Received: 22 August 2018 | Revised: 29 October 2019 | Accepted: 4 November 2019

DOI: 10.1002/pst.1992

MAIN PAPER

WILEY

Enrichment designs using placebo nonresponders

Norbert Benda[1,2] | **Britta Haenisch**[1,3,4]

[1]Research Department,Federal Institute for Drugs and Medical Devices (BfArM), Bonn, Germany

[2]Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

[3]Population Health Sciences, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

[4]Center for Translational Medicine, University of Bonn, Bonn, Germany

Correspondence
Norbert Benda, Federal Institute for Drugs and Medical Devices (BfArM), Bonn, Germany.
Email: norbert.benda@bfarm.de

Abstract

Enrichment designs that select placebo nonresponders have gained much attention during the last years in areas with high placebo response rates, eg, in depression. Proposals were made that re-randomize patients who did not respond to placebo during a first study phase as the sequential parallel design (SPD). This design uses in a second phase an enriched patient population where the treatment effect is expected to be more pronounced. This may be problematic if an effect in the overall population is claimed. Proposals were made to combine the treatment effects in the overall population from study phase 1 and the enriched population from study phase 2, alleviating but not solving the issue of a potential selection bias. This paper shows how this bias corresponding to the effect difference between the overall population and the enriched population depends on the variability of a potential subject-by-treatment interaction. Sample sizes are given, which lead to a significant result in the combining test with a given probability if actually the average effect in the overall population is zero. If, on the other hand, no subject-by-treatment interaction is given, the enrichment is shown to be inefficient. We conclude that enrichment designs using placebo nonresponders are not able to claim a positive average effect in the overall population if a subject-by-treatment interaction cannot be excluded. It cannot be used to demonstrate positive efficacy in the overall population in a pivotal phase III trial but may be used in early phases to demonstrate varying treatment effects between patients.

KEYWORDS

drug approval, enrichment design, placebo response, sequential parallel design

1 | INTRODUCTION

High placebo response in clinical trials has been quoted as a possible reason for unsuccessful clinical trials in the development of new drugs in specific therapeutic areas, as suggested by Fava et al.[1] Especially in trials that investigate new medicines intended to treat depressive disorders, placebo response was discussed by several authors, eg, Sonawalla and Rosenbaum[2] and Walsh et al,[3] and has been regarded as one obstacle in demonstrating the drug's efficacy in the light of many unsuccessful trials. Nevertheless, as pointed out by Otto and Nierenberg,[4] the (negative) result of a clinical trial should not be used to assess (reject) it's adequacy.

As noted by Ernst and Resch,[5] in the description of the placebo response, one must account for the natural course of the disease, besides other factors. In a randomized placebo-controlled trial, being the gold standard for the confirmatory proof of efficacy of a new medicine, as stated in the ICH E9 guideline on *Statistical Principles in Clinical Trials*,[6] placebo is

used to mimic the absence of a treatment, ie, to compare the effects of the new treatment with those in untreated patients during a predefined period of time. Considering that many trials are conducted as add-on trials, the comparison is often made between the new treatment on top of an existing background therapy as the standard of care to the background therapy only. Specific effects in addition to those attributed to the natural time course cannot be excluded in the clinical trial setting, but whether these additional effects, if present, are *differential* with respect to both groups and favor placebo is not obvious.

Different proposals were discussed that include a placebo run-in phase, eg, by Trivedi and Rush,[7] or enrichment designs that select patients who did not respond to placebo in a first study phase and re-randomize them to either placebo or the active drug in a second one. Referring to the latter designs, the initial proposal was made by Fava et al[1] and called by them sequential parallel comparison design or simply sequential parallel design (SPD) by others. It bases the statistical test on superiority over placebo on an (optimal) combination on the effect estimate from phase 1 (in the overall population) and that of phase 2 (in the enriched population). Several studies have been reported that were conducted using the proposal made by Fava et al[1]; see previous studies.[8-13]

Properties of the SPD design as well as several modifications and the resulting treatment effect estimation have been described by several authors; see other works.[14-21] Chi et al[22] proposed to use a combined effect estimate to estimate a true underlying effect size in the overall population implicitly assuming that the placebo effect measured in phase 1 and resulting from all patients is artificially increased as compared with clinical practice outside a clinical trial setting; see also Liu et al.[23]

Interestingly, although Fava et al[1] describes a statistical test on superiority to placebo, the corresponding null hypothesis to be tested is not mentioned or discussed. Therefore, discussions on a proper type 1 error are likely to be misunderstood by clinical and regulatory experts, since type 1 error control is obtained for a specific null hypothesis. As pointed out and discussed by Tamura et al[17] and Chi et al,[22] the null hypothesis to be tested by the test proposed by Fava et al[1] is given by the intersection of the null hypothesis of no effect in phase 1 and that of no effect in phase 2. Consequently, a statistical significant result allows for a conclusion of superiority over placebo in *either* the overall population *or* the enriched population but does not serve as a proof of superiority in the overall population.

In contrast to other enrichment designs, placebo nonresponder enrichment inhibits the issue that these patients cannot be identified in advance. In regulatory practice, a drug can only be approved if its benefit is shown in the population to be treated resulting in a proper drug labeling that usually corresponds to the patient population included in the pivotal phase III trials. If the intention of the trial is to predict the outcome in future patients, the effect estimated in the enriched population may be enlarged as compared with the overall population and would not reflect the effect in the indicated population unless a corresponding restriction on prescribing is imposed implying that patients are given placebo first for a certain phase of time. There are, however, many ethical and practical problems that may impede this procedure, as described by Senn.[24, section 6.2.1]

As described by Chi et al[22] and others, specific conditions can be given that suffice to conclude on a positive effect in the overall population if a positive effect is given in the enriched one. Especially, the monotonicity condition ensures this conclusion, basically stating that the effect in those patients who would respond to placebo would be at least as large if they were treated with the active drug as if they were treated with placebo. Simply speaking, a placebo responder must be a responder under active treatment as well. As pointed out by Chi et al,[22] the monotonicity condition is a sufficient condition only and may be relaxed. Nevertheless, specific assumptions are required to translate a positive effect in the enriched population in that in the overall population. Apparently, the implicit assumption by some authors is that enrichment can only enlarge an effect that is already present or, in other words, that the null hypothesis of no effect in the overall population is equivalent to the corresponding null hypothesis in the enriched population. Under this assumption, the statistical test in the enriched population, ie, in phase 2 of a SPD design, would also be a valid test with proper type 1 error control for the overall population. Consequently, under this assumption, the analysis of phase 2 alone would be sufficient, but a test combining data from both phases has the advantage of optimally using all data generated by the study design. However, it appears highly questionable, whether the assumption of a sole effect *enlargement* of an already positive effect can be justified.

