Received: 28 October 2023 | Revised: 21 February 2024 | Accepted: 1 April 2024

DOI: 10.1002/sim.10082

RESEARCH ARTICLE

Statistics in Medicine WILEY

Statistical considerations in model-based dose finding for binary responses under model uncertainty

Zhiwu Yan[1] | **Min Yang**[2]

[1]Biostatistics Department, 89bio, Inc., San Francisco, California, USA

[2]Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, Illinois,

Correspondence
Zhiwu Yan, Biostatistics Department, 89bio, Inc., San Francisco, CA, USA.
Email: andrew.yan@89bio.com

The statistical methodology for model-based dose finding under model uncertainty has attracted increasing attention in recent years. While the underlying principles are simple and easy to understand, developing and implementing an efficient approach for binary responses can be a formidable task in practice. Motivated by the statistical challenges encountered in a phase II dose finding study, we explore several key design and analysis issues related to the hybrid testing-modeling approaches for binary responses. The issues include candidate model selection and specifications, optimal design and efficient sample size allocations, and, notably, the methods for dose-response testing and estimation. Specifically, we consider a class of generalized linear models suited for the candidate set and establish D-optimal designs for these models. Additionally, we propose using permutation-based tests for dose-response testing to avoid asymptotic normality assumptions typically required for contrast-based tests. We perform trial simulations to enhance our understanding of these issues.

KEYWORDS

dose finding, MCP-Mod, model uncertainty, optimal design, permutation test

1 | INTRODUCTION

Drug development is a complex, time-consuming, and expensive process, and identifying the "right" dose is a critical component of this process. In spite of considerable efforts to improve dose finding efficiency throughout drug development, improper dose selection due to limited understanding of the dose-response relationship in phase II remains a key factor behind many failed phase III trials and post-marketing dose adjustments.

To address the need for more informed dose selection, recent statistical methodologies for dose finding have shifted the focus from traditional suboptimal approaches, such as pairwise comparisons, to more efficient hybrid approaches, like MCP-Mod (multiple comparison procedures and modeling).[1,2] A hybrid approach combines the advantages of both multiple testing and modeling, thereby overcoming the limitations of each approach alone. Both the European Medicines Agency (EMA)[3] and the US Food and Drug Administration (FDA)[4] have qualified MCP-Mod as an efficient statistical method for phase II dose finding studies. These regulatory endorsements have since promoted the use of hybrid approaches as a principled strategy for adequate dose exploration in drug development.

Despite the support of the regulatory agencies, several important statistical issues surrounding the hybrid approaches remain largely unaddressed in the literature. A common issue at the design stage, for example, is how to select and specify the candidate models with limited pharmacological data/dose-response information. Specifically, what types of models should be included in the candidate set, how many models need to be considered and how to specify each of them? There seems to be a lack of justified approaches to addressing these questions. A second issue associated with the hybrid approaches is the sample size allocations (to different dose groups). An efficient sample size allocation can

often be obtained based on the knowledge of optimal designs. While numerous exemplary works can be found in the field of optimal designs for dose finding studies,[5–9] there is a notable research gap in addressing optimal design issues related to model uncertainty. This gap is likely attributed to the mathematical complexities associated with the multiple dose-response models, particularly for non-normal responses. A third important issue is how to select appropriate methods for dose-response testing and estimation. The MCP-Mod approach offers a versatile framework for accommodating general parametric models and general study designs.[2] However, the contrast-based tests and generalized least squares (GLS) estimates underlying this approach require asymptotic normality assumptions, which can be severely violated in situations beyond the normal data. It is well known, for example, that the normal approximation to the binomial distribution may produce very poor results in cases where the sample size is small, or the binomial probability is too low or too high. In such scenarios, an alternative method not reliant on normality assumptions is likely to be more suitable for dose-response testing and estimation. An example of such an approach is the permutation testing and modeling (PTM) method introduced by Klingenberg.[10] This method entails performing multiple permutation tests, each based on a penalized likelihood ratio statistic. PTM has apparent advantages over MCP-Mod. First, the use of permutation tests ensures more robust and reliable small sample size inferences in non-normal situations. Second, the use of likelihood ratio based statistics can typically better inform on goodness of fit, potentially leading to more powerful significance testing. Although a direct comparison between these two approaches is currently not available, it is reasonable to expect that PTM is at least comparable to or, in certain situations, even superior to MCP-Mod for dose-response testing and estimation. This article aims to explore these issues, about which little is known.

In the subsequent sections, we will investigate the aforementioned design and analysis issues in the context of a phase II dose finding study with binary responses. We consider a class of generalized linear models studied by Pinheiro et al[2] and establish D-optimal designs for these models under mild regularity conditions. It turns out that the D-optimal designs, if they exist, are the same for all models with monotone dose-response shapes in this class. This implies the potential to identify a sample size allocation that is simultaneously efficient for multiple (or all) candidate models, a crucial consideration in designing an efficient dose finding study under model uncertainty. To deepen our understanding, we perform power simulations to illustrate the impact of different sample size allocations. Additionally, we conduct simulations to compare the operating characteristics (type I error and power in declaring statistical significance) between MCP-Mod and PTM. As anticipated, the simulation study yields clear and compelling evidence favoring the PTM approach for dose-response modeling. Lastly, we also assess the performance of three different methods for target dose estimation in the simulation study.

In Section 2, we introduce the phase II dose finding study that motivated this research. We briefly describe dose-response testing for MCP-Mod and PTM in Section 3. In Section 4, we provide the rationale and details of candidate model selection and specifications. D-optimal design results are given in Section 5, followed by trial simulations in Section 6. We conclude with a short discussion in Section 7. Technical proofs are provided in the Appendix.

2 | A MOTIVATING STUDY

This research was motivated by the statistical challenges presented in a phase II dose finding study in patients with compensated cirrhosis due to nonalcoholic steatohepatitis (NASH). The primary efficacy endpoint of the study was a binary indicator for liver histological improvements after 48 weeks of treatment, and the primary objective of the study was to obtain the dose-response information of the experimental drug and identify one or two dose levels for future phase III confirmatory trials. Four dose levels, including 0 (placebo), 0.3, 1, and 3 mg, were selected for this study based on the sponsor's historical data in noncirrhotic NASH patients. The planned total sample size of 150 patients (or 120 completers, assuming a 20% dropout rate) was determined primarily based on the study budget. However, the study team was also optimistic that this sample size would provide adequate power to demonstrate a drug effect, assuming a response rate of approximately 10% for placebo and 35% for the 3 mg dose. Given the small sample size and considerable uncertainty regarding the size and dose-response shape of the drug effect, the study team was highly motivated to develop efficient statistical methods to ensure the success of the trial.

