

RESEARCH ARTICLE

Combined assessment of early and late-phase outcomes in orphan drug development

Konstantinos Pateras¹ | Stavros Nikolakopoulos¹ | Kit C. B. Roes²

¹Department of Data Science and Biostatistics, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

²Department of Health Evidence, Section Biostatistics, Radboud University Medical Centre, Nijmegen, The Netherlands

Correspondence

Konstantinos Pateras, Department of Data Science and Biostatistics, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands
Email: kostas.pateras@gmail.com

In drug development programs, proof-of-concept Phase II clinical trials typically have a biomarker as a primary outcome, or an outcome that can be observed with relatively short follow-up. Subsequently, the Phase III clinical trials aim to demonstrate the treatment effect based on a clinical outcome that often needs a longer follow-up to be assessed. Early-phase outcomes or biomarkers are typically associated with late-phase outcomes and they are often included in Phase III trials. The decision to proceed to Phase III development is based on analysis of the early-Phase II outcome data. In rare diseases, it is likely that only one Phase II trial and one Phase III trial are available. In such cases and before drug marketing authorization requests, positive results of the early-phase outcome of Phase II trials are then likely seen as supporting (or even replicating) positive Phase III results on the late-phase outcome, without a formal retrospective combined assessment and without accounting for between-study differences. We used double-regression modeling applied to the Phase II and Phase III results to numerically mimic this informal retrospective assessment. We provide an analytical solution for the bias and mean square error of the overall effect that leads to a corrected double-regression. We further propose a flexible Bayesian double-regression approach that minimizes the bias by accounting for between-study differences via discounting the Phase II early-phase outcome when they are not in line with the Phase III biomarker outcome results. We illustrate all methods with an orphan drug example for Fabry disease.

KEYWORDS

Bayesian, bias correction, biomarker, borrowing strength, decision-induced bias, rare diseases, surrogate endpoint, trial combination

1 | INTRODUCTION

Drug development programs typically include exploratory (Phase II) and confirmatory (Phase III) randomized controlled trials (RCTs) to assess the efficacy, safety and appropriate dosages of an experimental (new) treatment. For regular “large disease” drug development programs decisions to conduct a Phase III trial are based on positive Phase II trials. If these trials are only retrospectively evaluated in combination, that is, during the drug marketing authorization request, the ad

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

TABLE 1 Main randomized studies described in the European Public Assessment Report of Galafold

Study number	Duration	Annualized rates of change in eGFR from baseline to month 6	Annualized rates of change in eGFR from baseline to month 18	Sample size	Start date
AT1001-011	6 months	Collected	Not collected	67	August 2009
AT1001-012	18 months	Collected	Collected	52	December 2010

hoc synthesis may induce a form of decision-induced bias (the succeeding trials are only conducted when the first trials were positive). Such a bias is not an issue if the early and late Phase trials are prospectively considered in the design phase (eg, a seamless approach).

However, it is not uncommon that in rare diseases, no more than two independent RCTs are conducted and available, one exploratory and one confirmatory.¹ Phase II primary endpoints are typically biomarkers or surrogate outcomes.² Phase III primary clinical outcomes are likely established endpoints and they may either require (1) larger sample sizes, (2) more costly collection, (3) to be observed after a considerable time, or (4) be more variable outcomes than early-phase outcomes, therefore, even if $N = N_2 + N_3$ number of patients participate in both trials, only N_3 patients will be available to provide responses for the primary clinical outcome of interest. Biomarkers (early-phase) and secondary clinical outcomes are often observed earlier and, therefore, easily included in both trials and, hence, available for all N patients. After both trials have been conducted, inference on the treatment efficacy is typically performed by evaluating the late-phase outcome responses of N_3 patients. In a rare disease setting, N_3 may not be large enough to solidly confirm treatment efficacy. In assessing the totality of evidence, the positive results from the Phase II trial could *retrospectively* be seen as supportive, even if the two clinical trials were designed/conducted independently, as typically the early-phase outcome would be assumed to be associated with the late phase primary clinical outcome. Throughout the article the terms “*retrospective (ly)*” denote the retrospective combination of the available Phase II and Phase III trial after both trials are completed and their final results are available.

For example, Galafold (migalastat) acquired marketing authorization as an orphan drug for the treatment of Fabry disease in 2016 within Europe. Fabry disease is a rare, progressive disorder with an estimated prevalence of 1:117 000 to 1:40 000.³ The condition affects major organs and may result in life-threatening events. Until then, standard treatment for Fabry disease consisted of Enzyme Replacement Therapy.³ Two main studies were submitted during the marketing authorization of migalastat; one randomized, placebo-controlled (AT1001-011, migalastat vs Placebo) superiority study and one active comparison randomized trial (AT1001-012, migalastat vs Enzyme Replacement Therapy), with a noninferiority design.

In trial 011 patients switched to migalastat 6 months postrandomization, while in trial 012 primary follow-up was considerably longer, with switching taking place 18 months postrandomization. In the first trial, the change in average globotriaosylceramide (GL-3) inclusions from baseline to 6 months was the primary outcome which produced nonconclusive evidence. The second trial utilized the annualized change in glomerular filtration rate (eGFR) at month 18 as primary clinical outcome (Table 1). Both GL-3 and the annualized change in eGFR at month 6 were collected in both trials (011 and 012). No strong correlation has been established in the literature between the GL-3 outcome and the change in glomerular filtration rate (eGFR).⁴

In study 011 after 6 months of treatment with migalastat 150 mg, eGFR values increased, whereas in the placebo treated group eGFR values declined.³ This outcome among other secondary results led to the conduct of study 012. In trial 011, all patients treatment switched to migalastat at 6 months, an action that restricts the observation of a treatment effect on the primary late-phase outcome. Given the limited available data, evidence from both trials were retrospectively (ad hoc) assessed for the final approval decision.

Analysis methods that use the relation between early and late-phase outcomes may be applied to retrospectively, but formally, synthesize the evidence on treatment efficacy across the two trials. Engel and Walstra⁵ formulated a *double-regression* (DR) approach, which can aid in more precise treatment effect estimation, by accounting for unobserved late-phase outcome responses via observed early-phase outcome responses. Their method utilizes the correlation to ultimately inform the mean and variance estimates of the treatment effect on the late-phase outcomes. For large samples their method has the potential to increase precision. However, for small sample sizes this is not necessarily true.⁶ Previously, in RCTs the DR approaches have been suggested mainly to inform treatment selection during interim analysis in

seamless Phase II/III designs.⁷⁻⁹ Double-regression methods can be even generally applied wherever there is possibility to include early outcome information in decision making during the course of a trial.¹⁰

A Bayesian *double-regression* (BDR) analogue can be readily constructed¹¹ which maintains similar limitations to the frequentist alternative but could flexibly model the two Phase III outcomes' data. Such a model can include historical trial data (ie, Phase II early-phase outcome data or external information on the early and late-phase outcome correlation) as a elicited prior distributions.¹² Furthermore, this Bayesian model accounts for the uncertainty around each parameter during the borrowing of information.

In this article, we investigate how to model and estimate the efficacy of a new treatment on the late-phase clinical outcome, using data on early-phase outcomes from both trials. Most literature on double-regression focuses on design aspects such as interim analysis or seamless design of phase II/III trials, though, in the present article we propose methods that would be applied retrospectively (ad hoc) only after the Phase III trial. We propose and investigate methods that either account or do not account for the potential decision-induced bias when combining retrospectively the Phase II and Phase III trials. We investigate the two proposed models, the bias corrected DR approach and the flexible Bayesian approach regarding their performance to estimate the treatment effect on the late-phase outcome. We focus on two related key problems: (1) the magnitude of the type 1 error inflation when retrospectively combining data from Phase II and III and (2) how to estimate the treatment effect on the late-phase outcome, using results from both studies and we assess this estimate in terms of bias and variance.

The article is organized as follows. First, we describe a bivariate linear model, we introduce its conditional form and we formalize the (often visual) retrospective pooling by utilizing DR with nonavailable Phase II late-phase outcome data, then briefly discuss specific model variations, for example, the *single-regression* (SR) approach. We introduce the problem of decision-induced bias moving from Phase II to Phase III based on the Phase II early-phase outcome in Section 3 and then provide an approximate analytical solution. In Section 4, we propose and formulate a Bayesian two-step solution to the estimation problem, a model that down-weights the impact of the biomarker data via a historical power prior. This prior dynamically accounts for the bias in estimating the same treatment effect across the two trials, by accounting for additional between-trial differences (variability) around the biomarker outcome effect. The article ends with a discussion and steps for further research.

2 | MODELS FOR THE JOINT PHASE II AND III DATA

Consider a Phase II trial of total sample size N_2 and a Phase III trial of total sample size N_3 . For both trials it is assumed that a number of patients ($N_k = n_{ck} + n_{ek}$, $n_k = N_k/2$, $k = 2, 3$) are randomized to the control and experimental treatment. Let us denote Y_{ik} the late-phase treatment response for patient i in trial k and X_{ik} the early-phase treatment response for patient i in trial k , $k = 2, 3$, $i = 1, 2, \dots, N_k$.

2.1 | Bivariate modeling for early-phase and late-phase outcomes between studies

When all late-phase $Y_i = (Y_{i2}, Y_{i3})$ and early-phase $X_i = (X_{i2}, X_{i3})$ outcomes are available where $i = 1, 2, \dots, N$, they are assumed to follow a bivariate normal distribution as

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim \text{BVN} \left[\begin{pmatrix} a_x + b_x \mathbf{t}_i \\ a_y + b_y \mathbf{t}_i \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_x^2 & \rho \sigma_x \sigma_y \\ \rho \sigma_x \sigma_y & \sigma_y^2 \end{pmatrix} \right], \tag{m1}$$

where σ_x^2 and σ_y^2 denote the true outcomes variances, ρ the true correlation between the two outcomes and \mathbf{t}_i a vector indicating whether the i th patient receives control or experimental treatment. For the remainder of the article we drop index i to aid readability.

