Statistics in Biosciences (2022) 14:237–258
https://doi.org/10.1007/s12561-021-09323-5

Leveraging Natural History Data in One- and Two-Arm Hierarchical Bayesian Studies of Rare Disease Progression

Arnaud Monseur, et al. ***[full author details at the end of the article]***

Received: 3 March 2021 / Revised: 17 September 2021 / Accepted: 20 September 2021 /
Published online: 15 October 2021

Abstract

The small sample sizes inherent in rare and pediatric disease settings offer significant challenges for clinical trial design. In such settings, Bayesian adaptive trial methods can often pay dividends, allowing the sensible incorporation of auxiliary data and other relevant information to bolster that collected by the trial itself. Previous work has also included the use of one-arm trials augmented by the participants' own natural history data, from which the future course of the disease in the absence of intervention can be predicted. Patient response can then be defined by the degree to which post-intervention observations are inconsistent with the predicted "natural" trajectory. While such trials offer obvious advantages in efficiency and ethical hazard (since they expose no new patients to a placebo, anathema to patients or their parents and caregivers), they can offer no protection against bias arising from the presence of any "placebo effect," the tendency of patients to improve merely by being in the trial. In this paper, we investigate the impact of both static and transient placebo effects on one-arm responder studies of this type, as well as two-arm versions that incorporate a small concurrent placebo group but still borrow strength from the natural history data. We also propose more traditional Bayesian changepoint models that specify a parametric functional form for the patient's post-intervention trajectory, which in turn allow quantification of the treatment benefit in terms of the model parameters, rather than semi-parametrically in terms of a response relative to some "null" model. We compare the operating characteristics of our designs in the context of an ongoing investigation of centronuclear myopathies (CNMs), a group of congenital neuromuscular diseases whose most common and severe form is X-linked, affecting approximately 1 in 50,000 newborn boys. Our results indicate our two-arm responder and changepoint methods can offer protection against placebo effects, improving power while protecting the trial's Type I error rate. However, further research into innovative trial designs as well as ongoing dialog with regulatory authorities remain critically important in rare disease research.

The members of the NatHis-MTM Study Group are listed in Acknowledgment.

1 Introduction

The field of rare disease is particularly challenging for professionals who design, lead, and manage randomized clinical trials (RCTs). The very small sample sizes associated with such rare disease patient populations make it difficult for traditional clinical trial designs to achieve the statistical power desired by trial sponsors, while at the same time maintaining limits on Type I error (false positive) rates required by regulatory authorities. Pediatric RCTs often present similar challenges.

Fortunately, recent developments in clinical trial methodology have contributed to decrease the number of patients needed for clinical development. In particular, historical data (often from previous RCTs) may help to appreciate not only the natural evolution of a population that closely matches the studied population, but also the economic and social burden of the disease [1]. More sensitive outcomes may capture a minimal change in patients [2], and surrogate endpoints or biomarker expression levels may demonstrate the proof of mechanism and engagement before a clinical benefit is measured [3]. Finally, the selection of specific populations that are expected to be best responders to treatment [4] may help to maximize the treatment effects.

Incorporating auxiliary information in the statistical plan offers the opportunity to augment statistical power. Hierarchical Bayesian statistical approaches [5, 6] are particularly well suited to this task, since they offer a formal framework for combining auxiliary information via prior distribution. Specific examples of methods often used in rare and pediatric disease include power priors [7], commensurate priors [8], and robust mixture priors [9]. These approaches attempt to weight the auxiliary information appropriately, often by judging how well its results agree with the direct information obtained inside the trial itself. Neuenschwander and Schmidli [10] provide a review of Bayesian hierarchical approaches to incorporating historical data, including their links to meta-analysis. Cooner et al. [11] offer a related review of methods useful in rare disease, while Basu and Carlin [12] focus on pediatric applications, where the approach is used to borrow strength from corresponding adult data. On this last point, Gamalo-Siebers et al. [13] also review Bayesian approaches in pediatric trials, as well as provide European and U.S. regulatory perspectives. Most recently, techniques incorporating propensity scores have been used to borrow from real-world (non-randomized) historical datasets. Zhao et al. [14] illustrate the use of commensurate priors for this purpose, while Wang et al. [15] and Li et al. [16] recommend a composite likelihood approach that has been suggested for use by US regulatory authorities.

Perhaps, the most common RCT usages of auxiliary data are designs that borrow strength from historical data on the control group, i.e., patients that would be assigned placebo or standard of care (SOC) in a traditional RCT. Unlike corresponding data on the novel treatment's effect in humans, historical control data will often be available, either from the sponsor's own previous research on the target population or through blinded control data sharing sources, such as the TransCelerate project (https://www.transceleratebiopharmainc.com/). In pediatric research, historical data on adult controls are often used to supplement or even replace pediatric controls, processes referred to as *partial* and *full extrapolation*, respectively [17]. While such extrapolation is often necessitated by small sample sizes and ethical concerns, it also carries the risk of biasing estimates of the drug effect, due to incommensurability between the auxiliary and primary data sources. For example, even if the patients in the historical study are roughly the same age as those in the current trial, temporal evolution in the SOC may mean that historical placebo success rates are too low, making the novel treatment appear more effective relative to the current SOC than it really is.

To combat this problem, many researchers in rare and pediatric disease are turning to a new auxiliary data source: historical observational data on the patients actually in the trial, referred to as *natural history data.* The basic idea here is reminiscent of a crossover study: each patient acts as their own control, with their natural history data providing the baseline estimate of the control effect. Each patient then "crosses over" to active treatment, and again the impact on the clinical endpoint of interest is measured. While not a true crossover study (since no patient "crosses back" from treatment to SOC), the use of natural history data offers a more ethical approach to rare and pediatric clinical trials, albeit one whose evidential value is lower than that of a traditional two-arm RCT.

In this paper, we propose a class of natural history study (NHS)-leveraging trials designed for evaluating rare or pediatric drugs, where randomizing a full complement of patients to placebo is both unethical and practically impossible. Our model operates on longitudinal continuous responses y_{ij}, where i indexes the patient and j indexes the time of observation. Our "base design" is a one-arm study and is reminiscent of one recently used by Fouarge et al. [18] in a study of centronuclear myopathies (CNMs), a group of rare congenital neuromuscular diseases. The design uses each patient's individual NHS-based disease trajectory to establish a baseline from which response to active treatment can be judged. Like most modern Bayesian adaptive trials, our design uses simulation to calibrate its operating characteristics, including Type I error and power. We then consider the impacts of various departures from the model's assumptions on these and other characteristics. In particular, we measure the impacts of the temporal length and number of patient's historical observations, as well as the presence of varying types of "placebo effect," a well-known threat to the validity of one-arm studies.

Next, we modify our design to a two-arm study that includes a very small placebo group, to assess whether such a change can assist in reducing bias without an unacceptable corresponding increase in variance (or ethical hazard). This work involves modestly extending the size of our model in carefully prescribed ways that minimize the additional estimation burden. We then compare the performance of this model to our initial one-arm design, again in both the presence and absence of various placebo effects. We also consider a hierarchical Bayesian changepoint model [19] that parametrizes the post-intervention trajectory, permitting the significance of the treatment effect to be judged using the posterior distribution of the change in slope. Our ultimate goal is a design that offers sufficient clinical evidence, protects patients, and will be acceptable to regulatory authorities in the United States, Europe, and elsewhere.

