

Optimal Serial Dilutions Designs for Drug Discovery Experiments

Alexander N. Donev & Randy D. Tobias

First version: 11 September 2010

Research Report No. 5, 2010, Probability and Statistics Group
School of Mathematics, The University of Manchester

Optimal Serial Dilutions Designs For Drug Discovery Experiments

Alexander N. Donev
School of Mathematics, University of Manchester, UK

and

Randy D. Tobias
Linear Models R&D, SAS Institute Inc., Cary, NC 27513, USA

Summary. Dose-response studies are an essential part of the drug discovery process. They are typically carried out on a large number of chemical compounds using serial dilution experimental designs. This paper proposes a method of selecting the key parameters of these designs (maximum dose, dilution factor, number of concentrations and number of replicated observations for each concentration) depending on the stage of the drug discovery process where the study takes place. This is achieved by employing and extending results from optimal design theory. Population D- and D_S -optimality are defined and used to evaluate the precision of estimating the potency of the tested compounds. The proposed methodology is easy to use and creates opportunities to reduce the cost of the experiments without compromising the quality of the data obtained in them.

Keywords: bioassay, population D- and D_S -optimality, dose response study, high-throughput screening, minimum significant difference

Corresponding author: A. N. Donev. E-mail: a.n.donev@manchester.ac.uk

1. Introduction

Modern technology allows for using automated robotic systems to simultaneously test and compare many compounds with respect to properties of scientific interest, e.g. potency or toxicity in pharmacology. Such studies are common in the long drug discovery process and are also important to the fields of biology and chemistry. This paper is concerned with the design of such studies.

The properties of the compounds of interest are assessed in dose-response studies where the response of interest is measured at different doses of the studied compounds. Usually the doses of each compound are obtained by serial dilution from a single stock solution. The same experimental design is used to study all compounds. Robots can be easily programmed to pipette different amounts of the compounds by serial dilution. Typically plates with 96, 384 or a larger number of wells are used, where each of the required experimental conditions is tested in one or more of the wells in the plates.

For example, early in the drug discovery process, chemical libraries are searched to identify compounds that have the potential to produce desirable biological effect, e.g. potency. Initially, all compounds could be tested at a single dose. Those that show potential are studied with a serial dilution design at the next stage where the focus is on identifying useful chemical series. Therefore compounds with a wide range of potencies are tested and only those that have negligible potencies are of no interest. This stage is known as high-throughput screening (HTS). New compounds are also introduced at this stage, as medicinal chemists iterate changes to structures in existing compounds in efforts to improve potency. This may lead to nominating a (still large) number of compounds to study further in the next and subsequent stages. However, the number of compounds is considerably reduced from stage to stage as the search focuses on looking for the most potent compounds that also

possess other desirable characteristics (e.g. selectivity and safety). Hence, ordering the compounds with respect to their characteristics becomes the highest priority. Clearly there is no one experimental design that can be good for all stages in the studies.

There are many reports in the literature of successful use of serial dilution assays. Specific issues arising in the analysis of data collected in such experiments are discussed by Lee and Whitmore (1999), and analysis software that is easy to use is presented by Ritz and Streibig (2005). The effect of choosing different serial dilution experimental designs on the accuracy of the estimation of the potencies of the compounds is discussed by Strijbosch, Does and Buurman (1988), and by Macken (1999), though in somewhat different and simpler experimental settings than those addressed in this paper. Mehrabi and Matthews (1998) discuss the drawbacks of using ad hoc considerations in designing assays. They propose the use of Bayesian optimal designs for immunological applications where the proportion of subtype cells has to be estimated and the criterion of design optimality requires this proportion to be estimated as precisely as possible. They question the practicality of the designs they construct to be optimal with respect to a particular criterion (e.g. D-optimality) with no restriction on the dose levels; they focus instead on easy to implement designs, like the serial dilution designs, and show that the number of their support points depends on the prior information about the parameter that has to be estimated.

Various aspects of compound screening experiments in industry are discussed in the literature, see for example Eastwood et al (2006), Fox et al (2006) and Woodward et al (2006). Examples of their use are also available (e.g. Parker et al (2000), Bhat et al (2006), and Fox et al (2006)). However, there is no clearly defined common general practice of doing such experiments at different stages of the drug discovery process and related research. Practices vary across companies and experimental laboratories. However, it appears that ad hoc considerations continue to play a decisive role in choosing the

experimental designs. In some cases this may lead to wasting resources; in others the results could be ambiguous and inaccurate.

As the aim of the experimenter changes through the stages of the drug discovery process, so should the designs used. We propose a unified systematic approach for their selection that takes into account the aims and objectives of the stage where the design will be used. Results from the theory of optimal experimental design are used and extended to choose the best serial dilution design for each plate. We focus on the choice of the serial dilution design used on each plate because it is repeatedly used in each study and therefore it is crucial for the statistical properties of the results and for the cost effectiveness of the study.

In Section 2 the main features of the serial dilution designs and the typical analysis of data collected by using them are described. In Section 3 we briefly review results of the optimal design theory. As serial dilution designs are usually used to study a set, or a population of compounds with different characteristics for their biological activity, we extend the classical criteria of optimality to assess the desired properties of the designs when such a population is studied. A method for carrying out power calculations is also shown. Thus, we extend the suggestions made by Eastwood et al (2006). In Section 4 we show how population criteria for design optimality can be used to select serial dilution designs for typical experimental scenarios. The paper concludes with a discussion about the usefulness of the proposed approach. Computer code implementing aspects of the method described in this paper is available from the corresponding author on request.

