Robust designs accounting for model uncertainty in longitudinal studies with binary outcomes

Jérémy Seurat, Thu Thuy Nguyen, France Mentré

▶ **To cite this version:**

Jérémy Seurat, Thu Thuy Nguyen, France Mentré. Robust designs accounting for model uncertainty in longitudinal studies with binary outcomes. Statistical Methods in Medical Research, 2019, pp.096228021985058. 10.1177/0962280219850588 . inserm-02146208

HAL Id: inserm-02146208

https://inserm.hal.science/inserm-02146208

Submitted on 16 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Robust designs accounting for model uncertainty in longitudinal studies with binary outcomes

Jérémy Seurat, Thu Thuy Nguyen and France Mentré

IAME, UMR 1137, INSERM, Université Paris Diderot, Sorbonne Paris Cité, Paris, France

Abstract

To optimize designs for longitudinal studies analyzed by mixed-effect models with binary outcomes, the Fisher information matrix (FIM) can be used. Optimal design approaches however require *a priori* knowledge of the model. We aim to propose, for the first time, a robust design approach accounting for model uncertainty in longitudinal trials with two treatment groups, assuming mixed-effect logistic models. To optimize designs given one model, we compute several optimality criteria based on FIM evaluated by the new approach based on Monte-Carlo/Hamiltonian Monte-Carlo (MC/HMC). We propose to use the DD_S-optimality criterion as it ensures a compromise between the precision of estimation of the parameters, and hence the Wald test power, and the overall precision of parameter estimation. To account for model uncertainty, we assume candidate models with their respective weights. We compute robust design across these models using compound DD_S-optimality. Using the FIM, we propose to predict the average power over these models. Evaluating this approach by clinical trial simulations, we show that the robust design is efficient across all models, allowing one to achieve good power of test. The proposed design strategy is a new and relevant approach to design longitudinal studies with binary outcomes, accounting for model uncertainty.

Keywords: Optimal design; Compound optimality; Fisher information matrix; Longitudinal binary data; Nonlinear mixed effect models

Present address for correspondence: J Seurat, IAME, UMR 1137, INSERM, University Paris Diderot Sorbonne Paris Cité, 16, rue Henri Huchard, 75018 Paris, France
e-mail: jeremy.seurat@inserm.fr

1. Introduction

Nonlinear mixed effect models (NLMEMs) are frequently used in model-based drug development to analyze pharmacokinetic/pharmacodynamic data [1]. The use of NLMEMs (*i.e.* the population approach) allows the estimation of mean parameters, inter and intra-subject variabilities as well as covariate effects, and is appropriate for exploiting the richness of longitudinal data [2–4]. NLMEMs can compensate for the lack of individual information by borrowing the strength from the data in the whole population and therefore allow for precise parameter estimation even with sparse designs, where few samples are collected from each subject.

NLMEMs are increasingly used for the analysis of longitudinal clinical studies, for both continuous and discrete data (binary, count or time-to-event). Binary outcomes are frequently encountered to characterize the clinical response in different therapeutic areas such as infectiology (virological success, bacterial carrying), and can be used for responder/non-responder analysis as well as for toxicity analysis. For instance, the carriage of *Streptococcus pneumoniae* was studied in children between 2 months and 2 years, by collecting nasopharyngeal swabs and aspirates over 10 visits to detect the impact of determinants as age or health status on the proportion of positive bacterial samples [5]. In another context, [6] recently proposed a binary outcome (corresponding to a minimal change in UDysRS Part III Impairment scale from baseline) to characterize the remission of dyskinesia in Parkinson's disease. The analysis of the longitudinal data in these kind of outcomes required discrete response generalized linear mixed effect models (GLMMs) or discrete response NLMEMs and adequate estimation methods.

Before modeling, it is crucial to choose an appropriate design in order to obtain good precision of parameter estimates. Indeed, the informativeness of a dataset for parameter estimation depends on the design choice. A design in NLMEMs, called a population design, is composed of the number of elementary designs, the specification of each elementary design and the associated number of subjects. In this settings, the term elementary design is used to describe a group of subjects with identical design characteristics such as the number of samples per subject and the allocation of informative times and doses. To evaluate and optimize designs, two approaches have been proposed. The first approach, based on clinical trial simulation (CTS) is very time-consuming and is therefore limited in term of designs that can be evaluated. Alternatively, the expected Fisher Information Matrix (FIM) can be used [7–9], as its inverse is

defined as the lower bound of the variance-covariance matrix of any unbiased estimated parameters, according to the Cramér-Rao inequality. However, the FIM has no analytical form in NLMEMs and its computation, which requires multiple integrations, can be challenging. A first method to evaluate the FIM in NLMEMs, based on first order (FO) linearization of the model around the expectation of random effects, was proposed [10] and implemented in several software programs [11]. Although efficient in general, FO presents limitations when used with complex models, large variability [12,13] and when using longitudinal models for discrete data [13]. Alternative to compute the FIM in continuous NLMEMs, without linearization, has been proposed using Monte Carlo (MC) integrations [14]. There are also a large body of work on the FIM for GLMMs with random intercepts, using marginal quasi-likelihood, penalized quasi-likelihood, complete enumeration-based or MC methods [15,16]. More recently, new methods have been developed to compute the FIM for both continuous and discrete NLMEMs using MC methods combined with Adaptive Gaussian quadrature (AGQ) [17] or Hamiltonian Monte Carlo (HMC) [13]. The latter approach (MC/HMC) was implemented in the R package *MIXFIM* [18], which uses functions written in the probabilistic language Stan [19], which was developed for Bayesian inference. This method, by efficiently drawing HMC samples and calculating partial derivatives of the log-likelihood, has been shown to be more suitable than MC/AGQ in complex models with a large number of random effects [13].

From expression of the FIM evaluated by these approaches, different optimality criteria can then be computed to optimize designs. The widely used D-optimality criterion consists in maximizing the determinant of the FIM *i.e.* to optimize the precision of estimation for all model parameters. As an alternative to the D-optimality, the D_S-optimality [20] can accommodate situations in which only a subset of the model parameters is of interest. This can be particularly useful to optimize the precision of a "treatment effect" parameter on the longitudinal evolution of biomarkers in clinical trials for example, which is directly linked to the power of the Wald test to detect this effect [21]. However, Ds-optimal designs can lead to problem in estimation of all the parameters of the model because of lack of experimental identifiability [22]. To find a balance between optimizing the precision of the parameters to be tested and the precision all the model parameters, a mixture of D- and D_S-optimality, the DD_S-optimality, can be used [20]. However, these methods require *a priori* knowledge of the model and its parameter values, which can usually be obtained from previous studies, but lead to designs that are locally optimal. If the final model is very different from the *a priori* assumed model, the design optimized with the wrongly assumed model might not be informative enough to precisely estimate parameters

of the final model. Therefore, there is need to develop and evaluate new methods to optimize robust designs accounting for model uncertainty. In order to propose optimal designs across a set of candidate models, we make use of both the theory of compound optimal design [7] and the principle of model averaging [23]. The compound optimal design theory was previously used to propose informative designs for both estimation and model discrimination [24], and recently to find a common design for cocktail of drugs [25]. Meanwhile model averaging is a promising alternative approach to model selection when modeling data, which uses model weights (computed from AIC or BIC) to calculate a weighted average of the predictions [26–28].

Combining these approaches, our main objective is to propose and evaluate a new methodological strategy for robust designs in longitudinal studies with discrete outcomes taking into account model uncertainty, based on the expected FIM in NLMEMs. We choose to use the MC/HMC method to compute the FIM without any linearization. From expression of the FIM, we predict the power of the Wald test to detect a treatment effect included in the model and to calculate the number of subjects needed for a required power. Our FIM-based strategy aims to find designs that are both informative to achieve good average power over several models and to precisely estimate the whole set of population parameters. We illustrate these methods through an example of designing a trial with repeated binary outcomes, inspired from a previous work within the team [29], considering several candidate models which include a treatment effect on the longitudinal response. The relevance of these approaches is evaluated by clinical trial simulations in terms of bias and imprecision of parameter estimates and power of the Wald test.

In Section 2, we detail the notations and present different optimality criteria for a given model or for taking into account model uncertainty. In Section 3, we show an application of these methods to design a longitudinal trial with binary outcomes. In Section 4, we present a simulation study to evaluate the relevance of the proposed design strategy. Finally, we discuss the results and perspectives of this work in Section 5.

2. Methods

2.1. Basic concepts and Notation

2.1.1. Population design

The elementary design ξ_i for the individual i ($i = 1, \dots, N$) is defined by the number n_i of samples and their allocation in time $(t_{i1}, \dots, t_{in_i})$. The population design $\Xi = \{N, (\xi_1, \dots, \xi_N)\}$ is

defined by the number of individuals N, and the set of elementary designs to be performed in each individual $(\xi_1, \dots, \xi_N)$, with a total number of observations $\sum_{i=1}^{N} n_i$.

2.1.2. Nonlinear mixed effect models for binary outcomes

This work considers a set of candidate NLMEMs $m = 1, \dots, M$ for binary data. Considering a given logistic model m, the logit of the conditional probability for observation y_{ij} of individual i at sample j $(j = 1, \dots, n_i)$ is written

$$logit\left(p(y_{ij} = 1|b_i)\right) = f_m\left(t_{ij}, g(\boldsymbol{\mu_m}, \boldsymbol{b_i}, \boldsymbol{z_i}, \boldsymbol{\beta_m})\right) \quad (1)$$

In equation (1), f_m is the function describing the logit-probability of y_{ij} given individual random effects b_i, for a time t_{ij} and a vector of subject-specific parameters modelled through g. The function g depends on the vector of fixed effects μ_m, b_i as well as the vector of covariates z_i and the vector of covariate effects β_m. The random effects are assumed to follow a normal distribution with mean zero and variance-covariance matrix Ω_m which accounts for the inter-individual variability, *i.e.* $b \sim N(0, \Omega_m)$. We denote by $\boldsymbol{\psi_m}$ the vector of population parameters *i.e.* $\psi_m = \left(\boldsymbol{\mu}^T, \boldsymbol{\beta_m}^T, \boldsymbol{\Omega_{m,u}}^T\right)^T$, where $\boldsymbol{\Omega_{m,u}}$ is a vector containing all unique elements of Ω_m. Let P_m denote the length of the vector $\boldsymbol{\psi_m}$. Observations are usually assumed to be independent conditionally upon the random effects, *i.e.* the joint conditional probability for the vector of observations for individual i $(y_i = \left(\boldsymbol{y_{i1}}^T, \dots, \boldsymbol{y_{in_i}}^T\right)^T)$ is:

$$p(y_i|b_i, \psi) = \prod_{j=1}^{n_i} h_m\left(y_{ij}, f_m\left(t_{ij}, g(\boldsymbol{\mu_m}, \boldsymbol{b_i}, \boldsymbol{z_i}, \boldsymbol{\beta_m})\right)\right)$$

2.1.3. Fisher Information Matrix

For a model m and its parameter values ψ_m, if the N subjects are independent, the population FIM $\mathcal{M}(\psi_m, \Xi)$, is the sum of N elementary FIMs, $\mathcal{M}(\psi_m, \xi_i)$, *i.e.* $\mathcal{M}(\psi_m, \Xi) = \sum_{i=1}^{N} \mathcal{M}(\psi_m, \xi_i)$. In this work, we assume the same elementary design for all individuals, *i.e.* $\xi_i = \xi$ for $i = 1, \dots, N$, then $\Xi = \{N, \xi\}$ and $\mathcal{M}(\psi_m, \Xi) = N\,\mathcal{M}(\psi_m, \xi)$. The elementary FIM for $\boldsymbol{\psi_m}$ can be expressed as

$$\mathcal{M}(\psi_m, \xi) = E_y\left(\frac{\partial \log\left(L(y|\psi_m)\right)}{\partial \psi_m} \frac{\partial \log\left(L(y|\psi_m)\right)}{\partial \psi_m}^T\right), \quad (2)$$

where the likelihood L of the observations vector $\boldsymbol{y}$ of an individual (subscript i is omitted for simplicity) is given by

$$L(y|\psi_m) = \int_b p(y|b,\psi_m)p(b|\psi_m)db, \quad (3)$$

with $p(y|b,\psi_m)$ the probability density function (p.d.f.) of y given the random effects b, and $p(b|\psi_m)$ the p.d.f. of b. To evaluate the FIM in equation (2), one needs to compute two integrations: one over the observations y and one in equation (3) over the distribution of random effects. We evaluate the former by MC integrations and the latter by HMC, as proposed in [13], using the package *MIXFIM* [18] in R v3.2.1. *MIXFIM* calls functions of the R package *rstan* written in the probabilistic language Stan [19] to draw HMC samples and calculate partial derivatives of the log-likelihood.