The above bivariate model can be conditionally expressed as

$$\begin{aligned} X|\mathbf{t} &\sim N(a_x + b_x \mathbf{t}, \sigma_x^2) \\ Y|\mathbf{t}, x &\sim N(a_0 + b_0 \mathbf{t} + \gamma x, \sigma_0^2), \end{aligned} \tag{m2}$$

where $\sigma_0^2 = \sigma_y^2 - \gamma^2 \sigma_x^2$, $a_0 = a_y - \gamma a_x$, $b_0 = b_y - \gamma b_x$, and $\gamma = \rho \sigma_y / \sigma_x$

2.2 | Double regression to estimate the effect of primary late-phase outcome

At the end of both trials early-phase outcome data X for $N = N_2 + N_3$ patients and late-phase outcome data Y_3 for only N_3 patients are observed. As Y_2 is not observed, $Y = Y_3$ and $X = (X_2, X_3)$ now denote the observed late-phase and early-phase outcome data which correspond to patients of Phase II and Phase III trials. Y corresponds to the outcome of interest related to which estimation and hypothesis testing will be performed in N_3 patients. The DR utilizes the relation between early-phase and late-phase outcomes and allows estimation of the main parameter of interest, the treatment effect on the late-phase clinical outcome, b_y (Figure 1).

Based on the DR method, parameters $a_x, b_x,$ and σ_x^2 are estimated via the regression of $X|\mathbf{t}$ on N patients, as $\hat{a}_x, \hat{b}_x, s_x^2$ and parameters $a_0, b_0, \gamma,$ and σ_0^2 are estimated via the regression of $Y_3|X_3, \mathbf{t}$ on N_3 patients, as $\hat{a}_0, \hat{b}_0, \hat{\gamma}, s_0^2$, while $s_y^2 = s_0^2 + \hat{\gamma}^2 s_x^2$, $\hat{a}_y = \hat{a}_0 + \hat{\gamma} \hat{a}_x, \hat{\rho} = \hat{\gamma} s_x / s_y$.^{5,8} The primary effect of interest b_y is then estimated as:

$$\hat{b}_y = \hat{b}_0 + \hat{\gamma} \hat{b}_x. \tag{eq1}$$

The variance of \hat{b}_y is shown in Reference 5 to be equal to

$$\text{var}(\hat{b}_0) + \hat{\gamma}^2 \text{var}(\hat{b}_x) + \hat{b}_x \text{var}(\hat{\gamma}) + 2\hat{b}_x \text{cov}(\hat{b}_0, \hat{\gamma})$$

estimates of the above can be obtained by using the individual estimates acquired from the regression analyses (m2). Under model (m2), hypothesis testing is performed as $H_0 : b_y \leq 0$ vs $H_1 : b_y > 0$ via $z_{1-\alpha_3} > \hat{b}_y / \sqrt{\widehat{\text{var}}(b_y)}$, where $z_{1-\alpha_3}$ is the $(1 - \alpha_3)$ th standard normal quantile and α_3 is the alpha level of the late-phase primary outcome of phase III trial. A direct Bayesian analogue to the conditional model (m2) has been discussed elsewhere.¹¹ Under diffuse “non-informative” priors, this Bayesian model has been shown to produce comparable posterior means for all parameters to the estimates produced by model (m2).

2.3 | Bayesian (double-) regression

We can model the Phase II biomarker data (X_2) via a Bayesian SR, $X_2|\mathbf{t} \sim N(a_x + b_x \mathbf{t}, \sigma_x^2), a_x, b_x \sim N(0, 10^2), \sigma_x^2 \sim \text{IG}(1, 1)$ of N_2 patients and we can utilize the posterior distribution Markov Chain Monte Carlo sample draws to construct a prior on a BDR model on the Phase III early-phase outcome data as follows.

Let us assume a bivariate normal model for the biomarker and the primary late-phase clinical outcome data X_3 and Y_3 corresponding to N_3 patients with a covariance matrix Σ as in model (m1). In our two-dimensional scenario, a bivariate normal likelihood could be specified on the early-phase and late-phase Phase III outcome data by conditional distributions as follows

$$\begin{aligned} X_3|\mathbf{t} &\sim N(a_x + b_x \mathbf{t}, \sigma_x^2) \\ a_x &\sim N(\mu_{ah}, \sigma_{ah}^2), b_x \sim N(\mu_{bh}, \sigma_{bh}^2) \\ \sigma_x^2 &\sim \text{IG}(\alpha_h, \beta_h) \\ Y_3|\mathbf{t}, x_3 &\sim N\left(a_y + b_y \mathbf{t} + \rho \frac{\sigma_y}{\sigma_x} x_3, (1 - \rho^2) \sigma_y^2\right) \\ a_y &\sim N(0, 10^2), b_y \sim N(0, 10^2) \\ \sigma_y^2 &\sim \text{IG}(1, 1) \\ \rho &\sim U(-1, 1). \end{aligned} \tag{m3}$$

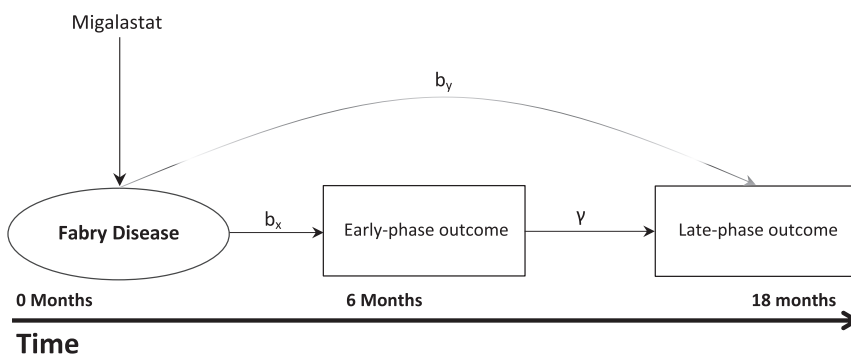


FIGURE 1 Relation between treatment vs early-phase outcome, treatment vs late-phase outcome and early-phase vs late-phase outcome in the example of Fabry disease

The prior on ρ uniformly weights our prior considerations around the correlation parameter. In order to mimic model (m2) we have set normal distribution priors based on Phase II posterior effect and variance mean estimates of the early-phase outcome parameters $(\mu_{ah}, \mu_{bh}, \sigma_{ah}^2, \sigma_{bh}^2)$. To further mimic model (m2) we inform the σ_x prior based on the posterior model variance samples from Phase II early-phase outcome data, that is, fitting them over an optimized gamma prior distribution, $\sigma_x^2 \sim G(\alpha_h, \beta_h)$. The above two-step procedure will allow for possible discounting of the Phase II trial by down-weighting the early-phase historical outcome data, which is further discussed in section 4.

In comparison to the direct Bayesian analogue of model (m2), where the strength of the relationship between early and late-phase endpoints becomes clear only after combining the posterior mean estimates via the γ parameter, model (m3) is more intuitive, as it directly models the correlation (ρ) between the two outcomes, and it directly produces posterior Markov Chain Monte Carlo draws from b_y . Therefore, under such a fully Bayesian approach there is no need for numerical addition of treatment effect mean estimates.¹¹ Posterior inference can be obtained via traditional Markov Chain Monte Carlo application software (ie, JAGS¹³) or even analytically under convenient prior distributions.¹² In this Bayesian model we assume that hypothesis testing for H_0 vs H_1 will be performed by utilizing posterior probabilities as $Pr(b_y > 0|Y) > \omega$ where $\omega = 0.95$.

If we set the correlation very close to zero; that is, $\rho \sim U(-0.01, 0.01)$, then, the Phase III trial late-phase outcome data are evaluated individually under a standard (Bayesian) linear SR model. In comparison to the SR models, the advantage of models (m2), Bayesian (m2) and (m3) rest in their ability to numerically calculate/imitate the impact of accounting for the Phase II early-phase outcome data in analyzing the late-phase outcome. Additional details of the (Bayesian) SR models can be found in Appendix A.

3 | TYPE 1 ERROR INFLATION AND BIAS DUE TO SELECTION BASED ON EARLY-PHASE OUTCOME RESULTS

The potential issues with retrospective combination of early and late phase results stem mostly from the fact that they are not independent. Usually, a Phase II decision leads to the initiation of a Phase III trial. This decision could be based on a test statistic for the early-phase outcome and an imposed critical value; that is, $z_{1-\alpha_2}$. This is clearly an oversimplification of the actual Phase II to Phase III transition decision, but used here to illustrate the potential impact on type 1 error and bias if the results are retrospectively combined. In this simplified model, the distribution of the available Phase II trial early-phase outcome $f(X_2|Z_{X_2} > z_{1-\alpha_2})$, will be truncated, where Z_{X_2} denotes the standardized difference of the early-phase Phase II trial outcome. If the analysis of Phase III data occurs independently from earlier Phase trial data, we expect no increase of Type I error and bias, though the power might remain low due to the limited trial sample size. In the retrospective assessment of the totality of evidence in this rare disease setting, however, positive results from both the Phase II trial and Phase III trial may well be seen as reinforcing. This informally combines evidence between trials which often results in positively biased inferences in favor of the late-phase treatment effect b_y , while an error inflation is observed in the double-regression late-phase outcome inference (models (m2) and (m3)) (Figure 2). In such situations, the bias on \hat{b}_y estimate, based on model (m2) is given by the following approximation (Appendix B),

$$\text{Bias}(\hat{b}_y) = \sigma'_y \frac{w_2 \rho \lambda \sigma_{x2}}{\sigma'_x \sqrt{n_2/2}}, \tag{eq2}$$

where $\sigma_y'^2 = \sigma_y^2 + \gamma^2 D$, $\sigma_x'^2 = \sigma_x^2 + D$, $\lambda = \frac{\phi(\omega)}{1-\Phi(\omega)}$, $\omega = \frac{z_{1-\alpha_2} - \mu_{x2}}{\sigma_{x2}/\sqrt{n_2/2}}$, $w_2 = n_2/n$. ϕ and Φ are probability density and cumulative functions of the standard normal distribution, respectively, $D = w_2 ((2\sigma_{x2}^2/n_2)\zeta + A^2(1 - w_2^2 - w_3^2) + 2A(\mu_{x2} - \mu_x))$, $A = (\sigma_{x2}/\sqrt{n_2/2})\lambda$, $\zeta = \alpha_2 \lambda - (\lambda)^2$.