2. Serial dilution experimental designs

A serial dilution design is specified by the maximum dose (MD) that is used, the number of doses (ND) and the dilution factor (DF). For example, if $MD = 81$, $ND = 5$ and $DF = 3$, the design requires the doses: 81, 27, 9, 3 and 1 to be used. Here, as well as in the

rest of the paper, the units of doses are omitted as the actual values depend on the nature of the experiment.

We assume that all doses are randomized on wells of a single plate and that there are no positional (e.g. edge) effects on the plates. Positive and negative controls are also included in the design as they allow for estimating the response when no compound is used, and when a very large amount is used; indicating when considering doses beyond the range of experimentation can only be expected to lead to insignificant changes in the response. As is common practice, we assume equal replication r at all ND doses. Situations when unequal number of replications could be useful are considered by Donev et al (2008). The ideas presented in this paper can be extended to such cases too.

The model

$$Y_i = \gamma + \frac{\delta - \gamma}{1 + 10^{(x_i - \alpha)\beta}} + \varepsilon_i \quad (2.1)$$

is then fitted to the data. In (2.1), Y_i is the response observed at a dose whose logarithm to base 10 is x_i , ε_i is the experimental error which we assume to be normally distributed with zero mean and variance proportional to the expectation of the response, as practical experience suggests. Also, β is known as the *Hill slope*, named after Hill (1913) who proposed model (2.1). The parameter α is the base 10 logarithm of the *IC50*, the dose required to achieve a response half way between the maximum and minimum possible responses, δ and γ respectively. Thus the parameter α is a natural measure of the potency of a compound, with smaller values indicating higher potency. This parameter is the primary focus of the experiments discussed in this paper; we shall use “ α ”, “potency”, and “LogIC50” interchangeably to refer to it. The use of logarithms is often seen as the natural scale on which to compare potencies of compounds. The difference $(\delta - \gamma)$ defines the assay window and usually it is estimated in a validation experiment prior to the main experiment. The data can be scaled and therefore without loss of generality we assume $(\gamma, \delta) = (0, 1)$.

Model (2.1) is nonlinear in the parameters and its parameters are estimated by numerical optimization for each of the compared compounds using the data obtained on each plate. Comprehensive descriptions of the available methods for estimating nonlinear models are given by Seber and Wild (2005) and Bates and Watts (2007). Statistical methods in bioassay are discussed by Finney (1979).

Highly potent compounds have small values for α . Usually considerable resources are allocated to study such compounds further. Some of them may later be taken forward to long and expensive clinical studies and eventually become drugs. Estimating the parameters of model (2.1) precisely, especially α , ensures that experimental resources are spent on studying the right compounds in subsequent studies. However, this is made difficult by the dependence of the precision on the values of the model parameters. Naturally the tested compounds are expected to have different values for α and β . Therefore, we recommend that the experimental design for each stage of the drug discovery process be chosen in such a way that it allows for estimating the properties of compounds of interest at that stage. In the majority of situations these can be defined as compounds with potencies within a chosen range, say between α_L and α_U . For example, there are two distinct possibilities about the experimenter's interest in these compounds:

- Case (A): all of similar importance;
- Case (B): importance increases as the potency gets closer to α_L .

It is clear from the discussion in Section 1 that Case (A) covers typical situations at the start of searching for useful compounds, while Case (B) is typical later in the drug discovery process, when ranking the tested compounds with respect to their characteristics is the focal purpose of the experiment. In Case B the difference between α_L and α_U is much smaller than in Case (A). In both cases while it is possible to specify $\alpha_L = -\infty$ it is not desirable to

do so, because such an unrealistic expectation would unnecessarily increase the cost of the experiment. This is illustrated with an example later in the paper.

Similarly, values characterizing the range of acceptable values for the Hill slope β , say β_L and β_U , can be specified. Then α_L , α_U , β_L and β_U define the set \mathfrak{R} of compounds of interest in the study. The statistical properties of the estimates of the model parameters $\hat{\alpha}$ and $\hat{\beta}$ depend on the experimental design that is used to obtain the data, as well as on the values of α and β . Therefore, an experimental design that allows for studying sufficiently well all compounds that belong to \mathfrak{R} has to be chosen.

We illustrate the typical statistical challenges and the methods of addressing them with two examples of real applications that arose in pharmaceutical research. For reasons of confidentiality, we cannot reveal the exact details of those experiments or the materials that were involved.

Example 1. A large number of compounds were studied in the early stages of drug development. While small values for IC50 are desirable, in this in vitro study it was recognized that it is unnecessary to spend experimental resources in order to find precisely the IC50 (or α) of compounds that are less than 1 μM . Also, compounds whose IC50 values were bigger than 10 μM were not of interest either. Similarly, it was decided that the desirable range for the Hill slope was between 0.5 and 1.5. Hence, the constraints $\alpha_L = \log(1)$, $\alpha_U = \log(10)$, $\beta_L = 0.5$ and $\beta_U = 1.5$ defined \mathfrak{R} . This is an example of a Case (A) scenario.

Example 2. A medicinal chemist has developed 4 alternative chemical structures for the molecule of a compound that has shown potency during HTS. These new molecules were to be compared with the initial one. For the initial compound, α and β had been estimated to be $\log(3)$ and 1.2, respectively. It was hoped that the potency of the new compounds would be

similar or better, but a possibility that it could be somewhat worse could not be ruled out. The value of β for the new compounds was expected to be similar to that of the standard compound. Hence, \mathfrak{R} was defined by $\alpha_L = \log(1)$, $\alpha_U = \log(3.5)$, $\beta_L = 0.9$ and $\beta_U = 1.4$. Testing whether the structure change has a significant effect on potency and obtaining the correct rank order of the five compounds with respect to their potencies would allow the experimenter to select the best compound to study further in vivo. This is an example of a Case (B) scenario.