2.2. Optimality criteria

The methods we propose to find robust designs are presented in this section. Different criteria to find an optimal design are presented, according to different purposes. For a given model, based on the determinant of the FIM, the D-optimality is used to obtain informative designs for a precise estimation of all parameters of a model, the D_S-optimality accommodates situations in which only a subset of parameters is of interest, and the compound DD_S-optimality offers a compromise between the D- and the D_S-criteria. Averaging over several models and weighting each one, the compound CD-, CD_S- and CDD_S-optimalities accounts for model uncertainty.

2.2.1. Optimality criteria for a given model

To optimize the estimation precision for the whole set of population parameters ψ_m, we maximize the widely used [7] D-optimality criterion $\Phi_{D,m}(\Xi)$ (equation (4), Table 1).

Let $\boldsymbol{\psi_{S,m}}$, a subset of $\boldsymbol{\psi_m}$, be the vector of parameters of interest of length S_m. To optimize the estimation precision for the S_m parameters, we use the D_S-optimality $\Phi_{D_S,m}(\Xi)$ (equation (5), Table 1) [7]. This criterion is particularly useful to find a design which maximizes the predicted power of the Wald test or minimizes the number of subjects needed to detect a significant treatment effect β_m, by focusing on the precision of estimation for β_m.

Let $\boldsymbol{\psi_{T,m}}$ be the remaining parameters of $\boldsymbol{\psi_m}$ beside $\boldsymbol{\psi_{S,m}}$. To find a compromise between precision of $\boldsymbol{\psi_{S,m}}$ and $\boldsymbol{\psi_{T,m}}$, the compound DD_S-optimality criterion [7,20], $\Phi_{DD_S,m}(\Xi,\alpha_m)$ is used (equation (6), Table 1). α_m is a term between 0 and 1 which quantifies the interest in the

estimation precision of $\boldsymbol{\psi_{S,m}}$. When $\alpha_m = 1/P_m$,the DD_S-criterion coincides with the D-criterion, thus, all the parameters of the model are of equal interest. When $\alpha_m = 1$, the compound DD_S-criterion coincides with the D_S-criterion. When $\alpha_m = 0$, the precision of the parameters $\boldsymbol{\psi_{T,m}}$ is optimized. Maximizing the DD_S-criterion with $1/P_m < \alpha_m < 1$ provides a compromise between the D- and the D_S-criteria.

For any given optimality criterion X, the relative efficiency $\mathrm{E}_{X,m}(\Xi)$ of a design Ξ, with respect to the X-optimal design $\Xi_{X,m}$ for model m, is computed as

$$\mathrm{E}_{X,m}(\Xi) = \frac{\Phi_{X,m}(\Xi)}{\Phi_{X,m}(\Xi_{X,m})}.$$

2.2.2. Robust optimality criteria accounting for model uncertainty

We propose to account for uncertainty in the model choice by computing optimality criteria based on the theory of compound optimality [7] and of model averaging [23]. We assume that a set of M candidate models is available, with each model associated to a weight w_m ($\sum_{m=1}^{M} w_m = 1$) quantifying the balance between the M models. To find a common optimal design for the M models, according to the X-optimality, we maximize the weighted product of efficiencies $\prod_{m=1}^{M} \mathrm{E}_{X,m}{}^{w_m}$, with X being the D-, D_S- or DD_S-optimality criterion. Maximizing the weighted product of efficiencies is equivalent to maximizing the weighted product of the criteria, hence the definition of compound optimality criteria [7], $\Phi_{\mathrm{CD}}(\Xi)$, $\Phi_{\mathrm{CD_S}}(\Xi)$ or $\Phi_{\mathrm{CDD_S}}(\Xi, \alpha_m)$ as provided (equations (7-9), Table 1). Of note, it is also equivalent to minimizing the weighted sum of minus-log-criteria, that is a convex function to which the General Equivalence Theorem can be applied [30].

3. Application to design a longitudinal trial with binary response

In this section, we present how the methods introduced in Section 2 are applied to design a longitudinal study over one year, with two balanced treatment groups and binary response. Candidate NLMEMs including a treatment effect and describing the response probability over time are considered, inspired by [29,31]. We aim to determine the optimal location of a limited number of measurement times chosen among a discrete design composed of **(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)** in month unit, according to different optimality criteria. We investigate the predicted efficiency of the various optimal designs obtained as well as the power of the Wald test to detect a treatment effect over all the candidate models.

3.1. Models and Parameters

The studied models describe binary responses (y_{ij} = 1, responder or 0, non responder) recorded over time (t_{ij} = 0 to 12 months) in two different treatment groups. The response probability at time t_{ij} is given by equation (1).

We denote $f_m(t_{ij}) = f_m\left(t_{ij}, g(\boldsymbol{\mu_m}, \boldsymbol{b_i}, \boldsymbol{z_i}, \boldsymbol{\beta_m})\right)$ and omit index ij for simplicity in the followings. We consider four candidate models f_m describing the evolution of the logit-probability of the response over time (Figure 1), which are the linear model (M1), loglinear model (M2), quadratic model (M3) and exponential model (M4).

M1: $f_1(t) = \theta_1 + \theta_2(1 + \beta \times 1_T)t$.

M2: $f_2(t) = \theta_1 + \theta_2(1 + \beta \times 1_T)\log(t + 1)$.

M3: $f_3(t) = \theta_1 + \theta_2(1 + \beta \times 1_T)t^2$.

M4: $f_4(t) = \theta_1 + \theta_2(1 + \beta \times 1_T)(\exp(\theta_3 t) - 1)$.

1_T is a treatment group indicator variable (with $1_T = 0$ if control group, $1_T = 1$ if treated group). We denote β the effect size of the treatment on parameter θ_2. Model M1 and its parameters are inspired by [29]. Values of $\boldsymbol{\psi_1}$ are given in Table 2, θ_1 and θ_2 follow a normal distribution: $\theta_p = \mu_p + b_p$ where $b_p \sim N(0, \omega_p^2)$ for p = {1,2}. Models M2 to M4 are alternative models with parameters normally distributed as in M1. Values of their fixed parameters are chosen to give the same mean value of the logit of the response probability as for M1 at the beginning (t = 0 month) and the end (t = 12 months) of the study in the two treatment groups. Similar treatment effect value and a standard deviation ω_2 equal to 188% of μ_2 for are assumed for all models. Assumed values of population parameters $\boldsymbol{\psi_m}$ for each model m ($m = 1, \ldots, 4$) are given in Table 2.

3.2. Design optimization and power evaluation

We aim to propose an informative design with 4 measurement times identical in all subjects, denoted $\xi = \{\boldsymbol{t_1}, \boldsymbol{t_2}, \boldsymbol{t_3}, \boldsymbol{t_4}\}$, first, for each candidate model separately, then, by averaging over the four candidate models, given the parameters values $\boldsymbol{\psi_m}$ given in Table 2.

3.2.1. Choice of MC-HMC samples for FIM evaluation

To perform design optimization for the presented models, we need to compute different optimality criteria, based on the FIM evaluated by MC/HMC. Since this is a stochastic approach, we can quantify uncertainty in the computation of the determinant of the FIM,

according to the MC and HMC samples. For instance, the 95% uncertainty intervals of the D-criterion can be computed from the 2.5th and 97.5th percentiles of the distributions generated using non-parametric bootstrap. According to [13] the FIM evaluation method implemented in *MIXFIM* performs well with 5000 Monte Carlo samples and 200 Hamiltonian Monte Carlo samples, for a longitudinal logistic model for binary response (similar to our model M1) and a rich design. To ensure that these settings are appropriate for all candidate models even for sparser design (4 measurement times), the convergence of the D-criterion is studied with respect to the MC samples increasing from 50 to 10000 (see Supplementary Material, Figure S1), with a non-optimized equispaced design $\xi_{ES} = \{\mathbf{0}, \mathbf{4}, \mathbf{8}, \mathbf{12}\}$. Two different configurations were studied: 1000 HMC samples and 1000 burn-in *vs.* 200 HMC and 500 burn-in. We found that the convergence plots of D-criterion shows a similar trend between the two configurations. Moreover, 5000 MC samples seems sufficient for a good convergence of the D-criterion. We therefore set the number of MC samples to 5000 and the number of HMC samples to 200 with 500 burn-in in *MIXFIM* to evaluate the FIM and to calculate the different optimality criteria with all the designs and the models of the study.

To ensure that these settings are satisfying, convergence of the D-criterion is also verified with an optimized design (Supplementary Material, Figure S2).

3.2.2. Design optimization

As in [29], we assumed that t_1 and t_4 are fixed to 0 and 12 months (*i.e.* the end of the study), only two times t_2 and t_3 are optimized among the following set: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 months. No repetition is considered, *i.e.* four measurement times must be different. Combinatorial optimization is performed for different optimality criteria. This corresponds to $\binom{11}{2} = 55$ possible elementary designs to be evaluated and the optimal one is chosen according to each considered criterion. We considered *N*=100 subjects equally distributed in the two treatment groups for design optimization, although this number as no influence on optimal sampling time or loss of efficiency in design comparisons, but only on the specific value of relative standard errors when reported.

3.2.2.1. D-, D_S- and DD_S-optimal designs for each model

Using the D-, D_S- or DD_S-optimality, the optimal location of measurement times are obtained and compared between the four models and between these optimality criteria. The relative

D- ,D_S- or DD_S-efficiencies of each optimal design are also computed assuming each of the four models to evaluate the impact of model misspecification on design performance.

The D_S- and the DD_S-criteria are computed with a particular interest on the treatment effect $(\psi_{S,m} = \beta)$. The preliminary step to compute compound DD_S-optimal designs is to find an appropriate α_m value for each model. For each model, we therefore compute the DD_S-optimal designs $\Xi_{DD_S,m}$ for α_m values from 0 to 1 by increment of 0.05. Then, we compare the DD_S-optimal allocations of measurement times and the D- and D_S-efficiencies of these DD_S-optimal designs over α_m. We consider that for a DD_S-design to be satisfactory, α_m should maximize $E_{D,m} \times E_{D_S,m}$. Moreover, both the D- and D_S-efficiencies should be above 0.8 (other threshold values are not investigated in this work). The final DD_S-optimal design is computed with the retained α_m value.

3.2.2.2. Robust optimal designs

To propose a robust design accounting for model uncertainty, different compound criteria (CD-, CD_S- and CDD_S-optimality) are evaluated for a combination of the four candidate models. The same weight w_m is assigned for each model (total uncertainty) *i.e.* $w_m = 1/4$.

3.2.3. Expected average power using FIM

We aim to evaluate the average power of the Wald test for a design $\Xi = \{N, \xi\}$ to detect a significant treatment effect over all the candidate models varying the total number of subject from 50 to 450. We also computed the number of subjects to achieve an average power of 0.9.

For that, first, we derive for each model m the SE (Standard Error) of β from the expected FIM and we compute the power $\pi_m(\Xi)$ to detect the treatment effect in that model, following what has been done in [21]. Then the average power $\pi_{average}(\Xi)$ over M candidate models is evaluated as

$$\pi_{average}(\Xi) = \sum_{m=1}^{M} w_m \times \pi_m(\Xi),$$

where w_m is the weight associated to each model m used in the computation of the optimality criteria (here $1/M$, with $M = 4$).

3.3. Results

Figure 2 displays, for each model, the efficiencies with respect to each criterion, obtained with the different combinations of the second and third samples out of four measurement times.

3.3.1. D- and D_S-optimal designs for each model

Different models lead to different optimal designs. Table 3 reports the D- and D_S-efficiencies respectively of the D-optimal designs $\Xi_{D,m}$ and D_S-optimal design $\Xi_{D_S,m}$, when the model used for data analysis is model m or another one. For instance, the design $(\mathbf{0},\mathbf{4},\mathbf{5},\mathbf{12})$ is D-optimal for M3 but underperforms for M4 with a low D-efficiency of only 0.646 and the design $(\mathbf{0},\mathbf{1},\mathbf{11},\mathbf{12})$ is D_S-optimal for M2 but conduces to a low D_S-efficiency of only 0.389 for M3.