The remainder of our paper evolves as follows. Section 2 describes our base one-arm NHS-leveraging design, and shows how it can be calibrated to have any desired Type I error and power performance. Section 3 then reevaluates this performance in the presence of various forms of model misspecification, including a placebo effect. In Sect. 4, we introduce our extended two-arm models, and again judge whether they can outperform the simpler one-arm model, as well as a traditional two-arm frequentist model. Finally, Sect. 5 discusses our findings and offers directions for future work in this area.

2 One-Arm Bayesian Natural History Data Model

Beginning with the one-arm Bayesian natural history model of Fouarge et al.[18],[1] the observed response y_{ij} for subject i and time j, scaled between 0 and 1, is modeled as

$$y_{ij} \sim Beta\left(a_{ij}, b_{ij}\right).$$

The beta distribution is defined by the parameters a_{ij} and b_{ij}, which are defined in terms of the mean μ_{ij} and “the sample size” v of the distribution as

$$a_{ij} = \mu_{ij} * v \text{ and } b_{ij} = \left(1 - \mu_{ij}\right) * v.$$

The parameter v is estimated from the data, and the mean μ_{ij} is defined as a random effects model with logit-link function,

$$\mu_{ij} = \frac{1}{1 + \exp(-\chi_{ij})}, \text{where} \chi_{ij} = \alpha_i + \beta_i * \left(T + t_j\right), \tag{1}$$

and T is a constant to center time. The variability in the model can be derived using properties of the beta distribution. The temporal evolution of the mean is linear on the logistic scale. This implies that except near the boundary, the evolution of the response is approximately linear. On a long time scale, the evolution of patients is probably not linear, but when reduced to a trial compatible time scale of 6 months or 1 year, linearity is a reasonable expectation. A similar assumption of linearity is regularly accepted in rare disease; for instance, when evaluating Duchenne muscular dystrophy patients with the 6-min-walk distance when the time window is on the order of a year, even though this is a measure with a non-linear inverse U-shaped evolution over the life-course [20]. Capitalizing on this model, the method can be summarized by the three following items:

1. *Individual prediction* Combining the NHS data with the run-in data from patients enrolled but not included in the NHS allows the derivation of individual predictive

[1] The difference in the sample size parameter between this paper and that in Fouarge et al. [18] is due to a typo in that older paper. The sample size parameter v need not be time- or subject-specific.

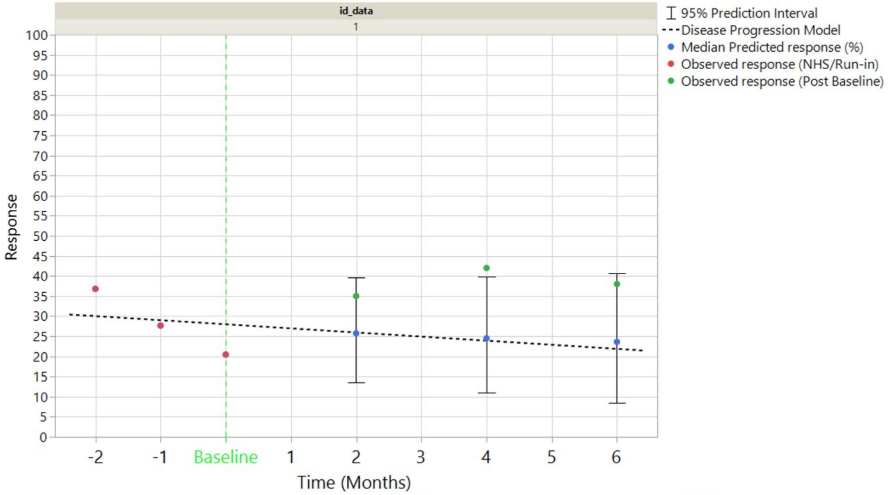


Fig. 1 Graphical Representation of the global strategy to define a responder

distributions of the endpoint or endpoints over time after treatment administration. For the time being, only pre-treatment data (i.e., from the NHS or a short "run-in" period) are used to fit our null model. All data gathered post-administration are used only to determine patient responder status, as follows:

2. *Identifying responders* The joint predictive probability of improvement (increase or decrease over time depending on the endpoint measured) is computed for each patient. If the joint predictive probability of the patient's observed improvement is smaller than some threshold (say, 0.05), then a patient is declared to be a responder. Formally, a patient is considered a responder if

$$P\left(y_{i,t+1} > \hat{y}_{i,t+1} \text{and} y_{i,t+2} > \hat{y}_{i,t+2} \text{and} y_{i,t+3} > \hat{y}_{i,t+3}\right) < 0.05$$

where $\hat{y}_{i,t+1}$ is the observed value for patient i at time $t + 1$, and $y_{i,t+1}$ is the random variable as derived from the model for patient i at time $t + 1$.

Figure 1 shows a graphical representation of the global strategy to define a responder. Given the observed natural history (red dots) of a patient for a given endpoint (here, bounded between 0 and 100), the evolution of the data can be predicted using the defined model (dotted line), which is of course also influenced by the responses of the other subjects in the dataset. The observed responses during the trial post-baseline (green dots) can then be compared to the predictive distributions readily generated using the Bayesian hierarchical model. If the joint probability of observing these values is sufficiently low, then the patient is classified as a responder. The intervals are 95% point-wise prediction intervals.

3. *Control of trial Type I error* Using this definition of responder, the predictive distribution of the rate of response is simulated under the null hypothesis of no treatment effect. A statistical significance threshold value is then determined to guarantee an overall Type I error of less than or equal to 5%.

This predictive probability-based definition of responder is reminiscent of a "Bayesian *p*-value" [21], and is particularly useful in settings where we seek a very generally applicable framework for detecting a treatment effect. Our hierarchical Bayesian model can be applied to potential primary and secondary endpoints by age class and by genotype. Patient-level random effects $\left(\alpha_i, \beta_i\right)$ enable the modeling of differences in level and progression of the disease. Since our responses are scores bounded by an upper and lower value, the beta distribution is useful to ensure that predictions do not fall outside the possible score range.

The model is readily fit using Proc MCMC in SAS 9.4,[2] using diffuse, non-informative prior distributions for all unknown parameters. More formally, the following priors were used:

- The random effect vectors $\left(\alpha_i, \beta_i\right)$ are assigned a bivariate normal prior with mean (α, β) and variance matrix V
- For the fixed effects (α, β), a normal distribution centered on zero with variance equal to 2
- The sample size parameter ν is assigned an improper uniform distribution, bounded below by zero
- The matrix for generation of the random effects is sampled from an inverse Wishart(2,S) where the hyperparameter S is given by

$$S = \begin{pmatrix} 1 & 0 \\ 0 & 0.01 \end{pmatrix}.$$

In addition, prior sensitivity was assessed. Multiple scenarios obtained by changing the hyperparameters of the priors were conducted, wherein the fixed effect parameters were given different means and variances, while the variance–covariance matrix of the random effects was also modified. For the fixed effects, the new hyperparameters allowed for priors that were more informative (reducing the prior variability of the parameters by 50%) and also encouraged a mean shift of 1 prior standard deviation. For the random effects, the hyperparameter matrix S above was replaced by 2S; small positive correlations between the two components were also added. Despite these changes, the resulting posterior model estimates never deviated by more than 5% from their original values, suggesting that the effect of priors on the parameters was minimal. This suggests our originally specified priors are indeed minimally informative, as desired.