3. Criteria of design optimality

There are many ways to characterize the optimality of experimental designs; for a comprehensive review of the theory of optimal experimental designs and their applications, see Atkinson et al (2007). Widely used is the D-optimality criterion. It requires the volume of the confidence ellipsoid for the estimates of the model parameters to be minimized. This is mathematically equivalent to maximization of the determinant of the information matrix of the design (page 136, Atkinson et al, 2007).

The 4×4 asymptotic information matrix for a design with n observations ξ_n is

$$\mathbf{M}(\xi_n, \alpha, \beta, \gamma, \delta) = (\mathbf{X}^T \mathbf{W} \mathbf{X}) \sigma_\varepsilon^{-2}, \quad (3.1)$$

where for model (2.1) \mathbf{W} is a diagonal matrix whose diagonal elements are the inverses of the expectations $1/f(x_i, \xi_n, \alpha, \beta, \gamma, \delta)$; \mathbf{X} is the $n \times 4$ Jacobian matrix for the design, so that the i th element of its four columns are

$$\frac{\partial f(x_i, \xi_n, \alpha, \beta, \gamma, \delta)}{\partial \alpha} = \frac{(\gamma - \delta) 10^{(x_i - \alpha)\beta} \beta \ln(10)}{(1 + 10^{(x_i - \alpha)\beta})^2}$$

$$\frac{\partial f(x_i, \xi_n, \alpha, \beta)}{\partial \beta} = \frac{(\gamma - \delta) 10^{(x_i - \alpha)\beta} (x_i - \alpha) \ln(10)}{(1 + 10^{(x_i - \alpha)\beta})^2}$$

$$\frac{\partial f(x_i, \boldsymbol{\theta})}{\partial \gamma} = \frac{10^{(x-\alpha)\beta}}{1+10^{(x-\alpha)\beta}}$$

$$\frac{\partial f(x_i, \boldsymbol{\theta})}{\partial \delta} = \frac{1}{1+10^{(x-\alpha)\beta}},$$

respectively; as before, x_i is the logarithm at base 10 of the dose used in the i th observation and $\ln(10)$ denotes the natural logarithm of 10.

As the elements of $\mathbf{M}(\xi_n, \alpha, \beta, \gamma, \delta)$ depend on the true values of the model parameters, its determinant

$$D = |\mathbf{M}(\xi_n, \alpha, \beta, \gamma, \delta)| \quad (3.2)$$

takes different values for different compounds studied with the design ξ_n .

There are several possible ways to assess the D -optimality of serial dilution experimental designs that will be used to study a set, or a *population*, of compounds with different characteristics defined by \mathfrak{R} . As a primary version of this criterion we propose a natural extension of the standard criterion where the D values are integrated over \mathfrak{R} . In practice, instead of integrating over a continuous space it is easier to calculate the D values, as defined by equation (3.2), over a fine grid Θ of possible values for α and β in \mathfrak{R} and use the weighted average

$$D_{\text{ave}} = \sum_{\alpha_i, \beta_i \in \Theta} w_i |\mathbf{M}_i(\xi_n, \alpha_i, \beta_i, \gamma, \delta)|$$

as a measure of the population \mathfrak{R} D -optimality of a design ξ_n . No constraints are needed for the values of γ and δ . The choice of weights $w_i, i = 1, \dots, n$ depends on the aim of the experiment. For example, in Case (A) described earlier setting equal weights $w_i = n^{-1}, i = 1, \dots, n$ would be appropriate, while in Case (B) the weights for small values of α_i could be chosen to be larger than when α_i is large.

Another useful characteristic of an experimental design ξ_n is the quantity

$$D_{\min} = \min_{\alpha_i, \beta_i \in \Theta} |\mathbf{M}_i(\xi_n, \alpha_i, \beta_i, \gamma, \delta)|,$$

as it identifies the worst properties with respect to the D-optimality criterion that a compound that belongs to \mathfrak{R} with parameters α_i^* and β_i^* may have.

The first two diagonal elements of the inverse of the information matrix (3.1) are the variances of the estimates of α and β , $\sigma_{\hat{\alpha}}^2 = \sigma_1^2 \sigma_{\varepsilon}^2$ and $\sigma_{\hat{\beta}}^2 = \sigma_2^2 \sigma_{\varepsilon}^2$, respectively, while the corresponding off diagonal element is their covariance. Note that σ_1^2 and σ_2^2 depend only on the experimental design, while σ_{ε}^2 is common for all compounds tested in the same experimental conditions and, hence, has to be estimated.

Example 1 (continued). Figure 1 shows a contour plot for σ_1^2 when a serial dilution design with MD=30, DF=2 and ND=10 has been used. Note that α will be estimated better for compounds with low potency and the precision varies considerably with the Hill slope. In fact, there are over ten fold differences between the values for σ_1^2 for compounds with the desired characteristics.

When the interest is in the precision for estimating a subset of the model parameters, the resulting criterion is called D_S -optimality (page 138, Atkinson et al, 2007). As discussed earlier, estimating α precisely for all compounds is most important. Therefore defining conditions for population D_S -optimality of serial dilution designs that focus on the precise estimation of α for the studied compounds is also useful. Similarly to the way the population D-optimality was defined above, we define the following measures for the D_S -optimality of a population of compounds defined by \mathfrak{R} :

$$S_{\text{ave}} = \sum_{\alpha_i, \beta_i \in \Theta} w_i \sigma_1^2$$

and

$$S_{\max} = \max_{\alpha_i, \beta_i \in \Theta} \sigma_1^2.$$

The weights $w_i, i = 1, \dots, n$, can be chosen as discussed earlier.

