standard error	t	$P > \|t\|$	[95% confidence interval]	
x_1	5·582023	1·183783	4·715	0·000	3·094988	8·069058
_cons	− 0·048719	0·8370607	− 0·058	0·954	− 1·807318	1·70988

(*d*) Regression analysis of Y_2 on X_2 (correct model – Less efficient allocation of X's)

Source	SS	d.f.	MS
Model	24·6805628	1	24·6805628
Residual	78·5551996	18	4·36417775
Total	103·235762	19	5·43346117

Number of obs = 20
$F(1, 18) = 5·66$
Prob > $F = 0·0287$
R-squared = 0·2391
Adj R-squared = 0·1968
Root MSE = 2·0891

y_2	Coefficient	Standard error	t	$P > \|t\|$	[95% confidence interval]	
x_2	2·974769	1·250913	2·378	0·029	0·3466994	5·602839
_cons	0·9110929	0·7806436	1·167	0·258	− 0·7289785	2·551164

(*e*) Regression analysis of Y_3 on X_3 (correct model – design less efficient than given by X_1 [(*c*) analysis])

Source	SS	d.f.	MS
Model	78·713809	1	78·713809
Residual	100·825839	18	5·6014355
Total	179·539648	19	9·44945516

Number of obs = 20
$F(1, 18) = 14·05$
Prob > $F = 0·0015$
R-squared = 0·4384
Adj R-squared = 0·4072
Root MSE = 2·3667

y_3	Coefficient	Standard error	t	$P > \|t\|$	[95% confidence interval]	
x_3	5·290287	1·411248	3·749	0·001	2·325364	8·255209
_cons	0·546681	0·7484274	0·730	0·475	− 1·025707	2·119069

Table I. (Continued)

Correct model: $E(Y|x) = 11x - 6x^2$

(f) Regression analysis of W_1 on X_1 and X_1^2 (correct model, but design prevents estimating quadratic terms)

Source	SS	d.f.	MS
Model	64·9821823	1	64·9821823
Residual	79·6091957	18	4·42273309
Total	144·591378	19	7·61007252

Number of obs = 20
$F(1, 18) = 14·69$
Prob > $F = 0·0012$
R-squared = 0·4494
Adj R-squared = 0·4188
Root MSE = 2·103

w_1	Coefficient	Standard error	t	$P > \|t\|$	[95% confidence interval]	
x_1	3·605057	0·9405034	3·833	0·001	1·629133	5·580981
X_1^2	(dropped)					
_cons	0·634291	0·6650363	0·954	0·353	− 0·7628985	2·03148

(g) Regression analysis of W_2 on X_2 and X_2^2 (correct model)

Source	SS	d.f.	MS
Model	91·3522269	2	45·6761134
Residual	126·624177	17	7·44848097
Total	217·976403	19	11·4724423

Number of obs = 20
$F(1, 17) = 6·13$
Prob > $F = 0·0099$
R-squared = 0·4191
Adj R-squared = 0·3508
Root MSE = 2·7292

w_2	Coefficient	Standard error	t	$P > \|t\|$	[95% confidence interval]	
x_2	13·30121	5·757084	2·310	0·034	1·154823	25·44759
X_2^2	− 8·105359	5·520267	− 1·468	0·160	− 19·7521	3·541387
_cons	− 0·0883785	1·188492	− 0·074	0·942	− 2·595878	2·419121

(h) Regression analysis of W_3 on X_3 and X_3^2 (correct model but design prevents estimating quadratic terms)

Source	SS	d.f.	MS
Model	127·606753	1	127·606753
Residual	70·9579032	18	3·94210573
Total	198·564656	19	10·4507714

Number of obs = 20
$F(1, 18) = 32·37$
Prob > $F = 0·0000$
R-squared = 0·6426
Adj R-squared = 0·6228
Root MSE = 1·9855

w_3	Coefficient	Standard error	t	$P > \|t\|$	[95% confidence interval]	
x_3	6·735821	1·183908	5·689	0·000	4·248523	9·223119
X_3^2	(dropped)					
_cons	− 0·540429	0·6278619	− 0·861	0·401	− 1·859518	0·7786599

The regressions computed from the three designs are given in Tables I(*c*) to I(*h*). The first response model, $E(Y|x) = 5x$, is the correct model for the computer output in Tables I(*c*), (*d*) and (*e*). The second response model, $E(Y|x) = 11x - 6x^2$ is the correct model for the output in Tables I(*f*), (*g*) and (*h*). The output is taken from STATA.[2] Most standard statistical software packages produce similar output.