Although a larger effect in phase 2 is not convincingly seen in the different studies that were conducted with the SPD design, the effort to be made is only justified—in the sense of a more efficient design—if indeed, the effect in placebo nonresponders is larger than in all patients, hence in placebo responders. Investigations on the power under the different assumed effect sizes in both groups have been conducted, eg, by Chen et al.[14] Hence, a certain subject-by-treatment interaction, ie, between-subject variability of the treatment effect is required for the efficiency of a SPD design. If the

trial can successfully demonstrate superiority over placebo by using the data from phase 2, it can be concluded that some patients indeed profit from the drug. However, the average effect over the whole population may still be zero or even negative if other patients experience a negative effect. In this situation, the expected benefit for a future randomly selected patient to be treated is not positive if it is unknown whether this patient belongs to the enriched population or not.

This paper shows that if the subject-by-treatment interaction is modeled in a rather natural way by adding a normally distributed random subject effect, any positive subject-by-treatment interaction will lead to an enlarged effect size in the enriched population from phase 2 as compared with the overall effect in the total population even under the null hypothesis of an average overall effect of zero or less. In other words, if the aim of the study is to predict the effect size in the overall population, the effect from the second phase or any combined effect with positive weights will be biased since the larger treatment effect from the second phase only results from the specific selection. If, on the other hand, the variance would be zero or even small, the design loses its efficiency, since the power would not increase as compared with the parallel group design but an additional effort has to be made by adding another treatment phase. We will give results on the bias under different scenarios. Combining the effect estimates from both phases will alleviate but not solve the issue. Under the null hypothesis of no effect in the overall population (ie, the population used in phase 1), the power of the statistical test using the data from phase 2 or of the combined test corresponds to an increased type 1 error for the overall population.

Considering the bias resulting from the between-subject variability of the treatment effect, a sample size that is required to obtain a successful enrichment trial can be given. In other words, if the average effect over the total population is zero, the enrichment trial could—*formally*—be made successful with a high probability by choosing a sufficient sample size as a function of the (unknown) proportion of the overall variance attributed to the subject-by-treatment interaction. To illustrate the resulting issue, we will give the required sample sizes and would call this the *number needed to cheat*, ie, the sample size needed to make the discussed enrichment study formally successful, if, on average, the drug does not work. Although a between-subject effect variability postulates that the drug works for some patients, a nonpositive average of the effect implies that expected effect for a given patient to which the drug is prescribed is not positive, since it is not known whether she or he belongs to the enriched population.

As pointed by Senn[25] in a paper on precision medicine, identification of the variance component related to subject-by-treatment interaction is rather difficult and requires repeated assessment of the outcome for each treatment; ie, neither a parallel group comparison nor a two-phase crossover trial is sufficient to identify subject-by-treatment interaction. Although heterogeneity between subgroups gives already some indication, a proper determination of the subject-by-treatment interaction requires studies that are usually not performed, especially in depression. Considering this, it will be difficult to appraise this variance component.

However, in the linear model framework as proposed in Section 3, placebo nonresponder enrichment would either be inefficient or lead to a biased result and type 1 error inflation when considering the null hypothesis in the overall population. Other models would be required, eg, models that explicitly model effect modulators by a positive subject specific factor. Nevertheless, this would require a mixture of additive and multiplicative terms, which could be questioned regarding its plausibility and can hardly be verified. Above all, it can be stated that if the identification of subject-by-treatment interaction is already difficult requiring specific designs, the discrimination of different models that involve subject-by-treatment interaction appears almost impossible for the envisaged phase III trial outcome. We therefore claim that no justification is given to properly conclude a positive treatment benefit in the population to be treated from a positive result in a SPD trial.

Instead, the use of the design could further be investigated in the explorative part of the development. A demonstration of a varying treatment effect between patients including the fact that some patients benefit from the drug could be of interest in early phases of drug development. If placebo nonresponders can further be identified by some characteristics, this information could, in principle, be used to refine the patient population to be treated and studied in a subsequent phase III trial that could potentially lead to an approval in the enriched and well-defined population.

The remainder of the article is organized as follows.

Section 2 briefly explains the different proposals for the design and analysis of placebo nonresponder enrichment studies. In Section 3, the expected effect in the enriched population is evaluated for a continuous outcome, where response required for enrichment is defined by an outcome value above a certain threshold. The asymptotic bias is derived as a function of between phase correlations and, in a more specific model, as a function of the variance proportion attributed to the subject-by-treatment interaction.

Under the null hypothesis of no effect in the overall population, the resulting type 1 error inflation is given, as well as the sample size required to obtain a positive study, if, in fact, the average effect is zero. Furthermore, alternative models are discussed that would be required for a placebo nonresponder enrichment to be both efficient and valid with respect to a conclusion for the overall population.

Section 4 concludes with a discussion and regulatory conclusions.