3 | DOSE-RESPONSE TESTING UNDER MODEL UNCERTAINTY

In this section, we provide a brief overview of dose-response testing for binary responses under model uncertainty for both MCP-Mod and PTM. We refer the reader to Pinheiro et al[2] and Klingenberg[10] for further details.

3.1 | MCP-Mod approach

We assume that the dose-response relationship for candidate model $m \in \{1, 2, \ldots, M\}$ can be expressed as

$$g(\pi(d)) = \alpha_m + \beta_m f_m(d, \boldsymbol{\theta}_m), \tag{1}$$

where $g(\cdot)$ is a link function (eg, logit or identity link), $\pi(d)$ is the response rate at a continuous dose level d, α_m and β_m are the unknown model parameters, $f_m(d, \boldsymbol{\theta}_m)$ denotes the so-called standardized model function[2] and $\boldsymbol{\theta}_m$ its parameter vector. It should be noted that the function $f_m(d, \boldsymbol{\theta}_m)$ (including the guesstimates for $\boldsymbol{\theta}_m$) needs to be completely specified a priori for each candidate model so Equation (1) actually defines a class of generalized linear models (GLMs).

To proceed, an analysis of variance (ANOVA) parametrization is also required to allow a separate parameter to represent the dose-response at each test dose (ie, dose is treated as a classification variable). Testing the dose-response trend under each candidate model is then accomplished by testing a linear contrast of the dose-response parameters under the ANOVA parametrization. This creates a mechanism for efficient multiplicity adjustment (associated with the multiple candidate models) since all linear contrasts are tested under the same ANOVA-type model. Specifically, let $d_1, \ldots, d_k$ denote the k distinct doses used in a trial, then the response model under the ANOVA parametrization can be written as

$$g(\pi(d_i)) = \mu_{d_i}, \tag{2}$$

where μ_{d_i} denotes the treatment effect for dose $d_i, i = 1, \ldots, k$.

Let $\widehat{\boldsymbol{\mu}}$ denote the vector of the ANOVA estimates of $\boldsymbol{\mu} = (\mu_{d_1}, \ldots, \mu_{d_k})'$ obtained using an appropriate estimation method (eg, maximum likelihood), where the prime symbol ($'$) indicates the transpose of a matrix (or vector). We assume that $\widehat{\boldsymbol{\mu}}$ follows an approximate normal distribution $N_k(\boldsymbol{\mu}, \boldsymbol{S})$, where $\boldsymbol{S}$ denotes the variance-covariance matrix of $\widehat{\boldsymbol{\mu}}$. The optimal contrast coefficients for candidate model m, denoted by a column vector $\boldsymbol{c}_m^{opt}$, can be obtained using the following proportional relationship

$$\boldsymbol{c}_m^{opt} \propto \boldsymbol{S}^{-1}\left(\boldsymbol{f}_m - \frac{\boldsymbol{f}_m' \boldsymbol{S}^{-1}\mathbf{1}}{\mathbf{1}'\boldsymbol{S}^{-1}\mathbf{1}}\mathbf{1}\right), \tag{3}$$

where $\boldsymbol{f}_m = (f_m(d_1, \boldsymbol{\theta}_m), \ldots, f_m(d_k, \boldsymbol{\theta}_m))'$ and $\mathbf{1}$ denotes the column vector of 1s. The matrix $\boldsymbol{S}$ is often unknown in practice but can be replaced by an estimate $\widehat{\boldsymbol{S}}$ from the observed data. The test statistic used for establishing an overall dose-response signal is $z_{(M)} = \max_m |z_m|$, the maximum of the individual model test statistics $|z_m|$, where

$$z_m = \frac{(\boldsymbol{c}_m^{opt})'\widehat{\boldsymbol{\mu}}}{\sqrt{(\boldsymbol{c}_m^{opt})'\widehat{\boldsymbol{S}}(\boldsymbol{c}_m^{opt})}}, m = 1, \ldots M.$$

The critical values for $z_{(M)}$ can be derived from the joint distribution of $z = (z_1, \ldots, z_M)'$, an approximate normal distribution $N_M(\mathbf{0}, (\boldsymbol{C}^{opt})'\boldsymbol{S}\boldsymbol{C}^{opt})$, where $\mathbf{0}$ denotes the column vector of 0s and $\boldsymbol{C}^{opt} = (\boldsymbol{c}_1^{opt}, \ldots, \boldsymbol{c}_M^{opt})$, under the null hypothesis of no dose-response effect.

3.2 | PTM approach

The PTM approach described in this section applies to general binary response models. For the purpose of this article, we are interested in the class of GLMs defined in Section 3.1 only. For each candidate model $m \in \{1, 2, \ldots, M\}$, we define a test statistic

$$T_m = -2\log(L_0/L_m) + 2(k_0 - k_m), \tag{4}$$

where L_0 and L_m denote the maximized binomial likelihood under the null model of no dose effect and model m, respectively, and k_0 and k_m are the corresponding number of unknown parameters in these two models. Unlike the one-sided test adopted by Klingenberg,[10] the statistic T_m defined here is two-sided.

With a random sample of B permutations of the observed data, the p-value for the observed test statistic $T_m^{obs}, m = 1, \ldots, M$, is given by

$$p_m^{obs} = \frac{1}{B}\sum_{b=1}^{B} I\left(T_m^{(b)} \geq T_m^{obs}\right), \tag{5}$$

where $I(\cdot)$ denotes the indicator function, and $T_m^{(b)}$ is the value of T_m obtained under the bth permutation, $b = 1, \ldots, B$. Likewise, the p-value for $T_m^{(b)}$ is given by

$$p_m^{(b)} = \frac{1}{B}\sum_{\ell=1}^{B} I\left(T_m^{(\ell)} \geq T_m^{(b)}\right) \tag{6}$$

The multiplicity adjusted p-value for an overall dose-response signal is then given by

$$p = \frac{1}{B}\sum_{b=1}^{B} I\left(\min_m p_m^{(b)} \leq \min_m p_m^{obs}\right). \tag{7}$$

Multiplicity adjusted p-values for the individual models can be obtained using the step-down procedure described in Westfall and Young.[11]

4 | CANDIDATE MODELS

The GLMs defined in Section 3.1 offer several compelling advantages, including (i) these models are not susceptible to overfitting or convergence problems, which is particularly relevant for dose finding studies with small to moderate sample sizes; (ii) these models are sufficiently flexible to accommodate a broad and diverse range of dose-response shapes – through proper selection and specifications of the standardized model function $f_m(d, \boldsymbol{\theta}_m)$; and (iii) the D-optimal designs for different models are the same under mild conditions. We will provide further elaboration on points (ii) and (iii) in the subsequent sections as they are critical in addressing model uncertainty issues.