From Table I(*c*), the coefficient B is estimated as 5·58 (1·18) where (1·18) is the standard error. From I(*d*), the estimate of B is 2·97 (1·25) and from I(*e*), the estimate is 5·29 (1·41). These are close to the correct value of 5 (none is significantly different from 5). From Tables I(*f*) and I(*h*), where the correct model is $E(Y|x) = 11x - 6x^2$, we note that the quadratic coefficient C cannot be estimated since there are only two doses which are given to the subjects. To fit a quadratic model at least three distinct values of x are needed. The estimates fit a straight line between dose $x = 0$ and dose $x = 1$ (or $x = 0{\cdot}75$). Since $E(Y|x = 1) = 5$, and $E(Y|x = 0) = 0$, the slope is again 5, and the estimates 3·61 (0·94) and 6·74 (1·18) reflect that (from Tables I(*f*) and I(*h*)). The only design of the three which allows us to estimate C, (Table I(*g*)) gives 13·30 (5·76) as the estimate for B and $-$ 8·11 (5·52) as the estimate for C. Neither estimate is significantly different from this parameter, which are $B = 11$ and $C = -6$, at the 0·05 significance level. The first and third designs do not permit estimation of some parameters. We note that the $N/2$ at 0 and $N/2$ at 1 is optimal for a simple linear regression with an intercept. It is not optimal for the model $E(Y|x) = 5x$ when the intercept is known to be 0. We see in these examples that the choice of design for a particular dose–response model is extremely important.

Determining patterns of dose responses is an important part of new drug evaluation. This may consist of a simple comparison of two levels (placebo and drug at some level), or may consist of placebo (dose $x = 0$) and multiple levels of the drug. We describe some of the design options the researcher has, and provide some guidance on choices. We consider first the equivalence of the two-group design and a two-dose design, and note some properties when the higher dose in the two-group design is less than the maximum dose in a dose–response design. We then review optimal designs and give the efficiencies of several candidate designs for the simple linear and quadratic regression models. For regression notation, we use capital letters (A, B, C) to denote the parameters, and lower case letters (a, b, c) to indicate their estimates. We assume normal errors with mean zero and variance σ^2 throughout. In all cases, we can use standard multiple regression software packages to estimate the parameters A, B, C and σ.

2. EQUIVALENCE OF TWO-GROUP DESIGN WITH DOSE–RESPONSE STUDIES

The two group design is equivalent to a linear regression with doses at two levels. The usual two-sample t-test for equality of means is equivalent to the test that the regression slope coefficient is zero. Assume that half of the observations are allocated to each dose (that is, equal sample sizes in each group). If the doses are coded 0 (placebo) and 1 (dose given) then the difference between treatments is $(\mu_1 - \mu_0)$, where μ_i is the mean response at the ith dose level. In the dose–response model, we assume the maximum dose given is coded as 1, with intermediate doses placed at values between 0 and 1. The usual analysis of this design is a regression (possibly polynomial). The two-group design is a special case of a dose response with only two levels. Denote the mean response in group i as $\bar{y}_i$. It is easy to show that the estimate of the slope is $b = (\bar{y}_1 - \bar{y}_0)$. This leads to the test of the coefficient being $\sqrt{N}(\bar{y}_1 - \bar{y}_0)/2s$, where N is the total sample size and s^2 is the regression mean squared error. This statistic is the two-sample t-test statistic. Simple modifications yield the unequal allocation case.

In the context of dose response the two-group design can be expanded. Thus far, we assumed that the maximum dose in the two-group design was 1. In such cases, the two-group design with

Table II. Sample size needed to detect $B/\sigma = 1$ ($\alpha = 0{\cdot}05$, $1 - \beta = 0{\cdot}9$)

K	2	3	4	5	6	7	8	9	10
N	43	64	76	85	91	95	99	101	104
n	22	22	19	17	16	14	13	12	11

n is the number needed per dose. Thus, $nK \geqslant N$
$B/\sigma = (\mu_1 - \mu_0)/\sigma$

half of the observations at each level is the optimum one for a linear regression $E(Y|x) = A + Bx$ (in the sense that var(b) is minimized). It may, however, be important to have fewer than half of the observations at the placebo or high dose levels (for example, for increasing the chance of receiving a hoped for effective treatment or for ethical reasons such as reducing the risk of side-effects), so the investigator might wish to place some observations at 0, some at 1, and the remainder at intermediate doses. The experimenter may even unbalance the design by making the number of observations at each dose unequal. Similarly, the 0 dose might be increased to x_0. This would not be a placebo controlled study, but would still be able to demonstrate effectiveness of x at $x = 1$ if there were an increasing response for increasing x. In the phase III drug approval context, regulatory agencies expect some of the dose to be at the level for which approval is sought. Thus, a dose–response study with doses at 0, 0·5 and 1·0 would not suffice for an approval at 0·75. Generally, dose–response studies are done in phase II. We will examine some examples of these when we consider the efficiencies of the designs in later sections.

Assume bivariate observations, (y, x), are taken, where x is the dose given and y is the response to the drug. The dose begins at 0 (placebo) and has a largest value of 1 (this can always be handled by appropriate scaling). With K equally spaced doses, we have values of x at 0, $1/(K-1)$, $2/(K-1), \ldots, (K-2)/(K-1)$, 1. If we assume a straight-line model, we have

$$E(Y|x) = A + Bx$$

where B is the amount of increase in $E(Y|x)$ for a one unit increase in the dose, x. That is, B is still $\mu_1 - \mu_0$, the mean difference between the drug at $x = 1$ and the placebo at $x = 0$. The variance of Y given x is σ^2, and $\text{var}(b) = \sigma^2/\Sigma(x_i - \bar{x})^2$.