2 | SEQUENTIAL PARALLEL DESIGN

Fava et al[1] proposed a design with two double-blind stages of the same duration. Patients that do not respond under placebo after the first stage are subsequently treated with placebo or the active drug in a second stage. The initial proposal made by Fava et al[1] required an upfront randomization of patients into three groups, those treated with placebo first followed by placebo (PP), those treated with placebo first followed by the active drug (PD), and those treated with the active drug (D) in the first phase. A variant of the design requires a re-randomization of the placebo nonresponders to placebo or active drug between first and second stage as shown in Figure 1. In both designs, patients that are treated with the active drug in the first stage may subsequently be treated with placebo or active drug in the second stage, but stage 2 data from these patients are not used for the treatment effect estimation or testing.

Several options to be investigated relate to the nature of the outcome measure (binary or continuous) and the definition of response used for the selection of placebo nonresponders as shown in Table 1.

3 | STATISTICAL PROPERTIES OF THE DESIGN

3.1 | Models and bias

In this section, we will describe the difference between the effect size in phase 2 and that in phase 1 resulting from the between-phase correlation. This difference can be considered as the selection bias of the phase 2 estimate resulting from

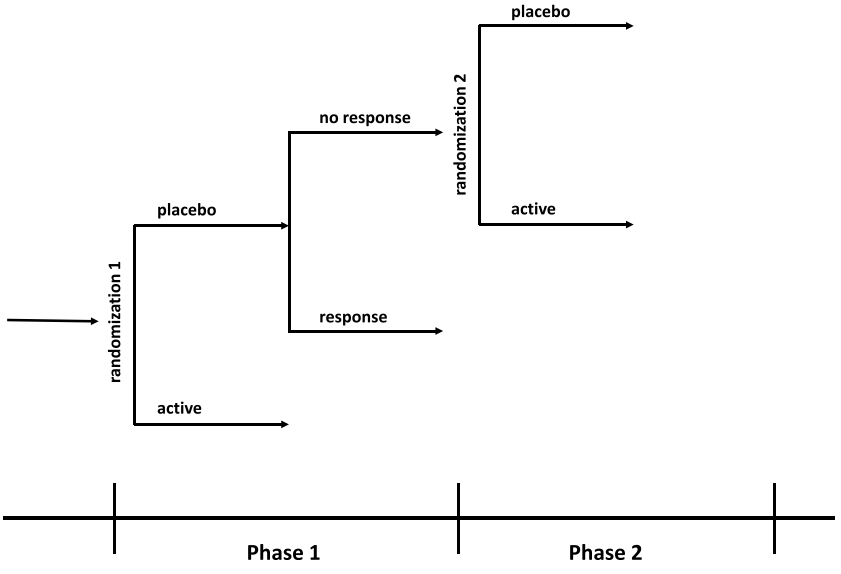

FIGURE 1 Sequential parallel design

TABLE 1 SPD options for outcome and selection criteria

Setting	Outcome	Selection by
1	Binary	Same response variable
2	Binary	Different response variable
3	Continuous	Absolute threshold
4	Continuous	Relative threshold
5	Continuous	Combination of absolute and relative threshold
6	Continuous	Different response variable

Abbreviation: SPD, sequential parallel design.

the enrichment if the target of estimation is the average effect in the overall population in the absence of an additional phase effect that is not related to the selection.

We assume a normally distributed outcome, as given by several outcome scores used in depression, eg, the Hamilton Depression Rating Scale (HDRS) and an additive treatment effect. An additive phase effect may be present; however, it is assumed that the treatment effect would not differ between both phases, if all subjects were observed in both phases. The latter condition is sometimes called the constancy assumption. Let T_i indicate the treatment given in phase i, where $T_i = 0$ refers to placebo and $T_i = 1$ to the active treatment. Then, the outcome Y_i in phase i given treatment $T_i = j$ is distributed as follows:

$$Y_i|T_i = j \sim \mathcal{N}(\mu_i + \delta j, \sigma^2),\ i = 1, 2,\ j = 0, 1,\ \sigma > 0. \tag{1}$$

We further assume that the response variable according to which placebo patients are selected for study phase 2 is based on the outcome measure itself and defined by the outcome falling below a given threshold c. Due to the symmetry of the assumed distribution, all subsequent conclusions are also applicable to setting, where response is given by the exceedance of the threshold.

The response probability for a placebo patient in phase 1 is given by

$$\pi_R = \mathrm{P}(Y_1 \leq c|T_1 = 0) = \Phi\left(\frac{c - \mu_1}{\sigma}\right)$$

with Φ being the cdf of the standard normal distribution.

Whereas the expected effect size in phase 2 is still the same as in phase 1, ie, δ, if all subjects were treated and observed in phase 2, the effect in phase 2 can be expected to be different if a selected population is used. Actually, the difference between the effect in phase 2 and that in phase 1, ie, the bias if the effect in the overall population would be described by the effect in placebo nonresponders, is given by

$$\begin{aligned} B := &\mathrm{E}(Y_2|T_1 = 0, Y_1 \leq c, Y_2 = 1) \\ &- \mathrm{E}(Y_2|T_1 = 0, Y_1 \leq c, Y_2 = 0) - \delta. \end{aligned} \tag{2}$$

Lemma 1. *Let $\pi_R := P(Y_1 \leq c|T_1 = 0)$ be the probability of a placebo response in phase 1. Then the bias of the effect in phase 2 relative to the effect in the overall population is given by*

$$B = \sigma \frac{f(\Phi^{-1}(1 - \pi_R)}{1 - \pi_R}(\rho_{00} - \rho_{01}) \tag{3}$$

with f being the pdf of the standard normal distribution and

$$\rho_{0j} = Corr(Y_1|T_1 = 0, Y_2|T_2 = j), \quad j = 0, 1,$$

being the correlation between the outcome observed in phase 1 under placebo and the outcome observed in phase 2 under either placebo ($T_2 = 0$) or the active treatment ($T_2 = 1$).

The proof is given in Appendix A.