4.1 | Points to consider

Emax models have a strong foundation in pharmacology research and are frequently used to model dose-response relationships. Empirical evidence from published literature, such as Thomas et al,[12] has demonstrated that dose-response relationships for drug compounds can often be adequately characterized by hyperbolic Emax models (ie, the Hill parameter = 1 in a sigmoid Emax model, thereafter referred to as "Emax models" for simplicity). While this suggests that Emax models can be used to account for a significant portion of model uncertainty, it is still advisable to include other types of models in the candidate set as a routine practice. Since Emax models only represent monotone and concave dose-response shapes, it may be beneficial to also consider models with other different shapes, such as exponential (convex shape), logistic (sigmoid or "S" shape) and quadratic (inverted "U" or bell shape) models, among others.

4.2 | Model selection and specifications

In practical applications, the function $f_m(d, \boldsymbol{\theta}_m)$ can be determined by initially selecting a suitable functional form, such as an exponential function, to effectively capture the desired dose-response shape. Following this, it is crucial to carefully specify the parameter value(s) within this function, guided by the assumed dose effect. In the subsequent discussion, we use the previously mentioned NASH trial to demonstrate how candidate models can be reasonably selected and specified, even with limited dose-response information.

Based on available pharmacological data, the study team had a strong believe that the dose-response function is monotonically increasing over the range of the test doses. Under this assumption, possible dose-response shapes can be

TABLE 1 Candidate dose-response models (logit link).

Scenario	Model type	Model specification
L-L	Exponential	$\alpha + \beta e^{(d/1.5)}$
L-M	Emax	$\alpha + \beta d/(5.3 + d)$
L-H	Logistic	$\alpha + \beta/[1 + e^{(3-5d)}]$
M-M	Emax	$\alpha + \beta d/(0.7 + d)$
M-H	Emax	$\alpha + \beta d/(0.4 + d)$
H-H	Emax	$\alpha + \beta d/(0.15 + d)$
Bell	Quadratic	$\alpha + \beta(d^2 - 3.6d)$

classified based on the magnitude of the response rate at the intermediate doses relative to the lowest and highest doses. For example, we can simply classify the response rate at each of the two intermediate doses (0.3 and 1 mg) as "L" (low), "M" (medium), or "H" (high) compared to the lowest dose (placebo) and the highest dose (3 mg). As a result, it suffices to consider six different dose-response scenarios: "L-L", "L-M", "L-H", "M-M", "M-H", and "H-H", where the two letters in each scenario represent the response magnitudes of the two intermediate doses. Apparently the classification method described here is by no means a strict one, as the response categories (low, medium and high) can only be loosely defined. Nevertheless, these dose-response scenarios provide a clear rationale for candidate model selection (including the number and types of models in the candidate set) and valuable insights for subsequent model specifications as well.

We observe that the "L-L" scenario is likely to display a convex shape and the "L-H" scenario an "S" shape, so we select an exponential model $f_m(d, \boldsymbol{\theta}_m) = e^{(d/\theta)}$ for the "L-L" scenario and a logistic model $f_m(d, \boldsymbol{\theta}_m) = (1 + e^{(\theta_1+\theta_2 d)})^{-1}$ for the "L-H" scenario. It is clear that each of the other four scenarios can be reasonably characterized by an Emax model $f_m(d, \boldsymbol{\theta}_m) = d/(\theta + d)$. As suggested in Section 4.1, we also include a quadratic model $f_m(d, \boldsymbol{\theta}_m) = d^2 + \theta d$ to safeguard against the possibility of a bell-shaped dose-response curve. Guesstimates for the parameters θ, θ_1 and θ_2 in these models can be obtained based on the initial efficacy assumptions (Section 2) and the "scenario-specific" response rates at the two intermediate doses (with necessary fine-tuning).

All together we consider seven candidate models for the NASH trial and the details are presented in Table 1, Figures 1 and 2.

5 | D-OPTIMAL DESIGN AND EFFICIENT SAMPLE SIZE ALLOCATIONS

Optimal designs play a critical role in model-based statistical inferences. It is well known that the optimality of a design often depends on the statistical model of interest. An optimal/efficient design for one specific model may perform poorly under other different models. This brings up an important question about model uncertainty: can we find a design that is simultaneously optimal or efficient for multiple candidate models? An affirmative answer to this question makes it possible to find efficient sample size allocations for a hybrid approach. Fortunately, this question can be addressed using the "complete class" framework studied in a series of papers.[13–17] In light of this tool, we establish a sufficient condition for the existence of a D-optimal design for the GLMs in Section 3.1, irrespective of the standardized model function $f_m(d, \boldsymbol{\theta})$ and of the unknown parameter values (in general, optimal designs for nonlinear models depend on the unknown model parameters). A D-optimal design minimizes the generalized variance (the expected volume of the confidence ellipsoid) of the parameter estimates. Specifically, we have the following result.

> **Theorem 1.** *Assume that the range of the response rate under model (1) with a logit link is* $[r_1, r_2]$, *where* $r_1 = \pi(D_1)$ *and* $r_2 = \pi(D_2)$. *If* $[r_1, r_2]$ *is a subset of the interval* $[(1 + e^{-x_1})^{-1}, (1 + e^{-x_2})^{-1}]$, *where* $x_2 < x^*$, $x^* \approx 1.5434$ *is the solution to the equation* $(1 - x)e^x + x + 1 = 0$ *and* x_1 *is the solution to the equation* $(x_2 - x)(1 - e^x) - 2(1 + e^x) = 0$, *then a design with one half of the observations at dose* D_1 *and the other half at dose* D_2 *is D-optimal for estimating* (α_m, β_m) *irrespective of the parameter values.*

The proof of Theorem 1 is given in the Appendix. This theorem states that a D-optimal design exists for model (1) as long as the range of the response rate $[r_1, r_2]$ under this model is within the interval $[(1 + e^{-x_1})^{-1}, (1 + e^{-x_2})^{-1}]$. For a