With an equal number of observations and equal spacing between them, the denominator of var(b) is $\sigma^2 N(K+1)/[12(K-1)]$ where N is the total sample size and $N = nK$, and n is the sample size per group. This is another example of a uniform design. For testing H_0: $B = \mu_1 - \mu_0 = 0$ against a two-sided alternative H_1: $B \neq 0$, the sample size needed to detect a slope of B is given by

$$N = 12(K-1)(Z_{1-\alpha/2} + Z_{1-\beta})^2\sigma^2/[B^2(K+1)]$$

where α is the significance level, and β is the probability of a type II error and Z_x is the upper $100 \times$ th percentile of the standard normal distribution.

One can use the above formula to compute the needed sample size. For example, to detect $B/\sigma = 1$ ($(\mu_1 - \mu_0)/\sigma = 1$) with $\alpha = 0{\cdot}05$, and $1 - \beta = 0{\cdot}9$, the required sample sizes are given in Table II. The sample sizes are rounded to integer values. In practice the sample size N should be increased so they are divisible by K.

When $K > 2$, the multiple x values permit us to examine curvature or a polynomial response. The equally spaced doses are not necessarily the optimally spaced x values for such fits. In practice, if there is confidence that the model is not much wrigglier than a quadratic, the value of K should not be much higher than 3 or 4 because of the considerably larger sample size requirements.

Table III. Equivalent maximum two-group dose for K equally spaced doses

K	2	3	4	5	6	10
Dose	1·0	0·816	0·745	0·707	0·683	0·638

For the case with $K = 2$ and $x = 0$ and 1, the above produces the sample size for a two-group design with equal number observations in each group,

$$N = 4(Z_{1-\alpha/2} + Z_{1-\beta})^2 \sigma^2/(\mu_1 - \mu_0)^2$$

with the required sample size in each group as $N/2$. This is the usual formula for the two-sample problem.

Given the above, the investigator can consider several issues. First, if the investigator is certain that the response is linear and the maximum dose is the one which will be administered to patients, the optimal two-group design allocates half of the sample to the 0 (placebo) dose and half to 1 and no multiple dosage design ($K > 2$) is more efficient in the sense that no other design can have a smaller variance of b. However, the high dose may have some toxicity which the investigators want to avoid. The alternatives are to reduce the high dose to a lower dose at $x < 1$ or to conduct a study with multiple doses so that fewer patients receive the highest dose. If we compare a dose–response study that has equal allocation to equally spaced doses between 0 and 1 (a uniform design) to a two-group design at 0 and x, can the dose–response design have a smaller variance of b than the two-group design where the slope B is still $\mu_1 - \mu_0$ and the total sample size N does not exceed that of the two-group design? (Answer: Yes). Where does the dose–response design begin to be better? (Answer: it depends on x). How does the number of doses relate to the maximum level (x) in the two-group case? (Partial answer: if $x < 1/\sqrt{3}$ any multiple point design wins). How do the variances of b and c compare when the doses are evenly spaced versus optimal design placement? (Answer: see below). It is sometimes proposed to have fewer observations at dose = 0 and dose = 1 for ethical reasons (fewer patients at dose = 0 to have more patients receiving something, fewer at dose = 1 to have less potential toxicity). What is the effect of reducing the number of observations at the extreme doses on the variances of the parameter estimates? (Answer: the increase in variance can be pretty bad).

If the corresponding maximum dose given in the two-group design is $x = 0{\cdot}5$ (for example, 500 mg in a 1000 mg maximum dose study), the change in mean response would be half that of the maximum. The sample size required (Table II for $K = 2$) would be multiplied by 4, and 172 patients (86 per group) would be required. This is larger than a dose–response design with 10 levels of drug, and one would clearly prefer a dose–response design. If the maximum dose for two-group dose design is 0·75, then the number of patients required would be 76 (38 per group) and the two group design would require about as many subjects as a dose–response design with four dosages (at 0, 0·33, 0·67 and 1). A two-group study and a dose–response study will have the same sample size if the formulae for N are equal. Some algebra shows that a two-group study where the drug is given at dose x will have the same sample size requirement as a dose–response study with K doses equally spaced from 0 to 1 if $x^2 = (K + 1)/[3(K - 1)]$. This leads to Table III as a table of equivalent sample size studies.

This can be interpreted to mean that if the maximum dose in the two-group study is 0·745, a four level study with doses at 0, 0·33, 0·67 and 1 will provide estimate of B with the same precision. Assuming equal spacing, we can show that the equivalent dose is never less than 0·577 ($1/\sqrt{3}$). These results suggest that 3 to 5 levels in a dose–response study will provide most of the gain when the two group dose is less than the maximum dose in the dose–response study. When

the dose to be studied in the two-group study is the maximum dose that would be administered in a dose–response study, the dose–response study does not provide a gain in power or precision of estimate. If the maximum dose in a two-group study is less than 0·8, an increase in power can usually be realized with a dose–response study.