Whenever the experimenter needs to rank compounds with respect to their potencies, or to compare a standard compound to one or more new candidate drugs, it is useful to define the minimum significant ratio (MSR) between IC50 values of different compounds, or the minimum significant difference (MSD) between LogIC50 values of different compounds that can be detected using each of the compared designs. Experience (Eastwood et al, 2006) with analysing such data arising from different research areas suggests that the estimates of α are usually approximately normally distributed. Then, the statistic

$$Z = \frac{d\sqrt{r \times ND}}{\sqrt{\hat{\sigma}_{1i}^2 + \hat{\sigma}_{1j}^2}}, \quad (3.3)$$

can be used to summarize the evidence whether or not compounds i and j have different potencies. In (3.3) $\hat{\sigma}_{1i}^2$ and $\hat{\sigma}_{1j}^2$ are the variances for the estimates of α for compounds i and j ,

respectively, and $d = \hat{\alpha}_i - \hat{\alpha}_j$, with standard error $\sqrt{\frac{\hat{\sigma}_{1i}^2 + \hat{\sigma}_{1j}^2}{r \times ND}}$ in the common case that

compounds i and j are tested independently. If $\hat{\sigma}_e^2$ is obtained using a large number of observations, Z follows approximately a normal distribution. Then, compound j will be declared more potent than compound i if $Z > Z_{1-\nu/2}$, where $Z_{1-\nu/2}$ is the critical value of the standard normal distribution at the specified significance level ν . If the number of the observations is small, Z follows approximately a t -distribution, and the critical value will be larger than $Z_{1-\nu/2}$.

In general σ_{1i}^2 and σ_{1j}^2 are different for any compounds i and j and for the purpose of comparing experimental designs it is easier to use the formula for the minimum significant difference between α_i and α_j that follows from (3.3), i.e.

$$MSD = Z_{1-\nu/2} \sqrt{\frac{\hat{\sigma}_{1i}^2 + \hat{\sigma}_{1j}^2}{r \times ND}}. \quad (3.4)$$

The statistical power θ to declare a difference $\hat{\alpha}_i - \hat{\alpha}_j$ significant, when $\alpha_i - \alpha_j = \lambda_0$, is approximately

$$\theta = \Pr \left(z > Z_{1-\nu/2} - \frac{\lambda_0 \sqrt{r \times ND}}{\sqrt{\hat{\sigma}_{1i}^2 + \hat{\sigma}_{1j}^2}} \right), \quad (3.5)$$

where z is a standard normally distributed variable. The choice of λ_0 depends on the aim of the experimenter.

Eastwood et al (2006) use a similar approach to define the minimum significant ratio (*MSR*) between IC50 values for any two compounds i and j for a particular design. They assume that the standard errors for the compared compounds are the same, say a , and for $\gamma = 0.05$ define $MSR = 10^{2\sqrt{2}a}$. However, as discussed before, in general the standard errors for all compounds are different (see Figure 1 for example) and therefore this quantity could be inaccurate. This approximation can still be useful when the compounds that are compared are believed to have similar potencies, or when the variability across different occasions is considerably bigger than that on different plates, or if estimation processes are used that make the variation of the potency estimate less dependent on the underlying potency.

4. Selecting a serial dilution design

Choosing an experimental design to test a compound on a single plate requires the experimenter to define: the population of compounds that is of interest, \mathfrak{R} ; the maximum dose MD; the dilution factor DF; the number of doses ND and the number of replications at each dose level r . The appropriateness of an experimental design for a particular study can be judged by a chosen statistical criterion (e.g. population D- or D_S-optimality), but it also depends on the cost and the practicalities of the experiment. We illustrate this using the

scenario of Example 1. We choose population D_S -optimality as a primary criterion, as usually the same design is used to study the compounds with different characteristics defined by \mathfrak{R} . This criterion is also easy to interpret and is directly related to the aim of the studies when such designs are to be used. We compare serial dilutions designs obtained for dilution factor $DF = 2, 3, 4$ and 5 , and number of doses $ND = 4, \dots, 20$.

Suppose \mathfrak{R} is defined in such a way that $\alpha_L = \log(1)$, $\alpha_U = \log(10)$, $\beta_L = 0.5$ and $\beta_U = 1.5$, and that $MD = 100$. Figures 2 (a) and 2 (b) show D_{ave} and S_{ave} for serial dilution designs with different dilution factors when there are no replications, i.e. $r = 1$. The values for the optimality criteria for bad designs which do not allow for α and β to be estimated are not shown. D_S -optimal designs also perform well with respect to the D-optimality criterion. Also, those with small number of doses are worse with respect to both criteria. The designs improve as ND and n increase, but both D_{ave} and S_{ave} converge to values depending on MD and DF . We refer to these values as *limits*. These limits can be found precisely for any values for MD and DF , but this has little practical importance and is beyond the scope of this paper.

Figures 2 (a) and 2 (b) show that if there are no limitations on the experimental resources, using $DF = 2$ and large values for ND will always lead to designs that are best with respect to the D- and D_S -optimality criteria. However, balance has to be struck between the design optimality and the cost of the experiment which increases with the number of doses tested. It can be seen that increasing ND beyond 15 does not lead to a noticeable improvement in the accuracy of estimating the potencies of the compounds. However, when $ND < 15$, the designs obtained using $DF = 3$ are better than those obtained with $DF = 2$.

Using small number of doses, say 7 or 8, may look attractive because it requires much smaller number of observations than when the number of doses is 15 or more. This is likely to shorten the experiment and therefore cut its cost. Such designs are possible to use if the precision of the results is considered satisfactory. This can be judged by looking at the values of the other two population criteria of optimality, D_{min} and S_{max} , as they are concerned

with the compounds for which the statistical properties of the estimates of their parameters are worse. Figure 3 shows the values of these criteria for the same designs as in Figure 2. Indeed, using smaller number of doses may give reasonable accuracy in earlier stages of drug discovery when the detection of biological activity is much more important than its accurate evaluation.