An approximate value for $\text{MSE}(\hat{b}_y)$ of the double-regression is equal to (Appendix B)

$$\text{MSE}(\hat{b}_y) = \underbrace{2\sigma_y'^2 \left(\frac{w_2 \rho \lambda \sigma_{x2}}{\sigma'_x \sqrt{n_2}} \right)^2}_{\text{Bias}(\hat{b}_y)^2} + \underbrace{2\sigma_y'^2 \left(\frac{1 - \rho^2}{n_3} + \frac{\rho^2}{n} \right)}_{\text{Var}(\hat{b}_y)}. \tag{eq3}$$

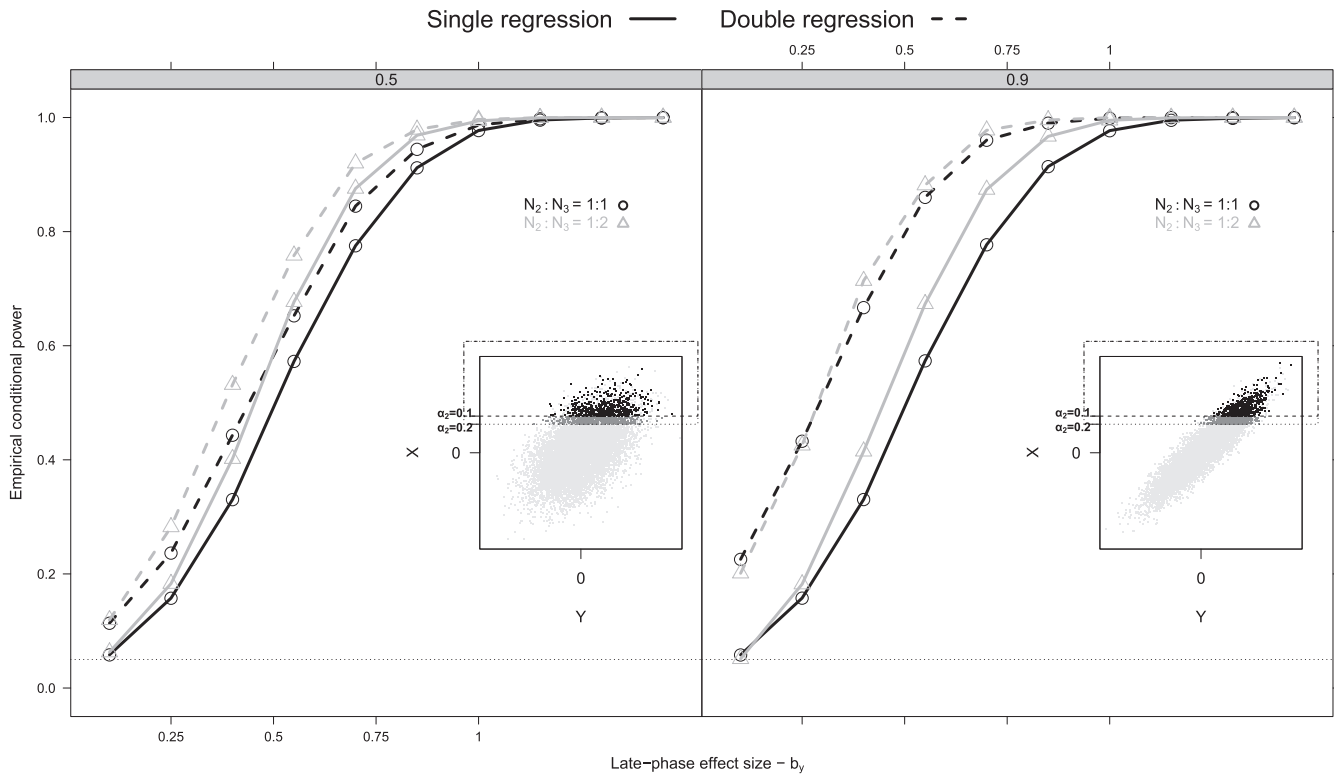


FIGURE 2 Conditional power curves comparing the performance of the single and double-regression for the following scenarios; $N_2:N_3 \in \{1:1, 1:2\}$, $\sigma_y^2 = \sigma_x^2 = 1$, $\rho_r \in \{0.1, 0.9\}$, $N = 120$, $\alpha_2 = 0.1$, and $b_y, b_x \in \{0, 0.1, 0.2, \dots, 1\}$. No between-trial outcome variation was introduced in this set up and each scenario was replicated 10 000 times. The inner figures serve as an explanation to the observed type I error increase, as they present the joint strict null hypothesis ($b_y = b_x = 0$) distribution of the early and late-phase treatment effect for the Phase III trials (light gray dots) and the truncated, based on a positive decision criteria, Phase II trials (black and dark gray dots). When utilizing the Phase II trials (darker dots in the inner Figures), larger critical levels result in an average overestimation of the treatment effect which consistently produces an average increase in error rates and on average larger bias is incorporated in the final inference. This mean increase can be observed in the expression of mean square error for the late-phase treatment effect estimate (eq3). As expected based on (eq3), all error rates increase with higher ρ and the power curve increases with lower σ . A similar behavior was observed between the equivalent Bayesian single-regression and Bayesian double-regression alternative

As we observe in (eq3), the inflation in MSE depends on (i) the decision threshold to initiate the Phase III trial through λ parameter, (ii) the Phase II early-phase outcome mean (μ_{x2}) and variance ($2\sigma_{x2}^2/n_2$), (iii) the number of patients in the Phase II trial (n_2) and (iv) and the magnitude of the correlation (ρ). An increase in σ_{x2} results in an increase of MSE, while as n_2 decreases, the MSE increases as well. A similar behavior is observed in terms of Type I error (Figure 2). More specifically, Type I error rates increase considerably with higher ρ , while the power curves, in general, increase with more patients being allocated to the Phase III trial (n_3) (Figure 2).

Based on the aforementioned bias and mean square error expressions and by replacing parameters with their estimates, the late-phase outcome effect and variance of a (bias) corrected *double-regression* (DRC) model are estimated as (Appendix B),

$$\hat{b}'_y = \hat{b}_y - \tilde{\text{Bias}}(\hat{b}_y) \tag{m4}$$

$$\text{var}(\hat{b}_y)' = 2(s_y^2 - \hat{\gamma}^2 \hat{D}) \left(\frac{1 - \hat{\rho}^2}{n_2} + \frac{\hat{\rho}^2}{n} \right).$$

The above expressions hold when treatment arms within studies are equal. Nonetheless, similar analytical expressions for unequal within study allocation ratios, can be acquired by appropriately changing the variances of \hat{b}_{x3} , \hat{b}_{x2} in Appendix B.1 based on the treatment arms sample sizes. For example, if the allocation ratio between arms in the Phase

II trial equals to 1:2, then the Phase II early-phase endpoint variance increases, $9\sigma_x^2/2N_2$ and the introduced bias could be reduced by half.

4 | BIAS REDUCTION BY ACCOUNTING FOR BETWEEN-TRIAL EARLY-PHASE OUTCOME VARIABILITY

All models above, including the bias corrected model, assume that the true overall treatment effect remains common between trials, no between-study variability on the early and late-phase outcomes exist and therefore, all N observations are derived from the same population. Phase II vs Phase III trials typically do not have similar protocols, as the Phase II trials are usually more restrictive in patient inclusions, therefore, exploring between-study variability becomes relevant.

The decision-induced bias discussed in Section 3, would materialize as difference in treatment effects between the two available trials as well. Therefore, accounting for between-study variability may act as a less rough approach to minimize this decision-induced bias. A proper estimation of the between-trial early-phase outcome variance is not feasible with just two available studies,¹⁴⁻¹⁷ therefore, in this article we choose not estimate but only account for this variance to aid towards the reduction of the bias.

To achieve this, we utilize a mechanism based on power priors to account for the between-study differences within a Bayesian framework.¹⁸ By estimating a power parameter $\hat{\eta}$ that represents conflict between the early-phase outcome data of the two available trials, model (m3) can be further extended to account for the early-phase outcome effect excess between-trial variability, along with any other biases.¹⁸⁻²⁰

4.1 | Bayesian flexible double-regression

Let us assume that data X_2 exist for the early-phase outcome from the Phase II study and \mathfrak{B} are a set of linear regression parameters. Given the definition of a power prior,²¹ the posterior distribution after observing the Phase II early-phase outcome data would be

$$\pi(\mathfrak{B}|X_2, \eta) \propto L(\mathfrak{B}|X_2)^\eta \pi_0(\mathfrak{B}).$$

Then, the posterior for \mathfrak{B} after observing the Phase III study's early-phase outcome data (X_3) would be

$$\pi(\mathfrak{B}|X, \eta) \propto L(\mathfrak{B}|X_3)L(\mathfrak{B}|X_2)^\eta \pi_2(\mathfrak{B}).$$