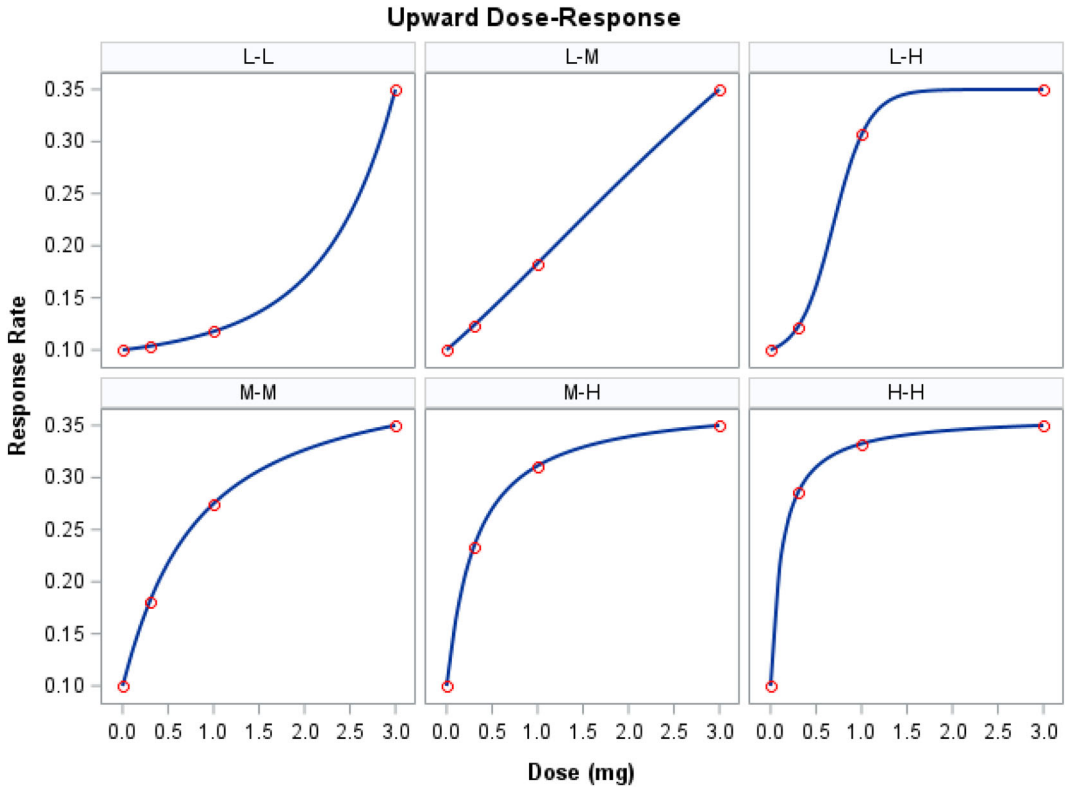

FIGURE 1 The six monotone dose-response models considered for the NASH trial. The red circles denote the responses at the four test doses.

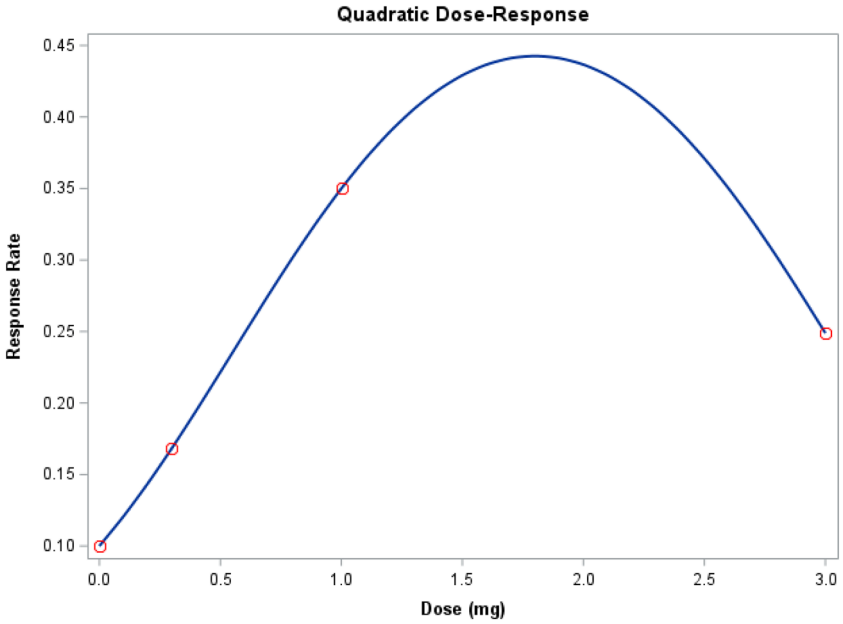

FIGURE 2 The quadratic dose-response model considered for the NASH trial. The red circles denote the responses at the four test doses.

TABLE 2 Numerical examples of interval $[(1+e^{-x_1})^{-1}, (1+e^{-x_2})^{-1}]$.

[0.05, 0.30]	[0.07, 0.40]	[0.09, 0.50]	[0.11, 0.60]	[0.14, 0.70]	[0.17, 0.80]

given x_2 value ($x_2 < 1.5434$), the value of x_1 can be approximated using the Newton-Raphson algorithm. Examples of the numerical results of the interval $[(1+e^{-x_1})^{-1}, (1+e^{-x_2})^{-1}]$ are shown in Table 2.

The optimality of the design in Theorem 1 depends on the range of the response rate, and the numerical results presented in Table 2 make the message more precise: a D-optimal design is likely to exist in many practical situations. For instance, a D-optimal design is applicable to a monotone dose-response model (1) when the efficacy assumptions outlined in Section 2 for the NASH trial hold true, since the range of the response rate [0.10, 0.35] lies within the interval [0.07, 0.40]. In addition, Theorem 1 implies that (i) if the dose-response is monotonic, then a D-optimal design allocates 50% of the observations to both the lowest and the highest doses; and (ii) if the dose-response has a bell shape, then a D-optimal (or D-efficient) design allocates 50% of the observations to the lowest or the highest dose (whichever has the minimum response rate) and another 50% to one of the intermediate doses (whichever has the maximum response rate).

An apparent issue for the D-optimal design in Theorem 1 is that it does not support goodness of fit assessments since the observations are assigned to the two doses D_1 and D_2 only. Practically, the optimal design can serve as a benchmark to facilitate the search for efficient sample size allocations. It is conceivable that the larger the sample size equally allocated to the two doses D_1 and D_2, the more efficient the design is expected to be. This will be verified through trial simulations in Section 6.

As already alluded to, a practical concern during candidate model selection is the potential for overfitting/convergence issues when dealing with small to moderate sample sizes in dose finding studies. Binary models with three or more unknown parameters often demand large sample sizes to ensure stable parameter estimates, particularly in the presence of nonlinear relationships with the model parameters. Moreover, optimal design issues are notoriously challenging for models with non-normal responses, mainly due to the nonlinear components of the unknown model parameters embedded within the Fisher information matrix. The presence of a large number of unknown model parameters further complicates the mathematics due to the increased dimension of the Fisher information matrix. In practice, optimal designs for binary responses are often obtained for particular models and specific values of the unknown model parameters, also known as "locally optimal designs" in the literature. The elegance of Theorem 1 lies in the fact that the D-optimal designs not only apply to the entire class of models (1) but also remain valid irrespective of the unknown parameter values. This ensures that both model and parameter uncertainties are accounted for in dose-response modeling from the optimal design perspective. However, for models with more than two unknown parameters, the lack of results similar to those in Yang and Stufken[13] makes it difficult to establish a counterpart to Theorem 1.