It can be shown that, when replicates are assumed to be independent, doubling the number of the observations of the response at each dose level halves S_{ave} . If replicates are positively correlated (as is admittedly often the case) then the value of replication will be less, but nevertheless the upper bound makes it easy to compare designs with different numbers of replications and different numbers of doses on the plate. For example, when $DF=2$ and $ND=20$, $S_{ave}=0.69$, while when $DF=2$ and $ND=10$, $S_{ave}=1.54$. Replicating the latter design reduces S_{ave} only to 0.77, hence, the former design is better. However, when $DF=3$ and $ND=10$, $S_{ave}=1.33$, and replicating this design gives $S_{ave}=0.67$. This means that, even if only marginally, replicating the design with $DF=3$ and $ND=10$ is better with respect to the population D_s -optimality criterion if 20 observations of the response can be permitted. However, while the design giving $S_{ave}=0.67$ requires the use of 300 units of each tested compound, that with $DF=2$ and $ND=20$ and $S_{ave}=0.69$ requires only 200 units. This difference is likely to significantly lower the cost of the experiment with a negligible loss of optimality; hence the latter design is likely to be preferred by the experimenters after all.

The two designs compared so far cover different dose ranges. If the experimenter would like to cover the dose range of the design with $DF=2$ and $ND = 20$, then the replicated design with $DF=4$ and $ND=10$ should also be considered as these designs have approximately the same dose range. The latter design has $S_{ave}=0.67$, hence it ensures more precise estimation of the compounds potencies. However, it is more expensive than the former design as it requires the use of 267 units of the tested compounds.

A similar approach can be used to decide what maximum dose to be used in an experiment. The properties of serial dilution designs with the same dilution factor but with different maximum doses then have to be compared. For example, Figures 4 (a) and 4 (b) show D_{ave} and S_{ave} for serial dilution designs when MD=30, while \mathfrak{R} is unchanged. Similar trends that were seen in Figures 2 (a) and 2 (b) can be seen again. Comparing various options when again 20 observations are allowed shows that replicating the design with DF=3 and ND=10 is best. For this design $S_{ave} = 0.62$, even smaller than S_{ave} for the best design using maximal dose of 100. Similarly, the best design with DF = 2 has $S_{ave} = 0.66$, but is potentially considerably cheaper than the former design. Hence, using smaller maximal dose in this case leads to a serial dilution design that would not compromise the quality of the data, while requiring 70% less of the studied compounds compared to the design where MD = 100 is used - a savings that cannot be ignored.

The design problem can be defined in ways different than those discussed here. For example, there might be situations when the total number of wells that can be used to study a compound are predetermined, while the experimenter is free to choose MD. Solution can be found in a similar way by redefining the optimization problem. Clearly, as the requirements for the experiment and the limitations change, so will the serial dilution design that will be most appropriate.

A clear idea of the aim of the study and good scientific knowledge are required to define \mathfrak{R} well. If \mathfrak{R} is unnecessarily large, the experimental resources required to achieve sufficient population precision of the results increase. For example, Figures 5 (a) and 5 (b) show D_{ave} and S_{ave} for serial dilution experimental designs when MD=30 but α_L is reduced from $\log(1)$ to $\log(0.01)$. Serial dilution designs with DF = 3 ensuring similar properties to the best designs seen so far have 11 or even 12 doses, compared to the required 10 doses when $\alpha_L=1$. Hence, 4 more observations are needed for each tested compound in order to obtain results with comparable accuracy. On the other hand, the design obtained with DF = 2

requires $ND = 24$ in order to match S_{ave} of the best design with the smaller \mathfrak{R} , i.e. again 4 observations more. As the design is likely to be used to study a number of compounds, such a difference may correspond to a considerable difference in the required experimental cost by both designs. Therefore, using knowledge about the likely properties of the studied compounds to define \mathfrak{R} accurately may ultimately reduce the cost of the study.

5. Discussion

The results presented in this paper were based on specific assumptions that will not always be satisfied. For example, as one of the referees pointed out, in binding assays the slope parameter β must in theory be equal to 1. Hence, model (2.1) can be simplified by setting $\beta=1$. The criteria of optimality will have to be modified accordingly. Once this modification has been made, though, the choice of a serial dilution design can be made in the same way.

There are situations when difficulties occur during the experiment and outliers have to be removed from the data before analysis. Clearly the statistical properties of estimates of the model parameters of interest will no longer be as intended. If such problems are anticipated before the experiment, further care is needed to ensure higher robustness of the experimental design, for example by increasing the number of replications.

Similarly, Eastwood et al (2006) pointed out that the estimates of the variance based on the model fit may seriously underestimate the true population variance. This problem can be considerably reduced by testing compounds of interest on different plates and on different occasions. The same serial dilution design can be used each time.

One may certainly wonder whether designs based on Bayesian D- and D_S - optimality criteria would outperform the serial dilution designs considered in this paper. Extensive empirical simulation studies carried out by the authors provided clear evidence that there is hardly anything to be gained in terms of design optimality. A complete description of such

studies is beyond the scope of the present research, but the authors have no doubts about its results - namely, a serial dilution design carefully selected according to the principles discussed in this paper will be negligibly inferior to a corresponding Bayesian D- or D_S-optimal design, while being much easier to implement in practice.