The posterior distribution of $\mathfrak{B}|X_2$ in the normal case²² is known to be equal to

$$\mathfrak{B}|X_2, \eta \sim N\left(\left(\mathbf{T}'_2\mathbf{T}_2\right)^{-1}\mathbf{T}'_2Y_2, \frac{\sigma_x^2}{\eta}\left(\mathbf{T}'_2\mathbf{T}_2\right)^{-1}\right), \quad (\text{eq4})$$

where \mathbf{T}_2 is the design matrix with column vectors $\mathbf{1}$, \mathbf{t} , and dimensions $N_2 \cdot 2$. We now consider the following conditional model for the early and late-phase outcome data of N_3 patients

$$\begin{aligned} X_3|\mathbf{t} &\sim N(a_x + b_x\mathbf{t}, \sigma_x^2) \\ a_x &\sim N(\mu_{ah}, \sigma_{ah}^2/\hat{\eta}), b_x \sim N(\mu_{bh}, \sigma_{bh}^2/\hat{\eta}) \\ \sigma_x^2 &\sim \text{IG}(\alpha_h, \beta_h) \\ Y_3|\mathbf{t}, x_3 &\sim N\left(a_y + b_y\mathbf{t} + \rho\frac{\sigma_y}{\sigma_x}x_3, (1 - \rho^2)\sigma_y^2\right) \\ a_y &\sim N(0, 10^2), b_y \sim N(0, 10^2) \\ \sigma_y^2 &\sim \text{IG}(1, 1) \\ \rho &\sim U(-1, 1). \end{aligned} \quad (\text{m5})$$

TABLE 2 Summary of aforementioned statistical methods

Abbreviation	Model	(F)requentist/ (B)ayesian	Early/late-phase	Phase (II/III)
(B)SR	(Bayesian) single-regression	F/B	Late phase	III
(B)DR	(Bayesian) double-regression	F/B	Early and late phase	II+III
DRC	Double-regression corrected	F	Early and late phase	II+III
BFDR	Bayesian flexible double-regression	B	Early and late phase	II+III

The conditional set-up of model (m5) remains similar to (m3). Now dynamic informative power priors parametrized through $\hat{\eta}$ are placed on the early-phase endpoint's parameters a_x and b_x . Such priors control the borrowing of the historical data and discount the early-phase prior in case of treatment effect's disagreement. We chose to model the parameters univariately to aid any formulation of elicited informative priors on a_y, b_y , and ρ , though, a wishart prior on the covariance matrix Σ (m1) could have jointly accounted for the association between the model parameters.

4.1.1 | Estimation of η

A number of power prior (guided-value) formulations has been suggested.¹⁸⁻²⁰ Among the above alternatives, we chose one that selects a guided-value that maximizes the marginal likelihood.²⁰ The guide value of η based on the marginal likelihood criterion has an estimate of

$$\hat{\eta} = \arg \min_{0 < \eta \leq 1} [-2 \log\{m(\eta)\}], \quad (\text{eq5})$$

where $m(\eta)$ is the marginal likelihood. Ibrahim et al²² provided an analytical expression of $-2 \log\{m(\eta)\}$ for the normal outcome case. Figure D1 (Appendix D) presents the empirically calculated relationship between η and varying levels of b_x .

In model (m5), similarly to model (m3), we are interested in the late-phase overall primary outcome effect b_y and we assume that hypothesis testing for H_0 vs H_1 will be performed by utilizing posterior probabilities as $Pr(b_y > 0|Y) > \omega$ where $\omega = 0.95$.

5 | SIMULATION STUDY

The main four approaches discussed are summarized in Table 2. The corrected double-regression approach as shown in Section 3 can be considered a rough (approximate) approach to minimize the decision-induced bias. The Bayesian flexible double-regression approach minimizes this bias by accounting for between-trial differences without ad hoc corrections. Their relative performance in the analysis of the Phase III late-phase outcome data, also in comparison to the two more trivial approaches (single and double-regression) is the main focus of the simulation study.

For illustrative purposes, we assume that the two available Phase II and Phase III trials had a similar control treatment, therefore, the Phase III trial would have been designed as a placebo-controlled trial. In this section, we assume that the decision to conduct the Phase III trial was taken on the basis of available evidence in the first Phase II trial on a single early-phase outcome. At the end of the Phase II trial, individual data of N patients are available on the early-phase and data of N_3 are available on the late-phase outcomes. The simulation study results were derived from a bivariate normal model simulation strategy as described in Appendix C.

The SR, DR, DRC methods ignore any between-study variability and therefore assume a different underlying data generating model in comparison to the Bayesian flexible *double-regression* (BFDR) approach. Even though, they are not directly comparable (Table 2), we empirically compared the four aforementioned statistical methods by generating at least 10 000 simulated combinations of the two available trials data. To do so, we simulated scenarios of the final trial analysis

on the late-phase primary endpoint assuming a variety of combinations between the early-phase (b_x) and late-phase (b_y) outcome treatment effects. The latter were varied as (Scenario I) $b_y = b_x = 0$, (Scenario II) $b_y = b_x = 0.6$, (Scenario III) $b_{x2}, b_{y2} = 0$, $b_{x3}, b_{y3} = 0.2$, and (Scenario IV) $b_y = 0.6, b_x = 0$, we assumed that $\rho = 0.9$, $\alpha_2 \in \{0.05, 0.1, 0.2\}$ the alpha level of the early-phase primary outcome of Phase II trial, while all within-study variances were set equal to 1. In the simulation setup we introduce a simulative parameter that place additional between-trial variance on the early-phase (τ_x) and late-phase (τ_y) outcomes (see Appendix C for details). Specific alternative versions of scenarios I and II were produced by varying ρ and τ_y, τ_x .

The first (I) scenario describes variations of the strict null ($\tau_y = \tau_x = 0$) and null hypothesis with additional between-trial variance ($\tau_y = \tau_x = 0.3$), while the second (II) scenario describes a common alternative hypothesis on both outcomes and trials. Scenario III can occur when heterogeneous populations are selected for the Phase II and Phase III trial, while the fourth (IV) scenario describes a situation where the late-phase outcome true effect exists but the early-phase outcome equals to 0. All remaining settings (ie. number of trials (k), total sample sizes N , sample size ratio between trials $N_2 : N_3$, within-study allocation ratios $n_{ck} : n_{ek}$) were reflective of a typical rare disease setting and based on the Galafold example (Table 1). All simulations were performed via R²³ and JAGS.¹³

5.1 | (Strict) null hypothesis scenario (I: $b_y = b_x = 0$)

The BFDR results in treatment effects closer to the SR estimates than the DR approach under the null hypothesis simulation (Scenario I—Table 3). The DRC approach presents a similar behavior producing late-phase effect estimates even closer to the SR than the BFDR approach. In the three null hypothesis scenarios I(b-d) ($b_y = b_x = 0$), DR results in the largest estimated treatment effect and produces the largest type I error inflation while DRC generally inflates the Type I error the least among the three investigated methods. An interesting exception that we further discuss in Section 7, is observed in scenario Ia, where the BFDR approach produces stricter error rates than the DRC approach. In general, the SR method controls type I error the most, while the DR method controls type I error the least. The DR and DRC methods consistently produce the smallest C(r)Is, while the BFDR method produces the largest C(r)Is among the investigated methods.

5.2 | Alternative hypothesis scenario (II: $b_y = b_x = 0.6$)

In scenario II ($b_y = b_x = 0.6$), all methods identified a treatment effect close to the true value (Table 4). The empirical power to identify a treatment effect is usually large for the BFDR, and considerably larger for the DRC than SR approach. Among the DRC and BFDR methods, BFDR produces treatment effect means closest to the true value. In scenario IIa ($\tau_y = \tau_x = 0$), DRC performs better in terms of 95% coverage whereas in scenario IIb where $\tau_y = \tau_x = 0.3$, BFDR results in coverage closest to 95%. The C(r)Is widths retained a similar behavior to the null hypothesis scenarios.

5.3 | Scenarios III and IV

In scenario III ($b_{y2} = 0, b_{y3} = 0.2, b_{x2} = 0, b_{x3} = 0.2$), the BFDR produces similar findings to the DR approach, while the DRC method discards most Phase II information and its results are close to the SR approach (Table 4). DRC retains a comparable behavior in scenario IV ($b_y = 0.6, b_x = 0$), where it discards most of the decision-induced bias and it produces results closer to the analysis of the Phase III study alone. In scenarios III, IV, as well as I, the naive pooling represented via the formal DR method, systematically and largely overstates our confidence in treatment efficacy.

5.4 | Summary of simulation results

Among the four methods, the *single regression* performed best in terms of type I error followed closely by the DRC. Similarly, the approach that led to the least bias was the SR, again followed closely by the DRC. The DRC and DR methods

TABLE 3 Late-phase conditional average treatment effect estimates (means, posterior means, confidence intervals, credible intervals) and average treatment efficacy P -values and probabilities of the four models (Table 2) given that $\rho = 0.9$, $\tau_x = \tau_y = 0.01$, and $\sigma_y = \sigma_x = 1$, except where noted otherwise, based on at least 10,000 simulations

Scenario	Model	Mean/Posterior mean b_y	Type I error	C(r)I widths
		$\alpha_2: (0.05 \cdot 0.1 \cdot 0.2)$	$\alpha_2: (0.05 \cdot 0.1 \cdot 0.2)$	$\alpha_2: (0.05 \cdot 0.1 \cdot 0.2)$
Ia. $b_y = b_x = 0$	SR	0.001 · 0.003 · 0.002	0.057 · 0.054 · 0.053	1.138 · 1.138 · 1.136
	DR	0.256 · 0.220 · 0.178	0.318 · 0.247 · 0.183	0.808 · 0.810 · 0.811
	DRC	0.087 · 0.075 · 0.063	0.079 · 0.066 · 0.060	0.810 · 0.812 · 0.813
	BFDR	0.170 · 0.156 · 0.133	0.054 · 0.037 · 0.022	1.343 · 1.330 · 1.319
b. $b_y = b_x = 0$ $\rho = 0.5$	SR	0.000 · 0.003 · 0.002	0.055 · 0.053 · 0.054	1.138 · 1.138 · 1.138
	DR	0.141 · 0.123 · 0.100	0.148 · 0.130 · 0.114	1.010 · 1.010 · 1.011
	DRC	-0.028 · -0.022 · -0.015	0.045 · 0.047 · 0.048	1.012 · 1.012 · 1.012
	BFDR	0.089 · 0.083 · 0.071	0.070 · 0.066 · 0.056	1.211 · 1.206 · 1.203
c. $b_y = b_x = 0$ $\tau_x = \tau_y = 0.3$	SR	0.002 · 0.004 · 0.002	0.058 · 0.054 · 0.054	1.188 · 1.187 · 1.187
	DR	0.246 · 0.211 · 0.171	0.267 · 0.211 · 0.164	0.896 · 0.898 · 0.899
	DRC	0.006 · 0.007 · 0.009	0.041 · 0.042 · 0.042	0.883 · 0.885 · 0.887
	BFDR	0.136 · 0.126 · 0.109	0.069 · 0.052 · 0.037	1.330 · 1.318 · 1.309
d. $b_y = b_x = 0$ $\tau_x = \tau_y = 0.3$ $\rho = 0.5$	SR	0.000 · 0.002 · 0.003	0.055 · 0.053 · 0.054	1.188 · 1.188 · 1.187
	DR	0.135 · 0.117 · 0.097	0.139 · 0.122 · 0.110	1.067 · 1.068 · 1.068
	DRC	0.002 · 0.004 · 0.006	0.059 · 0.058 · 0.059	1.064 · 1.065 · 1.065
	BFDR	0.073 · 0.069 · 0.060	0.076 · 0.072 · 0.065	1.209 · 1.205 · 1.201