6 | TRIAL SIMULATIONS

In this section, we compare the performance of MCP-Mod and PTM via simulations, with a particular focus on the assessments of type I error rate control and statistical power for dose-response testing. The impact of sample size allocations is also assessed along with the power simulations. In addition, we will touch upon commonly used simple methods for target dose estimation.

Simulation settings, including the total sample size, dose levels, and assumed response rates (for placebo and 3 mg) are based on the study information of the NASH trial described in Section 2. Additionally, different sample sizes and null response rates are also used for the type I error rate assessments. All simulation results are obtained based on 1,000 replications (simulated trials) and 10,000 random permutations for PTM. The statistical significance level is set to be $\alpha = 0.05$ for two-sided tests unless stated otherwise.

6.1 | Type I error rate

To assess the type I error rate control, we consider the null hypothesis of no dose effect $H_0 : \pi(d) = \alpha_0$, where α_0 is a constant (null response rate) for all dose d. Simulations are performed for scenarios based on equal allocations of 15, 30 and 60 patients to each dose group and null response rates of 0.1, 0.3 and 0.5, respectively. These sample sizes are realistic for typical phase II dose-finding studies. The results are presented in Table 3.

TABLE 3 Type I error rate for dose-response testing.

	MCP-Mod			PTM		
Sample size	$\alpha_0 = 0.1$	$\alpha_0 = 0.3$	$\alpha_0 = 0.5$	$\alpha_0 = 0.1$	$\alpha_0 = 0.3$	$\alpha_0 = 0.5$
15/Group	0.004	0.023	0.035	0.056	0.051	0.047
30/Group	0.011	0.028	0.050	0.053	0.041	0.055
60/Group	0.028	0.050	0.046	0.050	0.057	0.051

TABLE 4 Dose-response profiles (logit link).

P1: $\alpha + \beta d$	P2: $\alpha + \beta\sqrt{d}$	P3: $\alpha + \beta \log(1+d)$	P4: $\alpha + \beta/\sqrt{1+d}$
P5: $\alpha + \beta/(1+d)$	P6: $\alpha + \beta e^{e^{(d/max(d))}}$	P7: $\alpha + \beta(d - 2.6\sqrt{d})$	

TABLE 5 Power (in percent) for dose-response testing.

	MCP-Mod				PTM			
Profile	**Plan A**	**Plan B**	**Plan C**	**Plan D**	**Plan A**	**Plan B**	**Plan C**	**Plan D**
P1	63.0	58.9	74.8	80.4	68.3	72.2	79.5	85.1
P2	57.1	44.0	70.3	78.1	65.5	59.7	77.0	83.8
P3	59.2	47.1	72.2	78.1	68.6	62.7	78.9	83.0
P4	57.8	45.6	71.9	74.6	69.1	63.4	78.5	80.3
P5	56.6	42.4	73.7	78.9	70.3	60.8	78.4	83.7
P6	69.4	66.0	76.1	79.1	74.5	76.6	81.4	84.9
P7	50.2	20.4	67.5	73.8	70.7	49.2	77.8	81.6

We observe that PTM controls the type I error rate reasonably well around the nominal level of 0.05 for all scenarios, regardless of the sample size and null response rate. On the other hand, the performance of MCP-Mod demonstrates an apparent dependence on the sample size and response rate. That is, MCP-Mod tends to be overly "conservative" when the sample size per group is small or the response rate is low.

6.2 | Study power

We consider seven different dose-response profiles (models) for power simulations (Table 4). Models P1–P6 are selected from Klingenberg[10] and all have monotone dose-response shapes, while model P7 features a bell-shaped dose-response. In addition, four different sample size plans are used in conjunction with these dose-response profiles for the power assessments. These plans are denoted as follows: Plan A = (30, 30, 30, 30), Plan B = (15, 35, 35, 35), Plan C = (40, 20, 20, 40), and Plan D = (45, 15, 15, 45). The numbers in each plan represent the sample sizes for the four test dose levels: 0, 0.3, 1, and 3 mg, respectively. Plan A assigns an equal number of patients to the four dose levels and is probably the most commonly used sample size allocation method in practice. Plan B assigns an equal number of patients to the three active doses but fewer to the placebo group. This strategy is sometimes favored by the study team due to practical considerations, such as boosting patient enrollment or enhancing the safety profile of the experiment treatment. Plans C and D each assigns an equal number of patients to both the lowest and highest dose levels but considerably fewer to the two intermediate doses. These two plans closely align with the D-optimal design for monotone dose-response models.

Simulation results are presented in Table 5. We observe that (1) PTM consistently outperforms MCP-Mod in all 28 scenarios considered here, irrespective of the sample size plans and of the dose-response profiles; (2) Plan D stands out as the most efficient among these four plans, regardless of the dose-response profiles. This comes as no surprise, especially when considering the monotone dose-response profiles (P1–P6) – it is anticipated that the design efficiency will improve

TABLE 6 Bias and RMSE of the estimated response rate (in percent) for the estimated MED.

		Bias			RMSE		
Profile	**Sample size**	$\widehat{MED}_a$	$\widehat{MED}_s$	$\widehat{MED}_b$	$\widehat{MED}_a$	$\widehat{MED}_s$	$\widehat{MED}_b$
P1	Plan A	−2.5	−2.1	−0.4	4.1	4.3	5.6
	Plan B	−2.7	−2.4	−0.8	4.2	4.4	5.7
	Plan C	−2.9	−2.6	−0.4	4.3	4.4	5.9
	Plan D	−3.1	−2.9	−0.4	4.4	4.5	6.0
P2	Plan A	0.1	0.2	0.5	3.3	3.8	5.4
	Plan B	0.4	0.5	0.6	3.2	3.8	5.4
	Plan C	0.3	0.3	0.8	3.1	3.5	5.6
	Plan D	0.3	0.3	0.9	3.2	3.5	5.9
P3	Plan A	−1.0	−0.8	−0.4	3.8	4.2	5.9
	Plan B	−0.5	−0.3	0.1	3.8	4.4	6.0
	Plan C	−0.7	−0.6	0.4	3.9	4.2	6.5
	Plan D	−0.8	−0.7	0.6	3.8	4.2	6.6
P4	Plan A	−0.2	−0.2	−0.2	3.9	4.3	6.0
	Plan B	−0.3	−0.2	−0.5	3.7	4.2	5.9
	Plan C	0.2	0.2	0.6	3.8	4.2	6.3
	Plan D	0.3	0.3	1.0	4.0	4.4	7.0
P5	Plan A	0.7	0.6	0.2	4.1	4.5	5.8
	Plan B	0.9	0.8	0.3	3.9	4.4	6.0
	Plan C	0.9	1.0	0.7	4.3	4.6	6.4
	Plan D	1.4	1.3	1.3	4.4	4.7	7.0
P6	Plan A	−6.7	−6.4	−4.7	6.9	6.7	5.6
	Plan B	−6.8	−6.4	−4.8	7.0	6.8	5.7
	Plan C	−6.9	−6.5	−4.4	7.0	6.9	5.5
	Plan D	−7.2	−7.0	−4.5	7.4	7.2	5.6
P7	Plan A	5.9	5.2	4.1	6.7	6.2	5.6
	Plan B	6.0	5.1	4.4	6.8	6.1	5.7
	Plan C	6.8	6.2	4.7	7.6	7.0	6.1
	Plan D	7.4	6.9	4.9	8.2	7.7	6.4