Our motivating dataset arose in the context of a natural history study of 59 European patients suffering from CNM [22]. The patients were evaluated four times a year if they were younger than 2 years of age, twice a year if below 6 years old, and for older patients at 6 months and 12 months after enrollment and then once a year thereafter. Strict standard operating procedures were used to prospectively

[2] We note that alternative Bayesian computational approaches exist, including the increasingly popular BRMS package in R that calls the STAN software (mc-stan.org).

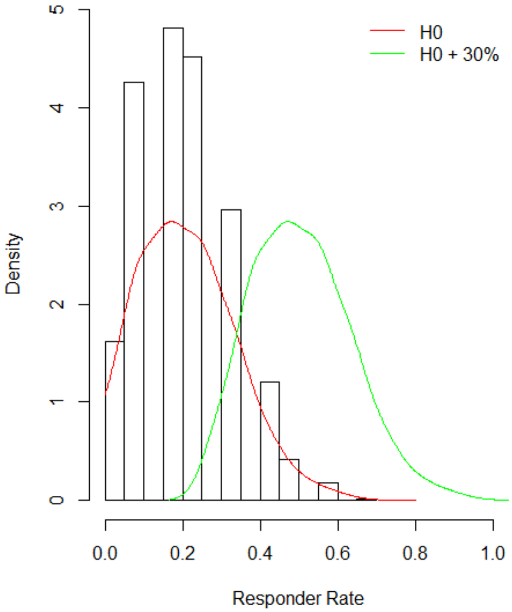


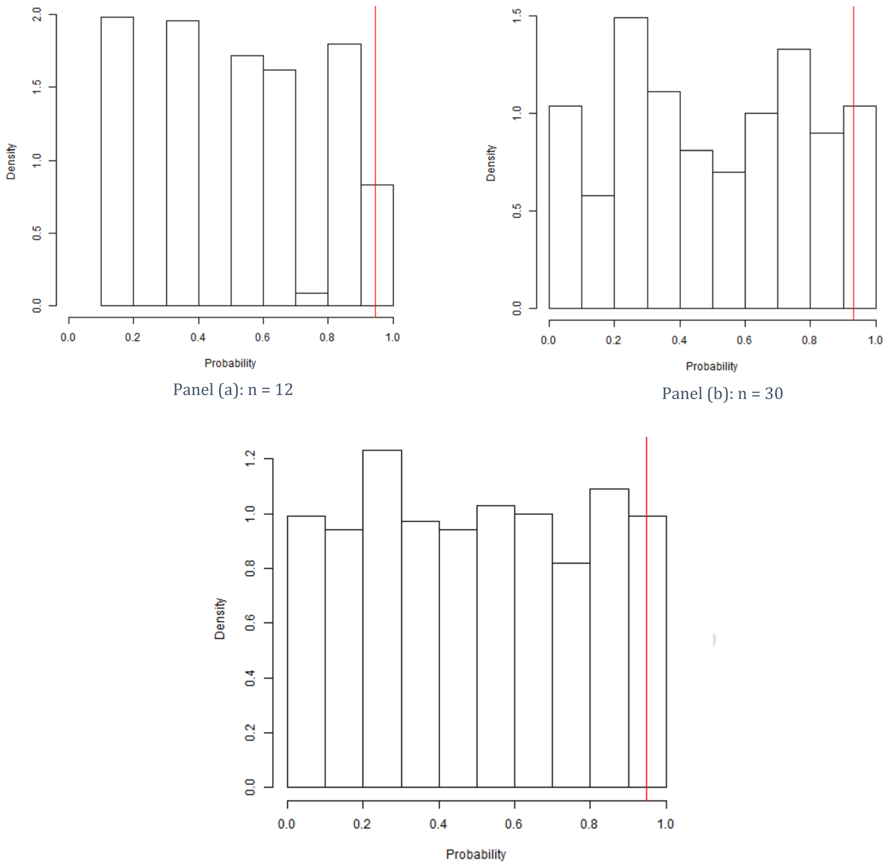
Fig. 2 Responder rate assuming no treatment effect and a responder rate under treatment of an additional 30%

acquire this data. The main clinical responses consist of quantitative endpoints which are bounded from below and above (scales, percentages, etc.). These bounds are the reason that Fouarge et al. [18] adopt the beta likelihood, since this makes the results more interpretable for clinical colleagues. As such, we too have adopted this approach, even though applying a normal model on a transformed (say, logit) scale could possibly offer statistical advantages. In future clinical trials, trial subjects who did not participate in the NHS will go through a brief pre-intervention "run-in" period, to collect data to be used in estimating the trajectory of their disease progression.

2.1 Type I Error Control

Assuming no treatment effect, it is possible to derive the naturally occurring responder rate in future trials. This responder rate depends only on the threshold to define a responder. A smaller (more stringent) threshold would imply fewer naturally occurring responders and may preclude detecting a change under treatment. On the other hand, a larger threshold may lead to an excessive rate of spurious responders.

Figure 2 shows the responder rates using the threshold of 5% under our model assuming no treatment effect (red curve) and a hypothetical increased responder rate under treatment (green curve). This curve was obtained by simulating studies of 12 patients. The red continuous shape is the corresponding beta distribution of the responder rate. To detect if the investigational product has a significant



Panel (a): n = 12

Panel (b): n = 30

Panel (c) : n=200

Fig. 3 Influence of the sample size on the distribution of the probabilities of success. Red vertical line represents the 95% quantile

effect, the responder rate in the future trial needs to be significantly greater than this responder rate. The red curve responder rate can be thought of as the “null responder rate” that needs to be compared to the one obtained in the trial.

Once the responder rate is obtained for the experimental treatment, the two responder rate posterior distributions can then be compared. If the difference between the two is sufficiently large, then the treatment can be considered a success. Typically, this is operationalized as requiring the proportion of the density of the difference lying to the right of zero to exceed some threshold (say, 95%). However, due to small sample sizes, this threshold may need to be adapted, as the usual asymptotic theory is not yet applicable. As such, simulations are used to better calibrate this threshold and ensure a false-positive rate of at most 5%.

To see this, note that the distribution of the probability that the difference in responder rates is greater than zero should follow a uniform distribution under

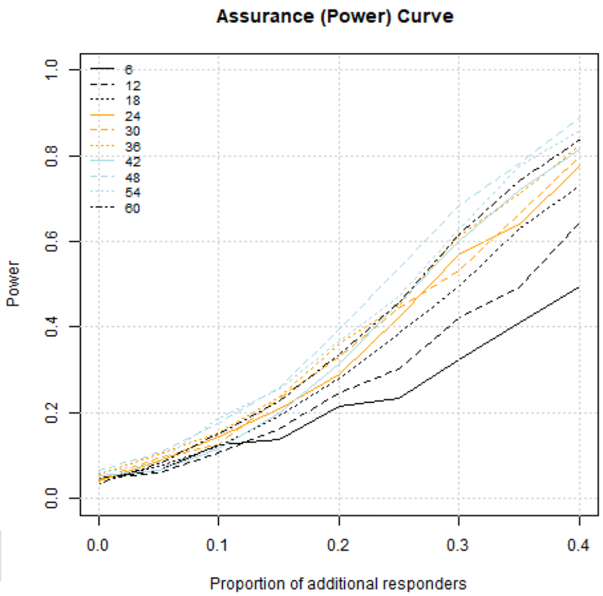


Fig. 4 Power curves

Table 1 Bayesian power for the one-arm model (no placebo effect)

n	6	18	30	42	54
0% Increase	0.042	0.049	0.056	0.035	0.054
40% Increase	0.505	0.741	0.817	0.834	0.87

the null assumption that there is no difference between the two responder groups. However, as can be seen from Fig. 3 this is approximately true for $n = 200$, but not for the smaller sample sizes $n = 12$ and 30. In each case, the red vertical line represents the 95% quantile. Under classical asymptotic theory this value should be 0.95, and the histogram should be nearly flat. However, for smaller sample sizes this value needs to be recalibrated to ensure a 5% Type I error rate.