Note: The first line SR of each scenario (I) presents a frequentist *single-regression* on the Phase III late-phase outcome data. DR correspond to the frequentist *double-regression*. Last, the DRC lines present the result for the bias corrected *double-regression* approach and the BFDR lines present the results for the Bayesian flexible *double-regression* approach. $\alpha_3 = 0.05$ and α_2 denotes the alpha level of the early-phase primary outcome of the phase II trial.

resulted in the narrowest intervals. The intervals of the BFDR were comparable or larger than these of the SR. In terms of power, the DR method showed the highest gain, closely followed by the DRC. Finally, the SR and DRC both attained coverage close to nominal levels.

Overall, the DRC resulted in similar operational characteristics to the SR but it demonstrated a large gain in empirical power under the alternative hypothesis scenarios in comparison to the SR (Tables 3 and 4).

6 | DISCUSSION

In a drug development procedure, it is not uncommon that positive Phase II results on early-phase (biomarker) outcomes are not predictive of a Phase III success on late-phase clinical outcomes. If Phase II and Phase III results are then assessed (perhaps informally) jointly to support efficacy, this retrospective (ad hoc) assessment may be subject to decision-induced bias and may increase uncertainty of the true primary late-phase treatment effect. Such an informal combination of results may increase to a great extent (more than three times) the Type I error rate of null hypothesis, rendering the retrospectively combined late-phase true treatment effect misleading. Especially in rare diseases, where the validation of early-phase surrogate endpoints can become problematic, due to the small and often heterogeneous populations, the small sample sizes and the insufficient number of available trials, only late-phase hard endpoints are usually appropriate to prove treatment efficacy.

In this article, in addition to identifying and investigating the above issue, we explored methods that can be utilized in order for early and late Phase trial data to be combined retrospectively (ie, right before drug marketing authorization request), while accounting for the underlying decision-induced bias. The flexible BDR includes the borrowing of historical

TABLE 4 Late-phase conditional average treatment effect estimates (means, posterior means, confidence intervals, credible intervals) and average treatment efficacy P -values and probabilities of the four models (Table 2) given that $\rho = 0.9$, $\tau_x = \tau_y = 0.01$, and $\sigma_y = \sigma_x = 1$, except where noted otherwise, based on at least 10,000 simulations

Scenario	Model	Mean/Posterior mean b_y	Power	95% coverage	C(r)I widths
		$\alpha_2: (0.05 \cdot 0.1 \cdot 0.2)$	$\alpha_2: (0.05 \cdot 0.1 \cdot 0.2)$	$\alpha_2: (0.05 \cdot 0.1 \cdot 0.2)$	$\alpha_2: (0.05 \cdot 0.1 \cdot 0.2)$
IIa. $b_y = b_x = 0.6$	SR	0.598 · 0.596 · 0.598	0.659 · 0.655 · 0.658	0.940 · 0.940 · 0.942	1.138 · 1.137 · 1.138
	DR	0.643 · 0.625 · 0.612	0.942 · 0.924 · 0.909	0.954 · 0.952 · 0.951	0.811 · 0.812 · 0.812
	DRC	0.634 · 0.621 · 0.611	0.935 · 0.920 · 0.907	0.956 · 0.954 · 0.952	0.812 · 0.812 · 0.813
	BFDR	0.632 · 0.617 · 0.607	0.663 · 0.634 · 0.612	0.997 · 0.997 · 0.997	1.304 · 1.304 · 1.305
b. $b_y = b_x = 0.6$ $\tau_x = \tau_y = 0.3$	SR	0.598 · 0.596 · 0.598	0.626 · 0.624 · 0.625	0.940 · 0.941 · 0.942	1.188 · 1.187 · 1.188
	DR	0.647 · 0.628 · 0.614	0.888 · 0.866 · 0.848	0.948 · 0.949 · 0.940	0.898 · 0.899 · 0.900
	DRC	0.634 · 0.621 · 0.612	0.876 · 0.859 · 0.845	0.950 · 0.949 · 0.946	0.896 · 0.898 · 0.899
	BFDR	0.629 · 0.615 · 0.607	0.648 · 0.622 · 0.610	0.989 · 0.989 · 0.990	1.292 · 1.293 · 1.293
III. $b_{x3}, b_{y3} = 0.2,$ $b_{x2}, b_{y2} = 0$ $\tau_x = \tau_y = 0.3$	SR	0.202 · 0.204 · 0.202	0.173 · 0.169 · 0.168	0.941 · 0.941 · 0.945	1.188 · 1.187 · 1.187
	DR	0.363 · 0.328 · 0.289	0.470 · 0.399 · 0.337	0.906 · 0.931 · 0.950	0.894 · 0.896 · 0.898
	DRC	0.226 · 0.221 · 0.214	0.244 · 0.232 · 0.223	0.961 · 0.963 · 0.968	0.883 · 0.886 · 0.889
	BFDR	0.315 · 0.296 · 0.271	0.194 · 0.158 · 0.125	0.985 · 0.987 · 0.991	1.307 · 1.299 · 1.296
IV. $b_y = 0.6, b_x = 0$ $\tau_x = \tau_y = 0.3$	SR	0.602 · 0.602 · 0.602	0.626 · 0.626 · 0.630	0.941 · 0.941 · 0.945	1.188 · 1.188 · 1.187
	DR	0.846 · 0.846 · 0.771	0.988 · 0.988 · 0.971	0.828 · 0.828 · 0.906	0.896 · 0.896 · 0.899
	DRC	0.606 · 0.606 · 0.609	0.870 · 0.870 · 0.869	0.960 · 0.960 · 0.967	0.883 · 0.883 · 0.887
	BFDR	0.735 · 0.735 · 0.708	0.736 · 0.736 · 0.743	0.970 · 0.971 · 0.985	1.329 · 1.330 · 1.309

Note: The first line SR of each scenario (II,III,IV) presents a frequentist *single-regression* on the Phase III late-phase outcome data. DR correspond to the frequentist *double-regression*. Last, the DRC lines present the result for the bias corrected *double-regression* approach and the BFDR lines present the results for the Bayesian flexible *double-regression* approach. α_2 denotes the alpha level of the early-phase primary outcome of the phase II trial. In Scenario III the correction for the DRC method is calculated based on that the true late-phase outcome effect is equal to 0.2.

information, while this model downgrades the historical prior upon early-phase outcome data conflict. The DRC method approximately corrects the biased late-phase mean effect and variance estimate.

In most scenarios, the DRC method better controls the Type I error and bias than the DR and BFDR methods. This is not observed in scenario Ia, where the BFDR controls better the Type I error than the DRC. This possibly happens because the BFDR approach completely downgrades the impact of Phase II trial when its early-phase treatment effect is different than the Phase III trial early-phase treatment effect. Therefore, on average the Bayesian approach becomes less prone to false-positive results based on possible very positive Phase II early-phase outcome trial effects when τ_x is low and/or ρ is high (see, black dots of inner right panel of Figure 2). On the contrary, the DRC corrects the Phase II effect and then utilizes both Phase II and Phase III effects without heavily downgrading the Phase II results data upon data conflict. The DRC requires a known α_2 but despite being approximate, it applies a more direct (decision-based) penalty to the Phase II effect than the Bayesian approach; which could explain its overall better performance in the simulation.

Both the BFDR and the DRC methods would be an attractive solution to the increased Type I error of the informal retrospective combination of two small available trials. The consideration of these methods was shown to be rather important when, (i) the preceding Phase II trial conservatively (ie, alpha level was small) resulted to the Phase III trial and/or (ii) the association of utilized early and late-phase outcomes is high. An informal combination of results across Phases often happens when both of the above hold, though, when neither holds then the complexity of suggested methods may outweigh the gains of their application.

Alternative versions of the BFDR model could be developed and they may perform more optimally in comparison to the current (ie, in terms of controlling the overall type I error) when applied on the flexible BDR via the use of an alternative guided value.¹⁸⁻²⁰ The power parameter is imposed on the early-phase endpoint and only indirectly affects the primary late-phase endpoint, therefore, inference on the late-phase endpoint via alternative guided values on the early-phase endpoints could be expected to be more comparable to some extent.