In this work, the D_S-optimality focuses on the estimation precision of the treatment effect β. We notice that using this criterion could lead to different optimal measurement times than those obtained with the D-criterion, as in the case of M2 or M4 (Figure 2). For M2, the allocations of informative measurement times leading to a D- and D_S-efficiency above 0.8 are very similar and the respective optimal designs are $\xi_{D,2} = (\mathbf{0},\mathbf{1},\mathbf{8},\mathbf{12})$ *vs* $\xi_{D_S,2} = (\mathbf{0},\mathbf{1},\mathbf{11},\mathbf{12})$. However we notice quite different efficient designs between D- and D_S-criterion for M4 ($\xi_{D,4} = (\mathbf{0},\mathbf{6},\mathbf{11},\mathbf{12})$ *vs* $\xi_{D_S,4} = (\mathbf{0},\mathbf{10},\mathbf{11},\mathbf{12})$). This could be explained by the shape of the model curve (Figure 1, the second measurement time should be more tardive to better distinguish the two treatment groups), and by the higher number of parameters of this model compared to others (the importance of estimating β is thus relatively reduced when using D-optimality).

3.3.2. DD_S-optimal design for each model

For each model, Figure 3 shows allocated measurement times as well as D- and D_S-efficiencies of the DD_S-optimal designs over α_m possible values varying between 0 and 1. The third measurement time tends to 11 with increasing α_m for all models. These results emphasize the necessity to get tardive measurements to better distinguish the responses between the control and the treated group, and thus to better estimate the treatment effect β. The product of D- and D_S-efficiencies are respectively maximal when $\alpha_1 \in [0.2,1]$, $\alpha_2 \in [0.3,1]$, $\alpha_3 \in [0.25,1]$ and $\alpha_4 \in [0.6,1]$. For these values of α_m, D- and D_S-efficiencies are over 0.8 excepted for M4. With M4, D- and D_S-efficiencies are over 0.8 when $0.5 \leq \alpha_4 \leq 0.55$. In order to conserve the same range of importance for estimating the precision of parameters between the four models, we

choose 0.5 as the same α_m value with each model. Thus, DD_S-optimal design for each model are computed with $\alpha_m = 0.5$, this value satisfying our two conditions.

We notice that the DD_S-criterion could lead to different optimal measurement times than those obtained with the D- or D_S-criterion (Figure 2), as in the case of M4 ($\xi_{D,4} = (\mathbf{0}, \mathbf{6}, \mathbf{11}, \mathbf{12})$, $\xi_{D_S,4} = (\mathbf{0}, \mathbf{10}, \mathbf{11}, \mathbf{12})$, $\xi_{DD_S,4} = (\mathbf{0}, \mathbf{9}, \mathbf{11}, \mathbf{12})$). These 3 designs lead to different predicted relative standard errors (RSE) of β, which for 100 patients, are 60% ,50%, 52%, respectively; illustrating the impact of using D_S- or DD_S-criterion on power to detect a treatment effect. For this model, we also studied the influence of the number of sampling times. We found that the DD_S-optimal design with 5 samples is $\xi_{DD_S} = (\mathbf{0}, \mathbf{6}, \mathbf{10}, \mathbf{11}, \mathbf{12})$), leading to a RSE of β of 50%, whereas the design with all 13 points lead to a RSE of 40%. For M1, M2 and M3, the DD_S-efficient designs coincide with the D_S-efficient designs. In the same way as for previous criteria (D- and D_S-efficiencies), different models lead to different DD_S-efficient designs. However, all the DD_S-optimal designs include the 11th month, which again emphasizes the importance of tardive measurement times (t_3 and t_4) in estimating the treatment effect. Table 3 reports the DD_S-efficiencies of the DD_S-optimal design $\Xi_{DD_S,m}$ when the true model is m or another one. For instance, the design $(\mathbf{0}, \mathbf{1}, \mathbf{11}, \mathbf{12})$ is DD_S-optimal for M2 but underperforms for M3 with a low DD_S-efficiency of 0.611. Therefore, this design provides poor precision on parameters of M3, especially β. With 100 individuals, the predicted relative standard error of β is 65% with this design vs. 41% with the DD_S-optimal for M3 $\left(\xi_{DD_S,3} = (\mathbf{0}, \mathbf{4}, \mathbf{11}, \mathbf{12})\right)$.

3.3.3. Robust optimal designs

The CD-, CD_S- and CDD_S-efficiencies of every possible designs as well as the corresponding optimal design are reported in the last column of Figure 2. First, we can note that efficient robust designs are nearly the same using CD_S- or CDD_S-optimality.

Assuming model uncertainty leads to a robust CD-optimal design ($\xi_{CD} = (\mathbf{0}, \mathbf{5}, \mathbf{11}, \mathbf{12})$) which is different from the four D-optimal designs for each model. The most CD-efficient designs are closer to the D-efficient designs for the quadratic model (M3) than for other models. Table 3 presents D-efficiencies for each model. The robust CD-optimal design conducts to a D-efficiency above 0.8 regardless of the selected model. A model misspecification can lead to loss of efficiencies of up to 0.354.

In the same way as using CD-optimality, assuming model uncertainty leads to CD_S-optimal design close to D_S-optimal design for model M3: $\xi_{CD_S} = (\mathbf{0}, \mathbf{4}, \mathbf{11}, \mathbf{12})$ and $\xi_{D_S,3} = (\mathbf{0}, \mathbf{5}, \mathbf{11}, \mathbf{12})$.

Furthermore, CD_S-efficient designs include mostly more tardive measurement times than the CD-efficient designs (Figure 2). The robust CD_S-optimal design also conducts to a D-efficiency above 0.8 regardless of the candidate model and reduces in average the maximal loss of D_S-efficiency as compared to the D_S-optimal designs obtained separately for each model (Table 3, 0.363 *vs* respectively 0.566, 0.619, 0.4 and 0.481 with $\xi_{D_S,1}$, $\xi_{D_S,2}$, $\xi_{D_S,3}$ and $\xi_{D_S,4}$). The CD_S-optimal design could be particularly appropriate to precisely estimate β, especially if the final model is M3 (D_S-efficiency of 0.906).

Assuming model uncertainty leads to a robust CDD_S-optimal design ($\xi_{CDD_S} = (\mathbf{0}, \mathbf{4}, \mathbf{11}, \mathbf{12})$) which is different from the 4 local DD_S-optimal designs for each model. The robust CDD_S-optimal design is the same as the CD_S- and thus conducts to a D-efficiency above 0.8 regardless of the candidate model. It also reduces the maximal loss of DD_S-efficiency as compared to the DD_S-optimal design obtained separately for each model (Table 3, 0.249 *vs* respectively 0.356, 0.389, 0.278 and 0.328 with $\xi_{DD_S,1}$, $\xi_{DD_S,2}$, $\xi_{DD_S,3}$ and $\xi_{DD_S,4}$), and shows DD_S-efficiencies above 0.8 for M1, M3 and M4.

3.3.4. Expected average power using FIM

Expected power over number of subjects *N*, under each model and in average, are presented with different designs in Figure 4: the CDD_S-optimal or robust design ($\xi_{CDD_S} = (\mathbf{0}, \mathbf{4}, \mathbf{11}, \mathbf{12})$), the DD_S-optimal design for M1 ($\xi_{DD_S,1} = (\mathbf{0}, \mathbf{2}, \mathbf{11}, \mathbf{12})$) and the non-optimized equispaced design ($\xi_{ES} = (\mathbf{0}, \mathbf{4}, \mathbf{8}, \mathbf{12})$). The exponential model (M4) presents the worst expected power among the 4 candidate models, especially with ξ_{ES}, mainly because this model involves one more estimated parameter than the three other models. Among these 3 designs, the ξ_{CDD_S} presents the closest curves of power under the four models. Moreover, to reach a $\pi_{average}$ of 0.9, the predicted NSN (Number of Subjects Needed) is 274 with ξ_{CDD_S} *vs.* 320 with $\xi_{DD_S,1}$ and 358 with ξ_{ES}. In our example, the CDD_S-criterion is able to propose design giving decent performances in terms of predicted power whatever the model chosen among the candidates.

4. Evaluation by simulation

In this section, we perform simulations to assess the relevance of the design strategy presented above. For this purpose, with simulated datasets of 274 subjects (*i.e.* the NSN to reach an average power of 0.9 with the robust design ξ_{CDD_S}), we 1) compare performances of ξ_{CDD_S} *vs.* a locally optimal design for the linear model M1 $\xi_{DD_S,1}$ and a non-optimized equispaced design

ξ_{ES} in terms of bias and precision of estimates and 2) evaluate the type I error and power as well as the adequacy between predictions and simulation results with the robust design.

4.1. Clinical trial simulation

4.1.1. Designs and models

Repeated binary response trials of 274 individuals (137 per treatment group) are simulated under each model m (M1 to M4) with their respective population parameters values indicated in Table 2.

For the first objective (to compare different designs performances), with ξ_{CDD_S}, $\xi_{DD_S,1}$ and ξ_{ES}, $K_{1,m} = 500$ datasets are simulated with each model under H_1 (treatment effect $\beta = 5$) and each one of these 3 designs.

To achieve the second objective, with the robust design ξ_{CDD_S}, $K_{0,m} = 500$ additional datasets are simulated with each model under H_0 (without treatment effect, $\beta = 0$), in order to evaluate the type I error.

4.1.2. Evaluation

For each dataset, population parameters are estimated by the SAEM algorithm [32] implemented in MONOLIX 2016R1 (www.lixoft.eu), a software devoted to maximum likelihood estimation of parameters in NLMEMs.

First, from the analysis of the $K_{1,m}$ simulated datasets, we calculate the inverse of the variance-covariance matrix computed from 500 estimated parameter vectors, for each of the three designs. We defined the observed relative D-efficiency between designs as the ratio of determinants of these matrices, normalized by the number of parameters. We also compare the distribution of the relative estimation error of parameters for each model (associated with the relative bias values) between the three designs.

Then, with the design ξ_{CDD_S}, the adequacy between FIM predictions and CTS observations is evaluated (in terms of parameter precision and power). The empirical standard errors (SE_{CTS}) of each parameter of each model is defined as the standard deviation of all estimated values. The empirical relative standard errors (RSE) are then calculated as the ratio of the SE_{CTS} to the parameter simulated values. The 95% confidence intervals for each empirical RSE are given as $\left[\sqrt{\frac{q_1 \times RSE^2}{K_{1,m}-1}} ; \sqrt{\frac{q_2 \times RSE^2}{K_{1,m}-1}}\right]$, with q_1 and q_2 respectively the 2.5% and 97.5% quantiles of the χ^2

distribution with $K_{1,m} - 1 = 499$ degrees of freedom. To evaluate the relevance of MC/HMC-based approach to evaluate the expected FIM (see Section 2.1.3.) and the parameter estimation method, predicted RSE obtained using the expected FIM are compared with empirical RSE and RRMSE (Relative Root Mean Square Error) obtained with CTS. Thus, the type I error and the power are evaluated. From the $K_{0,m}$ datasets simulated with $\beta = 0$, the type I error of the Wald test on the treatment effect is evaluated as the proportion of trials for which H_0 is rejected, and compared with 0.05. The 95% prediction interval (95% PI) for proportion π of H_0 rejection (type I error or power) is computed by binomial test [33]. As explained in Section 3.2.1., we can evaluate the uncertainty in the computation of the D-criterion and the expected power derived from FIM, according to the MC and HMC samples. Under H_1, we can thus give a confidence interval as 2.5th and 97.5th percentiles of bootstrap obtained with our FIM computation approach, around the FIM expected power value. The observed proportion of significant Wald tests is compared with the FIM expected power value and its (respectively prediction or confidence) intervals, by binomial test or by bootstrap.

4.2. Results

4.2.1. Comparison of robust design with non-optimized and non-robust design

Informativeness of the three designs is investigated here. For each model, ratios of observed normalized determinant of the variance-covariance matrix of these designs are computed with the robust design $\xi_{\mathrm{CDD_S}}$ as reference.

As presented in the Table 4, overall, the estimations of parameters are more precise with the robust design than with the equispaced design for all the models (observed D-efficiency with respect to $\xi_{\mathrm{CDD_S}}$ is below 1). With the optimal design for M1 $\xi_{\mathrm{DD_S},1}$, the parameters of M1 and M2 are slightly more precise than with the robust design, but this design induces important loss of precision for the parameters of M3 and M4 compared to the robust design.

