Received: 7 January 2021 | Revised: 4 June 2021 | Accepted: 15 July 2021

DOI: 10.1002/pst.2165

MAIN PAPER

WILEY

Using an interim analysis based exclusively on an early outcome in a randomized clinical trial with a long-term clinical endpoint

Leandro Garcia Barrado[1,2] | **Tomasz Burzykowski**[1,3] | **Catherine Legrand**[2] | **Marc Buyse**[1,3]

[1]International Drug Development Institute (IDDI), Louvain-la-Neuve, Belgium

[2]Institute of Statistics, Biostatistics, and Actuarial Sciences (ISBA), Louvain Institute for Data Analysis and Modeling, Louvain-la-Neuve, Belgium

[3]Data Science Institute, I-BioStat, Hasselt University, Hasselt, Belgium

Correspondence
Leandro Garcia Barrado, International Drug Development Institute (IDDI), Avenue Provinciale 30, 1340 Louvain-la-Neuve, Belgium.
Email: leandro.garciabarrado@iddi.be

Funding information
Service Public de Wallonie, Grant/Award Number: First Entreprice Docteur nr. 8162, jointly awarded to International Drug Development Institute (IDDI) SA and the Université Catholique de Louvain (UCLouvain)

Abstract

In RCTs with an interest in a long-term efficacy endpoint, the follow-up time necessary to observe the endpoint may be substantial. In order to reduce the expected duration of such trials, early-outcome data may be collected to enrich an interim analysis aimed at stopping the trial early for efficacy. We propose to extend such a design with an additional interim analysis using solely early-outcome data in order to expedite the evaluation of treatment's efficacy. We evaluate the potential gain in operating characteristics (power, expected trial duration, and expected sample size) when introducing such an early interim analysis, in function of the properties of the early outcome as a surrogate for the long-term endpoint. In the context of a longitudinal age-related macular degeneration (ARMD) ophthalmology trial, results show potentially substantial gains in both the expected trial duration and the expected sample size. A prerequisite, though, is that the treatment effect on the early outcome has to be strongly correlated with the treatment effect on the long-term endpoint, that is, that the early outcome is a validated surrogate for the long-term endpoint.

KEYWORDS

efficacy stopping, group-sequential design, interim analysis, randomized clinical trial, surrogate endpoint

1 | INTRODUCTION

In a RCT that uses a long-term efficacy endpoint, the follow-up time necessary to observe the endpoint may be substantial. This may result in long expected duration of the trial, with an increased risk of missing data issues, increasing costs, and deceleration of medical progress as consequences.[1] In addition, if the follow-up time exceeds the time to complete accrual of patients, it may not be possible to conduct any interim analysis of the treatment effect on the endpoint during the accrual period. As a result, it may be impossible to stop accrual early, even in case of a larger than expected beneficial treatment effect.

In such trials, a seemingly attractive option is to consider an early outcome that could be used to expedite the interim evaluation of treatment's efficacy. The idea of using early outcomes when evaluating the treatment effect on a long-term endpoint is not new. For instance, Marschner and Becker[2] investigated the case of binary endpoints.

Following their ideas, Galbraith and Marschner[3] considered the case of trials with longitudinally-measured continuous outcomes and extended the classical group-sequential design by combining the available information about early measurements with the available observations of the final (long-term) measurement in the interim analysis. In particular, they proposed to base the interim analysis on the predicted value of the treatment effect on the final (long-term) measurement of the endpoint. The prediction was obtained from the available information about the early and final (long-term) measurements of the endpoint. Galbraith and Marschner[3] showed that efficiency gains could be obtained even when only a few patients had final (long-term) measurements available at the time of the interim analysis. The gain was substantial, especially when the early and final (long-term) measurements were highly correlated. Moreover, Galbraith and Marschner[3] demonstrated that, in case of a longitudinally-measured continuous endpoint, most of the gain could be achieved already by considering a single early measurement.