An alternative approach that controls type I error on the late-phase outcome, while borrowing historical information, may also provide a more formal solution.¹⁹ Future research could compare these alternatives vis-à-vis each other or with other methods. More covariates could be included, and then their performance could be tested with ease as all presented models are readily generalizable to full regressions. In this article, we set independent informative priors on the model parameters, however, accounting for the correlation between these parameters could also be considered through a well-defined informative Wishart prior on the whole covariance matrix. Finally, in this work, we accounted for but did not estimate between-study variance. Due to the only two available studies, a proper estimation of the between-study outcome variability is currently known to be almost nonfeasible.¹⁴⁻¹⁷

In the motivating example we assumed that both trials were superiority trials, while if we had kept the initial designs, different strategies may have been more appropriate. Nonetheless, examples of two superiority trials, one Phase II and one Phase III, exist in the literature. For example, the drug development program of thalidomide for the treatment of multiple myeloma contained two randomized superiority clinical studies of similar design, a supportive (GISMM2001) and a main study (IFM 99-06), that compared melphalan-prednisone (control treatment) to thalidomide (experimental treatment).² The supportive study was shorter and it reported clinical response rates and event free survival as primary endpoints. The main study was longer in duration and it reported overall survival, as main endpoint and clinical response rates and event free survival, as secondary endpoints. The suggested methodology could be tailored to account for the possibility of decision-induced bias under survival and other types of outcomes and even to combine different study designs.

Throughout the article normality was assumed, an assumption that could be challenged with rare diseases sample sizes.¹⁻³ We approximated a truncated normal with a normal distribution with mean and variance equal to that of the former. This decision was made to aid calculations on the distribution mixture (Appendix B). Better approximations for the truncated normal distribution may exist, such as the chi-square distribution and their performance could be explored as well.²⁴ We should note that for moderately sized N_2 in comparison to N and small correlation between the two outcomes, a SR might be more efficient than a DR, due to the noise introduced by the early-phase outcome.⁵ In the simulation study we assumed that the Phase II trial had equal allocation between trial arms, while the Phase III trial had allocation equal to 1:2 between the control vs treatment arm. We expect that our findings would be comparable under different allocations between arm sample sizes, though further investigation could provide more insights between the relative performance of BDR and DRC methods.

In this article, we performed a post hoc (retrospective) combination of available information after the conduct of the Phase II and Phase III trial. However, it may be very relevant to (prospectively) plan to pool the data from both studies and to use the early-phase outcomes of the Phase II study to increase the precision, with which the efficacy on late-phase outcome is estimated overall.⁷⁻⁹ An alternative strategy could be to conduct one single trial with interim analysis, then, based on the observed treatment effects on the early-phase endpoints decide whether to follow-up the patients.⁸

To conclude, especially in a small population context, the often informal retrospective pooling of a single Phase II early-phase outcome data to support the true late-phase outcome data inference at the end of a single confirmatory Phase III trials could induce bias and it should be performed via formal numerical approaches. Such approaches should control this decision-induced bias, in order to avoid inflating the Type I error under the null hypothesis and prevent overestimating our beliefs on the primary treatment effect. We hope that this article, except for introducing possible solutions, raises awareness of potential mishaps with post hoc combinations of trial outcome results.


ACKNOWLEDGEMENTS


This work has been funded by the FP7-HEALTH-2013-INNOVATION-1 project Advances in Small Trials Design for Regulatory Innovation and Excellence (ASTERIX) Grant Agreement No. 603160.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in GitHub at <https://github.com/kpatera/data-earlylate>.

ORCID

Konstantinos Pateras  <https://orcid.org/0000-0002-6005-9798>

Stavros Nikolakopoulos  <https://orcid.org/0000-0002-9769-3725>

REFERENCES

1. Pontes C, Fontanet JM, Gomez-Valent M, et al. Milestones on orphan medicinal products development: the 100 first drugs for rare diseases approved throughout Europe. *Clin Ther.* 2016;37(8):132. <https://doi.org/10.1016/j.clinthera.2015.05.378>.
2. European Medicines Agency Thalidomide pharmion (thalidomide) - assessment report. Technical report; 2008.
3. European Medicines Agency Galafold (migalastat) - assesment report. Technical report; 2012.
4. Schiffmann R, Ries M, Blankenship D, et al. Changes in plasma and urine globotriaosylceramide levels do not predict Fabry disease progression over 1 year of agalsidase alfa. *Genet Med.* 2013;15(12):983-989. <https://doi.org/10.1038/gim.2013.56>.
5. Engel B, Walstra P. Increasing precision or reducing expense in regression experiments by using information from a concomitant variable. *Biometrics.* 1991;47(1):13. <https://doi.org/10.2307/2532491>.
6. Conniffe D, Moran MA. Double sampling with regression in comparative studies of carcass composition. *Biometrics.* 1972;28(4):1011. <https://doi.org/10.2307/2528637>.
7. Kunz CU, Friede T, Parsons N, Todd S, Stallard N. A Comparison of methods for treatment selection in seamless Phase II/III clinical trials incorporating information on short-term endpoints. *J Biopharm Stat.* 2015;25(1):170-189. <https://doi.org/10.1080/10543406.2013.840646>.
8. Stallard N. A confirmatory seamless Phase II/III clinical trial design incorporating short-term endpoint information. *Stat Med.* 2010;29(April 2009):959-971. <https://doi.org/10.1002/sim.3863>.
9. Hampson LV, Jennison C. Optimizing the data combination rule for seamless Phase II/III clinical trials. *Stat Med.* 2015;34(1):39-58. <https://doi.org/10.1002/sim.6316>.
10. Galbraith S, Marschner IC, Young DM. Interim analysis of continuous long-term endpoints in clinical trials with longitudinal outcomes. *Stat Med J.* 2003;22(11):1787-1805. <https://doi.org/10.1002/sim.1311>.
11. Manner D, Seaman JW, Young DM. Bayesian methods for regression using surrogate variables. *Biom J.* 2004;46(6):750-759. <https://doi.org/10.1002/bimj.200210073>.
12. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis.* Vol 2. Boca Raton, FL: Chapman & Hall/CRC Press; 2014.
13. Plummer M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. Paper presented at: Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Wien, Vienna; March, 20–22, 2003. <https://doi.org/10.1.1.13.3406>.
14. Pateras K, Nikolakopoulos S, Roes KCB. Data- generating models of dichotomous outcomes: heterogeneity in simulation studies for a random effects meta analysis. *Stat Med.* 2018;37:1115-1124. <https://doi.org/10.1002/sim.7569>.
15. Pateras K, Nikolakopoulos S, Roes KCB. Prior distributions for variance parameters in a sparse-event meta-analysis of a few small trials. *Pharm Stat.* Under revision. 2020;20:39-54.
16. Gonnermann A, Framke T, Großhennig A, Koch A. No solution yet for combining two independent studies in the presence of heterogeneity. *Stat Med.* 2015;34(16):2476-2480. <https://dx.doi.org/10.1002/sim.6473>.
17. Pateras K, Nikolakopoulos S, Mavridis D, Roes KCB. Interval estimation of the overall treatment effect in a meta-analysis of a few small studies with zero events. *Contemp Clin Trials Commun.* 2018;9:98-107. <https://doi.org/10.1016/j.conctc.2017.11.012>.
18. Ibrahim JG, Chen M-H, Sinha D. On optimality properties of the power prior. *J Am Stat Assoc.* 2003;98(461):204-213. <https://doi.org/10.1198/016214503388619229>.
19. Nikolakopoulos S, Tweel IT, Roes KCB. Dynamic borrowing through empirical power priors that control type I error. *Biometrics.* 2017;74:874-880. <https://doi.org/10.1111/biom.12835>.
20. Gravestock I, Held L. Adaptive power priors with empirical Bayes for clinical trials. *Pharm Stat.* 2017;16(5):349-360. <https://doi.org/10.1002/pst.1814>.
21. Ibrahim JG, Chen M-H. Power prior distributions for regression models. *Stat Sci.* 2000;15(1):46-60. <https://doi.org/10.1214/ss/1009212673>.
22. Ibrahim JG, Chen MH, Gwon Y, Chen C. The power prior: theory and applications. *Stat Med.* 2015;28(34):3724-3749. <https://doi.org/10.1002/sim.6728>.
23. R Core Team R: *A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. (2019). <http://www.R-project.org/>.
24. Donald B, Sherrill T. Mean and variance of truncated normal distributions. *Am Stat.* 2008;53(4):357-361. <https://doi.org/10.1080/00031305.1999.10474490>.
25. Malzahn U, Bohning D, Holling H. Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika.* 2000;87(3):619-632. www.jstor.org/stable/2673634.
26. Johnson, N.L, Kotz, S. Balakrishnan, N. *Continuous univariate distributions.* New York, NY: John Wiley & Sons; 1994;1, (Section 10.1).
27. Riley R, Held L. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics.* 2008;9(1):172-186. <https://doi.org/10.1093/biostatistics/kxm023>.

How to cite this article: Pateras K, Nikolakopoulos S, Roes KCB. Combined assessment of early and late-phase outcomes in orphan drug development. *Statistics in Medicine.* 2021;40:2957–2974. <https://doi.org/10.1002/sim.8952>

APPENDIX A. DETAILS OF (BAYESIAN) UNIVARIATE MODEL

The standard linear SR reference model to demonstrate late-phase treatment efficacy assumes $Y|\mathbf{t} \sim N(a_y + b_y\mathbf{t}, \sigma_y^2)$, where σ_y denotes the true outcome variance, \mathbf{t} denotes a vector of length n_3 indicating whether a patient receives control or experimental treatment.

A conjugate Bayesian analogue (BSR) of the model above can be expressed also as above where $a_y, b_y,$ and σ_y are random variables and need a prior distribution. This model offers the flexibility to directly impact inference via placing informative priors on parameters $a_y, b_y,$ and σ_y . Model B(SR) corresponds exactly to the aforementioned model SR under convenient noninformative priors on $a_y, b_y,$ and σ_y .¹²

In the above SR model, we are interested in \hat{b}_y and we assume that hypothesis testing for $H_0 : b_y = 0$ vs $H_1 : b_y > 0$ will be evaluated as $z_{1-\alpha_3} < \hat{b}_y / \sqrt{\text{var}(\hat{b}_y)} = \Phi\left(\hat{b}_y / \sqrt{\text{var}(\hat{b}_y)}\right)$, where $z_{1-\alpha_3}$ is the α_3 th standard normal quantile. In the Bayesian SR analogue, we are interested in b_y and we assume that hypothesis testing for H_0 vs H_1 will be performed by utilizing posterior probabilities as $Pr(b_y > 0|Y) > \omega$ where $\omega = 0.95$.