As shown in Section 4, using different serial dilution designs can lead to considerable differences in the statistical properties of the results and the cost of the experiment. The approach presented in this paper helps to choose the one that would ensure the results have certain desirable statistical properties. It has already been successfully used in various biological studies. We hope this paper will encourage more researchers to use it and, hence, collect better data in the future.

Acknowledgement

Alexander Donev is grateful to AstraZeneca for the opportunity to gain knowledge about the problem discussed in this paper during the time of his employment within the company.

References

Atkinson, A. C., Donev, A. N. and Tobias, R. D. (2007). *Optimum Experimental Designs, with SAS*. Oxford: Clarendon Press.

Bates, D. M. and Watts, D. G. (2007). *Nonlinear Regression Analysis and Its Applications*. Wiley, New York.

Bhat, J., Rane, R., Solapure, S., Surkar, D., Sharma, U., Harish, M.N., Lamb, S., Plant, D., Alcock, P., Peters, S., Barde, S. and Roy, R.K.: High-Throughput Screening of RNA Polymerase Inhibitors Using a Fluorescent UTP Analog: *J Biom Screen* 2006, 11, 968-976.

Donev, A. N., Tobias, R. D. and Monadjemi, F. (2008). Cost-cautious designs for confirmatory bioassay. *Journal of Statistical Planning and Inference*, 138, p. 3805-3812.

Eastwood, B. J., Farnen, M. W., Iversen, P. W., Craft, T. J., Smallwood, J. K., Garbison, K. E., Delapp, N. W. and Smith, G. F. (2006). The Minimum significant ratio: a statistical parameter to characterize the reproducibility of potency estimates from concentration-response assays and estimation by replicate-experiment studies. *Journal of Biological Screening*, **11**, 253-261.

Finney, D. J. (1978). *Statistical methods in biological assay*. 3rd ed. 69-147 Charles Griffin & Co. London, UK.

Fox, S., Farr-Jones, S. Sopchak, L., Boggs, A., Nicely, H., Khoury, R. and Biros, M.: High-Throughput Screening: Update on Practices and Success. *J Biom Screen* 2006; 11, 864-869.

Lee, M-L., T. and Whitmore, G. A. (1999). Statistical inference for serial dilution assay data. *Biometrics*, **55**, 1215-1220.

Macken, C. (1999). Design and analysis of serial limiting dilution assays with small sample sizes, *Journal of Immunological Methods*, **222**, 13-29.

Mehrabi, Y. and Matthews, J. N. S. (1998). Implementable Bayesian designs for limiting dilution assays. *Biometrics*, **54**, 1398 – 1406.

Parker, G.J., Law, T.L., Lenocho, F.J. and Bolger, R.E.: Development of high throughput screening assays using fluorescence polarization: nuclear receptor-ligand-binding and kinase/phosphatase assays: *J Biom Screen* 2000; 5, 77-88.

Ritz, C. and Streibig, J. C. (2005). Bioassay analysis using R. *Journal of Statistical Software*, Volume 12, Issue 5, 1-22.

Seber, G. A. F. and Wild, C. J. (2005). *Nonlinear regression*. Wiley, New York.

Strijbosch, L. W. G., Does, R. J. M. M. and Buurman, W. A. (1988). Computer aided design and evaluation of limiting and serial dilution experiments. *International Journal of Bio-Medical Computing*, **23**, 279-290.