Dose–response (regression) designs enable us to compare the response to different drug levels and evaluate the responses for possible curvature. If we assume a quadratic response, $E(Y|x) = A + Bx + Cx^2$, we can find optimal designs which minimize var(c), var($y(x)$) for a given x, or the generalized variance (the determinant of the covariance matrix of the estimates). Here, $y(x)$ is the predicted value of y given x. These designs will provide estimates of the curvature (that is, the quadratic coefficient) and also allow us to estimate the linear dose–response. We next discuss these concepts.

3. CONSIDERATIONS IN CHOOSING DESIGNS

A major advantage that dose–response studies have over two-group studies is their ability to examine departures from linear response. For example, by plotting the responses against the doses we can examine (visually) if there are large departures from the assumed linear dose–response relationship. It is also possible to test formally this relationship using the pure error term when there are replicated observations at the different doses. This is fully explored in such texts as Rawlings[3] or Neter *et al.*[4] (In submissions to regulatory agencies, this should usually be indicated in the analysis plan submitted with the study proposal.)

Similarly, all of the powerful diagnostic tools of regression analysis (for example, residual analysis, influence statistics, normal probability plots) are available to the investigator (see Rawlings[3] or Neter *et al.*[4]). While one could argue that these are also available in the two-group design (it is a special case of the dose–response study with $K = 2$), residual analysis is not able to detect curvature in this case, and influence is the same for all observations when equal allocation is used.

If we assume that the model is quadratic,

$$E(Y|x) = A + Bx + Cx^2$$

the variances of the coefficients are found by inverting the $\mathbf{X}'\mathbf{X}$ matrix where $\mathbf{X}$ is the design matrix (see, for example, Neter *et al.*[4]). The first column of $\mathbf{X}$ is a column of 1s. The next columns are the values of x and x^2. To find the $\mathbf{X}'\mathbf{X}$ matrix for the quadratic regression model, we define

$$F = (\Sigma x_i)/N, \quad S = (\Sigma x_i^2)/N, \quad T = (\Sigma x_i^3)/N, \quad Q = (\Sigma x_i^4)/N$$

(standing for first, second, third and fourth (quartic) powers, respectively). The covariance matrix is

$$\Sigma = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2/[N(SQ + 2FTS - S^3 - T^2 - QF^2)] \begin{bmatrix} SQ - T^2 & ST - FQ & FT - S^2 \\ ST - FQ & Q - S^2 & SF - T \\ FT - S^2 & SF - T & S - F^2 \end{bmatrix}$$

giving

$$\text{var}(a) = \sigma^2[SQ - T^2]/[N(SQ + 2FTS - S^3 - T^2 - QF^2)]$$

$$\text{var}(b) = \sigma^2[Q - S^2]/[N(SQ + 2FTS - S^3 - T^2 - QF^2)]$$

$$\text{var}(c) = \sigma^2[S - F^2]/[N(SQ + 2FTS - S^3 - T^2 - QF^2)].$$

From the covariance matrix, the optimal design for estimating any one of the parameters can be found by minimizing the variance of its estimate. However, the optimization problem is usually a complicated one. For example, if interest is only in B, the problem becomes how to allocate observations in the closed interval $[0, 1]$ so that the quantity

$$(Q - S^2)/[N(SQ + 2FTS - S^3 - T^2 - QF^2)]$$

is minimized subject to the constraint that the number of observations at the x_i's sum to N. The solution to this and related problems is generally difficult. For this reason, we use an alternative approach using approximate optimal design theory.

4. A BRIEF REVIEW OF APPROXIMATE OPTIMAL DESIGN THEORY

An optimal design is one which minimizes (or maximizes) some function of the covariances of the parameter estimates. There are many optimal design criteria. Often they lead to complicated (or intractable) expressions to optimize. As a way of dealing with the problem of complicated algebraic expressions (like the one above), Kiefer[5] proposed the concept of approximate designs. Although there was controversy at the time, it is now an accepted way of solving a design problem. Generally, a design is defined by specifying the number of points (locations) in the interval where observations are taken, and the number (or proportion) of observations to be taken at each of these points. For example, if N is the total number of observations in the experiment, a design is defined by taking n_i observations at specified points x, $i = 1, 2, \ldots, k$ with $\sum n_i = N$. Alternatively, if ξ_i is the proportion of observations to be taken at x_i this is the same as taking $N\xi_i = n_i$ observations at x_i. The concept of approximate designs extends this notion of design by allowing the $N\xi_i$ to be non-integers. Consequently, approximate designs may require taking a fractional number of observations at a point. This is not a problem if N is large. In practice, we round to the nearest integer, subject to $\sum n_i = N$.