From this, one can now power future studies. For each sample size assessed, one needs to recalibrate the threshold for significance between the two responder rate distributions. Assuming a particular increase in the proportion of additional responders in the treatment group, one can then obtain the power to succeed. Figure 4 shows the associated assurance (Bayesian power) curves. The Bayesian power is obtained by averaging the usual frequentist power over the prior distribution of the treatment effect. For each increase in the average treatment effect considered (x axis in Fig. 4), a prior distribution of the responder rate is obtained (as for the green curve in Fig. 2). The Bayesian power curve thus accounts for the uncertainties in the responder rates for all considered possible increases in average responder rates under treatment (x axis in Fig. 4). The corresponding

Table 2 Average (MC standard errors) of the 500 simulated prediction interval widths

3-visit run-in over	3 months	4 months	6 months
Average width	0.22 (0.096)	0.19 (0.079)	0.16 (0.072)
4-month run-in with	3 assessments	4 assessments	5 assessments
Average width	0.19 (0.079)	0.20 (0.080)	0.18 (0.072)

values for a 0% (null) and 40% (target) increase in responder rates are additionally detailed in Table 1. As can be seen from both the figure and the table, the Bayesian analog of Type I error has been appropriately controlled. We also see assurance greater than 80% for $n \geq 30$.

3 Robustness of the One-Arm Model

3.1 Effect of Run-In Length

For patients coming from a natural history study, a key component of the design is the time and number of assessments in the run-in period. This run-in needs to be appropriately defined as it will serve as the baseline against which the responder status is defined.

To assess the impact of the run-in length and number of assessments, the width of the prediction interval will serve as our main metric. Prediction interval width is a convenient measure that indicates the effect on the uncertainty and reflects the “ease” of detecting a responder under treatment. Indeed, the narrower the prediction interval, the more likely a response is to be detected.

Our simulation plan is designed to mimic a real-world situation where not all patients would go through a run-in phase in the trial. Indeed, those coming from the natural history study would not need a run-in. Since these patients have longer historical data trajectories that have high leverage on the results of the modeling, the simulations include such patients in order to account for such a leveraging effect. Our simulation algorithm thus proceeds as follows:

- Simulate 5 patients randomly according to the results of the model previously fitted to the data. The number of data points for the run-in length and the number of visits to be simulated is determined by the scenario at hand (see tables below);
- Include these patients in the database, and remodel the new database with 7 extra patients from the original data (NHS patients), to maintain a total number of 12 patients;
- Compute the interval width for the first predicted visit after 3 months;
- After conducting these steps 500 times, assess the average width of the prediction interval over all replications for all patients jointly.

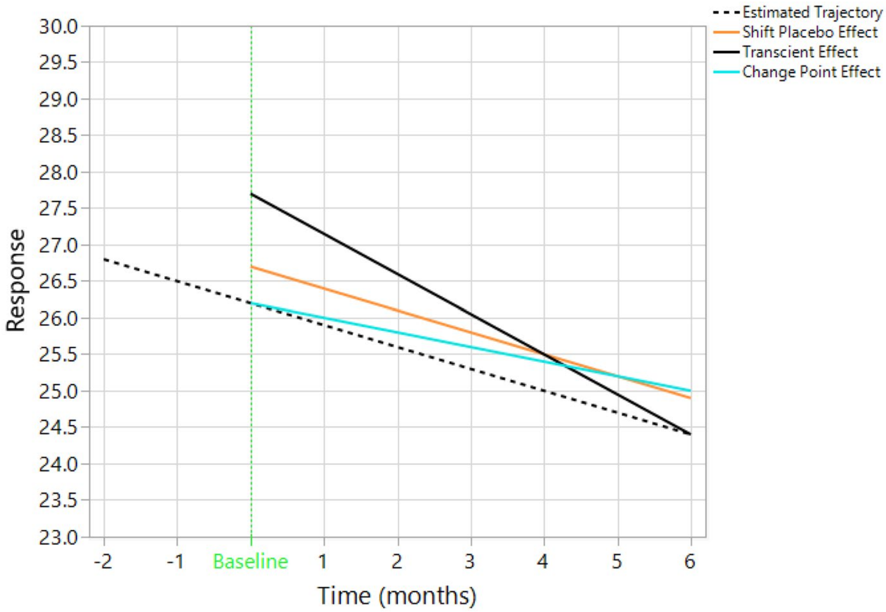


Fig. 5 Examples of various types of placebo effects and a potential changepoint model

Table 2 shows the average prediction interval widths and associated Monte Carlo standard errors for various run-in lengths, and for varying numbers of run-in observations. The first case is considered in the upper part of the table, where all three cases are conducted assuming 5 exchangeable patients with a 3-visit run-in, as in Fouarge et al. [18]. One can see that the intervals become narrower as the run-in period length increases, with the 6-month intervals being roughly 30% narrower than the 3-month intervals. This makes sense, as the longer observation window stabilizes the slope of the fitted regression line, leading to narrower prediction intervals. However, the lower part of the table reveals the that increase in the number of visits (from 3 to 4 or 5) in a fixed temporal window (here, 4 months) has little to no effect on interval width. This suggests that it is run-in *length* that is the key driver of improved model performance.

3.2 Impact of Placebo Effects

As mentioned above, a common concern with one-arm studies is they effectively assume no “placebo effect”; any deviation from the predicted trajectory is attributed to the treatment. As such, the impact of different placebo effects, and the probability that their responses will be mistakenly attributed to the treatment, needs to be assessed.

We consider three cases. In the first, it is assumed that the placebo increases the response rate by a constant proportion q across the time scale. That is, the drug delivers a static (time-constant) shift in the response, akin to an increase in the

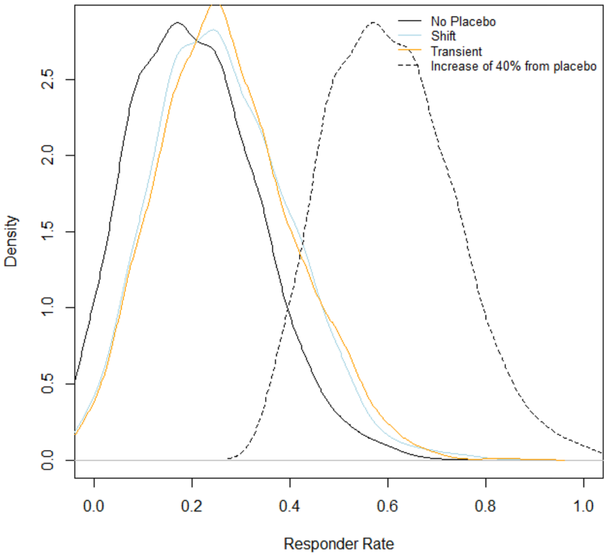


Fig. 6 Distribution of the responder rates used in the simulations for impact on the placebo effect

Table 3 Change in Power at a 40% increase in response rate when considering a placebo effect created by a shift of q units

Shift (q)	Sample size (n)				
	6	18	30	42	54
0.005	0.554	0.541	0.701	0.714	0.685
0.01	0.254	0.420	0.469	0.458	0.466
0.015	0.281	0.301	0.265	0.294	0.351
0.02	0.095	0.199	0.194	0.193	0.173
0.03	0.103	0.173	0.170	0.216	0.245

intercept α_i in Eq. (1). This corresponds to the orange line in Fig. 5, where a constant vertical shift is observed post baseline. The second case instead assumes the placebo effect is more dramatic initially, but is also transient, vanishing at the end of 6 months; this is the solid black line in the figure. Finally, the third case (shown as the light blue line) involves an improvement in slope starting at baseline. We defer a discussion of the impact of this “changepoint” model (and its potential not only as a placebo effect, but as a replacement for our responder model) to Sect. 4 below.