The three designs shows acceptable and similar distribution of relative estimation errors (REE, Figure 5) under M1, M2 and M3. REE under M4 are larger than other models with the three designs. The robust design performs better on M4 than the other designs, especially for the parameters μ_2 and ω_2 which present larger range of whiskers. Moreover, the maximal relative bias with $\xi_{\mathrm{CDD_S}}$ is 15% but is over 20% with $\xi_{\mathrm{DD_S},1}$ (37% for μ_2 and 35% for ω_2 on M4) and ξ_{ES} (47% for μ_2 and 35% for ω_2 on M4).

4.2.2. Adequation between FIM predictions and CTS results with the robust design

As shown in Figure 6 for all models, the predicted RSE by the FIM computation (RSE_{FIM}) are in the same range as empirical RSE from CTS (RSE_{CTS}) especially for the fixed effect parameters (μ) and for standard deviations of random effects parameters (ω), even when the model is more complex (M4, excepted μ_3). RSE of the covariate effect β are however slightly under-predicted. Nevertheless, the maximal differences between RSE_{FIM} and RSE_{CTS} are at most 4%. Moreover, the population parameters of all candidate models could be estimated with reasonable estimation error for the proposed robust design (RSE around 30% for the fixed effects and 50% for the random effects) excepted M4 which contains 6 parameters: θ_2 is quite difficult to be estimated with such a sparse design.

Table 5 displays the type I error of the Wald test on the treatment effect in each candidate model, using the estimated or empirical SE. For the Wald tests using estimated SE, it can be seen that the type I error lies in the 95% PI of the nominal level [0.033,0.073] for all models. Table 5 shows also the predicted power from FIM *vs* observed power from CTS under H_1. Two different kinds of intervals are built around the FIM predicted power: as under H_0 using the binomial test or as percentiles of bootstrap performed with FIM evaluation. We can note that the interval built by bootstrap is smaller than the binomial test one. The choice of MC and HMC samples seems therefore appropriate for power predictions (see convergence plots in Supplementary Material, Figure S2). Nevertheless, for each model, the observed power is slightly above the predicted power. These results are partly due to the positive bias on β (Figure 5) and developed further in the discussion.

5. Application to a real study

5.1. Presentation of the study

We applied the proposed approach to a clinical study conducted by the National Institute on Aging [34] using the published model, *i.e.* not assuming model uncertainty. The study EPESE (Established Populations for Epidemiologic Studies of the Elderly) was a longitudinal study with 4162 persons aged of 65 years or older involved. The outcome of interest was the instrumental activities of daily living (IADL). The IADL was measured once a year during 3 years. The evolution of IADL was compared between those who were cognitively impaired at baseline and those who were cognitively intact.

In [35], the IADL tasks (traveling, shopping, preparing meals, doing housework and managing finances without assistance) were dichotomized: individuals unable to perform 4 or 5 tasks were classified as disabled (y_{ij} = 1) and those able to perform at least 2 tasks were not classified as disabled (y_{ij} = 0). This dichotomous outcome was measured at 0, 1, 2 and 3 years. They studied the cognitive impairment at baseline on the outcome. The proposed model in [35] to describe probability to be disabled at time t_{ij} was

$$logit\left(p(y_{ij}=1|b_i)\right) = \theta_1 + \beta_1 \times 1_{BI} + (\theta_2 + \beta_2 \times 1_{BI})t.$$

where 1_{BI} is the cognitive impairment at baseline ($1_{BI} = 0$ if no impairment, $1_{BI} = 1$ if impairment). β_1 was the impairment effect on the intercept parameter θ_1 and β_2 the impairment effect on the slope parameter θ_2. θ_1 follow a normal distribution: $\theta_1 = \mu_1 + b_1$, where $b_1 \sim N(0, \omega_1^2)$. No interindividual variability was considered on θ_2 ($\theta_2 = \mu_2$).

5.2. Design evaluation and optimization

5.2.1. Methods

To evaluate the relevance of the FIM-based approach, using the EPESE design $\xi_{\text{EPESE}} = (\mathbf{0}, \mathbf{1}, \mathbf{2}, \mathbf{3})$ and the provided model described above, we compared the SE predicted using MC/HMC (see Section 2.1.3) and the SE reported in [35] after data fitting.

Then, based on the DD_S-criterion (with the subset of parameters of interests ψ_S composed of μ_2 and β_2), we propose an optimized design of 4 measuring times, assuming that t_1 and t_4 are fixed to 0 and 3 years (*i.e.* the end of the study), and that only two times t_2 and t_3 are optimized among 0.2 to 2.8 years by increment of 0.2 years. No repetition is considered, *i.e.* four measurement times must be different. Combinatorial optimization is performed for different optimality criteria. This corresponds to $\binom{14}{2}$ = 91 possible elementary designs to be evaluated and the optimal one is chosen according to each considered criterion.

Finally, we computed the power of the Wald test and the NSN are calculated (as explained Section 3.2.3) for the effect of cognitive impairment on the slope, β_2,. for the slope parameter μ_2.

5.2.2. Results

The comparison of the observed and the FIM computed SE of the parameters is reported in Table 6. SE of each fixed or random effect are close between FIM predictions and results reported in [35].

According to the D-criterion, the D-optimal design with the provided model is $\xi_D = (\mathbf{0}, \mathbf{1.2}, \mathbf{2.8}, \mathbf{3})$. The evolution of the probability to be disabled and the cognitive impairment effect on this evolution were particularly of interest in this study. Thus, we compute the D_S-criterion with μ_2 and β_2 in the subset of parameters ψ_S (see equation (5) in Table 1). The D_S-optimal design is $\xi_{D_S} = (\mathbf{0}, \mathbf{1.2}, \mathbf{2.8}, \mathbf{3})$, *i.e.* the same measuring times allocation of the D-optimal design. Moreover, using the DD_S-criterion with α between 0.4 (*i.e.* D-criterion) and 1 (*i.e.* D_S-criterion) leads to the same optimal design ($\xi_{DD_S} = (\mathbf{0}, \mathbf{1.2}, \mathbf{2.8}, \mathbf{3})$). As for the linear model M1 of Section 3, all the possible designs leads to D-efficiencies higher than 0.8. For instance, with the ξ_{EPESE}, the D-efficiency is 0.944.

The provided model in [35] shows that the slope μ_2 is non null. Moreover, only 31 individuals would have been needed with ξ_{DD_S} (*vs.* 33 with ξ_{EPESE}) to reach a power of 0.8. However, the covariate β_2 is not significant. With the ξ_{EPESE}, the number of subjects needed to detect this covariate with a power of 0.8 would be 484,518 subjects. With the optimized design $\xi_{DD_S} = (\mathbf{0}, \mathbf{1.2}, \mathbf{2.8}, \mathbf{3})$, it would be 437,087 subjects. Recruiting this number of individuals, even with the optimized design, would be not feasible. Nevertheless, this result highlights that the DD_S-optimal design can reduce the NSN compared to a non-optimized design.

6. Discussion

In this paper, we propose a new methodological strategy based on MC/HMC and compound optimality theory to design longitudinal trials with discrete outcomes. Our approach accounts for model uncertainty when assuming a set of candidate models and ensures a compromise between the overall precision of estimation and the power of the Wald test to detect a covariate effect. To our knowledge, this is the first time that a robust optimal design approach, based on MC/HMC to compute the FIM without any linearization, is evaluated by extensive clinical trial simulations. The relevance of using the FIM to efficiently predict the power of the Wald test has been evidenced before but only in the context of repeated continuous outcomes [21,36]. We have now confirmed the utility of this FIM-based method to predict average power across several candidate models, with an example of repeated discrete outcomes including two treatment groups. Although the considered example is quite theoretical here (Sections 3. and 4), the proposed method is illustrated with a published study analyzing real data (Section 5). In this study, the repeated binary outcome was the ability to undertake life tasks (based on the Instrumental Activities Daily Living tasks). The chosen settings can also easily be extended to clinical longitudinal studies with binary outcomes such as remission of dyskinesia in

Parkinson's Disease. In such cases, it is not possible to obtain dense measuring times because medical tests to evaluate this scale are very time-consuming. In another context, a method to design longitudinal studies was proposed when dynamics is modelled by a binary Markov process, using examples of infection by *Streptococcus pneumoniae* [37].

To determine informative designs, we study different optimality criteria, according to different purposes of the study. The D-optimality is used to optimize the precision of the whole set of population parameters. The D_S-optimality to accommodate situations in which only a subset of the model parameters is of interest, which is particularly useful to minimize the standard error on the treatment effect, thus to maximize the power of the Wald test of the study and therefore to reduce the number of subjects needed. A parameter which is a primary end-point (*i.e.* treatment effect) or a key secondary end-point (*i.e.* time effect) of a study should be included in the subset of parameters of interest. This subset could also be defined based on the future use of the model. For example, in clinical trial simulations, if outcomes are sensitive to some specific parameters, they should be included in the subset of parameters of interest. However, this criterion usually leads to designs that do not ensure the experimental identifiability [22]. To overcome this issue, the DD_S-optimality allows the experimenter to find a compromise between the D- and the D_S-criterion [20], by weighting each one according to the importance given to each model parameter. The weight for each criterion depends of the objective of the study, and its meaning is further discussed in [38]. In spite of its usefulness, the DD_S-optimality was mostly used with continuous standard nonlinear regression models [39,40] but has not been reported so far in the context of discrete NLMEMs, to our knowledge. Another novelty of this work is to extend these standard optimal criteria to account for model uncertainty, using their weighted product to propose compound CD-, CD_S- and CDD_S-criterion over a set of candidate models.

In the motivation example, we considered several candidate binary NLMEMs to describe the logit-probability of response over time: linear (M1), loglinear (M2), quadratic (M3) and exponential (M4). M4 contains one more parameter than other models and therefore provides information about the predicting abilities of the MC/HMC method with a more complex model. The non-optimized equispaced design in our example provides D-efficiencies > 0.8 under each candidate model. This design shows however very poor D_S-efficiency of 0.393 under M4 (Table 3). When optimizing designs, we find that accounting or not for uncertainty in models and focusing or not on the estimation precision of the covariate effect may lead to different optimal locations of measurement times. Moreover, misspecification of models in case of local design optimization (*i.e.* assuming a given model) can lead to D-, D_S- or DD_S-efficiency

respectively as low as 0.666, 0.389 or 0.611 for some models. The proposed compound CD-, CD_S- or CDD_S-optimal designs provides a better compromise for different candidate models and greatly reduces the loss of efficiency. These three compound optimal designs conduce all to D-efficiencies of at least 0.8 for each model. Furthermore, in terms of ability to detect the treatment effect, choosing the robust design ξ_{CDD_S} instead of the design optimal for the linear model $\xi_{DD_S,1}$ or either a non-optimized ξ_{ES} reduces the required sample size of the study. The NSN for an average power of 0.9 is indeed 274 with ξ_{CDD_S} *vs.* 320 with $\xi_{DD_S,1}$ and 358 with ξ_{ES}. The NSN is however close with either the CDD_S- or the CD-design (274 *vs.* 280). This could be explained by the fact that these two designs only differ by their respective second time point (respectively at 4^{th} and 5^{th} month). However, the gain in power could be larger in other examples, with more differences in allocation of sampling times between the CDD_S- and the CD-design. The three optimal designs using compound criteria CD-, CD_S- or CDD_S- are close, which is partly explained by the use of a combinatorial optimization, with only one possible measurement per month, limiting the total number of possible designs. Surprisingly, the CD-design includes a second measurement time which is latter than the one of the CD_S-design which maximizes the average power. It is explained by the loglinear model (M2) for which early measurement seems already informative to distinguish the logit probability of response between the two treatment groups. Overall, we notice that with this set of models, picking only tardive measurement could be very suboptimal to maximize the precision of the parameter β and thus the power to detect a treatment effect. The closeness of D- and D_S-optimal measurement times under each model and averaging over the 4 models (using CD- and CD_S-optimality) indicates that a precise estimation of all the parameters (and probably, even more the "trend" parameters μ_2 and ω_2 on which β acts) is necessary to reach a satisfactory power. In the optimization procedure, it would also be interesting to consider reducing the duration of the study.