APPENDIX B. DERIVATION OF MSE(\hat{b}_y)

The MSE(\hat{b}_y) of the late-phase outcome equals to

$$\text{MSE}(\hat{b}_y) = \text{Bias}(\hat{b}_y)^2 + \text{Var}(\hat{b}_y).$$

B.1 Derivation of Bias(b_y)

Let assume that σ_{x2}, σ_{x3} are known for the Phase II and Phase III trials, then the early-phase outcome treatment effect estimates are distributed as $\hat{b}_{x3} \sim N(\mu_{x3}, \frac{2\sigma_{x3}^2}{n_3})$ and $\hat{b}_{x2} \sim N(\mu_{x2}, \frac{2\sigma_{x2}^2}{n_2})$. In practice the Phase II early-phase outcomes would follow an one-sided truncated normal distribution. The adjusted mean (μ_{x2}) and variance (σ_{x2}^2) of this early-phase outcome one-sided truncated normal distribution $\hat{b}_{x2} \sim N_{a_2}(\mu'_{x2}, \frac{2\sigma_{x2}^2}{n_2})$ equal to

$$\mu'_{x2} = \mu_{x2} + \frac{\sigma_{x2}}{\sqrt{n_2/2}} \lambda \tag{eq3}$$

$$\sigma_{x2}^{\prime 2} = \sigma_{x2}^2 [1 + \zeta], \tag{eq4}$$

where $\lambda = \frac{\phi(\omega)}{1-\Phi(\omega)}$, $\zeta = a\lambda - (\lambda)^2$ and $\omega = Z_{1-\alpha_2}^x - \frac{\mu'_{x2}}{\sigma'_{x2}/\sqrt{n_2/2}}$ and ϕ and Φ are the probability density and the cumulative function of the standard normal distribution.

We assume that we can approximate a truncated normal with a normal distribution with updated mean and variance as follows $\hat{b}_{x2} \overset{\text{approx}}{\sim} N(\mu'_{x2}, \frac{2\sigma_{x2}^{\prime 2}}{n_2})$.²⁵ The overall \hat{b}_x would be a mixture of the above density functions.

Given the set of two densities and weights (w_1 and w_2), such that $w_i \leq 0$ and $\sum w_i = 1$ the mixture can be represented as

$$f(x) = \sum_{k=1}^2 w_k p_k(x).$$

The mean and variance of the above normal mixture of two distributions equal to $\mu_x = \sum_{k=1}^2 w_k \mu_{xk}$ and $\sigma_x^2 = \sum_{k=1}^2 w_k (\mu_{xk}^2 + \frac{2\sigma_{xk}^2}{n_i} - \mu_x^2)$ with $w_k = N_k/N$.²⁶ Therefore, $\hat{b}_x \sim N(\mu_x, \sigma_x^2)$.

Therefore, the updated mean and variance of \hat{b}_x , are equal to

$$\begin{aligned} \mu'_x &= w_2 \mu'_{x2} + w_3 \mu_{x3} \\ &= w_2 \mu_{x2} + w_3 \mu_{x3} + w_2 \lambda \frac{\sigma_{x2}}{\sqrt{n_2/2}} \\ &= \mu_x + w_2 \lambda \frac{\sigma_{x2}}{\sqrt{n_2/2}} \end{aligned} \tag{m6}$$

$$\begin{aligned} \sigma_x'^2 &= \sum_{k=1}^2 w_k \left(\mu_{xk}^2 + \frac{2\sigma_{xk}^2}{n_k} - \mu_x^2 \right) + D \\ &= \sigma_x^2 + D, \end{aligned}$$

where $D = w_1 ((2\sigma_{x2}^2/n_2)\zeta + A^2(1 - w_2^2 - w_3^2) + 2A(\mu_{x2} - \mu_x))$, $A = (\sigma_{x2}/\sqrt{n_2/2})\lambda$ and $\zeta = a\lambda - (\lambda)^2$.

A bias is introduced after combining the Phase II and III trial early-phase outcome effect estimates as $\frac{\sigma_{x2}\lambda \cdot w_2}{\sqrt{n_2/2}}$.²⁶ Then based on (eq1) and assuming that $\sigma_x = \sigma_y = 1$, the bias of b_y equals to

$$Bias(B) = \frac{w_2\lambda \rho\sigma_{x2}}{\sqrt{n_2/2}}. \tag{m7}$$

B.2 Derivation of Var(\hat{b}_y)

The variance of late-phase outcome b_y is equal to Reference,⁵

$$Var(b_y) = var(\hat{b}_0) + \hat{\gamma}^2 var(\hat{b}_x) + \hat{b}_x^2 var(\hat{\gamma}) + 2\hat{b}_x cov(\hat{b}_0, \hat{\gamma}). \tag{eqA1}$$

An estimate of $Var(\hat{b}_y)$ can be obtain via estimates of the relevant parameters which can be obtained directly via the regression of $X|\mathbf{t}$ and the regression of $Y|X, \mathbf{t}$ on N and N_3 patients, respectively.

Assuming that \mathbf{t} is an indicator variable and n_k corresponds to the total sample size per treatment arm of the k th trial, the q-dependent variance of (\hat{a}_x, \hat{b}_x) can be derived as $\sigma_x'^2(\mathbf{T}'\mathbf{T})^{-1}$, where T is the design matrix of $X|\mathbf{t}$ on N patients as follows

$$(\mathbf{T}'\mathbf{T})^{-1} = \begin{pmatrix} 2n & n \\ n & n \end{pmatrix}^{-1} = \begin{pmatrix} (1/n) & -(1/n) \\ -(1/n) & (2/n) \end{pmatrix} = \frac{1}{n} \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \tag{eqA2}$$

and as a mixture of two distributions $\sigma_x'^2 = \sigma_x^2 + D$

From (eqA2), $var(b_x) = \frac{2\sigma_x'^2}{n}$, an estimate of which can be derived as $\hat{v}ar(\hat{b}_x) = \frac{2s_x'^2}{n}$, where $s_x'^2$ follows from the regression of $X|\mathbf{t}$ on N patients.

Subsequently, the variance of $(\hat{a}_0, \hat{b}_0, \hat{\gamma})$ can be derived as $\sigma_o^2(1 - \rho^2)(\mathbf{T}_3'\mathbf{T}_3)^{-1}$, where \mathbf{T}_3 is the design matrix of $Y|X, \mathbf{t}$ on N_3 patients.

$$E(\mathbf{T}_3'\mathbf{T}_3) = E \begin{pmatrix} 2n_3 & n_3 & \sum_{C,E} x_3 \\ n_3 & n_3 & \sum_E x_3 \\ \sum_{C,E} x_3 & \sum_E x_3 & \sum_{C,T} x_3^2 \end{pmatrix} = N_3 \begin{pmatrix} 2 & 1 & \mu_C + \mu_E \\ 1 & 1 & \mu_E \\ \mu_C + \mu_E & \mu_E & 2\sigma_{y_3}^2 + (\mu_C^2 + \mu_E^2) \end{pmatrix}. \tag{eqA3}$$

The variance estimates are derived by inverting matrix (eqA3) and replacing σ_o^2 with $\sigma_y'^2 = \sigma_o^2 + \gamma^2\sigma_x'^2$.⁸

$$var(\hat{\gamma}) = \frac{\sigma_y'^2(1 - \rho^2)}{2n_3\sigma_x'^2}, \tag{eqA4}$$

$$var(\hat{b}_0) = \frac{2\sigma_y'^2(1 - \rho^2)}{n_3} + \frac{\sigma_y'^2(1 - \rho^2)}{2n_3\sigma_x'^2} b_x^2, \tag{eqA5}$$

$$cov(\hat{b}_0, \hat{\gamma}) = -\frac{\sigma_y'^2(1 - \rho^2)}{2n_3\sigma_x'^2} b_x. \tag{eqA6}$$

Replacing (eqA4) to (eqA6) in (eqA1) we obtain $var(\hat{b}_y)$

$$Var(\hat{b}_y) = var(\hat{b}_0) + \hat{\gamma}^2 var(\hat{b}_x) + \hat{b}_x^2 var(\hat{\gamma}) + 2\hat{b}_x cov(\hat{b}_0, \hat{\gamma})$$

$$\begin{aligned}
& \times \frac{2\sigma_y'^2(1-\rho^2)}{n_3} + \frac{\sigma_y'^2(1-\rho^2)}{2n_3\sigma_x'^2} b_x^2 + \frac{2\sigma_y'^2\rho^2}{n} \\
& + b_x^2 \frac{\sigma_y'^2(1-\rho^2)}{2n_3\sigma_x'^2} + 2\hat{b}_x \left(-\frac{\sigma_y'^2(1-\rho^2)}{2n_3\sigma_x'^2} b_x \right) \\
& = 2\sigma_y'^2 \left(\frac{(1-\rho^2)}{n_3} + \frac{\rho^2}{n} \right) \\
& \sigma_y'^2 = \sigma_y^2 + \gamma^2 D.
\end{aligned}$$

B.3 Derivation of MSE(\hat{b}_y)

Based on the calculated alternative variance of the overall late-phase effect $\text{Var}(\hat{b}_y)$ and the method of moments, the MSE(\hat{b}_y) is given by

$$\begin{aligned}
\text{MSE}(\hat{b}_y) &= \text{Bias}(\hat{b}_y)^2 + \text{Var}(\hat{b}_y) \\
&= \left(\frac{w_2\lambda}{\sigma_x'\sqrt{n_2/2}} \frac{\rho\sigma_y'\sigma_{x2}}{\sigma_x'} \right)^2 + 2\sigma_y'^2 \left(\frac{1-\rho^2}{n_3} + \frac{\rho^2}{n} \right) \\
&= 2\sigma_y'^2 \left(\frac{w_2\rho\lambda\sigma_{x2}}{\sigma_x'\sqrt{n_2}} \right)^2 + 2\sigma_y'^2 \left(\frac{1-\rho^2}{n_3} + \frac{\rho^2}{n} \right) \\
& \sigma_y'^2 = \sigma_y^2 + \gamma^2 D.
\end{aligned}$$

In Appendix D, Figure D2 presents a short simulation demonstrating the association between the approximate analytical bias and the bias introduced by the use of the DR method. Equivalent simulations were performed for the updated variance parameters, all scenarios resulted in less than 10% difference between the approximate and analytical derived variances.