Hence, using an additive model, the difference in the effect sizes can be described as a simple function of the placebo response rate multiplied by the difference in the between-phase correlations given placebo or active treatment in phase 2 (without selection). Defining

$$g(x) := \frac{f(\Phi^{-1}(1 - x))}{1 - x}, 0 < x < 1 \tag{4}$$

with f being the pdf and Φ the cdf of the standard normal distribution, the difference in standardized effect sizes, ie, the standardized bias, is given by

$$B/\sigma = B(\pi_R)/\sigma = g(\pi_R)(\rho_{00} - \rho_{01}). \tag{5}$$

It can easily be seen that g is a monotonically increasing function with $\lim_{x\to 0} g(x) = 0$ and $\lim_{x\to 1} g(x) = \infty$; ie, the bias is 0 if all subjects are placebo nonresponders and increasing and unbounded with increasing proportion of responders.

Hence, the treatment effect measured in the selected population is biased in favor of the active treatment as compared with the effect in the overall population if two subsequent placebo observations are more strongly correlated than a placebo observation followed by an observation under active treatment.

3.2 | Design efficiency

If, on the other hand, one assumes equal effect sizes in both study phases, eg, by arguing that treatment sequence dependent correlations are unlikely, the efficiency of the design may be questioned. Chen et al[14] derived the asymptotic power of the weighted ordinary least squares (OLS) statistic given by

$$Z_{\mathrm{OLS}} = \frac{w\hat{\theta}^{(1)} + (1-w)\hat{\theta}^{(2)}}{\sqrt{w^2 \mathrm{Var}\left(\hat{\theta}^{(1)}\right) + (1-w)^2 \mathrm{Var}\left(\hat{\theta}^{(2)}\right)}}, \tag{6}$$

where $\hat{\theta}^{(i)}, i = 1, 2$ are the treatment effect estimates in phases 1 and 2 for the corresponding treatment effects $\theta^{(i)}, i = 1, 2$, and w the weight attributed to the effect in phase 1. Furthermore, the final sample properties were investigated in extensive simulations leading to similar results as given by formula (6). The empirical power was given under the assumption of equal effects in both phases. Actually, a small power gain was observed as compared with the conventional design using equal randomization in a single study phase only. However, under the assumption of equal effects in both phases, it would be most efficient to use all placebo patients in a second phase (declaring all patients to be nonresponder using an infinite threshold for response), as it can be seen from the asymptotic power given by

$$\Phi\left(\Phi^{-1}(\alpha/2) - \frac{\left(w\theta^{(1)} + (1-w)\theta^{(2)}\right)\sqrt{n}}{\sigma\sqrt{\frac{w^2}{2r(1-2r)} + \frac{2(1-w)^2}{r\pi_N}}}\right), \tag{7}$$

assuming equal variances σ in both phases and a placebo nonresponder probability of $\pi_N = 1 - \pi_R$ in both placebo groups with $2r$ being the proportion of patients allocated to placebo (ie, r to each placebo group) and a two-sided significance level α.

Assuming equal effect sizes $\theta = \theta^{(1)} = \theta^{(2)}$ in both phases, the power is given by

$$\Phi\left(\Phi^{-1}(\alpha/2) - \frac{\theta\sqrt{n}}{\sigma} k_{\pi_N}(w, r)\right) \tag{8}$$

with

$$k_{\pi_N}(w, r) = \frac{1}{\sqrt{\frac{w^2}{2r(1-2r)} + \frac{2(1-w)^2}{r\pi_N}}}. \tag{9}$$

Accordingly, the asymptotic power is an increasing function of $k_{\pi_N}(w, r)$. Given the probability of being a placebo nonresponder π_N, the factor k and hence the power can be maximized by setting $\nabla k_{\pi_N}(w, r) = 0$ leading to

$$r^* = r^*(\pi_N) = 1/4 + \pi_N/16 \tag{10}$$

$$w^* = w^*(\pi_N) = \frac{4 - \pi_N}{4 + \pi_N} \tag{11}$$

$$k_{\pi_N}(w^*, r^*) = 2r^* = 1/2 + \pi_N/8. \tag{12}$$

The power of the conventional design is given by $k = 1/2$. Note that the efficiency related to $k_{\pi_N}(w^*, r^*) > 1/2$ can only be obtained with a correct prior guess of the placebo nonresponder probability, optimized weights and randomization ratio and if efficiency gain is measured by the number of subjects to be recruited not accounting for the prolonged study duration in the SPD design. Using, however, a conventional choice of $w = 1/2$ and $r = 3/8$ as realized in the ADAPT-A Study, see Fava et al,[8] eg, a 50% placebo responder rate would lead to $k_{0.5}(w, r) = 1/2$, ie, a sample size requirement equal to the conventional design, if the phases 1 and 2 effects were expected to be equal.

Under this assumption, since

$$k_{\pi_N}(w^*, r^*) < k_1(w^*(1), r^*(1)) = 5/8, \forall \pi_N < 1, \tag{13}$$

the asymptotic efficiency of the SPD design is always less than that of the design that uses all placebo patients in a second study phase, which would lead to a sample size reduction as compared with the conventional design of

$100\left(1-\left(\frac{1/2}{5/8}\right)^2\right)\% = 36\%$ as compared with $100\left(1-\left(\frac{1/2}{9/16}\right)^2\right)\% = 21\%$ if a response probability of 0.5 is chosen and the weights and randomization ratio are optimized accordingly.

This means that any potential efficiency gain of the SPD design that exceeds that of simply continuing all placebo patients is based on an enlarged effect size in phase 2, ie, a bias with respect to the effect in the overall population.

3.3 | Subject-by-treatment interaction and bias

Assuming a more specific model with fixed phase effect α, random subject intercept b, mean treatment effect δ in the overall population, and a random treatment effect a (eg, the subject-by-treatment interaction), the outcome measure Y can be described by

$$Y = \mu + \beta(P-1) + (a+\delta)(T-1/2) + b + \epsilon, \tag{14}$$

where P corresponds to the phase (1 and 2), T to the treatment (1 for the active treatment and 0 for placebo). Random effects are given by normally distributed a with mean 0 and variance σ_T^2 and normally distributed b with mean 0 and variance σ_S^2. Measurement error is given by $\epsilon \sim \mathcal{N}(0, \tau^2)$, intercept by μ, and phase effect by β.