The main advantages of considering approximate designs are: (i) the optimization problem is simplified; (ii) frequently, the approximate optimal design is close to the optimal design.[5] Further background readings in this area are (in ascending order of difficulty) Atkinson and Donev,[6] Fedorov,[7] and Pazman.[8]

Define Ω as the space of the design points, x_i. In the context of this article, Ω is the closed interval $[0, 1]$. Let ξ be an approximate design (or simply a design from now on) giving probability mass ξ_i at the point x_i, $i = 1, 2, \ldots, K$. The analogue of the $\mathbf{X'X}$ matrix for ξ is its information matrix defined by

$$\mathbf{M}(\xi) = \Sigma \mathbf{f}(\mathbf{x}_i)\,\mathbf{f}(\mathbf{x}_i)^t \xi_i$$

where x_i is the set of predictor variables. For example, consider the simple linear regression $\mathbf{f}(\mathbf{x})^t = (1\ \mathbf{x})$. If ξ is a design with equal allocation at $x_1 = 0$ and $x_2 = 1$, we have $\xi_1 = 0{\cdot}5$ and $\xi_2 = 0{\cdot}5$, and

$$\mathbf{M}(\xi) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} 0{\cdot}5 + \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} 0{\cdot}5 = \begin{pmatrix} 1 & 0{\cdot}5 \\ 0{\cdot}5 & 0{\cdot}5 \end{pmatrix} \tag{1}$$

which is $1/N$ times the usual $\mathbf{X'X}$ matrix.

Many practical optimality criteria are formulated in terms of $\mathbf{M}(\xi)$. As an illustration, Table IV lists some common criteria as convex functions, $\mathbf{H}(\mathbf{M}(\xi))$ of $\mathbf{M}(\xi)$ and reasons for using them. For ease of notation, we assume the underlying model for the first two criteria is quadratic, that is, $\mathbf{f}(\mathbf{x})^t = (1\ \mathbf{x}\ \mathbf{x}^2)$. Generalizations to other models are straightforward. An optimal approximate

Table IV. Criteria for optimal designs

Optimality criterion	Interest	$H(\mathbf{M}(\xi))$
var(b)	Estimate B accurately	$\{\mathbf{M}(\xi)^{-1}\}_{2,2}$
var (c)	Estimate C accurately	$\{\mathbf{M}\{\xi)^{-1}\}_{3,3}$
L-optimality	Estimate response at $x = x_0$	$\mathbf{f}(\mathbf{x}_0)^t\mathbf{M}(\xi)^{-1}\mathbf{f}(\mathbf{x}_0)$
D-optimality	Estimate all parameters (A, B, C)	$-\log\lvert M(\xi)\rvert$
G-optimality	Minimize max var($\hat{y}(x)$)	$\max_{\mathbf{x}\in\Omega}\mathbf{f}(\mathbf{x})^t\mathbf{M}(\xi)^{-1}\mathbf{f}(\mathbf{x})$

The notation $\{\ \}_{2,2}$ and $\{\ \}_{3,3}$ refer to minimization of var(b) and var(c), respectrively, which are the second and third diagonal elements of $\mathbf{M}^{-1}$

design ξ° is one for which min $H(\mathbf{M}(\xi)) = H(\mathbf{M}(\xi^\circ))$, where the minimization is taken over the set of all approximate designs on Ω.

The verification of an optimal approximate design is straightforward in many cases (Fedorov[7]). For example, if interest is in estimating all parameters, a D-optimal design is appropriate (see below). For linear models one can check if a given design, ξ, is D-optimal by verifying

$$\mathbf{f}(\mathbf{x})^t\mathbf{M}^{-1}(\xi)\mathbf{f}(\mathbf{x}) \leqslant p \quad \text{for all } \mathbf{x}\in\Omega \tag{2}$$

where p is the number of parameters in the model. For simple and quadratic linear regression models p is equal to 2 and 3, respectively. This condition is easily verified in practice. Corresponding checking conditions for the other criteria are available in Fedorov.[7] As an illustration how (2) might be used, return to the simple linear regression example with $H(\mathbf{M}(\xi)) = -\log(|\mathbf{M}(\xi)|)$, then if ξ° assigns equal numbers of observations at 0 and 1 the information matrix $\mathbf{M}(\xi^\circ)$ is as given in (1). After some algebra, the left hand side of (2) is seen to be $2(1-2x+2x^2)$. Since $p = 2$, the condition (2) is satisfied and ξ° is D-optimal. Thus, the equal allocation design is D-optimal. Other optimal designs can be similarly verified.

Under the assumption of normality of the errors, the D-optimality criterion seeks to minimize the volume of the confidence ellipsoid for the parameters. This is achieved by maximizing the determinant of the information matrix, $|\mathbf{M}^{-1}(\xi)|$ or minimizing the generalized variance, that is, $|\mathbf{M}(\xi)|$ over the set of all approximate designs. When interest is in only one parameter (or the response at a particular point), minimizing the variance of the estimator is reasonable. This is the rationale behind the first three criteria in Table IV. Minimizing the variance of b is the most relevant goal for simple linear regression. Minimizing the variance of c is important in quadratic regression. Minimizing the variance of the response at a point is a goal which is obviously met by placing all observations at the point, but this strategy would eliminate all possibility of estimating regression coefficients. G-optimality minimizes the maximum variance of a predicted value over the interval [0, 1]. Consequently, this criterion may be useful for estimating the response curve.