Note, however, that in the approaches of Marschner and Becker[2] and Galbraith and Marschner,[3] some observations of the long-term endpoint have to be available at the time of interim analysis. Thus, those approaches still do not allow conducting the interim analysis earlier than the follow-up time necessary to observe the endpoint. To allow for that, one would have to conduct the analysis based solely on the early outcome. However, as pointed out by Marschner and Becker,[2] this requires assumptions regarding the association between treatment effects on the early outcome(s) and on the long-term endpoint.

Similar considerations can be encountered in the context of multi-arm multi-stage (MAMS) designs[4] and adaptive seamless phase II/III designs.[5] In those designs, an interim-analysis is introduced to stop inefficacious treatments at the end of the phase II part and selecting the most promising treatment(s) to proceed to the phase III part. To decrease development costs and shorten the time-to-market even further, the futility interim-analysis may be based only on early-outcome data.[6,7] Friede et al[6] concluded that operating characteristics of such a design depend, among others, on the correlation between the early outcome and the long-term endpoint. In addition, adhering to a selection rule that only promotes the single best treatment to the phase III part of the study is discouraged in the absence of a validated association between the effect of the treatment on the early outcome and the long-term endpoint.

In the framework proposed by Buyse et al,[8] the association is used to evaluate the validity of the early outcome as trial-level surrogate for the long-term endpoint. Buyse et al[8] defined a validated surrogate endpoint as an outcome that is causally linked to the final endpoint (individual-level surrogacy) and for which treatment effect is strongly correlated with treatment effect on the final endpoint (trial-level surrogacy). Typically, individual-level surrogacy is evaluated by using a coefficient capturing the strength of association between the individual-patient values of the surrogate and final endpoints. Trial-level surrogacy is assessed by using the correlation between the treatment effects on the surrogate and final endpoints. Based on a validated surrogate, treatment effect on the final endpoint can be reliably predicted from the treatment effect on the surrogate.

Li and Taylor[9] focussed on the use of an early outcome as an auxiliary variable in helping to predict the treatment effect on the long-term endpoint. They showed that even when only a small fraction of observations on the long-term endpoint is available, a large individual-level correlation leads to substantial gains in the efficiency of prediction. This observation can be linked to the conclusions drawn by Galbraith and Marschner.[3] However, Li and Taylor[9] showed that, in the absence of information on the long-term endpoint, the trial-level association drives the potential gain in prediction efficiency.

In the current manuscript, we investigate the question about using an interim analysis based only on an early outcome to expedite the evaluation of treatment's efficacy. We focus on the case of a continuous early outcome and a continuous long-term endpoint. In particular, we investigate an extension of the approach considered by Galbraith and Marschner[3] by adding an interim analysis before any observation of the long-term endpoint is collected. This additional interim analysis is based solely on the early outcome. We evaluate the potential gain in operating characteristics (power, expected trial duration, and expected sample size) when introducing this early interim analysis, in function of the properties of the early outcome as a surrogate for the long-term endpoint. In general, the considered methodology applies to any continuous early outcome and long-term endpoint. However, since a longitudinally-measured endpoint is a natural candidate in this respect, we investigate the extension in the context of a longitudinal age-related macular degeneration (ARMD) trial.

2 | METHODS

Denote by S_j and T_j the measurement of continuous early outcome S and continuous long-term endpoint T, respectively, for patient $j = 1,...,N$. Define a binary indicator Z_j to indicate the treatment randomly assigned to patient j, with $Z_j = 0$ and $Z_j = 1$ for the control and experimental treatment, respectively.

We assume that

$$\begin{pmatrix} S_j \\ T_j \end{pmatrix} \mid Z_j \sim N\left(\begin{pmatrix} \mu_{S,Z_j} \\ \mu_{T,Z_j} \end{pmatrix}, \Sigma = \begin{bmatrix} \sigma_S^2 & \rho_{ST}\sigma_S\sigma_T \\ \rho_{ST}\sigma_S\sigma_T & \sigma_T^2 \end{bmatrix}\right),$$

where $\mu_{S,Z}$ and $\mu_{T,Z}$ denote the mean values of S_j and T_j, respectively, for treatment group Z, and σ_S^2 and σ_T^2 indicate the variances of S_j and T_j, respectively. Finally, ρ_{ST} expresses the correlation between S and T. The variances σ_S^2 and σ_T^2 and correlation ρ_{ST} are assumed to be the same for both treatment groups.