with an increased number of patients equally assigned to both the lowest dose (ie, D_1 in Theorem 1) and the highest dose (ie, D_2 in Theorem 1); and (3) Plans A and B both exhibit inefficiency, primarily due to the "waste" of a considerable portion of resources (sample size) when compared to the more efficient Plans C and D. For example, Plan B allocates only 15 patients to the lowest dose (D_1), significantly fewer compared to the other three plans, making it the least efficient plan overall.

6.3 | Dose estimation

Once a statistically significant dose-response effect is established, one proceeds to target dose estimation. We consider estimating the minimum effective dose (MED), defined as the lowest dose that produces a clinically relevant efficacy response.

TABLE 7 Power (in percent) of identifying the restricted MED.

Profile	Sample size	$\widehat{MED}_a$	$\widehat{MED}_s$	$\widehat{MED}_b$
P1	Plan A	48.0	48.0	48.5
	Plan B	43.6	44.5	44.7
	Plan C	49.4	49.9	51.8
	Plan D	49.4	49.4	56.2
P2	Plan A	44.7	42.2	30.2
	Plan B	36.8	33.2	23.5
	Plan C	48.5	47.3	32.5
	Plan D	50.7	49.3	30.5
P3	Plan A	28.2	28.4	29.6
	Plan B	30.0	32.2	32.3
	Plan C	37.6	38.0	39.4
	Plan D	36.6	36.9	41.1
P4	Plan A	44.1	41.8	32.4
	Plan B	41.9	38.6	29.7
	Plan C	49.7	47.9	34.2
	Plan D	47.7	46.8	31.7
P5	Plan A	47.0	44.6	35.0
	Plan B	46.8	42.6	31.7
	Plan C	53.0	51.4	36.2
	Plan D	54.3	53.3	33.9
P6	Plan A	61.5	61.7	62.7
	Plan B	63.9	64.3	63.5
	Plan C	69.1	69.4	70.6
	Plan D	64.2	64.4	68.4
P7	Plan A	33.4	41.5	46.3
	Plan B	23.0	30.1	30.8
	Plan C	27.5	37.5	47.1
	Plan D	24.0	29.9	48.3

The MED is expected to provide statistically significantly superior response compared to placebo.[18] In the framework of the PTM approach, the MED for candidate model $m \in \{1, \ldots, M\}$ can be estimated by[10]

$$\widehat{MED}_m = \arg \min_{d \in (d_1, d_k]} \{\hat{\pi}_m(d) > \hat{\pi}_m(d_1) + \Delta, \hat{\pi}_m^L(d) > \hat{\pi}_m(d_1)\}, \tag{8}$$

where Δ is the clinically relevant effect, $\hat{\pi}_m(d)$ and $\hat{\pi}_m(d_1)$ are the maximum likelihood (ML) estimates of $\pi(d)$ and $\pi(d_1)$ respectively, and $\hat{\pi}_m^L(d)$ is the lower limit of a $100(1-\gamma)\%$ confidence interval for $\pi(d)$. In practice, the significance level γ may be chosen to be the same as or higher than that for dose-response testing.

An overall estimate of the MED can be obtained based on three different simple methods: (i) $\widehat{MED}_b$, the estimated MED from the best-fitting model; (ii) $\widehat{MED}_s$, a weighed average of the estimated MEDs from all statistically significant models; or (iii) $\widehat{MED}_a$, a weighed average of the estimated MEDs from all candidate models. We will compare the performance of these estimation methods.

Simulations in this section are still based on the seven dose-response profiles and the four sample size plans considered in Section 6.2. Additionally, we assume that $\Delta = 0.1$, $\gamma = 0.1$, and use a weight of $\exp(T_m/2)$ for each candidate model,[10] where T_m is defined by Equation (4). Assessments of estimation performance include the bias and root mean squared error (RMSE) of the estimated response rate for the estimated MED.

Simulation results are presented in Table 6. We observe that (1) there appears to be no clear winner/loser between the three estimation methods, that is, none of these methods is uniformly better/worse than the other two in all scenarios considered here; (2) the two model-averaging methods (ie, $\widehat{MED}_a$ and $\widehat{MED}_s$) show a higher level of comparability in terms of both bias and RMSE in contrast to the best-fitting model method ($\widehat{MED}_b$); and (3) overall, the best-fitting model method tends to generate results with lower bias (P1, P3, and P5–P7) but greater variability (P1–P5) compared to the two model-averaging methods.

Note that the MED can be any dose within the interval $(d_1, d_k]$, which is also referred to as an "unrestricted" MED in the literature. Practically, the MED is often identified as one of the test doses, which is then referred to as a "restricted" MED. If an unrestricted MED is between the $(i-1)$th and ith dose levels, then the corresponding restricted MED would be the ith dose level.[19] It is interesting to know how well the three estimation methods can identify the restricted MED.

Table 7 shows the simulated "power" of identifying the restricted MED (ie, the probability that the restricted MED is correctly identified) by each estimation method. We find that, overall, the two model-averaging methods remain comparable and, in general, they are either as effective as (P1, P3, and P6) or even more powerful than (P2, P4, and P5) the best-fitting model method in identifying the restricted MED.

When examining the impact of different sample size strategies on estimating the unrestricted MED, there appears to be no substantial differences among the four plans. Overall, Plan A demonstrates slightly better performance, whereas Plans C and D show slightly inferior performance. This is likely because an equal sample size allocation tends to assign more patients to a dose near the MED. In estimating the restricted MED, Plans C and D show slightly better overall performance, especially when compared to Plan B. The explanation for this is not straightforward due to the discrete nature of the restricted MED. It is important to highlight that optimal designs for dose-response testing and estimation can be very different. In other words, a design that is optimal or efficient for dose-response testing may not necessarily perform well in target dose estimation, and vice versa. Since dose estimation is typically a secondary objective in phase II dose finding studies, sample size allocations are often justified solely based on the power for dose-response testing.