We also evaluate our approach by simulating clinical trials under each candidate model. To assess the relevance of the robust design, we compare performances between three different designs: a non-optimized design with a measurement time every 4 months, a design optimal for a linear model (M1, maximizing the DDs-optimality) and a robust design accounting for model uncertainty (maximizing the compound DDs-optimality). In our example, choosing the robust design avoids important loss of information on parameters, especially when the right model was exponential (M4). Other scenarios could have been examined *e.g.* simulating data under another model M5 and fit under each candidate model M1 to M4. We also could evaluate the power

performances of model averaging [23] *vs.* model selection as in [28] but it is not the purpose of the present work. Furthermore, a method has been recently proposed to optimize designs when the analysis is done by model averaging [41].

With the CDD_S-design and 274 individuals, in spite of a slight under or over-prediction of RSE for some parameters, the discrepancy is at most 4%, thus, the prediction using the FIM evaluated by MC/HMC is accurate. This approach avoids extensive CTS and can efficiently help to easily detect non-informative designs in case of large predicted RSE values. Using the predicted SE given by FIM, we predict the power of the Wald test to show significant difference between treatment groups. Under H_0, we note a good control of the type I error by CTS, despite the asymptotic properties of the Wald test (explained in [42]) and our sparse measurement design. Under H_1, we found differences in predicted power between the different models: respectively 0.967, 0.889, 0.973 and 0.769 under M1, M2, M3 and M4. Whereas the SE for treatment effect (β) are not over-predicted, powers are under-predicted under all the models M1 to M4 (especially M2 and M4): observed power are respectively 0.988, 0.988, 0.996 and 0.86. However, we note that the power is never over-predicted. Moreover, the FIM predicted Wald statistic is close to the median of the observed Wald statistic distribution, even with M2 and M4 (Supplementary Material, Figure S2). We investigated the possibility that the number of MC/HMC samples would be insufficient under our design to adequately predict the precision on parameter β (and thus of the power). However the convergence plots of the determinant of the FIM and of the SE on β showed acceptable results (Figure S3, Supplementary Material). Discrepancies between predicted and observed power are mainly explained by the positive bias on β in CTS. If β would be estimated at its true simulated value (5), the estimated power would be close to the predicted values (Supplementary Material, Table S1). Other settings of the SAEM algorithm of MONOLIX were investigated, but none provided more satisfactory relative bias. To our knowledge, an evaluation of different algorithms for repeated binary responses has never been published, as it has been done for ordered categorical [43] or continuous data [44]. The analysis of this kind of data is challenging, requires informative design to avoid problem of bias. Further research works on the estimation and design method for this kind of data are necessary. It could thus lead to fair measures of discrepancies between FIM prediction and CTS observations, and also to a fair comparison of different designs in terms of average power.

We apply the described FIM based methods, using a published logistic model built on real data [35]. This model is describing the evolution of the disability probability over time, including parameters as the baseline (fixed and random effect), the time effect or the cognitive impairment

effect. With the model and the design elements provided in the study, we find that the SE predicted by FIM are close to those provided in the study, showing that the FIM evaluation method by MC/HMC is relevant. The provided model is the only one mixed model which is reported. If other possible models would have been proposed, we would be able to compute compound criteria, with each model weight depending on its likelihood, in accordance with the model averaging principle [23]. Therefore, a robust design based on the CDD_S-criterion would have been proposed. However, with the given model, we propose to optimize the choice of the measuring times according to the DD_S-criterion. We also predict the number of subjects that would be needed to reach a power of 0.8 to detect the time effect and the cognitive impairment. We show that, instead of using a non-optimized equispaced design, optimizing the design using the DD_S-criterion can reduce the number of subjects needed.

In this work, we perform combinatorial optimization by evaluating FIM for every possible design, an alternative approach would consist in implementing an optimization algorithm to improve computing time as suggested in [13]. More efficient algorithms could also be explored, such as stochastic approximation, annealing methods, randomized exchange algorithm [45], multiplicative method [46] or particle swarm optimization [47,48]. It is also important to account for uncertainty in parameters in addition to uncertainty in models [49,50], by evaluating the expectation of the FIM over the distribution of population parameters instead of assuming known values of these parameters. We also plan to integrate these robust design methods (over the parameter distribution and over the models) in a next version of PFIM, which is the software program for designing longitudinal studies developed by our team (www.pfim.biostat.fr) [51]. Furthermore, robust design methods studied in this paper could be combined with adaptive designs [52,53], by using the accumulated information at each stage of the study to update knowledge of the candidate models and of the parameter distributions to be taken into account in design optimization.

The proposed design strategy based on the expected FIM evaluated by MC/HMC and the compound optimality theory is a relevant approach, which enables, for the first time, optimization of informative sparse designs for longitudinal binary data, over several candidate models. This approach can also be applied to other type of discrete data models, such as multi-category or Poisson count data [54]. A robust design method and application using *MIXFIM* for Poisson models was studied in [50] but not for the comparison of two treatment groups hence using CD-optimality. The approach described in this article also aims to propose design

providing good power to detect the treatment effect and acceptable precision of all parameters while accounting for model uncertainty.

References

1. Lalonde RL, Kowalski KG, Hutmacher MM, et al. Model-based drug development. *Clin Pharmacol Ther* 2007; 82: 21–32.

2. Mould DR, Upton RN. Basic concepts in population modeling, simulation, and model-based drug development. *CPT Pharmacometrics Syst Pharmacol* 2012; 1: e6.

3. Mould DR, Upton RN. Basic concepts in population modeling, simulation, and model-based drug development—Part 2: Introduction to pharmacokinetic modeling methods. *CPT Pharmacomet Syst Pharmacol.* 2013; 2: e38.

4. Upton RN, Mould DR. Basic concepts in population modeling, simulation, and model-based drug development: Part 3—Introduction to pharmacodynamic modeling methods. *CPT Pharmacomet Syst Pharmacol.* 2014; 3: e88.

5. Syrjänen RK, Kilpi TM, Kaijalainen TH, et al. Nasopharyngeal carriage of Streptococcus pneumoniae in Finnish children younger than 2 years old. *J Infect Dis* 2001; 184: 451–459.

6. Mestre TA, Beaulieu-Boire I, Aquino CC, et al. What is a clinically important change in the Unified Dyskinesia Rating Scale in Parkinson's disease? *Parkinsonism & Related Disorders* 2015; 21: 1349–1354.

7. Atkinson A, Donev A, Tobias R. *Optimum Experimental Designs, with SAS*. Oxford, New York: Oxford University Press, 2007.

8. Pronzato L, Pázman A. *Design of Experiments in Nonlinear Models: Asymptotic Normality, Optimality Criteria and Small-Sample Properties*. Springer Science & Business Media, 2013.

9. Fedorov V, Leonov S. Optimal Design for Nonlinear Response Models. *CRC Press*.

10. Mentré F, Mallet A, Baccar D. Optimal design in random-effects regression models. *Biometrika* 1997; 84: 429–442.

11. Nyberg J, Bazzoli C, Ogungbenro K, et al. Methods and software tools for design evaluation in population pharmacokinetics-pharmacodynamics studies. *Br J Clin Pharmacol* 2015; 79: 6–17.

12. Nguyen TT, Mentré F. Evaluation of the Fisher information matrix in nonlinear mixed effect models using adaptive Gaussian quadrature. *Computational Statistics & Data Analysis* 2014; 80: 57–69.

13. Riviere M-K, Ueckert S, Mentré F. An MCMC method for the evaluation of the Fisher information matrix for non-linear mixed effect models. *Biostatistics.* 2016; 17: 737–50.

14. Mielke T. Approximation of the Fisher information and design in nonlinear mixed effects models (Ph.D thesis). Otto-von-Guericke-Universitat Magdeburg, Germany. 2012.

15. Waite TW, Woods DC. Designs for generalized linear models with random block effects via information matrix approximations. *Biometrika* 2015; 102: 677–693.

16. Waite TW. Design of experiments with mixed effects and discrete responses plus related topics (Ph.D thesis). University of Southampton, United Kingdom. 2012.

17. Ueckert S, Mentré F. A new method for evaluation of the Fisher information matrix for discrete mixed effect models using Monte Carlo sampling and adaptive Gaussian quadrature. *Computational Statistics & Data Analysis* 2017; 111: 203–219.

18. Riviere M-K, Mentré F. R package MIXFIM, version 1.0 http://mc-stan.org/.

19. Stan Development Team. Rstan: the R interface to stan, version 2.12.0. http://mc-stan.org/.

20. Atkinson AC, Bogacka B. Compound D- and Ds-optimum designs for determining the order of a chemical reaction. *Technometrics.* 1997; 39: 347–56.

21. Retout S, Comets E, Samson A, et al. Design in nonlinear mixed effects models: optimization using the Fedorov-Wynn algorithm and power of the Wald test for binary covariates. *Stat Med* 2007; 26: 5162–5179.

22. Sölkner J. Choice of optimality criteria for the design of crossbreeding experiments. *J Anim Sci* 1993; 71: 2867–2873.

23. Buckland ST, Burnham KP, Augustin NH. Model selection: an integral part of inference. *Biometrics.* 1997; 53: 603–18.

24. Atkinson AC. DT-optimum designs for model discrimination and parameter estimation. *J Stat Plan Infer* 2008; 138: 56–64.

25. Nguyen TT, Bénech H, Delaforge M, et al. Design optimisation for pharmacokinetic modeling of a cocktail of phenotyping drugs. *Pharm Stat* 2016; 15: 165–177.

26. Schorning K, Bornkamp B, Bretz F, et al. Model selection versus model averaging in dose finding studies. *Stat Med* 2016; 35: 4021–4040.

27. Aoki Y, Röshammar D, Hamrén B, et al. Model selection and averaging of nonlinear mixed-effect models for robust phase III dose selection. *J Pharmacokinet Pharmacodyn* 2017; 44: 581–597.

28. Buatois S, Ueckert S, Frey N, Retout S, Mentré F. Comparison of model averaging and model selection in dose finding trials analyzed by nonlinear mixed effect models. *AAPS J*. 2018; 20: 56.

29. Lestini G, Ueckert S, Mentré F. Model-based optimal robust design in pharmacometrics. PODE, Uppsala, Sweden; 2016. http://www.maths.qmul.ac.uk/~bb/PODE/PODE2016_JLestini.pdf

30. Kiefer J, Wolfowitz J. The equivalence of two extremum problems. *Canadian Journal of Mathematics* 1960; 12: 363–366.

31. Ogungbenro K, Aarons L. Population Fisher information matrix and optimal design of discrete data responses in population pharmacodynamic experiments. *J Pharmacokinet Pharmacodyn* 2011; 38: 449–469.

32. Kuhn E, Lavielle M. Maximum likelihood estimation in nonlinear mixed effects models. *Comput Stat Data Anal.* 2005; 49: 1020–1038.

33. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika.* 1934; 26: 404–13.

34. Blazer D, Burchett B, Service C, et al. The association of age and depression among the elderly: an epidemiologic exploration. *J Gerontol* 1991; 46: M210-215.

35. Landerman LR, Mustillo SA, Land KC. Modeling repeated measures of dichotomous data: testing whether the within-person trajectory of change varies across levels of between-person factors. *Soc Sci Res* 2011; 40: 1456–1464.

36. Nguyen TT, Bazzoli C, Mentré F. Design evaluation and optimisation in crossover pharmacokinetic studies analysed by nonlinear mixed effects models. *Stat Med* 2012; 31: 1043–1058.

37. Mehtälä J, Auranen K, Kulathinal S. Optimal designs for epidemiologic longitudinal studies with binary outcomes. *Stat Methods Med Res* 2015; 24: 803–818.

38. Cook RD, Wong WK. On the equivalence of constrained and compound optimal designs. *J Am Stat Assoc.* 1994; 89: 687–92.

39. Yeatts SD, Gennings C, Crofton KM. Optimal design for the precise estimation of an interaction threshold: the impact of exposure to a mixture of 18 polyhalogenated aromatic hydrocarbons. *Risk Anal* 2012; 32: 1784–1797.

40. Winkens B, Schouten HJA, van Breukelen GJP, et al. Optimal designs for clinical trials with second-order polynomial treatment effects. *Stat Methods Med Res* 2007; 16: 523–537.

41. Alhorn K, Schorning K, Dette H. Optimal designs for frequentist model averaging. *arXiv:180705234 [stat]*, http://arxiv.org/abs/1807.05234 (2018, accessed 1 August 2018).