The response rate histogram corresponding to a constant shift of $q = 0.005$ is shown as the light blue curve in Fig. 6 below. We see higher responder rates than those under a completely inactive placebo (black curve), but not as high as under the 40% increase in efficacy we hope to see from the drug (black dotted curve). The probability that a study would be declared a success under this modified placebo effect with 18 patients is 54.1%, down from 74.1% in the no placebo effect case (see Fig. 4, Tables 3 and 1).

Table 4 Change in Power at a 40% increase in response rate with a transient placebo effect of r%

Transient shift (r)	Sample size (n)				
	6	18	30	42	54
0.01	0.298	0.547	0.638	0.670	0.631
0.02	0.285	0.316	0.376	0.431	0.473
0.03	0.079	0.200	0.172	0.196	0.193
0.04	0.056	0.042	0.057	0.056	0.060
0.05	0.098	0.038	0.049	0.048	0.040

Next, we consider the case where the placebo effect is transient, delivering an initial increase in response probability, but one that fades to zero over time (solid black line in Fig. 5). This is what has been seen empirically in several past studies of neuromuscular diseases; see e.g., Mercuri et al. [23] and Goemans et al. [20]. We assume the placebo effect is high at the start of the study (say, an initial increase of proportion r on the vertical scale), but then attenuates back linearly to no increase (i.e., the expected trajectory) at 6 months. In this scenario, the probability that the drug will be declared a success assuming $r = 0.01$ is 54.7% with 18 patients (see Table 4). The orange curve in Fig. 6 shows the distribution of the responder rate under this placebo effect. Again, we see responder rates slightly higher than those under a completely inactive placebo (black curve), but not nearly as high as those seen under the hoped-for 40% increase in efficacy (black dotted curve).

Replacing the 40% increase in response rate with a 0% increase, we can see the Bayesian Type I error that results from ignoring the impact of the placebo effect in the calibration of the threshold if indeed there is one (Step 3 of the general methodology described in Sect. 2). We reconsidered different magnitudes of the static (shift) and transient placebo effects to assess their impact. In all cases, the Type I error drops to 0 fairly quickly, due to the lack of opportunities to reject the null. Small placebo effects therefore do have a meaningful impact on false-positive rates if they are not accounted for. In cases where placebo effects are expected, alternative methods are therefore required, especially if the nature of the placebo effect is not known up front. Such methods are the subject of the next section.

4 Two-Arm Bayesian Natural History Data Models

In this section, we propose two alternative methods to incorporate a placebo arm in the trial design. First, a "responder" approach like that used in the one-arm model above will be considered. The responder rates obtained in each of the treatment arms will then be compared, and significance will be assessed as in the one-arm design. Second, a more standard Bayesian approach based on parametric modeling of the placebo effect is considered. Treatment effects will then be assessed by looking at the posterior distribution of the change in slope between the two treatment arms.

Method 1 (*Responder Method*) Let us assume that the response rates in the treated and control arms are p_{trt} and p_{plac}, respectively. Define $\Delta = p_{\text{trt}} - p_{\text{plac}}$, and reject H_0 if $P(\Delta > 0|\text{data}) \geq \delta$, where δ is determined to ensure a 5% Bayesian Type I error rate, as in Sect. 2.

Method 2 (*Parametric Changepoint Method*) Without loss of generality (though possible loss in computational efficiency), take $T = 0$ in model (1), and let the intervention take place at $t_j = 0$ for each patient. This implies that negative times indicate natural history or run-in observations, while positive times refer to post-intervention (in-trial) observations. Now extend model (1) to a linear *changepoint* model,

$$\chi_{ij} = \alpha_i + \beta_i t_j + \gamma_i t_j^+, \tag{2}$$

where the "+" superscript denotes positive part, i.e., $t_j^+ = t_j$ for $t_j > 0$, and $t_j^+ = 0$ for $t_j < 0$. Thus, γ_i is the change in slope after the intervention. Our original model assumes $\gamma_i = 0$ for patients receiving placebo, and defines response relative to the (α_i, β_i) straight line that continues after intervention. However, with observations on patients receiving placebo, we can estimate a change in slope for *both* drug and placebo patients. Note, however, that, unlike Method 1, this approach assumes we use *all* the data (not just the pre-intervention data) to estimate the parameters, since some of them are now unique to the post-intervention period.

Next, we define treatment effect as the difference between the drug and placebo post-intervention slopes. That is, we assume $\gamma_i \sim N\left(\gamma_{\text{trt}}, \sigma^2_{\text{trt}}\right)$ for treated patients, $\gamma_i \sim N(\gamma_{\text{plac}}, \sigma^2_{\text{plac}})$ for placebo patients, and place vague hyperpriors on the 4 hyperparameters. We then fit the model, and base our test on the posterior distribution of $\Delta = \gamma_{\text{trt}} - \gamma_{\text{plac}}$, rejecting H_0 if $P(\Delta > 0|\text{data}) \geq 0.95$.

We compare these two methods by assessing the power obtained for the same treatment and placebo effects. One can already speculate that the parametric modeling scheme will lead to less uncertainty in the estimates, and thus detect treatment effects more efficiently when the model is correct. However, this parametric method is strongly dependent on the assumed treatment and placebo effects in the model (2). This model dependency is alleviated by the responder method, which does not specify the form of the post-intervention model.

For comparison purposes, a frequentist method was also applied to the two-arm clinical trial. This test looks for a change in response from baseline to 6 months that is significantly higher for the treatment effect than for the placebo effect. Figure 7 plots traditional (frequentist) power for this method versus sample size. Each curve corresponds to a true treatment effect. Total sample size is plotted on the horizontal axis, and a 2:1 patient allocation scheme is assumed. One can see that, for a treatment effect of a 3-point increase (which corresponds to a 25% increase in the responder rate in the previous section), the power never exceeds 30% even with 60 patients. No uncertainty is assumed on the fixed treatment effects in these simulations, so these are curves of the evolution of power as a function of the sample size for different types of effects in the frequentist sense and *not* assurance (Bayesian power).