7 | CONCLUSIONS

Hybrid testing-modeling approaches are well suited for model-based dose finding under model uncertainty. Statistical considerations such as candidate model selection and specifications, sample size allocations, methods for dose-response testing and estimation are paramount for the success of such approaches. These issues become prominent and tend to be more challenging in the presence of non-normal responses. This article attempts to address these issues in the context of a dose finding study with binary responses. Our conclusions are summarized as follows.

The class of GLMs (1) proposed by Pinheiro[2] is suitable for a hybrid approach due to its practical, computational and statistical advantages. The flexibility in choosing different types of the standardized model function $f_m(d, \boldsymbol{\theta}_m)$ ensures adequate coverage of plausible dose-response relationships in practice. The computational advantage of such models is well understood due to the nature of GLMs and the small number of unknown model parameters in Equation (1). The existence of simple D-optimal designs for these models is a great advantage for designing an efficient dose finding study. We recommend using Emax, exponential and logistic models to capture monotone dose-response relationships. Additionally, it is a good practice to include other types of models, such as quadratic models, to account for unexpected dose-response scenarios. The standardized model function $f_m(d, \boldsymbol{\theta}_m)$ for these models can be reasonably specified as described in Section 4.2.

Sample size allocations are an important design aspect for a hybrid approach. The D-optimal design results in Section 5 make the search for efficient sample size allocations an easy and straightforward task. Since monotone dose-response relationships are fairly common for most drug compounds, an efficient sample size allocation often requires more subjects in both the lowest and the highest dose groups. It is worth noting that, however, sample size allocations should also take into account important practical issues, such as the goodness of fit assessments and safety/tolerability concerns of the high dose level(s).

For dose-response testing, permutation tests using likelihood ratio based statistics (eg, PTM) are expected to produce more robust and reliable small sample size results compared to contrast-based tests (eg, MCP-Mod). This is supported by the simulations in Sections 6.1 and 6.2. For model fitting, the full likelihood approach (the ML estimate) for PTM is also an advantage compared to the proposed two-stage GLS approach for MCP-Mod, although the latter is computationally more efficient.[2] We have observed somewhat mixed results for MED estimation based on the three different methods.

Estimates of the unrestricted MED based on the best-fitting model ($\widehat{MED}_b$) tend to be less biased but more variable compared to the two model-averaging methods ($\widehat{MED}_a$ and $\widehat{MED}_s$). The two model-averaging methods exhibit an overall better performance than the best-fitting model method in identifying the restricted MED. Since model-averaging based on all candidate models provides a natural and coherent mechanism for addressing model uncertainty, we feel that such a method ($\widehat{MED}_a$) is likely a better choice for dose estimation.

Finally, we point out that the statistical issues discussed here are not necessarily limited to binary outcomes. We believe that the messages delivered in this article would provide useful insights into the design and analysis of model-based dose finding studies, particularly those with non-normal responses.

DATA AVAILABILITY STATEMENT

No data is used to support the findings in this paper.

ORCID

Zhiwu Yan https://orcid.org/0009-0000-3824-8034
Min Yang https://orcid.org/0000-0001-6208-3751

REFERENCES

1. Bretz F, Pinheiro J, Branson M. Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*. 2005;61(3):738-748.
2. Pinheiro J, Bornkamp B, Glimm E, Bretz F. Model-based dose finding under model uncertainty using general parametric models. *Stat med*. 2014;33(10):1646-1661.
3. EMA. Qualification opinion of MCP-Mod as an efficient statistical methodology for model-based design and analysis of phase-II dose finding studies under model uncertainty. https://www.ema.europa.eu/en/human-regulatory-overview/research-and-development/scientific-advice-and-protocol-assistance/opinions-and-letters-support-qualification-novel-methodologies-medicine-development 2014.
4. FDA. Drug Development Tools: Fit-for-Purpose Initiative on MCP-Mod. https://www.fda.gov/drugs/development-approval-process-drugs/drug-development-tools-fit-purpose-initiative 2016.
5. Dette H, Bretz F, Pepelyshev A, Pinheiro J. Optimal designs for dose finding studies. *J Am Stat Assoc*. 2008;103(483):1225-1237.
6. Bretz F, Dette H, Pinheiro J. Practical considerations for optimal designs in clinical dose finding studies. *Stat med*. 2010;29(7-8):731-742.
7. Dette H, Kiss C, Bevanda M, Bretz F. Optimal designs for the emax, log-linear and exponential models. *Biometrika*. 2010;97(2):513-518.
8. Bornkamp B, Bretz F, Dette H, Pinheiro J. Response-adaptive dose-finding under model uncertainty. *Ann Appl Stat*. 2011;5(2B):1611-1631.
9. Biedermann S, Dette H, Zhu W. Optimal designs for dose-response models with restricted design spaces. *J Am Stat Assoc*. 2006;101(474):747-759.
10. Klingenberg B. Proof of concept and dose estimation with binary responses under model uncertainty. *Stat med*. 2009;28(2):274-292.
11. Westfall P, Young S. *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*. Wiley Series in Probability and Statistics. New York: John Wiley and Sons; 1993.
12. Thomas N, Sweeney K, Somayaji V. Meta-analysis of clinical dose-response in a large drug development portfolio. *Stat Biopharmaceut Res*. 2014;6(4):302-317.
13. Yang M, Stufken J. Support points of locally optimal designs for nonlinear models with two parameters. *Ann Stat*. 2009;37(1):518-541.
14. Yang M, Stufken J. Identifying locally optimal designs for nonlinear models: a simple extension with profound consequences. *Ann Stat*. 2012;40(3):1665-1681.
15. Yang M. On the de la Garza phenomenon. *Ann Stat*. 2010;38(4):2499-2524.
16. Dette H, Melas VB. A note on the de la garza phenomenon for locally optimal designs. *Ann Stat*. 2011;39(2):1266-1281.
17. Dette H, Schorning K. Complete classes of designs for nonlinear regression models and principal representations of moment spaces. *Ann Stat*. 2013;41(3):1260-1267.
18. Filloon TG. Estimating the minimum therapeutically effective dose of a compound via regression modelling and percentile estimations. *Stat med*. 1995;14(9):925-932.
19. Zhou Y, Chen S, Sullivan D, et al. Dose-ranging design and analysis methods to identify the minimum effective dose (MED). *Contemp Clin Trials*. 2017;63(12):59-66.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Yan Z, Yang M. Statistical considerations in model-based dose finding for binary responses under model uncertainty. *Statistics in Medicine*. 2024;1-14. doi: 10.1002/sim.10082

APPENDIX

We need the following lemma to prove Theorem 1.