42. Dubois A, Lavielle M, Gsteiger S, et al. Model-based analyses of bioequivalence crossover trials using the stochastic approximation expectation maximisation algorithm. *Stat Med* 2011; 30: 2582–2600.

43. Savic RM, Mentré F, Lavielle M. Implementation and evaluation of the SAEM algorithm for longitudinal ordered categorical data with an illustration in pharmacokinetics–pharmacodynamics. *AAPS J*. 2010; 13:44–53.

44. Plan EL, Maloney A, Mentré F, et al. Performance comparison of various maximum likelihood nonlinear mixed-effects estimation methods for dose-response models. *AAPS J* 2012; 14: 420–432.

45. Harman R, Filová L, Richtárik P. A randomized exchange algorithm for computing optimal approximate designs of experiments. ArXiv180105661 Stat [Internet]. 2018 [cited 2018 Jun 12]. Available from: http://arxiv.org/abs/1801.05661

46. Yu Y. Monotonic convergence of a general algorithm for computing optimal designs. *Ann Statist* 2010; 38: 1593–1606.

47. Wong WK, Chen R-B, Huang C-C, Wang W. A modified particle swarm optimization technique for finding optimal designs for mixture models. *PloS One*. 2015; 10: e0124720.

48. Kim S, Wong WK. Extended two-stage adaptive designs with three target responses for phase II clinical trials. *Stat Methods Med Res* 2017; 962280217709817.

49. Foo LK, McGree J, Eccleston J, et al. Comparison of robust criteria for D-optimal designs. *J Biopharm Stat* 2012; 22: 1193–1205.

50. Loingeville F, Nguyen TT, Riviere M-K, Mentré F. Using Hamiltonian Monte-Carlo to design longitudinal count studies accounting for parameter and model uncertainties. PAGE, Budapest, Hungary; 2017. Available from: https://www.page-meeting.org/pdf_assets/3612-FLoingeville_PAGE_2017.pdf

51. Bazzoli C, Retout S, Mentré F. Design evaluation and optimisation in multiple response nonlinear mixed effect models: PFIM 3.0. *Comput Methods Programs Biomed* 2010; 98: 55–65.

52. Lestini G, Dumont C, Mentré F. Influence of the size of cohorts in adaptive design for nonlinear mixed effects models: an evaluation by simulation for a pharmacokinetic and pharmacodynamic model for a biomarker in oncology. *Pharm Res.* 2015; 32: 3159–69.

53. Strömberg EA, Hooker AC. The effect of using a robust optimality criterion in model based adaptive optimization. *J Pharmacokinet Pharmacodyn* 2017; 44: 317–324.

54. Fang X, Li J, Wong WK, et al. Detecting the violation of variance homogeneity in mixed models. *Stat Methods Med Res* 2016; 25: 2506–2520.

Figures

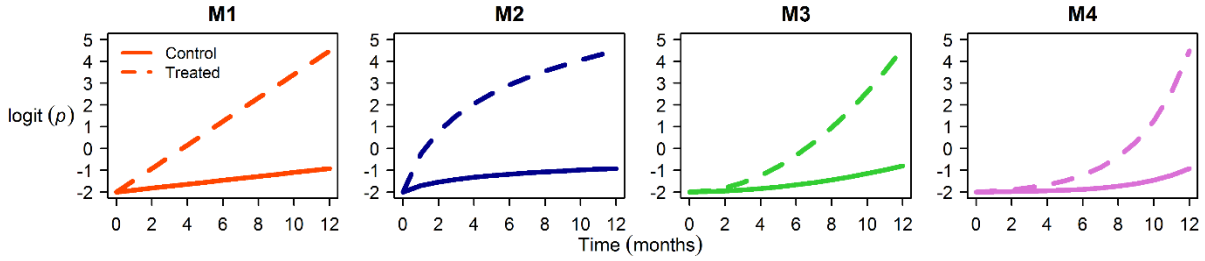

Figure 1. Plots of the logit of the response probability (p) over 12 months in each treatment group for the four candidate models: M1 linear (in orange), M2 loglinear (in blue), M3 quadratic (in green) and M4 exponential (in purple), using the fixed effect values indicated in Table 2. Solid lines represent the control group and dashed lines the treated group.

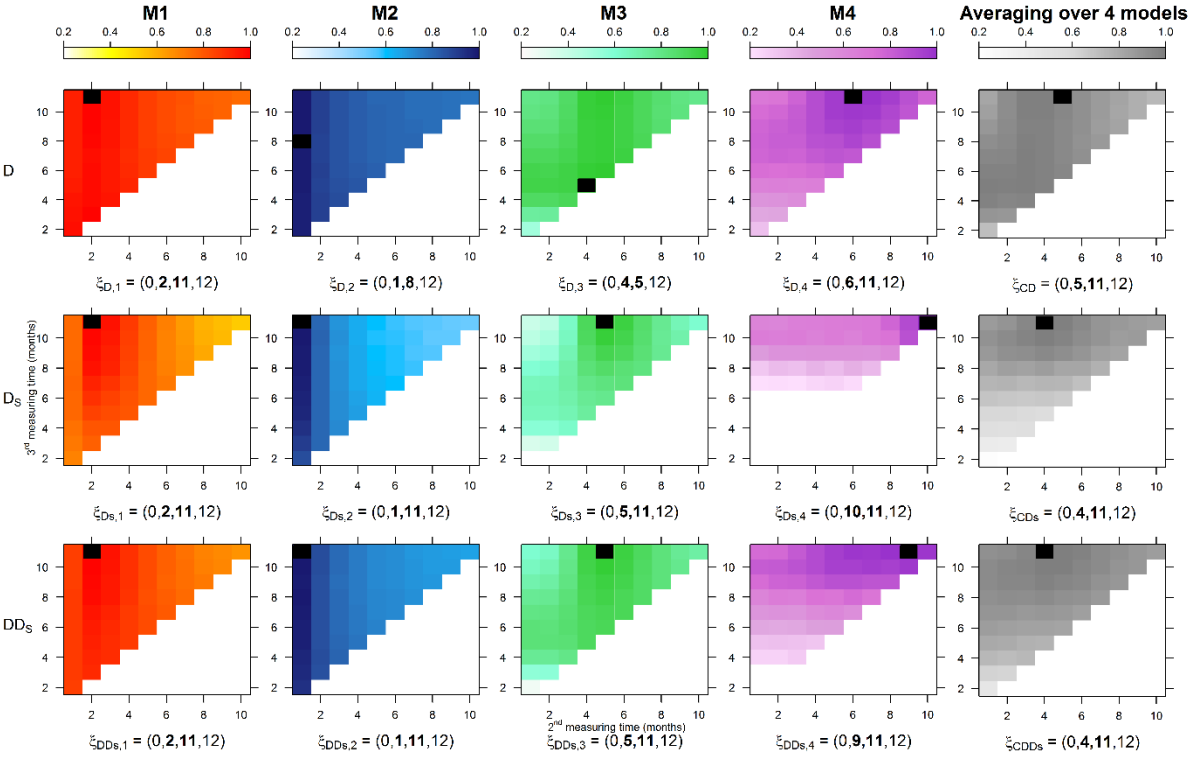

Figure 2. Heatmaps of design efficiencies, for each candidate model (M1 in red, M2 in blue, M3 in green, M4 in purple, by column) according to different considered optimality criterion (D-, D_S- or DD_S-optimality, by row). The fifth column (in grey) represents the robust criteria CD-, CD_S- or CDD_S-efficiencies. For each heatmap, the second measurement time is in abscissa and the third in ordinate. The filled black square shows the optimal design (*i.e.* efficiency = 1), which is also specified below each heatmap.

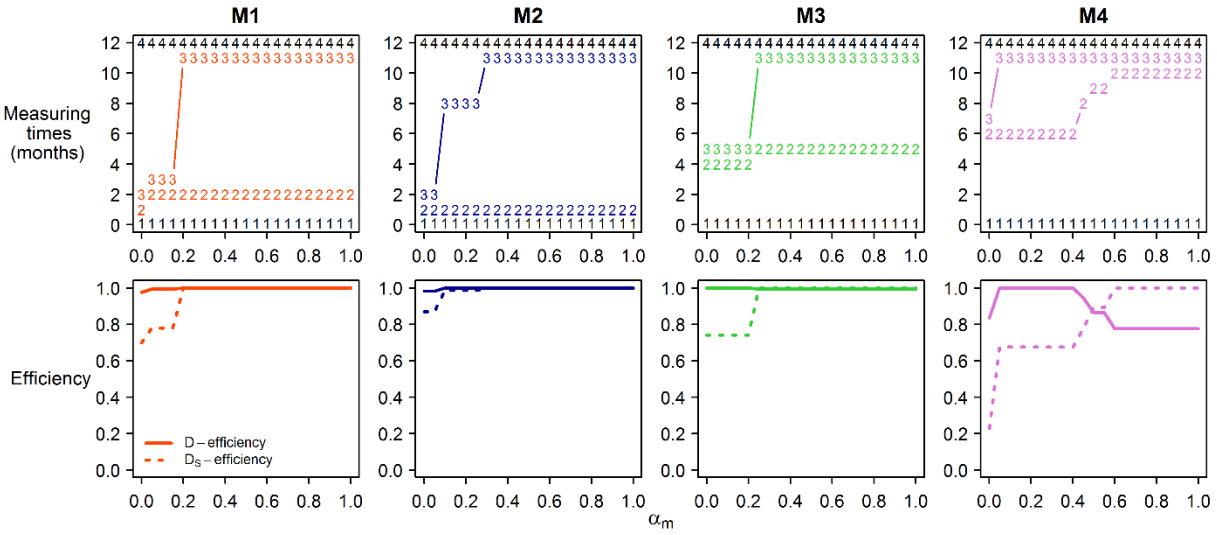

Figure 3. Results of the DD_S-optimal designs as function of the weight α_m, which quantifies the balance between the D- and the D_S-criteria, varying between 0 and 1. Results for each candidate model are represented by column (M1 in orange, M2 in blue, M3 in green, M4 in purple). First row, location of DD_S-optimal measurement times (in months): the symbol x represents the x^{th} measurement in each design. Second row, corresponding D- (solid line) and D_S-efficiency (dotted line), with respect to the D- and D_S-optimal design respectively.

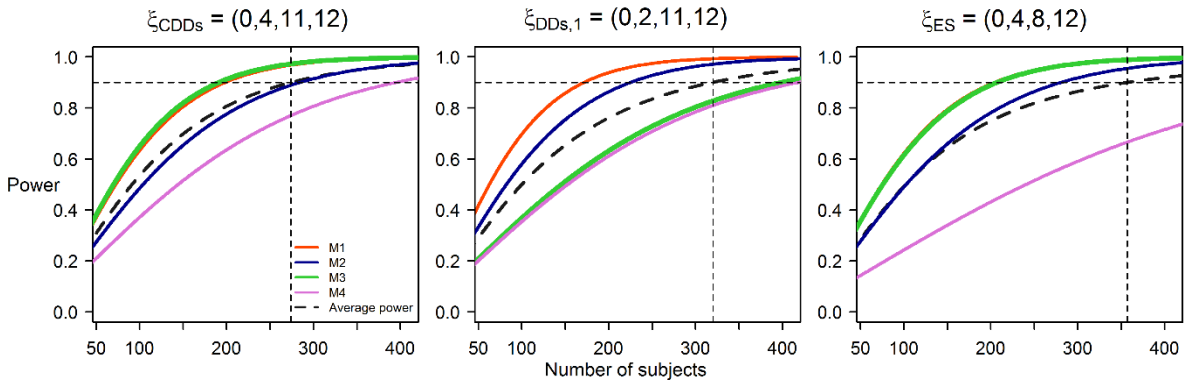

Figure 4. Predicted power for each model and in average with different designs, for varying total number of subjects from 50 to 450. From the left to the right, with the robust design ξ_{CDD_S}, the design optimized for model M1 $\xi_{DD_S,1}$, and the non-optimized design ξ_{ES}, the expected power from FIM for varying number of subjects is displayed in orange for M1, blue for M2, green for M3 and purple for M4. The average power over the 4 models is in dashed line. The dotted line crossing the x-axis shows the number of individuals needed to obtain an average power of 0.9.