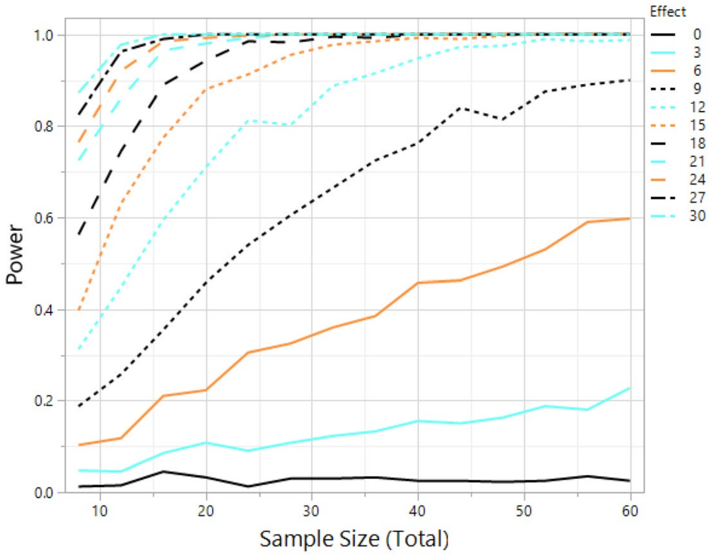


Fig. 7 Evolution of power in terms of sample size obtained for 2:1 randomized trial simulations using a frequentist method

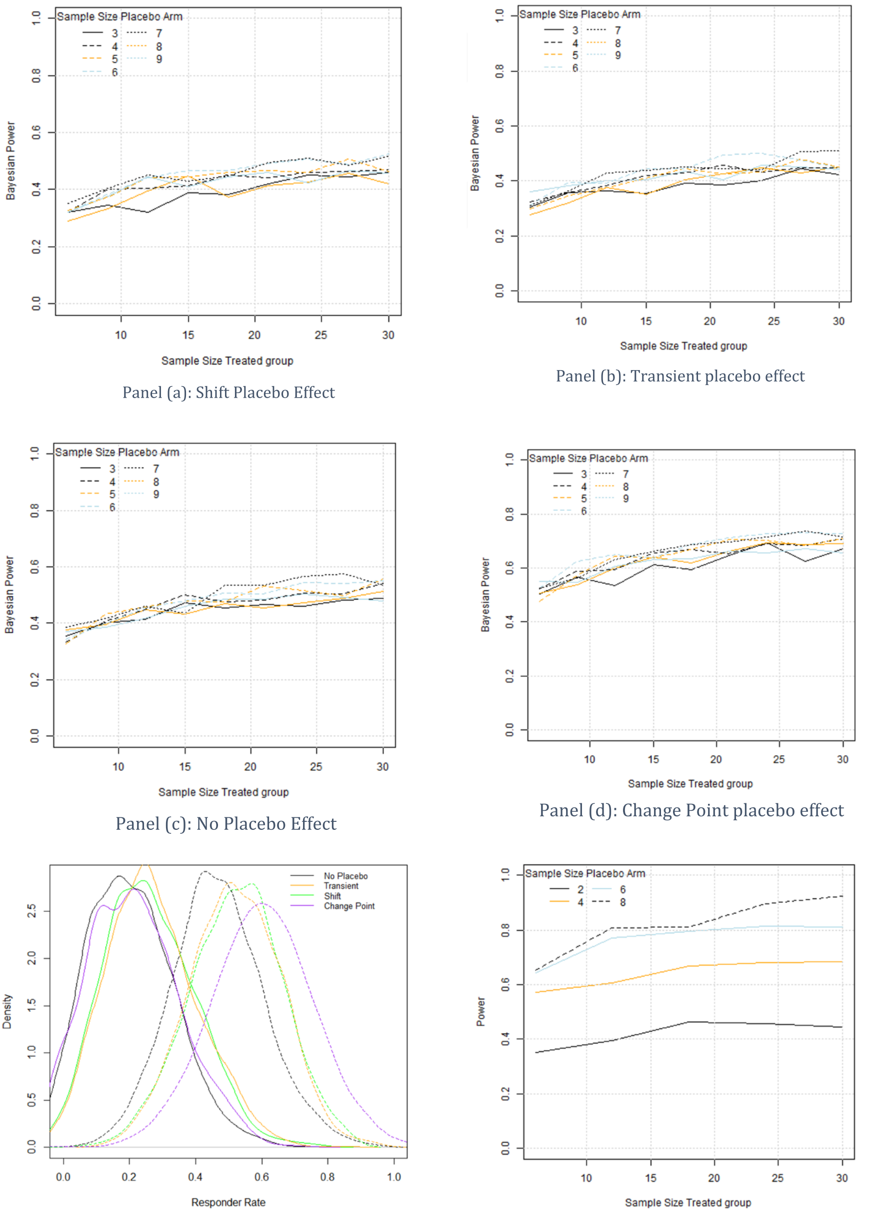
4.1 Responder Method Performance

Recognizing the negative impact on Type I error and power that ignoring the placebo effect may have, incorporating it into a two-arm model seems reasonable. Two different placebo effects will be considered in this section:

1. The constant shift placebo effect (q) introduced above (orange curve in Fig. 5);
2. The transient placebo effect (r) introduced above (solid black curve in Fig. 5); and
3. The changepoint placebo effect introduced above but not yet investigated (light blue curve in Fig. 5).

These three effects will allow us to compare the one-arm method's performance with those of our two-arm responder models. The assumed treatment effect will always be superimposed on the different placebo effects, and corresponds to a 3% increase in the response score at 6 months.

Panels (a) through (d) in Fig. 8 show the evolution of assurance (Bayesian power) in terms of the sample size of each of the treatment arms. One can directly see that power plateaus well below 1, with the changepoint placebo effect (panel d) adding the most artificial inflation of power. This is to be expected, since this type of effect (while atypical of most placebo effects) mimics the anticipated effect of the drug itself. Bayesian assurance accounts for the fact that, given our prior knowledge, many clinical trials will fail (placebo response will be higher than the treated response; this can be seen in Fig. 8e, in which solid lines



Panel (a): Shift Placebo Effect

Panel (b): Transient placebo effect

Panel (c): No Placebo Effect

Panel (d): Change Point placebo effect

Panel (e) : Densities of the responder rates

Panel (f): Bayesian power for the changepoint model

Fig. 8 **a** to **d** show the different Bayesian powers for a shift placebo effect, a transient placebo effect, no placebo effect, and the changepoint placebo effect using the responder methodology, respectively. **e** shows the densities used in these simulations, while **f** shows the Bayesian power for different sample sizes using the parametric changepoint model

Table 5 Comparison of power between the different modeling strategies for a given set of placebo effects and a total sample size of 30

	Frequentist	Two-Arm responder			One-Arm
	20 treated, 10 control	21 treated, 9 control	24 treated, 6 control	27 treated, 3 control	30 treated
Placebo effect types					
None	0.14	0.53	0.53	0.49	0.46
Shift effect	0.14	0.50	0.46	0.44	0.39
Transient effect	0.14	0.45	0.43	0.41	0.38

correspond to placebo groups and dashed lines to treatment groups). For further information on the distinction between Bayesian and frequentist power, the reader is referred to O'Hagan et al. [24] or Kruschke [25]. The priors used in this example for the responder rates were obtained from scenarios with 12 patients from the natural history study subject to each placebo effect.

To understand the benefit of considering this two-arm study, one must compare the one-arm model with a similar increase as considered here. The considered increase is a 25% increase in the responder rate with respect to the null. This 25% increase in the responder arm in a one-arm model approximately corresponds to the treatment effect used for the two-arm model on top of the placebo effect.

We next come to the issue of model comparison. Unfortunately, a fair comparison between the responder and parametric models using traditional Bayesian tools (Bayes factors, DIC, WAIC, or some other predictive or cross-validatory criterion) is not possible since the models do not take the same datasets as input. Specifically, the responder models use only the pre-intervention data, while the latter use all of the data, and thus their likelihoods are not directly comparable. Indeed, the responder models are not "predictive" models at all, using the post-intervention data only to determine a patient's response status.