We consider a trial aimed at verifying the null hypothesis

$$H_0: \beta = \mu_{T,1} - \mu_{T,0} \leq 0$$

against the alternative

$$H_1: \beta > 0.$$

As a starting point, we consider a group-sequential design with one interim analysis and the final analysis organized according to the Galbraith and Marschner approach.[3]

The final analysis is conducted when information about the long-term endpoint of interest T has been collected for all $N = 2n$ patients, where n is the sample size per group (assuming an equal randomization ratio). The test statistic is based on the difference of sample means of T between the two treatment groups. We denote this difference by $\hat{\beta}_F$. It follows that $\hat{\beta}_F \sim N\left(\beta, 2\frac{\sigma_T^2}{n}\right)$.

At the interim analysis, the predicted mean difference on T is used as the test statistic, with the prediction based on the available observations of the early outcome S and of the long-term endpoint T. We denote this predicted treatment effect as $\hat{\beta}_{ST}$. Let $n_{ST,S}$ and $n_{ST,T}$ denote the number of available observations (per treatment group) on S and T, respectively. Note that $n_{ST,S} \leq n$ and $n_{ST,T} \leq n$. Moreover, as we do not assume any dropout, $n_{ST,T} \leq n_{ST,S}$, and it follows that $\hat{\beta}_{ST}$ has the following sampling distribution:

$$\hat{\beta}_{ST} \sim N\left[\beta, 2\frac{\sigma_T^2}{n_{ST,T}}\left\{1 - \rho_{ST}^2\left(1 - \frac{n_{ST,T}}{n_{ST,S}}\right)\right\}\right].$$

More information about the definition and sampling distribution of this estimator can be found in Appendix A of the Supplementary Materials.

Next, we consider modification of the afore-described group-sequential design by including an earlier interim analysis conducted when only observations on S are available ("Interim 1" in Figure 1). The number of the available observations (per treatment group) at the interim analysis will be denoted by n_S; note that $n_S \leq n_{ST,S} \leq n$. In that case, prediction of the mean difference of T solely from the observed data becomes impossible without making untestable (in the absence of observations on T) assumptions about the joint distribution of S and T.

To overcome the issue, we assume that information about the validity of S as a surrogate for T is available. In particular, we assume that surrogacy of S was evaluated by using the meta-analytic framework proposed by Buyse et al.[8] Within this framework, treatment effects on S and T, observed in a clinical trial and denoted by α and β, respectively, are regarded as random variables. Moreover, the effects are assumed to follow a bivariate normal distribution, given by

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \sim N\left(\begin{pmatrix} \overline{a} \\ \overline{b} \end{pmatrix}, \mathrm{D} = \begin{bmatrix} d_S^2 & Rd_Sd_T \\ Rd_Sd_T & d_T^2 \end{bmatrix}\right), \tag{1}$$

where $\overline{a}$ and $\overline{b}$ are the expected treatment effect values, d_S^2 and d_T^2 are the variances, and R is the correlation coefficient. In what follows, we assume that the values of those parameters are known. In practical situations, these parameters are often estimated using a meta-analysis of previously conducted randomized trials with observations of both S and T.