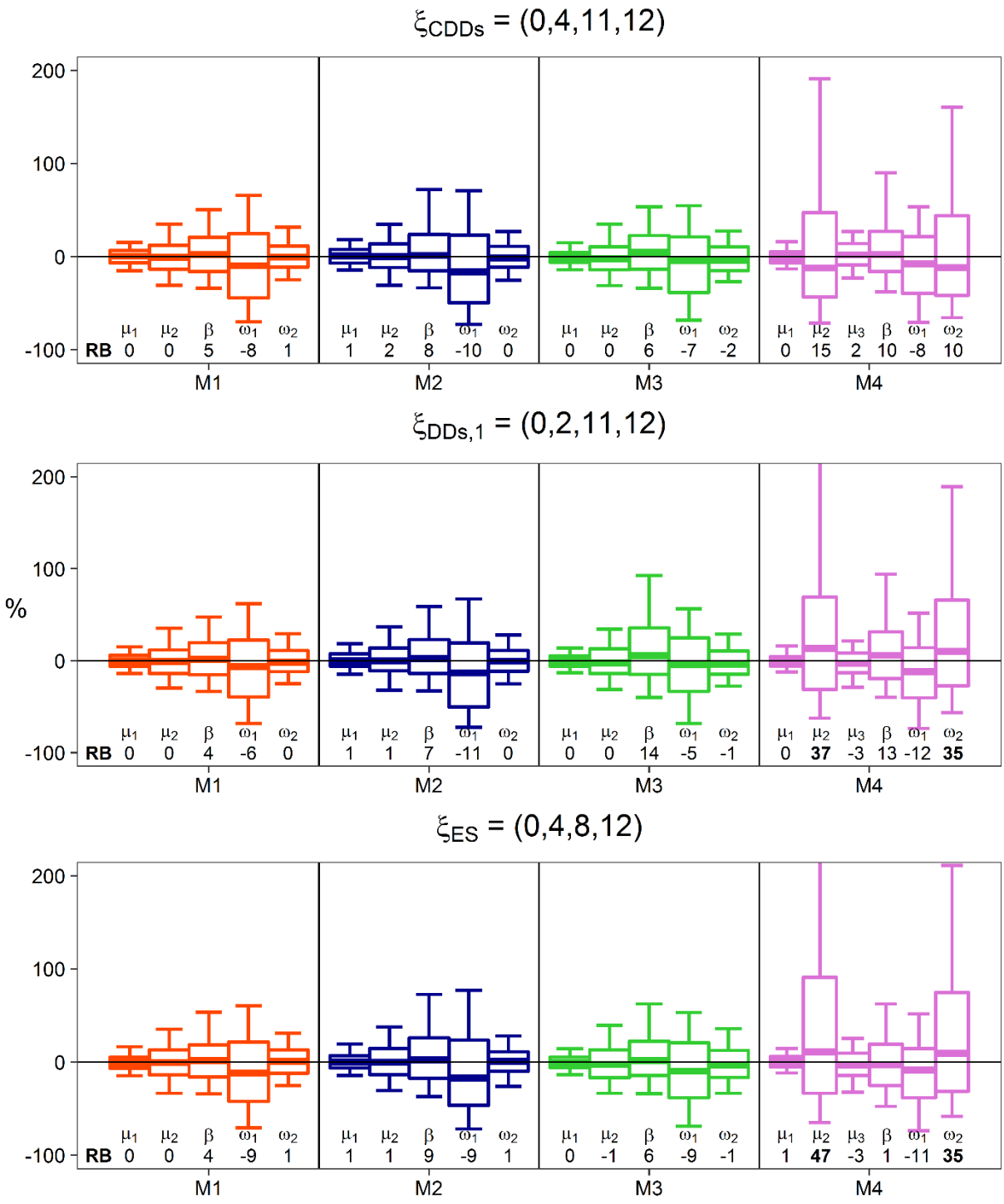

Figure 5. Boxplot of the relative estimation error (REE, in %, whiskers from 5th to 95th percentile) for each parameter of the four models: M1 in orange, M2 in blue, M3 in green and M4 in purple. From the top to the bottom with 500 datasets and 274 subjects by dataset, the robust design ξ_{CDD_S}, the optimal design of M1 $\xi_{DD_{S,1}}$, and the equispaced design ξ_{ES}. The relative bias (RB, in %) of each parameter are positioned under each corresponding boxplot, the values in bold font highlight the RB over 20%. μ are the fixed effects, β the covariate effect and ω the standard deviation of random effects.

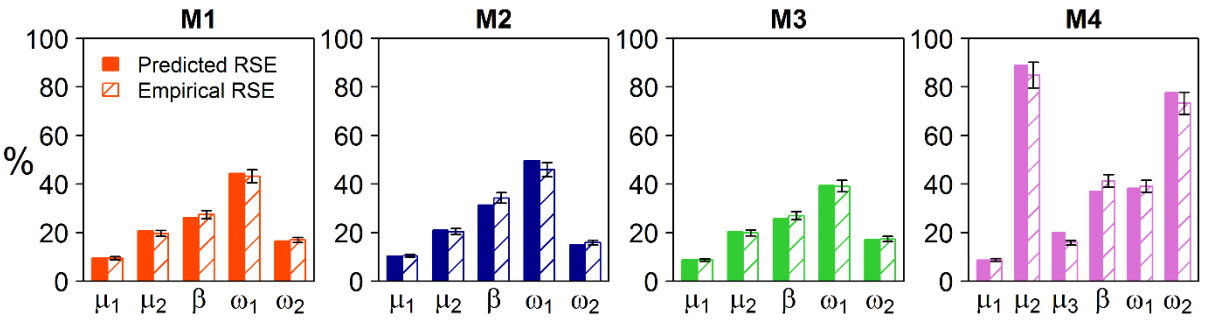

Figure 6. Comparison of predicted and observed estimation errors on all parameters of the four candidate models with 274 subjects and the robust design ξ_{CDD_S}. Relative standard error RSE (%) for the candidate models M1 to M4 are predicted using the expected FIM (full bar). Empirical RSE (hatched bar) are obtained by clinical trial simulations (CTS) with 500 datasets. μ are the fixed effects, β the covariate effect and ω the standard deviation of random effects.

Tables

Table 1. Various optimality criteria used depending on parameters of interest and accounting or not for model uncertainty.

Parameters of interest	**Given model m**	**Set of candidate models $m = 1, \dots, M$**
All the parameters ψ_m	D-optimality $\Phi_{D,m}(\Xi) = Det\left(\mathcal{M}(\psi_m, \Xi)\right)^{1/P_m}$ (4)	CD-optimality [7,24,25] $\Phi_{CD}(\Xi) = \prod_{m=1}^{M}\left(\Phi_{D,m}(\Xi)\right)^{w_m}$ (7)
Subset of parameters $\psi_{S,m}$	D_S-optimality $\Phi_{D_S,m}(\Xi) = \left(\frac{Det\left(\mathcal{M}(\psi_m,\Xi)\right)}{Det\left(\mathcal{M}(\psi_{T,m},\Xi)\right)}\right)^{1/S_m}$ (5)	CD_S-optimality $\Phi_{CD_S}(\Xi) = \prod_{m=1}^{M}\left(\Phi_{D_S,m}(\Xi)\right)^{w_m}$ (8)
Compromise between $\psi_{S,m}$ and $\psi_{T,m}$	DD_S-optimality $\Phi_{DD_S,m}(\Xi,\alpha_m) = \left(Det\left(\mathcal{M}(\psi_{T,m},\Xi)\right)\right)^{\frac{1-\alpha_m}{P_m-S_m}}\left(\frac{Det\left(\mathcal{M}(\psi_m,\Xi)\right)}{Det\left(\mathcal{M}(\psi_{T,m},\Xi)\right)}\right)^{\frac{\alpha_m}{S_m}}$ (6)	CDD_S-optimality $\Phi_{CDD_S}(\Xi) = \prod_{m=1}^{M}\left(\Phi_{DD_S,m}(\Xi)\right)^{w_m}$ (9)

ψ_m is the vector of population parameters (of size P_m) for a model m: $\psi_{S,m}$ the subset of parameters of interest (of size S_m) and $\psi_{T,m}$ which are not of interest. $\mathcal{M}(\psi_m, \Xi)$ is the Matrix containing the Fisher information for all the parameters ψ_m and $\mathcal{M}\left(\psi_{T,m}, \Xi\right)$ is obtained by truncation of $\mathcal{M}(\psi_m, \Xi)$ and keeping only the rows and columns corresponding to $\psi_{T,m}$. α_m $(0 \leq \alpha_m \leq 1)$ is a term which quantifies the balance between the D- and the D_S-optimality criteria, and expresses the interest in the precision of estimation for $\psi_{S,m}$. w_m is the weight for the model m $(\sum_{m=1}^{M} w_m = 1)$.

Table 2. Population parameter values ψ_m for each candidate model m ($m = 1, \ldots, 4$): M1 is the linear model, M2 the loglinear model, M3 the quadratic model and M4 the exponential model.

	M1	**M2**	**M3**	**M4**
μ_1	-2	-2	-2	-2
μ_2	0.09	0.42	7.50×10^{-3}	2.01×10^{-2}
μ_3	-	-	-	0.33
β	5	5	5	5
ω_1	0.70	0.70	0.70	0.70
ω_2	0.17	0.79	1.41×10^{-2}	3.79×10^{-2}

μ are the fixed effects, ω the standard deviation of random effects, and β the effect of the treatment covariate.

Table 3. D-, D_S- and DD_S-efficiencies for different optimal designs, accounting or not for model uncertainty.

Design Ξ $\Xi = \{N = 100, \xi\}$	$\mathbf{E_{D,1}}$	$\mathbf{E_{D,2}}$	$\mathbf{E_{D,3}}$	$\mathbf{E_{D,4}}$
$\xi_{ES} = (0, 4, 8, 12)$	0.908	0.926	0.975	0.869
$\xi_{D,1} = (0, 2, 11, 12)$	1	0.898	0.812	0.706
$\xi_{D,2} = (0, 1, 8, 12)$	0.932	1	0.875	0.787
$\xi_{D,3} = (0, 4, 5, 12)$	0.916	0.839	1	0.646
$\xi_{D,4} = (0, 6, 11, 12)$	0.829	0.802	0.960	1
$\xi_{CD} = (0, 5, 11, 12)$	0.860	0.805	0.994	0.958
$\xi_{CD_S} = (0, 4, 11, 12)$	0.910	0.828	0.977	0.882
$\xi_{CDD_S} = (0, 4, 11, 12)$	0.910	0.828	0.977	0.882
	$\mathbf{E_{D_S,1}}$	$\mathbf{E_{D_S,2}}$	$\mathbf{E_{D_S,3}}$	$\mathbf{E_{D_S,4}}$
$\xi_{ES} = (0, 4, 8, 12)$	0.841	0.646	0.835	0.393
$\xi_{D_S,1} = (0, 2, 11, 12)$	1	0.806	0.434	0.619
$\xi_{D_S,2} = (0, 1, 11, 12)$	0.751	1	0.389	0.627
$\xi_{D_S,3} = (0, 5, 11, 12)$	0.774	0.600	1	0.667
$\xi_{D_S,4} = (0, 10, 11, 12)$	0.529	0.519	0.611	1
$\xi_{CD} = (0, 5, 11, 12)$	0.774	0.600	1	0.667
$\xi_{CD_S} = (0, 4, 11, 12)$	0.865	0.637	0.906	0.655
$\xi_{CDD_S} = (0, 4, 11, 12)$	0.865	0.637	0.906	0.655
	$\mathbf{E_{DD_S,1}}$	$\mathbf{E_{DD_S,2}}$	$\mathbf{E_{DD_S,3}}$	$\mathbf{E_{DD_S,4}}$
$\xi_{ES} = (0, 4, 8, 12)$	0.882	0.754	0.924	0.723
$\xi_{DD_S,1} = (0, 2, 11, 12)$	1	0.863	0.644	0.765
$\xi_{DD_S,2} = (0, 1, 11, 12)$	0.859	1	0.611	0.755
$\xi_{DD_S,3} = (0, 5, 11, 12)$	0.827	0.722	1	0.947
$\xi_{DD_S,4} = (0, 9, 11, 12)$	0.683	0.672	0.761	1
$\xi_{CD} = (0, 5, 11, 12)$	0.827	0.722	1	0.947
$\xi_{CD_S} = (0, 4, 11, 12)$	0.893	0.751	0.954	0.894
$\xi_{CDD_S} = (0, 4, 11, 12)$	0.893	0.751	0.954	0.894

D-, D_S- and DD_S-efficiencies (respectively $E_{D,m}, E_{D_S,m}$ and $E_{DD_S,m}$) and their corresponding compound efficiencies (E_{CD}, E_{CD_S} and E_{CDD_S}) computed for each optimal design when ignoring model uncertainty ($\xi_{D,m}, \xi_{D_S,m}, \xi_{DD_S,m}$) or accounting for model uncertainty (ξ_{CD}, ξ_{CD_S} and ξ_{CDD_S}) with each model m ($m = 1, \ldots, 4$). The efficiencies are highlighted in lightgrey if between 0.5 and 0.8 and in darkgrey if below 0.5.