Fortunately, a meaningful "global" comparison between the methods is still available by calculating the different powers obtained for a given sample size. If two methods deliver roughly the same Type I error, the method with the higher power would be preferable. Table 5 below shows this comparison for a total of 30 patients. The table clearly shows a superiority of the two-arm model over both the frequentist and the one-arm model. The gain in power from the two-arm model over the one-arm model is due to a better characterization of the responder rate for untreated patients, and thus makes the distinctions between treated and untreated easier than with a noisy hypothesized distribution under H_0 with the one-arm model. Further, the superimposition of the treatment effect on top of a small placebo effect makes accurate estimation of the responder rate slightly more important. There is also some indication that a more balanced randomization (21 treated, 9 control) leads to better power. Finally, the two-arm method's power also persists better under either the shift or transient placebo effects. The gain over the frequentist method is likely linked to the fact that in both the one-arm and two-arm models, all the available information are used to make the

Table 6 Power for different types of models for the change point placebo effect

	Model types			
	Method 2: Two-Arm Parametric Model (24 treated, 6 control)	Method 1: Two-Arm responder (24 treated, 6 control)	One-Arm responder Model (30 treated)	Frequentist
Placebo effect types				
Change point Placebo	0.820	0.699	0.765	0.140
Type-1 error	0.050	0.056	0.059	0.050

decision with respect to treatment effect (joint probabilities), rather than focusing solely on the data collected at the last visit.

Throughout all these examples, one can see that the power to detect a 3-point increase is considerably higher than in the classical frequentist setting. Furthermore, in comparison to the one-arm model, assuming no placebo effect to adequately compare, this scenario corresponds to an increase of approximately 25% in the responder rate. Indeed, when assessing the average increase in the responder rate with a linear increase of up to 3% at 6 months, this value increases by 25% from the naturally occurring average amount of responders.

4.2 Parametric Changepoint Model Performance

In this subsection, we consider a third type of placebo effect, namely a variety of effects arising from prespecified changepoint models. This will in turn allow us to compare the two two-arm methods (responder versus parametric changepoint) considered in this paper. This effect was plotted in the blue curve in Fig. 5. Figure 8f shows that straying away from the response models and using parametric changepoint methods when they are justified permits a further increase in Bayesian power, even to values in the 80–90% range.

Table 6 compares the two two-arm methods introduced in this paper and the classical frequentist two-arm method. As can be seen, the two-arm parametric method outperforms the two-arm "semiparametric" (responder) modeling strategy. However, as previously mentioned, these results are only valid under the assumption that the model is well specified. Our current changepoint model assumes a linear trajectory post intervention for both the placebo and treated groups. Of course, more sophisticated models can be considered-say, ones that include a discontinuous "jump" in response right after the intervention, mimicking our shift placebo effect. But any parametric model will be at risk of model misspecification. The second row of Table 6 compares the Type I error of the three procedures, and finds that, as expected, it is well calibrated for two-arm semiparametric methodology. This is also the case for the parametric model (5% error rate).

An interesting feature of such methods (both two-arm models) is the observed gain in power when changing the randomization scheme. Indeed, an increase of two patients in the placebo group may yield a greater increase in power than increasing the sample size of the treated group from 6 to 30. This can be seen in the difference between the black and blue curves (or the blue and pink curves) in Fig. 8f. This feature is especially true for the small sample sizes, as a small increase in the placebo arm induces a better characterization of the slope and a greater reduction in the uncertainty linked to its estimation. Table 6 also shows that the One-Arm model performs slightly better than the two-arm model in this case. This feature was not seen in the previous paragraph. An explanation for this feature is linked to the type of placebo response considered and especially its magnitude. Indeed, in Fig. 8f, the increase in responder rate due to the placebo effect is rather small, while the treated effect on the other hand is rather important. The drop in power due to model misspecification observed in the previous case is therefore not impacting the models here.

5 Discussion and Future Work

Randomized clinical trials in the rare disease setting offer multiple challenges. We have proposed one-and two-arm Bayesian alternatives to classical statistical methods enabling the design of clinical trials with a limited number of patients. Since rare neuromuscular disease exhibits a wide range of phenotypic reactions, the data are also subject to a large heterogeneity in behaviors. Venturing away from semiparametric modeling where a typical post-treatment parametric trajectory can be sensibly imposed may also permit better accounting for these differences.

Our methodology allows calibration of an appropriate threshold for significance to control the Type I error (due to small sample size and deviations from asymptotic theory). Our simulations showed the design can yield sufficient power, and substantially more than comparable two-arm frequentist designs.

Our semiparametric response models yield slightly less power than corresponding hierarchical parametric Bayesian changepoint models. However, this feature is compensated by the fact that the classical hierarchical parametric Bayesian model is valid only if the model is appropriate. Simple parametric forms for placebo and treatment effects may be difficult to justify in small and heterogeneous populations. We therefore believe that the proposed response model methodology allows us to conduct a clinical trial with few patients having very different phenotypic behaviors that are hard to parametrically characterize without excessive loss of power. This is because more complex behaviors will typically require estimation of additional model parameters, with an associated loss in the estimation of degrees of freedom.

Future research looks to extending the methodology used in this paper to various other rare disease settings. Further, a more in-depth investigation of the gain or loss with respect to parametric models may be considered. More complex designs could also be included into this framework. For example, the trial could add a “crossover” aspect, wherein each placebo patient eventually also receives the treatment, further reducing the overall sample size required. Proper handling of the placebo effect and detection of response in the presence of a crossover requires further investigation.

In addition, future research might also further “stress test” the proposed responder analysis by continuing to challenge its hypotheses. This might include assessing the impact of model misspecifications in the early run-in periods, the impact and handling of missing visits, or changes to the definition of responder. All of these features may be tested, changed, and investigated further to fit the desired purpose and lead to a valid approach for future clinical trials in the rare disease setting. As these trials sometimes necessitate various endpoint measures that may be dependent on the patient's status, a hybrid approach considering two or more measures that jointly determine responder status may also be considered in order to improve decision-making in this context.

Finally, it is worthwhile noticing that these methods are independent of the chosen endpoint. In clinical trials, they may be applied to various types of endpoints simultaneously. Given the information at hand when designing the trial, one method may be preferred over another depending on, for example, the placebo effect, the treatment effect, the trajectory type, and other factors.

Appendix

In this brief appendix, we provide the SAS computer code used to fit our models:

```
proc mcmc data = data nmc = 500000 nbi = 10000 thin = 20 outpost = outpost;

      array b[2] b0 b1;
      array mub[2] beta0 beta1;
      array Sigma[2,2];
      array S[2,2] (1 0 0 0.01);
      array parBeta[2] a bbeta;

      parms beta0 0.06 beta1 0.01;
      parms nu 0.1;
      parms Sigma;

      prior beta0 ~ normal(1, var= 1);
      prior beta1 ~ normal(1, var= 1);
      prior nu ~ general(0,lower=0);

      prior Sigma ~ iWish(2,S);

      random b ~ mvn(mub, Sigma) subject = ID;
      mu  = b0 + b1*(time+11.92);
      mu2 = 1/(1+exp(-mu));
      a = mu2*nu;
      bbeta = (1-mu2)*nu;
      model y ~ beta(a,bbeta);
run;
```

Acknowledgements This work was supported by a grant from Dynacure. The authors gratefully acknowledge the contributions of the NatHis-MTM study group for curating and allowing us access to their motivating dataset on CNM. the NatHis-MTM Study Group: Teresa Gidaro, Elena Gargaun, Virginie Chê, Ulrike Schara, Andrea Gangfuß, Adele D'Amico, James J. Dowling, Basil T. Darras, Aurore Daron, Arturo Hernandez, Capucine de Lattre, Jean-Michel Arnal, Michèle Mayer, Jean-Marie Cuisset, Carole Vuillerot, Stéphanie Fontaine, Rémi Bellance, Valérie Biancalana, Ana Buj-Bello, Jean-Yves Hogrel, Hal Landy, Kimberly Amburgey, Barbara Andres, Enrico Bertini, Ruxandra Cardas, Séverine Denis, Dominique Duchêne, Virginie Latournerie, Nacera Reguiba, Etsuko Tsuchiya, and Carina Wallgren-Pettersson.