Knowledge of distribution (1) allows predicting one treatment effect from the other. For instance, according to distribution (1), the conditional distribution of α, given $\beta = \beta_1$, is as follows:

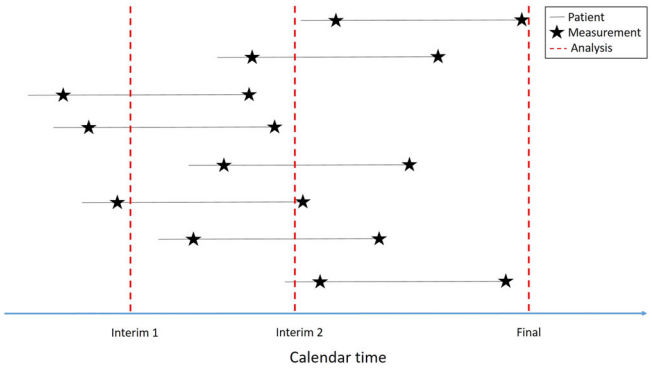

FIGURE 1 Structure of the proposed design, with an additional interim analysis based on only early-outcome data (Interim 1)

$$\alpha \mid \beta = \beta_1 \sim N\left(\overline{a} - R\left(\frac{d_S}{d_T}\right)\overline{b} + R\left(\frac{d_S}{d_T}\right)\beta_1, \left(1-R^2\right)d_S^2\right). \tag{2}$$

Suppose that at the first interim analysis, with only observations on S available, $\hat{\alpha}$ is the sample mean difference obtained based on n_S observations per treatment group. By considering distributions (1) and (2), the sampling distribution of the prediction $\widetilde{\beta}_S$ of the treatment effect on T, based on $\hat{\alpha}$, can be derived (see Appendix B of the Supplementary Materials) as follows:

$$\widetilde{\beta}_S \sim N\left(\beta_1, \left(\frac{1}{R}\frac{d_T}{d_S}\right)^2\left(\left(1-R^2\right)d_S^2 + 2\frac{\sigma_S^2}{n_S}\right)\right). \tag{3}$$

In distribution (3), the term $\left(\frac{1}{R}\frac{d_T}{d_S}\right)^2$ is related to the prediction variability introduced by predicting β from $\hat{\alpha}$, $(1-R^2)d_S^2$ is the residual variability introduced by the relationship between α and β in distribution (2), and $2\frac{\sigma_S^2}{n_S}$ is the estimation error in $\hat{\alpha}$.

The second interim analysis, which uses available information on S and T, is conducted based on the predicted treatment effect $\hat{\beta}_{ST}$, as in the Galbraith and Marschner[3] design. The final analysis is conducted using the difference of sample means $\hat{\beta}_F$.

To calculate the operating characteristics of the two group-sequential designs, we need the joint distribution of tests statistics $\widetilde{\beta}_S$, $\hat{\beta}_{ST}$, and $\hat{\beta}_F$. It is given by the following tri-variate normal distribution:

$$\begin{pmatrix}\widetilde{\beta}_S\\ \hat{\beta}_{ST}\\ \hat{\beta}_F\end{pmatrix} \sim N\left(\begin{pmatrix}\beta_1\\ \beta_1\\ \beta_1\end{pmatrix}, \begin{bmatrix}\left(\frac{1}{R}\frac{d_T}{d_S}\right)^2\left(\left(1-R^2\right)d_S^2 + 2\frac{\sigma_S^2}{n_S}\right) & 2\frac{1}{R}\frac{d_T}{d_S}\frac{\sigma_S\sigma_T\rho_{ST}}{n_{ST,S}} & 2\frac{1}{R}\frac{d_T}{d_S}\frac{\sigma_S\sigma_T\rho_{ST}}{n}\\ 2\frac{1}{R}\frac{d_T}{d_S}\frac{\sigma_S\sigma_T\rho_{ST}}{n_{ST,S}} & 2\left(\frac{\sigma_T^2}{n_{ST,T}}\left\{1-\rho_{ST}^2\left(1-\frac{n_{ST,T}}{n_{ST,S}}\right)\right\}\right) & 2\frac{\sigma_T^2}{n}\\ 2\frac{1}{R}\frac{d_T}{d_S}\frac{\sigma_S\sigma_T\rho_{ST}}{n} & 2\frac{\sigma_T^2}{n} & 2\frac{\sigma_T^2}{n}\end{bmatrix}\right). \tag{4}$$