APPENDIX C. BIVARIATE NORMAL SIMULATION

Regarding the bivariate normal simulation strategy, we generated a series of parallel-group design randomized trials with two treatment groups (control (C) and treatment (E)). We assume that the outcome values for i th control individual and k th trial, for the early-phase m_{Cxik} and late-phase m_{Cyik} outcome are generated by a bivariate normal distribution as follows

$$\begin{pmatrix} m_{Cxik} \\ m_{Cyik} \end{pmatrix} \sim \text{BVN} \left[\begin{pmatrix} \mu_{Cxk} \\ \mu_{Cyk} \end{pmatrix}, \Sigma_C = \begin{pmatrix} \sigma_{Cxk}^2 & \rho_C \sigma_{Cxk} \sigma_{Cyk} \\ \rho_C \sigma_{Cxk} \sigma_{Cyk} & \sigma_{Cyk}^2 \end{pmatrix} \right],$$

where μ_{Cik} are the true treatment means for each endpoint in the control arm and Σ_C is their covariance matrix, σ_{Ck}^2 are the variances of the early and late-phase endpoints and ρ_C is the correlation between these endpoints.

In a similar fashion we generate data outcome values for the i th treatment individual in the k th trial as follows

$$\begin{pmatrix} m_{Exik} \\ m_{Eyik} \end{pmatrix} \sim \text{BVN} \left[\begin{pmatrix} \mu_{Ezk} = \mu_{Cxk} + \theta_x \\ \mu_{Eyik} = \mu_{Cyk} + \theta_y \end{pmatrix}, \Sigma_T = \begin{pmatrix} \sigma_{Exk}^2 & \rho \sigma_{Exk} \sigma_{Eyik} \\ \rho \sigma_{Exk} \sigma_{Eyik} & \sigma_{Eyik}^2 \end{pmatrix} \right].$$

In order to incorporate between-study variability τ^2 , we can further assume that $m_i \sim N(\Delta, \tau^2)$.²⁷

$$\begin{pmatrix} m_{Exik} \\ m_{Eyik} \end{pmatrix} \sim \text{BVN} \left[\begin{pmatrix} \mu_{Exk} = \mu_{Cxk} + \Delta_x \\ \mu_{Eyik} = \mu_{Cyk} + \Delta_y \end{pmatrix}, \Sigma_T = \begin{pmatrix} \sigma_{Exk}^2 + \tau_x^2 & \rho \sigma_{Exk} \sigma_{Eyik} + \tau_x \tau_y \rho_B \\ \rho \sigma_{Exk} \sigma_{Eyik} + \tau_x \tau_y \rho_B & \sigma_{Eyik}^2 + \tau_y^2 \end{pmatrix} \right].$$

ρ_B parameter indicates how the early-phase and late-phase outcome are related across all available studies. In our framework we only have available summary value on both early and late-phase outcomes from only a single Phase III trial. Therefore, for simplicity in the simulation study we assume that the between-study correlation equals to zero $\rho_B = 0$.

We applied an alternative model that generates data in two stages to check for results' robustness with no observed noticeable variations in relative performances.

APPENDIX D. FIGURES AND TABLES

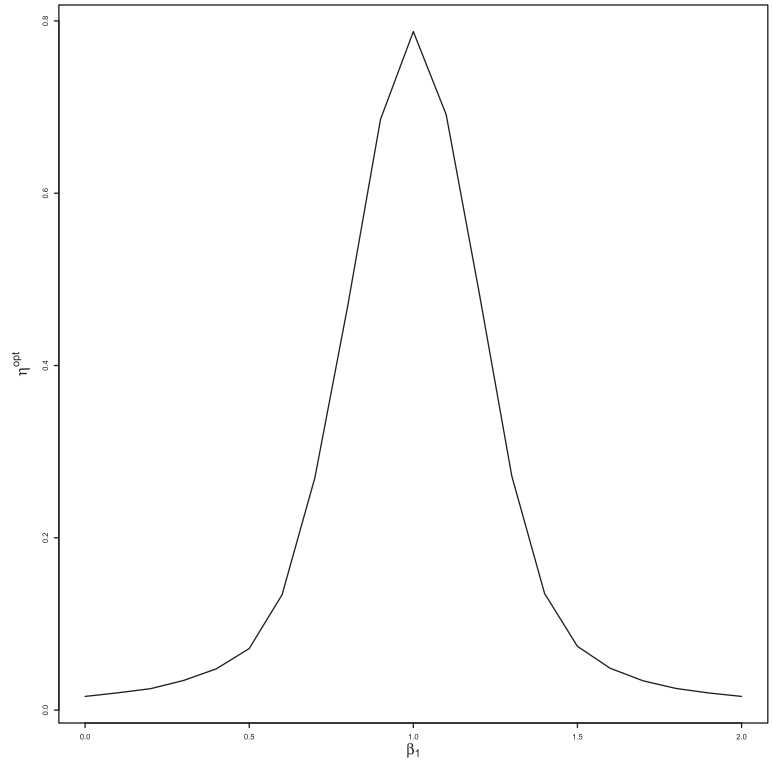


FIGURE D1 Relation between η^{opt} and varying true b_{x2} when $b_{x3} = 1$ and $\sigma_x = \sigma_y = 1$

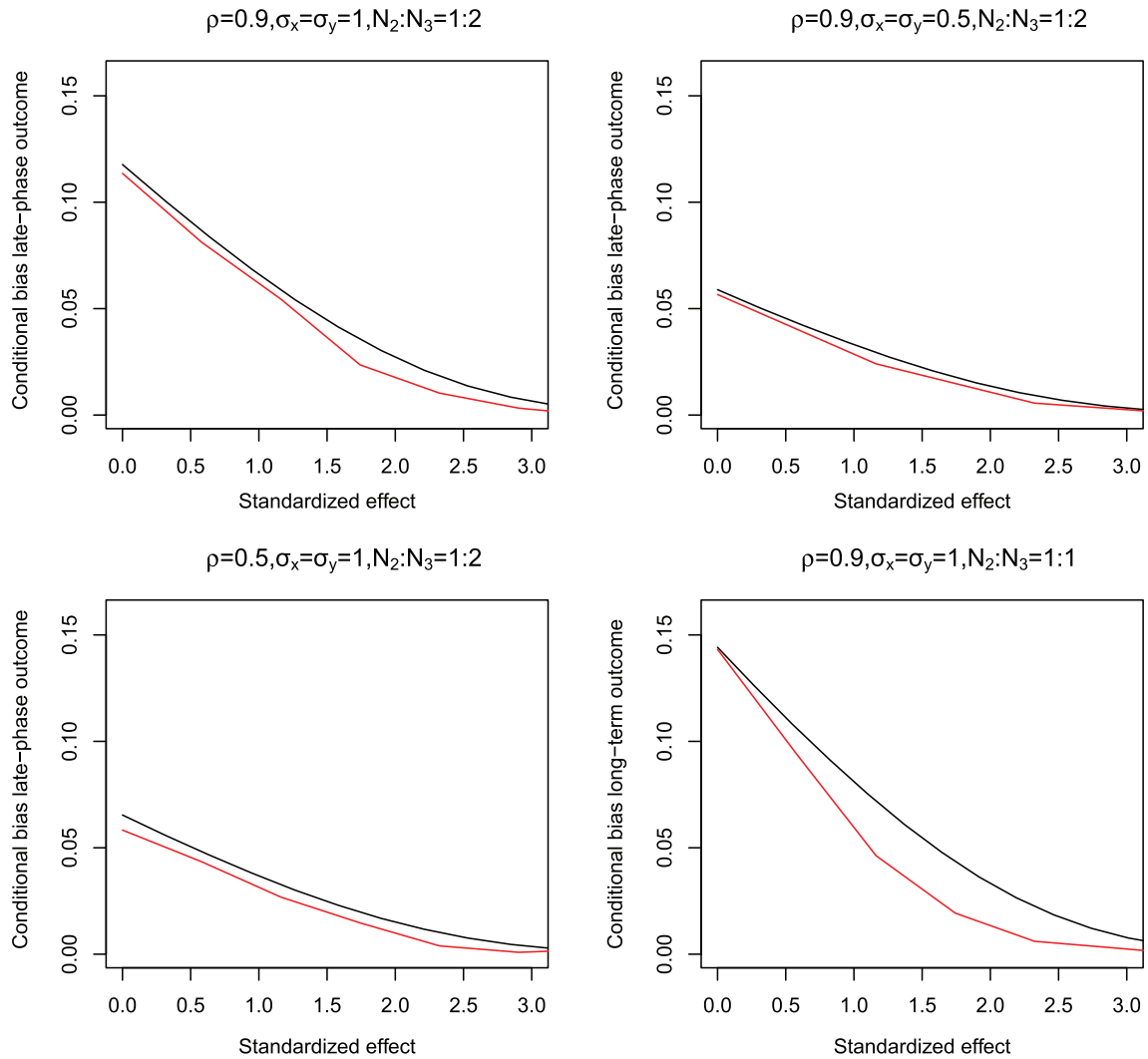


FIGURE D2 Approximation of a truncated normal distribution with a normal distribution ($\alpha_2 = 0.1$). The black lines represent the approximate analytical solution to the bias, while the red lines represent the simulated values based on the *double-regression* model [Colour figure can be viewed at wileyonlinelibrary.com]