To evaluate the usefulness of a design, we use the idea of design efficiency. This is a number between 0 and 1 and has the interpretation that its reciprocal measures the number of times the design has to be replicated for it to do as well as the optimal design. The efficiency is the ratio of the criterion of the optimal design to the value of the criterion for the proposed design. For example, if we want to estimate the parameter B using design ξ, the efficiency of ξ is given by $\text{var}_{\text{opt}}(b)/\text{var}_{\xi}(b)$. For the quadratic model, the optimal design (ξ_3 in Table V) to minimize var(c) places $N/4$ points at 0, $N/4$ points at 1 and $N/2$ at 1/2. The D-optimal design (ξ_2 in Table V) places $N/3$ at each of these points.[5] Thus, for ξ_3 the variances of a, b and c are $4\sigma^2/N$, $72\sigma^2/N$, and $64\sigma^2/N$, respectively, and for ξ_2 the variances are $3\sigma^2/N$, $78\sigma^2/N$ and $72\sigma^2/N$. These are σ^2/N

times the diagonal elements of the inverse of the information matrices for ξ_2 and ξ_3. Therefore, the efficiency of the design ξ_2 for the estimation of C is $\text{var}_{\xi 3}(c)/\text{var}_{\xi 2}(c) = 64/72 = 0{\cdot}889$. These calculations are illustrated more fully in Section 5.

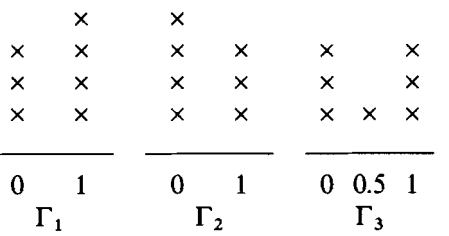
Sometimes this simple definition of efficiency must be modified to maintain its interpretation. For instance, consider D-optimality for a linear model with p parameters. If the D-optimal design is ξ_D, we have $|\mathbf{M}(\xi_D)| \geqslant |\mathbf{M}(\xi)|$ for all ξ, so that in order for ξ to do as well as ξ_D, ξ must be replicated, say r times. Since $|\mathbf{M}(\xi_D)| = |r\,\mathbf{M}(\xi)| = r^p|\mathbf{M}(\xi)|$, this implies $r = \{|\mathbf{M}(\xi)|/|\mathbf{M}(\xi_D)|^{-1/p}$ and so the D-efficiency of ξ is defined by $\{|\mathbf{M}(\xi)|/|\mathbf{M}(\xi_D)|^{1/p}$. For the two designs, ξ_2 and ξ_3, it can be verified that the D-efficiency of the design ξ_3 relative to ξ_2 is

$$\{|\mathbf{M}(\xi_3)|/\mathbf{M}(\xi_2)|\}^{1/3} = \{432/512\}^{1/3} = 0{\cdot}945$$

As a further example consider the case when $N = 2k + 1$, $\mathbf{f}(\mathbf{x})^t = (\mathbf{1}\ \mathbf{x})$, $\Omega = [0, 1]$ and there is interest in three designs:

Γ_1 places $k/(2k + 1)$ of its mass at 0 and the rest at 1;
Γ_2 places $k/(2k + 1)$ of its mass at 1 and the rest at 0;
Γ_3 places $1/(2k + 1)$ of its mass at 1/2, $k/(2k + 1)$ at 0 and $k/(2k + 1)$ at 1.

With $N = 7$ ($k = 3$) graphically these three designs look like



It is easy to verify that $|\mathbf{M}(\Gamma_1)| = |\mathbf{M}(\Gamma_2)| = k(k + 1)/(2k + 1)^2$ and $|\mathbf{M}(\Gamma_3)| = k/[2(2k + 1)]$, so that $|\mathbf{M}(\Gamma_1)| - |\mathbf{M}(\Gamma_3)| = k/[2/(2k + 1)^2] > 0$ for all positive integers. As k becomes larger, the D-efficiency of Γ_3 approaches that of Γ_1 or Γ_2. For example, if $k = 3$, the efficiency is $[(2k + 1)/(2k + 2)]^{1/2} = 0{\cdot}9354$. For $k = 5$ ($N = 11$), the efficiency is 0·9574.

Kiefer[9] gave a general bound for the D-efficiency of a design which has a non-singular information matrix, $\mathbf{M}$. He showed that if $\mathbf{M}^{-1}(\Gamma)$ exists, then the D-efficiency is always greater than or equal to

$$\exp(1 - \max_{x \in \Omega}(\mathbf{f}(\mathbf{x})^t\mathbf{M}^{-1}\Gamma\,\mathbf{f}(\mathbf{x})/p).$$

Applying this to Γ_2 we have its D-efficiency is at least $\exp(-1/2k)$. For $k = 10$ ($N = 21$) this lower bound is 0·95. The D-efficiency approaches 1 quickly.

Our consideration of D-efficiency has been limited to the case where interest is in all the parameters. When nuisance parameters are present, techniques exist for estimating a subset of the model parameters. The analogous expressions for the checking condition (2) are invariably more complicated.[5]

5. CASE STUDIES FOR SIMPLE LINEAR AND QUADRATIC REGRESSION

We now study five dose–response designs and evaluate their efficiencies (see Table V). The dosage levels are 0, 1/3, 1/2, 2/3, 1 and the five designs offer different allocations to each. Because some dosages have no observations allocated to them, these are not five level designs in the usual sense.