By using distribution (4), it is possible to define a vector of critical values $(C_{I,S}, C_{I,ST}, C_F)^T$ for the three test statistics such that the overall type-I error probability of the trial is at most α_{Trial}. In particular, toward this aim, we select significance levels α_S, α_{ST}, and α_F such that $\alpha_S + \alpha_{ST} + \alpha_F = \alpha_{Trial}$. The choice of values of α_S, α_{ST}, and α_F can be made by using, for example, any suitable alpha-spending paradigm. Then, we define critical value $C_{I,S}$ for the test at the first interim analysis such that $P\left(\frac{\widetilde{\beta}_S}{SE\left(\widetilde{\beta}_S\right)} > C_{I,S} \mid H_0\right) = \alpha_S$. Subsequently, we select critical value $C_{I,ST}$ at the second interim analysis such that $P\left(\frac{\widetilde{\beta}_S}{SE\left(\widetilde{\beta}_S\right)} \le C_{I,S}, \frac{\hat{\beta}_{ST}}{SE\left(\hat{\beta}_{ST}\right)} > C_{I,ST} \mid H_0\right) = \alpha_{ST}$. Finally, we choose C_F for the final analysis to ensure that $P\left(\frac{\widetilde{\beta}_S}{SE\left(\widetilde{\beta}_S\right)} \le C_{I,S}, \frac{\hat{\beta}_{ST}}{SE\left(\hat{\beta}_{ST}\right)} \le C_{I,ST}, \frac{\hat{\beta}_F}{SE\left(\hat{\beta}_F\right)} > C_F \mid H_0\right) = \alpha_F$.

After defining the critical values, distribution (4) can be used to compute the sample size necessary to reach a particular power under alternative hypothesis H_1 with β equal to a particular targeted treatment effect, say β_1.

2.1 | Evaluation of operating characteristics

We investigate operating characteristics of the proposed design by considering an ARMD trial similar to the one described by the CATT Research Group.[10] ARMD is a condition in which patients progressively lose vision. The generally accepted endpoint in ARMD trials is visual acuity (VA), measured by the number of letters read by a patient on the eye chart. In particular, the change in VA (cVA) at 1 year, as compared to baseline, is usually of primary interest. This implies that any analysis aimed at evaluation of treatment's efficacy can only take place more than 1 year after the start of the enrolment to the trial. However, in ARMD trials, VA is usually measured longitudinally, with measurements available at baseline and then every, for instance, 4 weeks. By using earlier measurements, one could shorten the time to the first evaluation of treatment effect on cVA.

The long-term endpoint of interest, T, is the change in VA (cVA) from baseline at 52 weeks. We consider using the early outcome at 4 weeks, S, to allow for an earlier assessment of treatment's efficacy. The variance–covariance structure of the two measurements is defined by using the results reported by Burzykowski and Buyse.[11] In particular, they found that cVA measurements may follow a multivariate normal distribution with an ante-dependence variance–covariance structure. Based on those results, we assume the variance of $\sigma_S^2 = \sigma^2 \times 4^\gamma = 5.73^2 \times 4^{0.28} = 48.4$ for cVA at 4 weeks and $\sigma_T^2 = 5.73^2 \times 52^{0.28} = 99.26$ for cVA at 52 weeks. For the correlation between the two measurements, we assume values of $\rho_{ST} \in \{0.1, 0.31, 0.9\}$, with $\rho_{4,52} = \rho_{4,12} \times \rho_{12,24} \times \rho_{24,52} = 0.60 \times 0.66 \times 0.78 = 0.31$ resulting from the ante-dependence structure.[11]

Regarding model (1), no real-life results for ARMD are available. Thus, we may consider d_S^2 and d_T^2 to be equal to 25% and 10% of the individual-level variances σ_S^2 and σ_T^2. Moreover, we assume $R \in \{0.5, 0.85, 0.95\}$. Note that we assume that we know the values of those parameters without any (estimation) error.