Accordingly, we can describe the outcome for each subject in phases 1 and 2, respectively, by

$$Y_1 = \mu + (a+\delta)(T_1 - 1/2) + b + \epsilon_1, \tag{15}$$

$$\begin{aligned} Y_2 = \mu + \beta + (a+\delta)(T_2 - 1/2) \\ + b + \epsilon_2, \text{ if } Y_1 \le c \text{ and } T_1 = 0, \end{aligned} \tag{16}$$

where $(a, b, \epsilon_1, \epsilon_2) \sim \mathcal{N}(0, \text{diag}(\sigma_T^2, \sigma_S^2, \tau^2, \tau^2))$.

The difference between the expected effects in phases 2 and 1, ie, the bias if the effect in the overall population is estimated by the effect in phase 2, is given by

$$\begin{aligned} B := \text{E}(Y_2 | T_1 = 0, Y_1 \le c, T_2 = 1) \\ - \text{E}(Y_2 | T_1 = 0, Y_1 \le c, T_2 = 0) - \delta. \end{aligned} \tag{17}$$

Lemma 2. *Let $\pi_R := P(Y_1 > c | T_1 = 0)$ be the probability of a placebo response in phase 1, $\sigma^2 := 1/4\sigma_T^2 + \sigma_S^2 + \tau^2$ the total variance of a single observation, and $q := \frac{1/4\sigma_T^2}{\sigma^2}$ the proportion of the total variance attributed to the subject-by-treatment interaction. Then, the bias of the effect in phase 2 relative to the effect in the overall population is given by*

$$B = 2q \times \sigma \times g(\pi_R), \tag{18}$$

where g is given by

$$g(x) = \frac{f(\Phi^{-1}(1-x))}{1-x}, x > 0$$

with f being the pdf and Φ the cdf of the standard normal distribution.

The proof is given in Appendix A.

As an example consider the case where half of the subjects respond to placebo and a quarter of the overall variance can be attributed to the subject-by-treatment interaction. In this case, the bias is given by $\sigma/\sqrt{2\pi} \approx 0.4\sigma$ corresponding to a medium effect size. Table 2 gives the effect size in placebo nonresponders, if the mean effect in the overall population is 0. In other words, the efficiency gain of the SPD design is counterbalanced by a relevant bias and, consequently, an increased probability to claim superiority in case the average effect in the overall population is 0.

Combining the effect estimate from both phases reduces the bias accordingly. Using a weight w for the estimate from phase 1 reduces the bias by $1 - w$. Whereas the weight w (in combination with the phase 1 randomization ratio) may be chosen to maximize the power of the resulting test (depending on assumed alternative hypothesis) in most of the trials using the SPD design, a weight of 0.5 is chosen due to the uncertainty of the assumed effect size and—perhaps—for practical reasons and an allegedly facilitated interpretation.

TABLE 2 Standardized effect size in placebo nonresponders if mean overall effect = 0

Variance Proportion Attributed to Subject-by-treatment Interaction	Placebo Response Rate 0.2	0.4	0.6	0.8
0.1	0.07	0.13	0.19	0.28
0.2	0.14	0.26	0.39	0.56
0.3	0.21	0.39	0.58	0.84
0.5	0.35	0.64	0.97	1.40

TABLE 3 Asymptotic one-sided H_1 type 1 error using weighted OLS: $w = 0.5$, $r = 0.375$, $\alpha = 0.025$

Variance Proportion Attributed to Subject-by-treatment Interaction	Placebo Response Rate 0.3	0.5	0.7
0.05	0.043	0.053	0.061
0.1	0.069	0.102	0.130
0.2	0.158	0.282	0.386

Abbreviation: OLS, ordinary least squares.

3.4 | Hypothesis testing and type 1 error control

Type 1 error control relates to a specific null hypothesis. In the initial paper by Fava et al,[1] the null hypothesis is not mentioned, which may have led to some confusion regarding a proper type 1 error control. As mentioned by 17. Tamura et al[17] and Chen et al,[14] the null hypothesis to be tested is the intersection $H_1 \cap H_2$ of the null hypothesis of a nonpositive effect in phase 1 (H_1) and that of nonpositive effect in phase 2 (H_2). Hence, a significant result is indicative of a positive effect either in the overall population *or* the enriched one, but not necessarily of a positive effect in the overall population, which would refer to H_1 only. Type 1 error control is therefore ensured only for the intersection hypothesis. However, the type 1 error of the hypothesis test in the enriched population with respect to the null hypothesis in the overall population (H_1) corresponds to its power assuming an effect size that is equal to the bias compared with the treatment effect δ in all patients. The type 1 error of the combined test can be approximated using the asymptotic power formula for the weighted OLS test derived by Chen et al[14] and using the effect sizes 0 for phase 1 and B for phase 2. Under the model given above, the overall variance in phase 2 is even slightly reduced due to the involved truncated distributions leading to further inflation. The one-sided type 1 error for a nominal alpha level of 2.5% is given in Table 3 for a total sample size of $n = 300$ choosing weight and randomization by $w = 0.5$ and $r = 0.375$ as in the ADAPT-A Study.[8] Even for a small treatment-by-subject interaction $q = 0.05$, type 1 error would be doubled for a response probability of 50%.