Table V. Candidate designs for dose–response studies

Design	Dosage				
	0	1/3	1/2	2/3	1
ξ_1	$N/2$	0	0	0	$N/2$
ξ_2	$N/3$	0	$N/3$	0	$N/3$
ξ_3	$N/4$	0	$N/2$	0	$N/4$
ξ_4	$N/4$	$N/4$	0	$N/4$	$N/4$
ξ_5	$N/6$	$N/3$	0	$N/3$	$N/6$

Table VI. Design efficiencies for some designs for the linear and quadratic models on [0, 1]

Design	var (b) e_1	var(c) e_2	L-optimal var($y \mid x_0 = 1$) e_3	D-optimal Generalized variance e_4	G-optimal max(var($y(x)$)) e_5
Simple linear regression: $y = A + Bx + \varepsilon$					
ξ_1	1·0000	NA	0·5000	1·0000	1·0000
ξ_2	0·6667	NA	0·4000	0·8165	0·8000
ξ_3	0·5000	NA	0·3333	0·7071	0·6667
ξ_4	0·5556	NA	0·3571	0·7454	0·7143
ξ_5	0·4074	NA	0·2895	0·6383	0·5790
Quadratic regression: $y = A + Bx + Cx^2 + \varepsilon$					
ξ_1	1·0000	0·0000	0·5000	0·0000	0·0000
ξ_2	0·0128	0·8889	0·3333	1·0000	1·0000
ξ_3	0·0139	1·0000	0·2500	0·9449	0·7500
ξ_4	0·0113	0·7901	0·2632	0·9048	0·7895
ξ_5	0·0099	0·7023	0·1833	0·7845	0·5500

As mentioned above, design ξ_1 is the optimal design for estimating B for the simple linear regression model, and is, in fact, optimal for other purposes as well (see Table VI). Design ξ_2 is optimal for jointly estimating all parameters in the quadratic model (that is, D-optimal). It is appealing also because it assigns an equal number of subjects to three dosages uniformly spaced between 0 and 1. Design ξ_3 is optimal for minimizing the variance of c in the quadratic model. Design ξ_4 has equal allocation to dosages uniformly spaced between 0 and 1, and is motivated by its simplicity and ease of explanation to clients. Designs ξ_2 and ξ_4 are special classes of uniform designs mentioned earlier. Besides being easy to construct, uniform designs also are robust when there is uncertainty in the regression model (Wiens[10]). Design ξ_5 is motivated by the ethical considerations of having fewer observations at the placebo and maximum doses.

Table VI shows the efficiencies of the five designs for the five optimality criteria. For any design, ξ, we will refer to these efficiencies as $e_1(\xi), e_2(\xi)$, etc. We note that none of these designs is L-optimal (that is, minimizes the variance of y at $x_0 = 1$). It is immediate that the L-optimal design places all observations at the point $x_0 = 1$. This design is useless for the other criteria, since it has zero efficiency for them.

Several interesting results are evident. If the simple linear regression model holds, design ξ_1 has design efficiencies of 1 except for $e_3(\xi_1)$. Thus, design ξ_1 is useful as it achieves several goals in the study, including, but not limited to, the first, fourth and fifth criteria. Unfortunately, this design is inefficient when the quadratic model holds. Its efficiencies are $e_2(\xi_1) = e_4(\xi_1) = e_5(\xi_1) = 0$ and

Table VII. Dosage levels for optimal estimation of polynomials (naive choices in parentheses)

Degree	X-values				
3	0	0·2764	0·7236	1	
	(0	0·3333	0·6667	1)	
4	0	0·1727	0·5000	0·8273	1
	(0	0·2500	0·5000	0·7500	1)
5	0	0·1175	0·3574	0·6426	0·8825	1
	(0	0·2000	0·4000	0·6000	0·8000	1)

$e_3(\xi_1) = 0{\cdot}5$. While design ξ_1 estimates B with the smallest variance (in the quadratic model) it estimates C and A with maximal variance (Preitschopf and Pukelsheim[11]). The practical implication is serious since design ξ_1 provides no information about C when the quadratic model holds. Interestingly, if we modify design ξ_1 to, say, ξ_1^*, so that half of the observations are taken at 0 and half at x $(0 < x \leqslant 1)$, then

$$e_1(\xi_1^*) = x^2$$

$$e_3(\xi_1^*) = x^2/(2x^2 - 4x + 4)$$

$$e_4(\xi_1^*) = x$$

$$e_5(\xi_1^*) = x^2/(x^2 - 2x + 2)$$

under the simple linear regression model.

The other entries in Table VI can be interpreted similarly. For the quadratic model, design ξ_2 gives a variance of c 1·125 (= 1/0·8889) times that of design ξ_3. This was the example noted earlier. This means that design ξ_2 requires 12·5 per cent more observations to attain the same variance for c as design ξ_3.