Lemma 1. *Let* $g(x,y) = (y-x)e^{(x+y)/2}(1+e^x)^{-1}(1+e^y)^{-1}$ *and* $x^* > 0$ *be the solution to the equation* $(1-x)e^x + x + 1 = 0$. *Then we have*

(i) *for any* $x \geq -x^*$, $g(x,y)$ *is an increasing function of* $y \in (x, x^*]$;
(ii) *for any* $y \leq x^*$, $g(x,y)$ *is a decreasing function of* $x \in [\phi(y), y)$, *where* $\phi(y)$ *is the solution to the equation* $\frac{\partial g(x,y)}{\partial x} = 0$.

Proof. It can be shown that

$$\frac{\partial g(x,y)}{\partial y} = \frac{e^{(x+y)/2}}{2(1+e^x)(1+e^y)^2} h(x,y),$$

where $h(x,y) = (1-e^y)(y-x) + 2(1+e^y)$.

So $\frac{\partial g(x,y)}{\partial y} = 0$ is equivalent to $h(x,y) = 0$, which implies that $x = y - 2 - 4(e^y - 1)^{-1}$. Note that the function $y - 2 - 4(e^y - 1)^{-1}$ is unbounded and strictly increasing when $y > 0$, thus for any given x value the equation $h(x,y) = 0$ always has a solution $y = \varphi(x) > 0$, which is a strictly increasing function of x. It can be verified that $\varphi(-x^*) = x^*$, thus $\varphi(x) \geq x^*$ for any $x \geq -x^*$.

Since $\frac{\partial^2 h(x,y)}{\partial y^2} = -e^y(y-x) < 0$ for $y > x$, then $h(x,y)$ is a concave function of y when $y > x$. Note that $h(x,y) > 0$ when $y = x$ and $h(x,y) = 0$ when $y = \varphi(x)$, thus $h(x,y) \geq 0$ for any $y \in (x, \varphi(x)]$. Consequently, $h(x,y) \geq 0$ when $x \geq -x^*$ and $y \in (x, x^*]$. This proves (i). The proof of (ii) is similar to that of (i) and thus omitted here. ∎

Proof of Theorem 1. For simplicity, we will omit the subscript m in model (1) as this is not expected to cause any confusion in subsequent proof. We consider an approximate design $\xi = \{(d_i, \omega_i), i = 1, \ldots, k\}$ with $\Sigma_{i=1}^k \omega_i = 1$. The information matrix for (α, β) under model (1) with a logit link can be written as

$$I_\xi = \begin{pmatrix} 1 & 0 \\ -\frac{\alpha}{\beta} & \frac{1}{\beta} \end{pmatrix} \left[\sum_{i=1}^k \omega_i \begin{pmatrix} \Psi(z_i) & z_i\Psi(z_i) \\ z_i\Psi(z_i) & z_i^2\Psi(z_i) \end{pmatrix} \right] \begin{pmatrix} 1 & -\frac{\alpha}{\beta} \\ 0 & \frac{1}{\beta} \end{pmatrix}, \tag{A1}$$

where $\Psi(z) = e^z(1+e^z)^{-2}$ and $z_i = \alpha + \beta f(d_i, \boldsymbol{\theta})$. The assumption about the response rate implies that $z_i \in [\alpha + \beta f(D_1, \boldsymbol{\theta}), \alpha + \beta f(D_2, \boldsymbol{\theta})]$, where $\pi(D_1) = r_1$ and $\pi(D_2) = r_2$. In the context of a locally optimal design, a D-optimal design maximizes the determinant of

$$\sum_{i=1}^k \omega_i \begin{pmatrix} \Psi(z_i) & z_i\Psi(z_i) \\ z_i\Psi(z_i) & z_i^2\Psi(z_i) \end{pmatrix}. \tag{A2}$$

It suffices to prove the conclusion for the following four scenarios:

(i) $r_1 \geq 0.5$;
(ii) $r_1 < 0.5 < 1 - r_1 \leq r_2$;
(iii) $r_1 < 0.5 < r_2 < 1 - r_1$;
(iv) $r_2 \leq 0.5$.

Scenario (i): $\pi(D_1) = r_1 \geq 0.5$ implies that $\alpha + \beta f(D_1, \boldsymbol{\theta}) \geq 0$. By Theorem 2 of Yang and Stufken,[13] an optimal design is based on two support points and one of them is $\alpha + \beta f(D_1, \boldsymbol{\theta})$. Moreover, the weights of the two support points are equal under the D-optimality. It remains to find the second support point. Notice that maximizing the determinant of matrix (A2) is equivalent to maximizing the function $g(x,y)$ defined in Lemma 1, where x and y $(x < y)$ are the two support points and $x = \alpha + \beta f(D_1, \boldsymbol{\theta})$. Since $[r_1, r_2]$ is a subset

of $[(1+e^{-x_1})^{-1}, (1+e^{-x_2})^{-1}]$, then $\alpha + \beta f(D_2, \boldsymbol{\theta}) \leq x^*$. By (i) of Lemma 1, the second support point must be $\alpha + \beta f(D_2, \boldsymbol{\theta})$. Consequently, a design ξ with half of the observations at D_1 and the other half at D_2 is D-optimal.

Scenario (ii): $r_1 < 0.5 < 1 - r_1 \leq r_2$ implies that $\alpha + \beta f(D_1, \boldsymbol{\theta}) < 0$, $\alpha + \beta f(D_2, \boldsymbol{\theta}) > 0$ and $|\alpha + \beta f(D_1, \boldsymbol{\theta})| < \alpha + \beta f(D_2, \boldsymbol{\theta})$. By Theorem 3 of Yang and Stufken,[13] an optimal design is based on two support points and one of them is $\alpha + \beta f(D_1, \boldsymbol{\theta})$. The rest of the proof follows the same arguments as in Scenario (i).

Scenario (iii): $r_1 < 0.5 < r_2 < 1 - r_1$ implies that $\alpha + \beta f(D_1, \boldsymbol{\theta}) < 0$, $\alpha + \beta f(D_2, \boldsymbol{\theta}) > 0$ and $|\alpha + \beta f(D_1, \boldsymbol{\theta})| > \alpha + \beta f(D_2, \boldsymbol{\theta})$. By theorem 3 of Yang and Stufken,[13] an optimal design is based on two support points and one of them is $\alpha + \beta f(D_2, \boldsymbol{\theta})$. In light of (ii) of Lemma 1, the rest of the proof follows the same arguments as in Scenario (i).

Scenario (iv): $r_2 \leq 0.5$ implies that $\alpha + \beta f(D_2, \boldsymbol{\theta}) \leq 0$. By theorem 2 of Yang and Stufken,[13] an optimal design is based on two support points and one of them is $\alpha + \beta f(D_2, \boldsymbol{\theta})$. The rest of the proof follows the same arguments as in Scenario (iii). ■