Since the H_1 type 1 error of the weighted OLS statistic corresponds to its power under the alternative hypothesis of an effect size equal to its bias, it increases with the sample size. Table 4 gives the sample size that would result in a H_1 type 1 error of the weighted OLS test of 0.3, 0.5, and 0.7. In other words, the lower part of Table 4 gives the patient number needed to obtain a significant result in favor of the active drug with a power of 80%, if, in fact, it is, on average, not different from placebo, which may be seen as the *number needed to cheat* with placebo nonresponder enrichment. For example, a placebo response probability of 50% and a variance proportion of 20% attributed to the subject-by-treatment interaction would lead to a sample size of about 1200. Although the variance proportion may be expected to be rather low leading to larger sample sizes, this table illustrates that sole sample size increase would make an ineffective drug successful, which is certainly an undesirable design property. Choosing a large response threshold, ie, a large responder probability, would further facilitate "formal success."

It follows from Lemma 2 that in the additive model (14), a test on the null hypothesis $H_1 : \delta = 0$ will not be valid if it is based on the effect estimate in phase 2 or on a linear combination the effect estimates in phases 1 and 2 and the variance of the individual treatment effects is greater than 0; ie, any treatment-by-subject interaction would invalid a conclusion from the second phase, also if the result from the second phase is combined with that of the first one using a linear combination. On the other hand, in the absence of a treatment-by-subject interaction, the selection of placebo nonresponders would be inefficient as shown in Section 3.3.

TABLE 4 Sample size corresponding to a given H_1 type 1 error using weighted OLS: $w = 0.5$, $r = 0.375$

H_1 Type 1 Error	Variance Proportion Attributed to Subject-by-treatment Interaction	Placebo Response Rate 0.3	0.5	0.7
0.2	0.1	1642	786	538
	0.2	411	197	135
0.5	0.1	5042	2414	1653
	0.2	1261	604	414
0.8	0.1	10302	4932	3377
	0.2	2576	1233	845

Abbreviation: OLS, ordinary least squares.

3.5 | Alternative models

Under the assumption of an additive model with a responder definition based on the outcome itself, the SPD design was shown to be either invalid for a conclusion in the overall population or inefficient, depending on whether a subject-by-treatment interaction is assumed or not. A multiplicative model with a relative responder definition would usually be handled by a log-transformation leading to corresponding properties on the log scale as described above. Using a relative responder definition or a combination of an absolute threshold and a relative change (eg, defined by a decrease of at least 50% or below a given limit) together with an additive model (AN(C)OVA on original values) would lead however to similar issues.

Two options may be considered that would justify the procedure, either by assuming two different subpopulations A and B, where subjects in population A are expected to be unresponsive under both treatments with a treatment effect of 0, and the outcome in subjects in population B are expected to follow the usual distributional assumptions with an unknown treatment effect δ. Since the used responder definitions are rather arbitrary without clear biologically plausible cutoff, it appears unlikely that the patient population can indeed be divided into such kind of subpopulations. Furthermore, the subpopulations may not fully be identified due to measurement errors, temporal within-subject fluctuations, and regression-to-the-mean effects.

The second option would assume an additive model with a multiplicative subject-specific effect modulator, eg,

$$Y = \mu + \beta(P-1) + \exp(a)\delta(T-1/2) + b + \epsilon \tag{19}$$

with

$$a \sim \mathcal{N}(0, \sigma_T^2). \tag{20}$$

Clearly, this model would allow for a positive overall effect if the effect in the enriched population is positive, since the algebraic sign of the treatment effect is the same for all subjects. Whatever subpopulation is targeted, a positive effect in the enriched population would always allow to conclude on a positive effect in all patients. However, it appears highly questionable whether the required mixture of additive and multiplicative effects is plausible. In any case, both models (14) and (19) can hardly be discriminated unless the outcome is measured in more than two phases using a high sample size and confounding effects, as crossover or treatment-by-phase interaction, can be excluded. The generation of such a database appears highly unrealistic at least in the targeted indications, as depressive disorders.

4 | DISCUSSION AND REGULATORY CONCLUSION

Enrichment designs that select patients to show a treatment benefit in the selected population are intended to either identify a population of patients who benefit from the medicine under development or strengthen the sensitivity of the trial by selecting the most sensitive part of the patient population. If a positive effect in the selected population can be shown in a confirmatory way, marketing authorization can be granted for the new medicine in the selected population provided the benefit is considered to outweigh the risks. However, placebo nonresponder enrichment inhibits the issue that the selected patient population cannot be identified in advance.

In order to overcome high placebo response rates in clinical trials—especially for antidepressants—placebo nonresponder enrichment was proposed using several but similar approaches. Treatment effect estimates as well as the statistical hypothesis test were proposed to combine the effects from both phases of the study, whereas the first one is conducted

in the overall population, and in the second, the treatment effect is estimated in those patients that do not respond to placebo. Since the statistical test is based on the null hypothesis of no effect in either of the two phases, it is not capable to conclude on a positive effect in the overall population without further assumptions.

We showed that under a conventional additive model, the treatment effect estimate from the second phase is biased in the presence of subject-by-treatment interaction. On the other hand, in the absence of such a bias, the proposal to select placebo nonresponders is always less efficient with respect to the number of subjects to be recruited than continuing all placebo patients and would not decrease the number of patients needed for a given power considerably as compared with the conventional one-phase design in all subjects at the cost of a doubled study duration. Hence, the design would be efficient only in the presence of a larger effect in placebo nonresponders. Consequently, the treatment effect estimate either from phase 2 or from the combined estimate is biased with respect to the effect in the overall population and the type 1 error of the corresponding hypothesis test is inflated with respect to the null hypothesis of no mean effect in the overall population. Since this type 1 error corresponds to the power of the tests assuming a certain effect in phase 2 only, the study could be made formally successful even in case of a mean effect of 0 in all patients if only a sufficient sample size is chosen. In that sense, the SPD design and analysis can be considered as either inefficient or invalid.

Different models than a conventional additive model would have to be assumed to justify the design. We argue that these models lack plausibility and that they can hardly be discriminated from the additive model proposed in the setting of depression trials.