For both ρ_{ST} and R, wide ranges of potential values are considered. The reasoning for the particular choice of range is twofold. First, in practice, for example, in oncology, a wide range of values for both ρ_{ST} and R may be encountered.[12] Second, the wide range of considered values allows to investigate the importance of the individual-level correlation as well as to evaluate the minimum required trial-level correlation for the particular setting at hand.

Finally, we focus on the alternative hypothesis H_1 with $\beta_1 = 3$. This case translates to a 50% increase in cVA from baseline, comparable to the findings of the CATT Research Group.[10] A fixed accrual rate of six patients per week per treatment arm was specified, comparable to the one reported by the CATT Research Group.[10]

Under the specified assumptions, a fixed-sample-size trial would require about 137 patients per treatment arm to reach 80% power at the 5% significance level. This would mean a total trial duration of about 75 weeks, including about 23 weeks of accrual and 52 weeks of follow-up.

3 | RESULTS

To find the minimum expected trial duration and sample size as a function of interim analysis timing, consider maximum sample size per treatment arm $n \in \{135, 136, ..., 175\}$. Considering $n < 137$ allows to investigate whether a gain in terms of maximum sample size is possible as well. In order to compare any potential gains of the proposed extended design, the operating characteristics of the design introduced by Galbraith and Marschner[3] are considered as a reference.

For the Galbraith and Marschner design,[3] select $\alpha_{Trial} = \alpha_{ST} + \alpha_F = 0.025 + 0.025 = 0.05$. With respect to the expected trial duration, when accruing six patients per week per treatment arm and $\rho_{ST} = 0.1$, the minimum expected trial duration for a trial with a power of at least 80% is equal to 72 weeks. With a corresponding maximum sample size of 153 patients per treatment arm and the interim analysis taking place at week 61 (with 35% of final measurement available), the trial has a 35% probability of stopping early for efficacy after 61 weeks. For $\rho_{ST} = 0.31$, the minimum expected trial duration for a trial with at least 80% power is equal to 71 weeks. With a maximum sample size of 149 patients per treatment arm and the interim analysis planned at week 63 (with 44% of final measurements available), this trial has a 43% probability of stopping early for efficacy after 63 weeks. Finally, for $\rho_{ST} = 0.9$, the minimum expected trial duration is equal to 67 weeks. With $n = 146$ and the interim analysis planned at week 57 (with 21% of

final measurements available), the trial has 50% probability of stopping early for efficacy after 57 weeks (for additional details see also Appendix C of the Supplementary Materials).

Given an accrual rate of six patients per week per treatment arm, the total accrual time for the smallest considered maximum sample size per treatment arm ($n = 135$) is equal to about 23 weeks. It is substantially shorter than the time necessary for observing the final cVA measurement of the first accrued patient, that is, 52 weeks. As the interim analysis in the Galbraith and Marschner design[3] has to be planned after having obtained final measurements for at least some patients, the expected sample size will be equal to the considered maximum sample size. A gain in the expected sample can thus only be established when the design can reach a power of at least 80% for a maximum sample size per treatment arm $n < 137$. Moreover, even in the case of $\rho = 0.9$, the required maximum sample size of the Galbraith and Marschner design[3] to reach at least 80% power, irrespectively of expected trial duration, is $n = 137$. So, in the considered setting, the Galbraith and Marschner design[3] does not lead to any improvement over the fixed-sample-size design in terms of sample size (see also Appendix C of the Supplementary Materials).

Consider now extending the design by adding an interim analysis based solely on available observations of the early (4 week) outcome. Select $\alpha_{Trial} = \alpha_S + \alpha_{ST} + \alpha_F = 0.016 + 0.016 + 0.018 = 0.05$ and plan the second interim analysis when 35% of the final measurements per treatment arm are available, that is, $n_{ST,T} = 0.35n$.