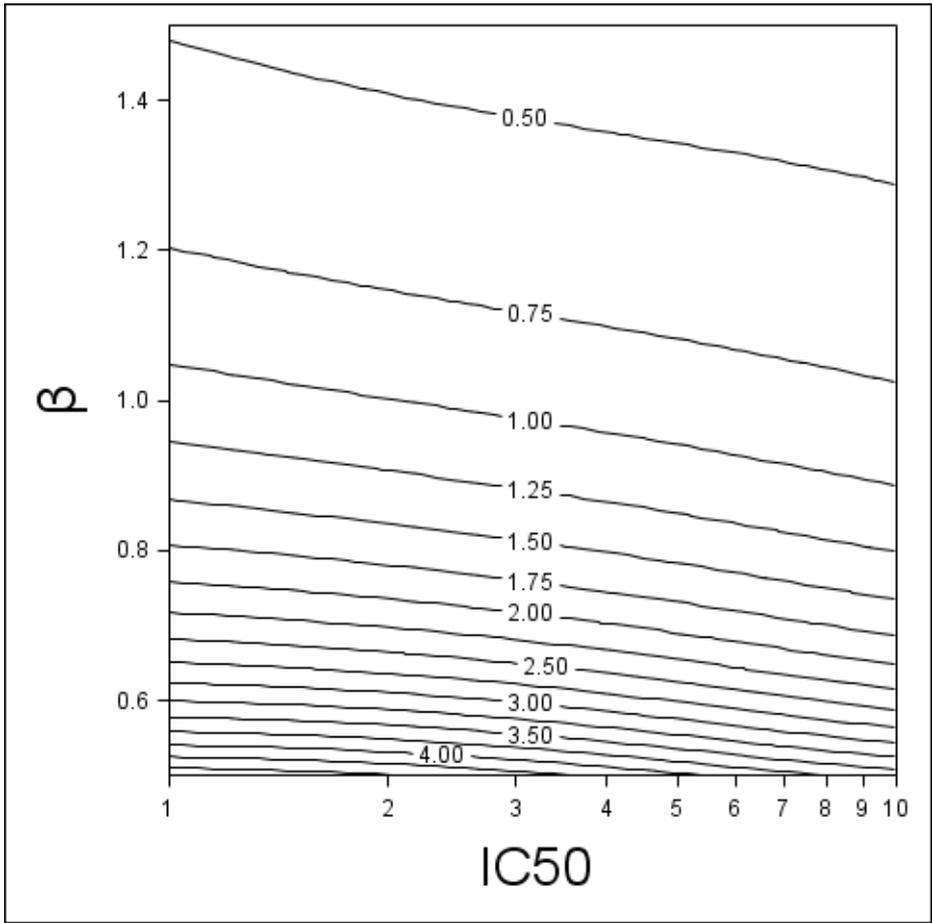
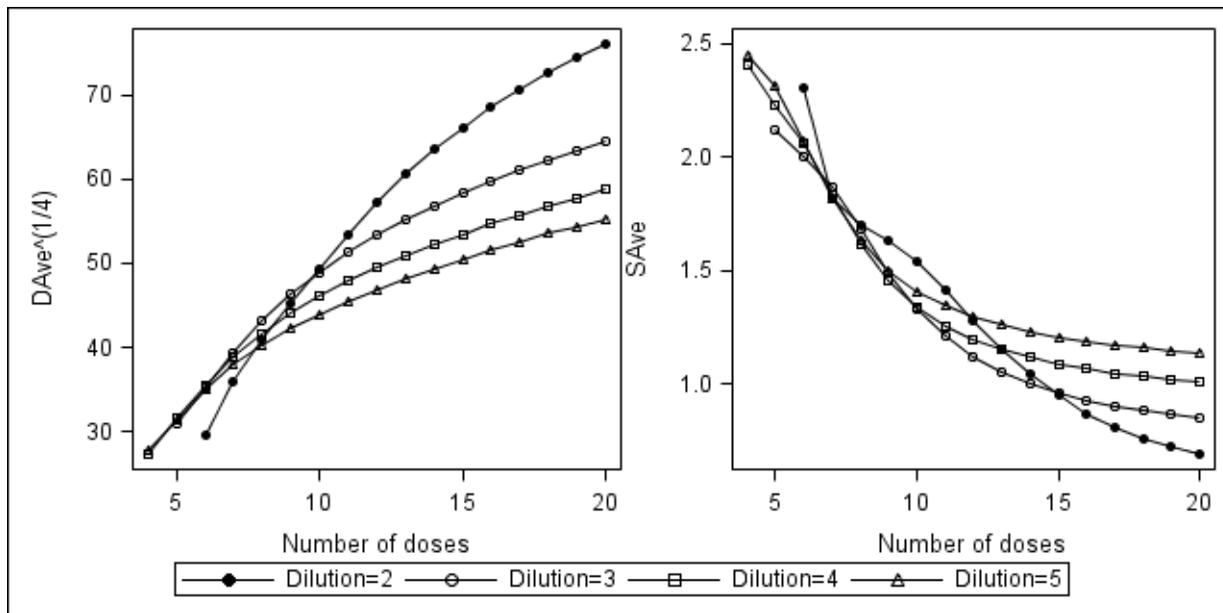


Figure 1. Contour plot for σ_1^2 when a serial dilution design with MD=30, DF=2 and ND=10 has been used.



(a)

(b)

Figure 2. (a) D_{ave} , (b) S_{ave} for serial dilution experimental designs with different dilution factors and a maximum dose of 100. $\alpha_L = \log(1)$, $\alpha_U = \log(10)$, $\beta_L = 0.5$ and $\beta_U = 1.5$.

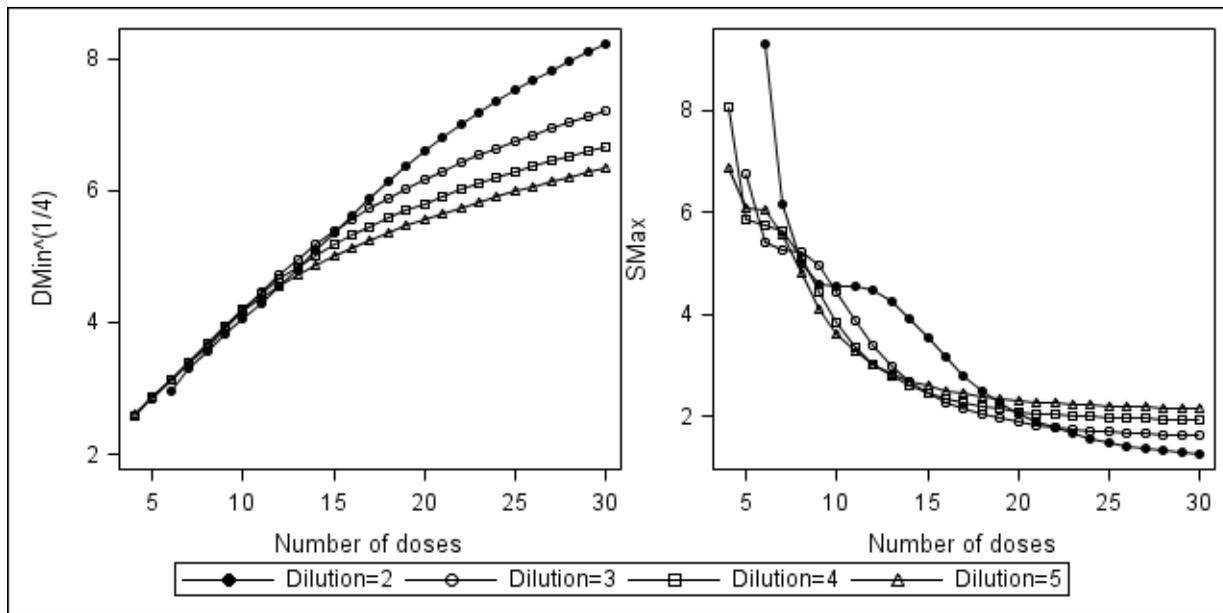
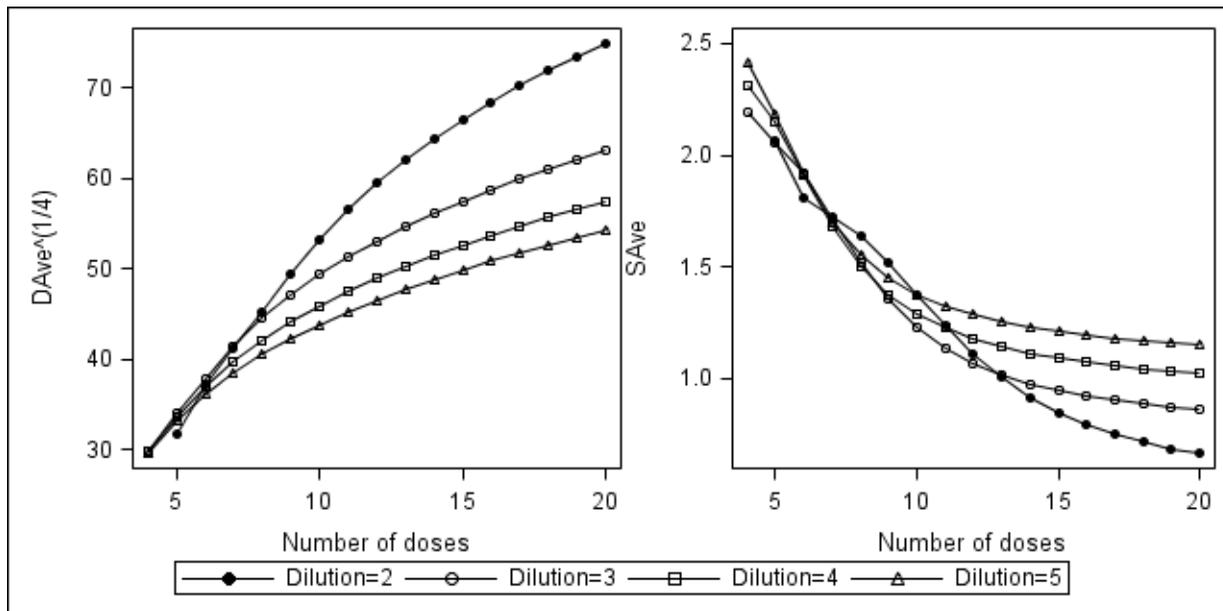