References

1. Announssamy M, Seferian AM, Daron A et al (2021) (2021) Natural history of Type 2 and 3 spinal muscular atrophy: 2-year NatHis-SMA study. Ann Clin Transl Neurol 8(2):359–373. https://doi.org/10.1002/acn3.51281
2. Lilien C, Gasnier E, Gidaro T et al (2019) Home-based monitor for gait and activity analysis. J Vis Exp. https://doi.org/10.3791/59668
3. Frank DE, Schnell FJ, Akana C et al (2020) Increased dystrophin production with golodirsen in patients with Duchenne muscular dystrophy. Neurology 94(21):e2270–e2282. https://doi.org/10.1212/WNL.0000000000009233
4. Dangouloff T, Servais L (2019) Clinical evidence supporting early treatment of patients with spinal muscular atrophy: current perspectives. Ther Clin Risk Manag 15:1153–1161. https://doi.org/10.2147/TCRM.S172291
5. Berry SM, Carlin BP, Lee JJ, Muller P (2011) Bayesian adaptive methods for clinical trials. Chapman and Hall/CRC Press, Boca Raton, FL
6. Carlin BP, Louis TA (2009) Bayesian methods for data analysis, 3rd edn. Chapman and Hall/CRC Press, Boca Raton, FL
7. Ibrahim JG, Chen M-H (2000) Power prior distributions for regression models. Stat Sci 15(1):46–60
8. Hobbs BP, Carlin BP, Mandrekar S, Sargent DJ (2011) Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. Biometrics 67:1047–1056
9. Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B (2014) Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics 70(4):1023–1032
10. Neuenschwander B, Schmidli H (2020) Use of historical data. In: Lesaffre E, Baio G, Boulanger B (eds) Bayesian methods in pharmaceutical research. Taylor and Francis/CRC Press, Boca Raton, FL, pp 111–137
11. Cooner F, Williamson F, Carlin BP (2020) Bayesian frameworks for rare disease clinical development programs. In: Lesaffre E, Baio G, Boulanger B (eds) Bayesian methods in pharmaceutical research. Taylor and Francis/CRC Press, Boca Raton, FL, pp 243–257
12. Basu C, Carlin BP (2020) Bayesian hierarchical models for data extrapolation and analysis in pediatric disease clinical trials. In: Lesaffre E, Baio G, Boulanger B (eds) Bayesian methods in pharmaceutical research. Taylor and Francis/CRC Press, Boca Raton, FL, pp 259–270
13. Gamalo-Siebers M, Savic J, Basu C, Zhao X, Islas CD, Gopalakrishnan M, Guo A, Song G, Baygani S, Thompson L, Xia HA, Price KL, Tiwari RC, Carlin BP, for the DIA Bayesian Statistics Working Group (2017) Statistical modeling for Bayesian extrapolation of adult clinical trial information in pediatric drug evaluation. Pharm Stat 16:232–249
14. Zhao H, Hobbs BP, Ma H, Jiang Q, Carlin BP (2016) Combining non-randomized and randomized data in clinical trials using commensurate priors. Health Serv Outcomes Res Method 16:154–171
15. Wang C, Lu N, Chen WC, Li H, Tiwari R, Xu Y, Yue LQ (2020) Propensity score-integrated composite likelihood approach for incorporating real-world evidence in single-arm clinical studies. J Biopharm Stat 30(3):495–507
16. Li H, Chen W-C, Lu N, Song C, Wang C, Tiwari R, Xu Y, Yue LQ (2021) Mitigating study power loss caused by clinical Trial Disruptions Due to the COVID-19 pandemic: leveraging external data via propensity score-integrated approaches. Statistics in Biopharmaceutical Research. https://doi.org/10.1080/19466315.2020.1860813
17. Dunne J, Rodriguez W, Murphy M, Beasley B, Burckhart G, Filie J, Lewis J, Sachs H, Sheridan P, Starke P (2011) Extrapolation of adult data and other data in pediatric drug-development programs. Pediatrics 128(5):e1242-1249

18. Fouarge E, Monseur A, Boulanger B, Annoussamy M, Seferian AM, De Lucia S, Lilien C, Thielemans L, Paradis K, Cowling BS, Freitag C, Carlin BP, Servais L, the NatHis-MTM Study Group (2021) Hierarchical Bayesian modelling of disease progression to inform clinical trial design in centronuclear myopathy. Orphanet J Rare Dis 16(1):1–11
19. Carlin BP, Gelfand AE, Smith AFM (1992) Hierarchical Bayesian analysis of changepoint problems. Appl Stat 41:389–405
20. Goemans N, Signorovitch J, Sajeev G, Yao Z, Gordish-Dressman H, McDonald CM et al (2020) Suitability of external controls for drug evaluation in Duchenne muscular dystrophy. Neurology 95(10):1381–1391
21. Gelman A, Meng X-L, Stern HS (1996) Posterior predictive assessment of model fitness via realized discrepancies (with discussion). Stat Sin 6:733–807
22. Annoussamy M, Lilien C, Gidaro T, Gargaun E, Chê V, Schara U et al (2019) X-linked myotubular myopathy: a prospective international natural history study. Neurology 92(16):e1852–e1867. https://doi.org/10.1212/WNL.0000000000007319
23. Mercuri E, Darras BT, Chiriboga CA et al (2018) Nusinersen versus sham control in later-onset spinal muscular atrophy. N Engl J Med 378(7):625–635
24. O’Hagan A, Stevens JW, Campbell MJ (2005) Assurance in clinical trial design. Pharm Stat J Appl Stat Pharm Ind 4(3):187–201
25. Kruschke J (2014) Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan, 2nd edn. Academic Press, New York

Authors and Affiliations

Arnaud Monseur[1] · Bradley P. Carlin[2] · Bruno Boulanger[1] · Andreea Seferian[3] · Laurent Servais[3,4,5,6] · Chris Freitag[7] · Leen Thielemans[7,8] · the NatHis-MTM Study Group

✉ Bradley P. Carlin
Bradley.P.Carlin@pharmalex.com

[1] Pharmalex Belgium, Mont-Saint-Guibert, Belgium

[2] Pharmalex US, Burlington, MA, USA

[3] Institute I-Motion, Hôpital Armand Trousseau, Paris, France

[4] Division of Child Neurology, Centre de Référence Des Maladies Neuromusculaires, Department of Paediatrics, University Hospital of Liège and University of Liège, Liège, Belgium

[5] MDUK Oxford Neuromuscular Centre, Department of Paediatrics, University of Oxford, Oxford, UK

[6] Department of Paediatrics, Level 2, John Radcliffe Hospital, Headley Way, Headington, Oxford OX3 9DU, UK

[7] Dynacure, 67400 Illkirch, France

[8] 2 Bridge, Rodendijk 60/X, 2980 Zoersel, Belgium