These computations demonstrate the importance of considering the efficiency of a design. A poor choice of a design wastes resources, while a carefully designed experiment can furnish more information with fewer resources. A poorly designed experiment could also, in an extreme case, produce little or no information after the experiment is run. Such would be the case if one wished to estimate C and design ξ_1 was used.

We have restricted attention to simple linear regression and quadratic regression models. Similar ideas apply to polynomial models of degrees higher than 2. For example, when a polynomial of degree 3, 4 or 5 is used to model the dose–response relationship and interest is in all the parameters, the approximate D-optimal designs take equal proportions at the dosage levels given in Table VII. The uniform design on $K + 1$ points are chosen as i/K, $i = 0, \ldots, K$. The optimal points are close to the naive choices and so the naive choices should pay only a slight penalty in the design criteria costs. Optimal designs for polynomials of higher degrees and optimal designs for estimating subsets of the parameters are available in Fedorov.[7]

Finally, we comment that if we consider approximate designs uniformly spaced on K points, there is no advantage in considering values of K greater than 4 for the simple linear regression model and K greater than 7 for the quadratic regression model (see Fedorov[7] for details). Table VIII shows that large values of K can result in substantial decreases in efficiency. For both these criteria, the efficiency decreases as K increases. Where possible one should avoid designs

Table VIII. Efficiency of a uniform design on K points for a quadratic model

Criterion	K				
	3	4	5	6	7
var(b)	0·8889	0·7901	0·7000	0·6372	0·5926
Generalized variance	0·9449	0·9048	0·8390	0·7946	0·7631

with excessive numbers of points. This is true both from the viewpoint of optimal design and from the logistics of conducting the experiment.

6. SUMMARY AND RECOMMENDATIONS

We have described dose–response designs and illustrated them with simple (artificial) examples. The dose–response design generally provides more information than the two-group design (with half of the points at 0 and half at 1) unless the response is linear. It places fewer points at the highest and lowest dose levels and provides information about non-linear response. There is only a small drop in efficiency in the designs as we move from the optimal, symmetric design to the uniform design. Designs which deliberately reduce the number of points at the extremes (dose = 0 and dose = 1) lose efficiency rapidly, and we do not generally recommend them. The approach and analysis we adopted here is based on optimal approximate design. The primary reason for doing this is that the design problem is simplified and the optimal approximate design that results provides a useful guide to the practitioner. There are many applications of the theory of optimal approximate design in practice and they are increasing. Some recent biomedical applications of these ideas can be found in Hoel and Jennrich,[12] Dunn,[13] Hatzis and Larntz,[14] Atkinson *et al.*,[15] and Kitsos *et al.*[16]

REFERENCES

1. Kotz, S. and Johnson N. L. (eds) *Encyclopedia of Statistical Sciences vol.* 2, Wiley, New York, 1982, p. 418.
2. StataCorp. *Stata Statistical Software: Release 4.0*, Stata Corporation, College Station, TX, 1995.
3. Rawlings, J. *Applied Regression Analysis: A Research Tool*, Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, CA, 1980.
4. Neter, J., Wasserman, W. and Kutner, M. *Applied Linear Statistical Models*, 3rd edn, Richard D. Irwin, Homewood, IL, 1990.
5. Kiefer, J. *Jack Carl Keifer Collected Papers III: Design of Experiments*, Springer-Verlag, New York, 1985.
6. Atkinson, A. C. and Donev, A. N. *Optimum Experimental Designs*, Clarendon Press, Oxford, 1992.
7. Fedorov, V. V. *Theory of Optimal Experiments*, translated and edited by Studden, W. J. and Klimko, E. M., Academic Press, New York, 1972.
8. Pazman, A. *Foundations of Optimum Experimental Design*, D. Reidel Publishing Company, Dordrecht, Boston, Lancaster and Tokyo, 1986.
9. Kiefer, J. 'Optimum experimental designs V, with application to systematic and rotatable designs', *Proc. 4th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley, 1960, pp. 381–405.
10. Wiens, D. P. 'Designs for approximately linear regression: two optimality properties of uniform designs', *Statistics and Probability Letters*, **12**, 217–221 (1991).
11. Preitschopf, F. and Pukelsheim, F. 'Optimal designs for quadratic regression', *Journal of Statistical Planning and Inference*, **16**, 213–218 (1987).
12. Hoel, P. G. and Jennrich, R. I. 'Optimal designs for dose response experiments in cancer research', *Biometrika*, **66**, 307–316 (1979).

13. Dunn, G. 'Optimal designs for drug, neurotransmitter and hormone receptor assays', *Statistics in Medicine*, **7**, 805–815 (1988).
14. Hatzis, C. and Larntz, K. 'Optimal design in nonlinear multiresponse estimation: Poisson model for filter feeding', *Biometrics*, **48**, 657–671 (1992).
15. Atkinson, A. C., Chaloner, K., Hertzberg, A. and Juritz, J. 'Optimum experimental designs for properties of a compartmental model', *Biometrics*, **49**, 325–337 (1993).
16. Kitsos, C. P., Titterington, D. M. and Torsney, B. 'An optimum design problem in rhythmometry', *Biometrics*, **44**, 657–671 (1988).