Showing a positive treatment effect only in a population that cannot be identified and well described in the drug labeling is insufficient to grant marketing authorization since in the presence of a subject-by-treatment interaction (which is the basis of the design's efficiency), a positive effect in one part of the population can be counterbalanced by a negative effect in the complementary group. This means that in this case, the expected treatment effect would be zero or could even be negative for a given randomly selected patient to be treated.

Provided that the drug effect at the first stage is significant, claiming the drug effect for the overall population would be possible (if properly pre-specified) but based on the results of the first phase only. However, the results from the combination test and estimate that incorporates the second-stage results would be based on a mixture of the effect in the overall population and that of the enriched population, where the latter could be enlarged due to the selection made.

The design is, however, certainly capable to show that *either* the drug is beneficial in all patients *or* the treatment effect varies across subjects. This may indeed be of further interest in a phase II setting as a proof of concept and to further define the precise indication and population, similar to a setting where a predictive biomarker is searched for to define a suitable patient population. If placebo nonresponders can be described by their patient characteristics, which may further be used to define the patient population to be treated, this group of patients may then be studied in a phase III trial that could lead to an approval in the studied population.

If, on the other hand, a claim is intended for the enriched population and the trial is to predict the effect in the enriched population (that may be enlarged as compared with that of the overall population), a corresponding restriction on prescribing would be required implying that patients are given placebo first for a certain phase of time. In this case, the effect estimate of the second phase only would be required. There are, however, many ethical and practical problems that may impede this procedure.

Certainly, external validity of a clinical trial in terms of generalizability to the wider population may often be questioned in the absence of a true random sampling. Still, prediction of the effect in the overall population could be exaggerated even more by placebo nonresponse enrichment. On the other hand, if the purpose of the trial is to show any causal relation between treatment and effect, one may argue that a positive effect in an enriched population already proofs such a causal (positive) effect in at least some patients. Assuming, however, varying treatment effects between patients, placebo nonresponse enrichment may falsely conclude or suggest a positive average treatment effect. It is acknowledged, though, that the presence and the kind of modeling of a (random) subject-by-treatment interaction could be a matter of debate similar to the long lasting and ongoing discussion about fixed versus random effects meta-analysis, see, eg, Senn. [26]

ACKNOWLEDGEMENT

We would like to thank Debora Parker for her collaboration and discussion on this topic during her stay at the BfArM.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Norbert Benda https://orcid.org/0000-0001-5605-2414

REFERENCES

1. Fava M, Evins A, Dorer D, Schoenfeld D. The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remediesand a novel study design approach. *Psychiotherapy and Psychosomatics*. 2003;72(3):115-127.
2. Sonawalla SB, Rosenbaum JF. Placebo response in despression. *Dialogues in Clinical Neuroscience*. 2002;4(1):105-113.
3. Walsh BT, Seidman SN, Syako R, Gould M. Placebo response in studies of major depression: variable, substantial and growing. *Journal of the American Medical Association*. 2002;287(14):1840-1847.
4. Otto MW, Nierenberg AA. Assay sensitivity, failed clinical trials, and the conduct of science. *Psychiotherapy and Psychosomatics*. 2002;71(5):241-243.
5. Ernst E, Resch KL. Concept of true and perceived placebo effects. *British Medical Journal*. 1995;311(7004):551-553.
6. ICH E9 Expert Working Group. Statistical Principles for Clinical Trials: ICH Harmonized Tripartite Guideline. *Statistics in Medicine* 1999; 18:1905-1942.
7. Trivedi MH, Rush J. Does a placebo run-in or a placebo treatment cell affect the efficacy of antidepressant medications? *Neuropsychopharmacology*. 1994;11(1):33-43.
8. Fava M, Mischoulon D, Iosifescu D, et al. A double-blind, placebo-controlled study of aripiprazole adjunctive to antidepressant therapy among depressed outpatients with inadequate response to prior antidepressant therapy (ADAPT-A Study). *Psychotherapy and Psychosomatics*. 2012;81(2):87-97.
9. Heo JY, Jeon HJ, Fava M, et al. Efficacy of ziprasidone monotherapy in patients with anxious depression: a 12-week, randomized, double-blind, placebo-controlled, sequential-parallel comparison trial. *Journal of Psychiatric Research*. 2015;62:56-61.
10. Papakostas GI, Vitolo OV, IsHak WW, et al. A 12-week, randomized, double-blind, placebo-controlled, sequential parallel comparison trial of ziprasidone as monotherapy for major depressive disorder. *The Journal of Clinical Psychiatry*. 2012;73(12):1541-1547.
11. Papakostas GI, Shelton RC, Zajecka JM, et al. L-Methylfolate as adjunctive therapy for SSRI-resistant major depression: results of two randomized, double-blind, parallel-sequential trials. *The American Journal of Psychiatry*. 2012;169(12):1267-1274.
12. Papakostas GI, Shelton RC, Zajecka JM, et al. Effect of adjunctive L-methylfolate 15 mg among inadequate responders to SSRIs in depressed patients who were stratified by biomarker levels and genotype: results from a randomized clinical trial. *The Journal of Clinical Psychiatry*. 2014;75(08):855-863.
13. Jeon HJ. Fava M, Mischoulon D, jBaer L, Clain A, Doorley J, DiPierro M, Cardoos A, Papakostas GI. Psychomotor symptoms and treatment outcomes of ziprasidone monotherapy in patients with major depressive disorder: a 12-week, randomized, double-blind, placebo-controlled, sequential parallel comparison trial. *International Clinical Psychopharmacology*. 2014;29(6): 332-338.
14. Chen YF, Yang Y, Hung HM, Wang SJ. Evaluation of performance of some enrichment designs dealing with high placebo response in psychiatric clinical trials. *Contemporary Clinical Trials*. 2011;32(4):592-604.
15. Chen YF, Zhang X, Tamura RN, Chen CM. A sequential enriched design for target patient population in psychiatric clinical trials. *Statistics in Medicine*. 2014;33(17):2953-2967.
16. Tamura RN, Huang X. An estimation of treatment effect for the sequential parallel design in psychiatrie clinical trials. *Clinical Trials*. 2007;4(4):309-317.
17. Tamura RN, Huang X, Boss DD. Estimation of treatment effect for the sequential parallel design. *Statistics in Medicine*. 2011;30: 3496-3506.
18. Ivanova A, Qaqish B, Schoenfeld DA. Optimality, sample size, and power calculations for the sequential parallel comparison design. *Statistics in Medicine*. 2011;30(23):2793-2803.
19. Ivanova A, Tamura RN. A two-way enriched clinical trial design: combining advantages of placebo lead-in and randomized withdrawal. *Statistical Methods in Medical Research*. 2015;24(6):871-890.
20. Rybin D, Doros G, Pencina MJ, Fava M. Placebo non-response measure in sequential parallel comparison design studies. *Statistics in Medicine*. 2015;34(15):2281-2293.
21. Doros G, Pencina MJ, Rybib D, Meisner A, Fava M. A repeated measures model for analysis of continuous outcomes in sequential parallel comparison design studies. *Statistics in Medicine*. 2013;32(16):2767-2789.
22. Chi GYH, Li Y, Liu Y, Lewin D, Lim P. On clinical trials with a high placebo response rate. *Contemporary Clinical Trials Communications*. 2016;2:34-53.
23. Liu Q, Lim P, Singh J, Lewin D, Schab B, Kent J. Doubly randomized delayed-start design for enrichment studies with responders or nonresponders. *Journal of Biopharmaceutical Statistics*. 2012;22(4):737-757.
24. Senn SJ. *Statistical Issues in Drug Development*. Hoboken: Wiley; 2007.
25. Senn SJ. Mastering variation: variance components and personalised medicine. *Statistics in Medicine*. 2016;35(7):966-977.
26. Senn SJ. Controversies concerning randomization and additivity in clinical trials. *Statistics in Medicine*. 2004;23:93729-93753.
27. Mukhopadhyay P. *Multivariate Statistical Analysis*. Singapore: World Scientific; 2009.
28. Johnson NL, Kotz S, Balakrishnan N. *Continuous Univariate Distributions, Volume 1*. 2nd ed. New York: Wiley; 1994.