Table 4. Observed D-efficiency computed from clinical trial simulations for each model, with respect to the robust design.

	M1	**M2**	**M3**	**M4**
ξ_{CDD_S} : *reference*	*1*	*1*	*1*	*1*
$\xi_{DD_{S,1}}$	1.11	1.06	0.79	0.74
ξ_{ES}	0.98	0.94	0.98	0.84

For each model M1 to M4, the inverse of the variance-covariance matrix of parameter estimates is calculated for each of the three designs: the DD_S-optimal design for M1 ($\xi_{DD_{S,1}}$), the non-optimized equispaced design (ξ_{ES}) and the robust design (ξ_{CDD_S}). The observed D-efficiency between designs is defined as the ratio of determinants of these matrices, normalized by the number of parameters.

Table 5. Wald tests predictions and simulations on the covariate effect, under H_0 and under H_1.

	Under H_0		Under H_1			
	Nominal type I error	Observed type I error	FIM predicted power			Observed power
				Binomial PI_{95}	Bootstrap CI_{95}	
M1	0.05 [0.033,0.073]	0.048	0.967	[0.949,0.982]	[0.963,0.971]	0.988
M2		0.060	0.889	[0.859,0.916]	[0.873,0.902]	0.988
M3		0.068	0.973	[0.959,0.986]	[0.969,0.976]	0.996
M4		0.036	0.769	[0.731,0.806]	[0.741,0.792]	0.860

Under H_0, the 95% prediction interval calculated by binomial test is [0.033,0.073] for 500 simulations. For each model M1 to M4, the predicted power to detect a covariate effect β under H_1 using the Fisher information matrix (FIM) is calculated with the design $\Xi = \{N = 274, \xi_{\mathrm{CDD}_S} = (\mathbf{0}, \mathbf{4}, \mathbf{11}, \mathbf{12})\}$. Two intervals are presented around the predicted power: a prediction interval obtained by binomial test (PI) and an uncertainty interval (CI) using bootstrap from MC/HMC method for FIM evaluation. The observed power from clinical trial simulations (CTS) is given as the observed proportion of significant Wald tests performed on the covariate effect β with estimated standard error from 500 simulated clinical trials with a type I error of 5%.

Table 6. Parameter estimation precision in EPESE analysis (Observed) in [35] and FIM evaluation (Predicted).

	Parameter values	**Observed SE**	**Predicted SE**
μ_1	-7.205	0.372	0.270
μ_2	1.677	0.107	0.074
β_1	5.604	0.384	0.367
β_2	0.034	0.140	0.157
${\omega_1}^2$	15.017	1.757	1.316

μ are the fixed effects, ω the standard deviation of random effect, and β the effect of the cognitive imparment.

Supplementary Material

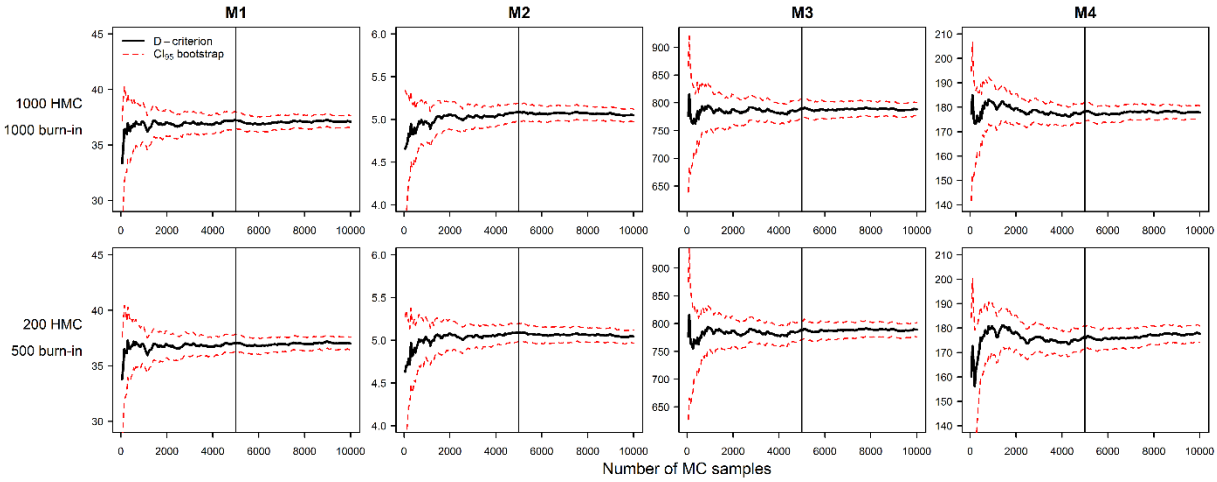

Figure S1. Convergence plots of D-criterion with a non-optimized design. The D-criterion is plotted in black as a function of the number of Monte Carlo samples, with the equispaced design $\xi_{ES} = (\mathbf{0}, \mathbf{4}, \mathbf{8}, \mathbf{12})$, for each model M1 to M4 by column. For the computation of the Fisher information matrix, two different algorithm settings are displayed by row (1000 Hamiltonian Monte-Carlo (HMC) samples with 1000 burn-in on the left or 200 HMC with 500 burn-in on the right). The red dotted lines represent the 2.5th and the 97.5th percentiles of the D-optimality values using bootstrap.

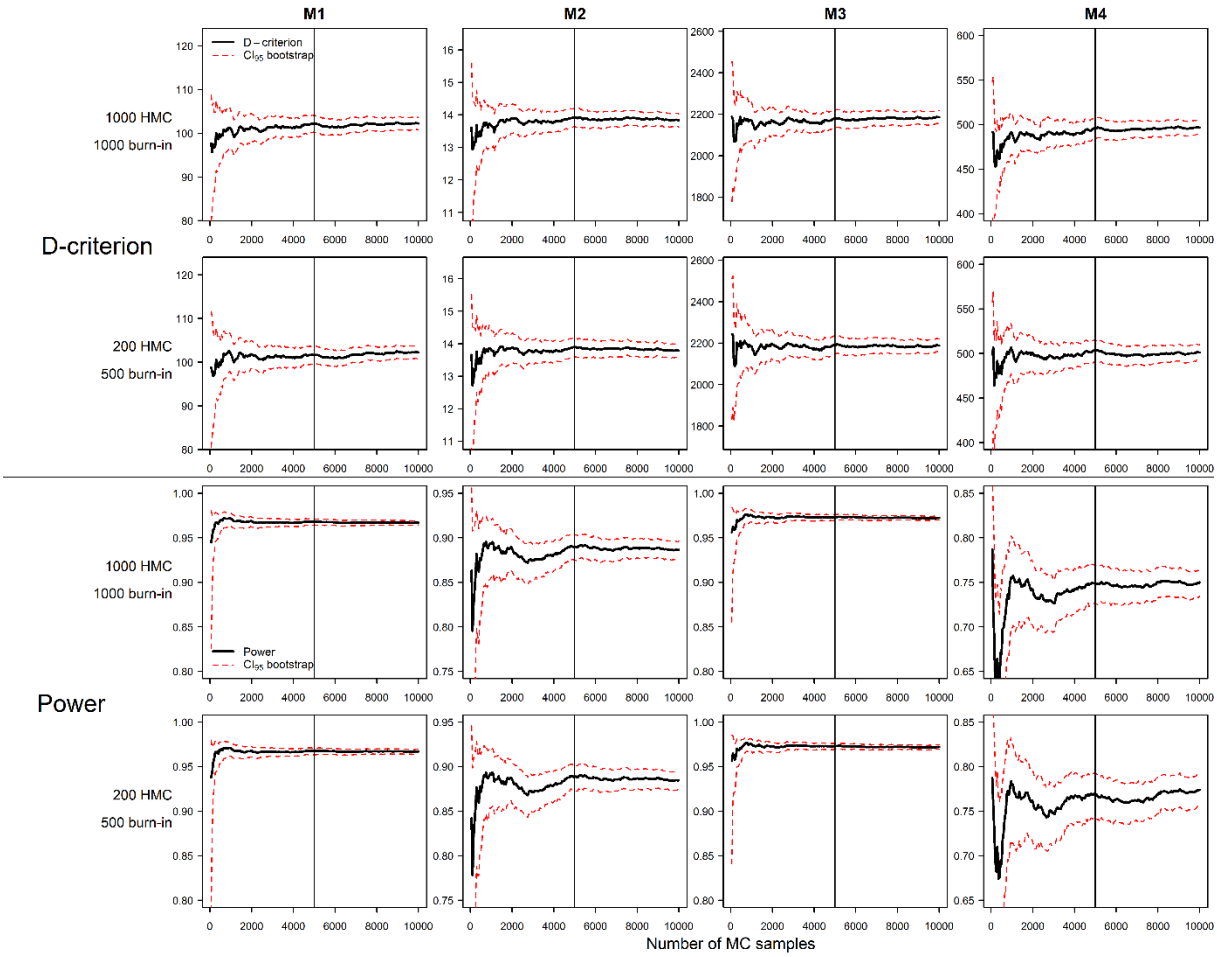

Figure S2. Convergence plots of D-criterion and power with the robust design. For each model M1 to M4 by column, the D-criterion on the top part and the power on the bottom part, are plotted in black as a function of the number of Monte Carlo samples with the robust design $\xi_{\mathrm{CDD_S}} = (\mathbf{0}, \mathbf{4}, \mathbf{11}, \mathbf{12})$. Different algorithm settings for the computation of the Fisher information matrix are represented (1000 Hamiltonian Monte-Carlo (HMC) samples and 1000 burn-in on the first row of each part or 200 HMC and 500 burn-in on the second row of each part), The red dotted lines represent the 2.5$^{\mathrm{th}}$ and the 97.5$^{\mathrm{th}}$ percentiles of the bootstrap values

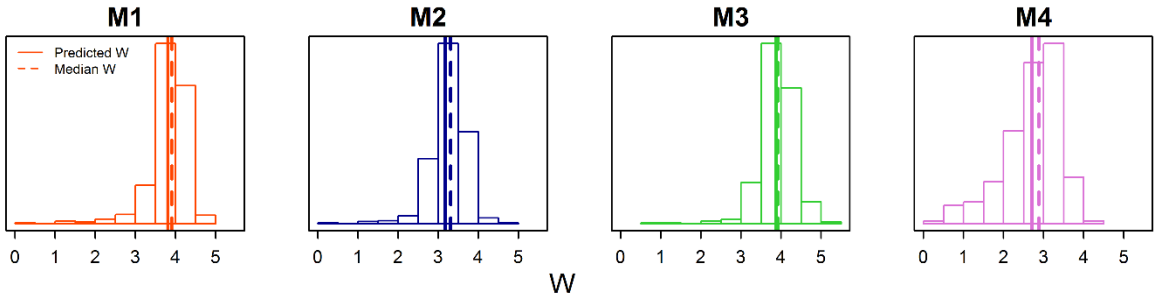

Figure S3. Distribution of observed Wald test statistic with the robust design. For each model M1 to M4, the distribution of the 500 observed Wald test statistic W is represented as histograms. For each model, the solid line is the FIM predicted W and the dashed line is the median of observed W.

Table S1. Observed power of the Wald test without bias on β with the robust design.

	M1	**M2**	**M3**	**M4**
FIM predicted power	0.968	0.889	0.973	0.77
CTS observed power	0.988	0.988	0.996	0.86
CTS observed power using simulated β	0.976	0.914	0.992	0.746

FIM predicted power are obtained using the true value of β (5) and the SE(β) calculated from FIM evaluation. CTS observed power are obtained using estimated β and their SE(β) as given by the SAEM algorithm. CTS observed power using simulated β are obtained using the true value of β (5) and SE(β) as given by the SAEM algorithm.