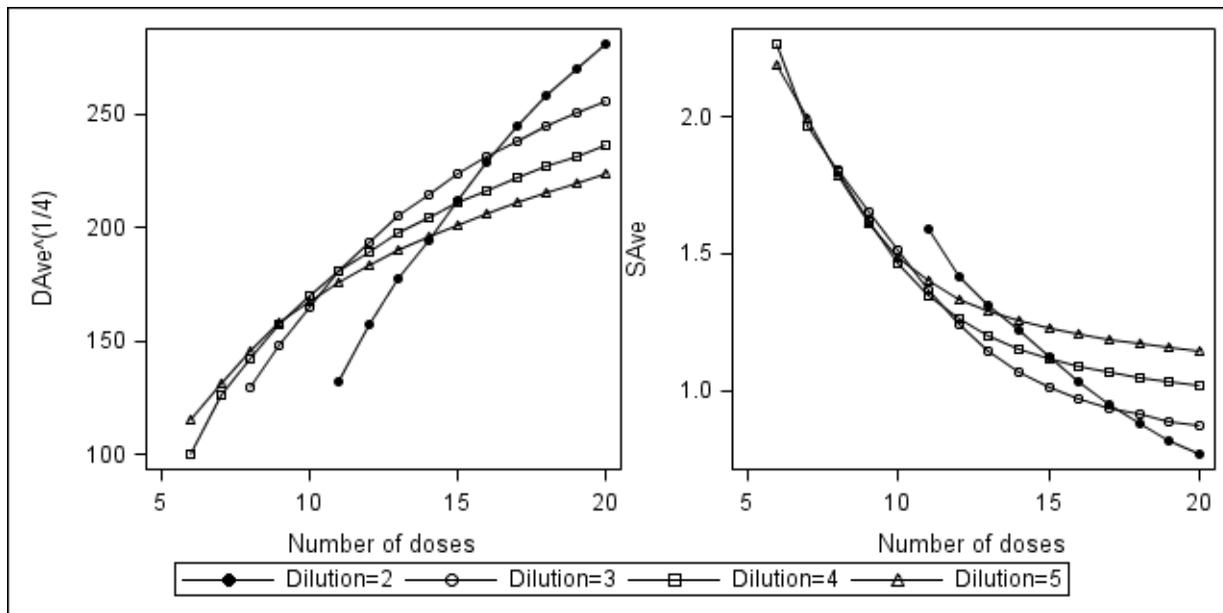
Figure 3. (a) D_{\min} , (b) S_{\max} for serial dilution experimental designs with different dilution factors and a maximum dose of 100. $\alpha_L = \log(1)$, $\alpha_U = \log(10)$, $\beta_L = 0.5$ and $\beta_U = 1.5$.



(a)

(b)

Figure 4. (a) D_{ave} , (b) S_{ave} for serial dilution experimental designs with different dilution factors and a maximum dose of 30. $\alpha_L = \log(1)$, $\alpha_U = \log(10)$, $\beta_L = 0.5$ and $\beta_U = 1.5$.



(a)

(b)

Figure 5. (a) D_{ave} , (b) S_{ave} for serial dilution experimental designs with different dilution factors and a maximum dose of 30. $\alpha_L = \log(0.01)$, $\alpha_U = \log(10)$, $\beta_L = 0.5$ and $\beta_U = 1.5$.