How to cite this article: Benda N, Haenisch B. Enrichment designs using placebo nonresponders. *Pharmaceutical Statistics*. 2020;19:303–314. https://doi.org/10.1002/pst.1992

APPENDIX A

Let f_Y the pdf of a random variable Y, f be the pdf, and Φ the cdf of the standard normal distribution and

$$\mathrm{Cov}_{0j} := \mathrm{Cov}(Y_1, Y_2 | T_1 = 0, T_2 = j), j = 0, 1.$$

Proof of Lemma 1

Without loss of generality, we assume $\mu_i = 0$, $i = 1, 2$. The bias is given by

$$B = A_1 - A_2 - \delta$$

with

$$\begin{aligned} A_1 &:= \mathrm{E}(Y_2 | T_1 = 0, T_2 = 1, Y_1 \le c) \\ A_2 &:= \mathrm{E}(Y_2 | T_1 = 0, T_2 = 0, Y_1 \le c). \end{aligned}$$

Since for any bivariate normally distributed vector $(X, Y) \sim \mathcal{N}\left((\eta_1, \eta_2)^T, \Sigma\right)$, with positive semi-definite covariance matrix $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$

$$\mathrm{E}(X | Y = y) = \eta_1 + \frac{\sigma_{12}}{\sigma_{22}}(y - \eta_2)$$

see Mukhopadhyay,[27, theorem (3.2.7)] it holds that

$$\begin{aligned} A_1 &= \int_\infty^c \frac{\mathrm{E}(Y_2 | T_1 = 0, T_2 = 1, Y_1 = y)}{\mathrm{P}(Y_1 \le c)} f_{Y_1}(y) \mathrm{d}y = \frac{1}{\mathrm{P}(Y_1 \le c)} \int_\infty^c \frac{\mathrm{Cov}_{01}}{\sigma^2} y f_{Y_1}(y) \mathrm{d}y + \delta \\ &= \frac{\mathrm{Cov}_{01}}{\sigma^2} \mathrm{E}(Y_1 | T_1 = 0, Y_1 \le c) + \delta = -\rho_{01} \frac{f(c/\sigma)}{\Phi(c/\sigma)} \sigma + \delta \\ &= -\rho_{01} \frac{f(\Phi^{-1}(1 - \pi_R))}{1 - \pi_R} \sigma + \delta \end{aligned}$$

using the expected value of a truncated normal distribution; see Johnson et al.[28, page 156, (13.134)] Accordingly,

$$A_2 = -\rho_{00} \frac{f(\Phi^{-1}(1 - \pi_R))}{1 - \pi_R} \sigma$$

Hence,

$$B = (\rho_{00} - \rho_{01}) \frac{f(\Phi^{-1}(1 - \pi_R))}{1 - \pi_R} \sigma.$$

□

Proof of Lemma 2

Lemma 2 follows from Lemma 1, since a, b , ϵ_1, and ϵ_2 are independent and

$$\begin{aligned} \mathrm{Cov}_{00} &= \mathrm{Cov}(-a/2 + b + \epsilon_1, -a/2 + b + \epsilon_2) = \mathrm{Var}(b) + \mathrm{Var}(a/2) \\ \mathrm{Cov}_{01} &= \mathrm{Cov}(-a/2 + b + \epsilon_1, a/2 + b + \epsilon_2) = \mathrm{Var}(b) - \mathrm{Var}(a/2). \end{aligned}$$

Hence,

$$\rho_{00} - \rho_{01} = \frac{\mathrm{Cov}_{00} - \mathrm{Cov}_{01}}{\sigma^2} = \frac{2\mathrm{Var}(a/2)}{\sigma^2} = 2q$$

which completes the proof. □