First, consider the setting where d_S^2 and d_T^2 are 25% of the individual-level variances σ_S^2 and σ_T^2, resulting in $d_S^2 = 12.10$ and $d_T^2 = 24.82$, respectively. The results of the extended design are summarized in Figures 2 and 3. In particular, Figure 2 presents the expected trial duration for different combinations of $\rho_{ST} \in \{0.1, 0.31, 0.9\}$ (rows) and $R \in \{0.5, 0.85, 0.95\}$ (columns). Points in each panel represent trials with a maximum sample size per treatment arm from the set $n \in \{135, 136, ..., 175\}$ and the first interim analysis taking place at a particular week (indicated on the horizontal axis) between week 5 and 52. Points for trials leading to a power of 80% or higher are marked in green. For those trials, the black cross indicates the trial with the minimum expected trial duration at the earliest interim timing. The red solid line presents the minimum expected trial duration for the Galbraith and Marscher design[3] equal to 72, 71, and 67 weeks for ρ_{ST} equal to 0.1, 0.31, and 0.9, respectively.

The plots in Figure 2 indicate a marked effect of the trial-level correlation R on the expected trial duration. In particular, the larger R, the shorter the expected trial duration. The plots in Figure 2 show that, as compared to $R = 0.85$, the minimum expected trial duration for $R = 0.95$ is about 6 weeks shorter for the same value of the individual-level correlation ρ_{ST}. For a particular value of R, increasing the value of the individual-level correlation ρ_{ST} also reduces the expected trial duration. For instance, as compared to $\rho_{ST} = 0.31$, the minimum expected duration for $\rho_{ST} = 0.9$ is shorter by about 4 weeks. Globally, the smallest value of the expected trial duration is obtained when the early measurement is a very good individual-level ($\rho_{ST} = 0.9$) and trial-level ($R = 0.95$) surrogate. In this case, the expected trial duration is equal to 59 weeks for a trial with a maximum sample size of 155 patients per treatment arm and with the first interim analysis planned at week 16. Moreover, for this trial, the probability of early stopping for efficacy is equal to 19% and 44% at 16 weeks (Interim 1) and 60 weeks (Interim 2), respectively. This implies a gain of 8 weeks as compared to the minimum expected trial duration of the Galbraith and Marschner design[3] and a gain of 16 weeks as compared to the fixed-sample-size design.

Figure 3 presents the expected sample size for different combinations of $\rho_{ST} \in \{0.1, 0.31, 0.9\}$ (rows) and $R \in \{0.5, 0.85, 0.95\}$ (columns). The meaning of the points in the panels is the same as in Figure 2. The black cross indicates the trial with the minimum expected sample size reaching a power of at least 80% at the earliest interim timing. The red solid line corresponds to the minimum expected sample size for the Galbraith and Marschner design,[3] that is, 137 patients per treatment arm.

The plots in Figure 3 show a marked effect of both the individual- and trial-level correlation ρ_{ST} and R, respectively, on the minimum expected sample size. It appears that, for the considered conditions, reaching an expected sample size smaller than 137 patients per arm is not possible. Approaching this sample size is only possible for $R = 0.95$, that is, if the early outcome is an excellent trial-level surrogate for the long-term endpoint. In that case, combined with a large individual-level correlation $\rho_{ST} = 0.9$, the minimum expected sample size is about 141 patients per treatment arm for a trial with a maximum sample size of 153 patients per treatment arm and the first interim analysis taking place at week 11. In this specific case, the trial has 14% probability of stopping early for efficacy after only 66 patients per treatment arm are enrolled in the trial.

Consider now that d_S^2 and d_T^2 are 10% of the individual-level variances σ_S^2 and σ_T^2, resulting in $d_S^2 = 4.84$ and $d_T^2 = 9.93$. In this case, for $R = 0.95$ and $\rho_{ST} = 0.9$, the gain in minimum expected trial duration would increase to 13 weeks as compared to the minimum expected trial duration of the Galbraith and Marschner design[3] and to 21 weeks as compared to the fixed-sample-size design. In particular, the gain is obtained for a trial with a maximum sample size of $n = 147$ and the first interim analysis at 21 weeks, for which the expected trial duration is equal to 54 weeks. For this

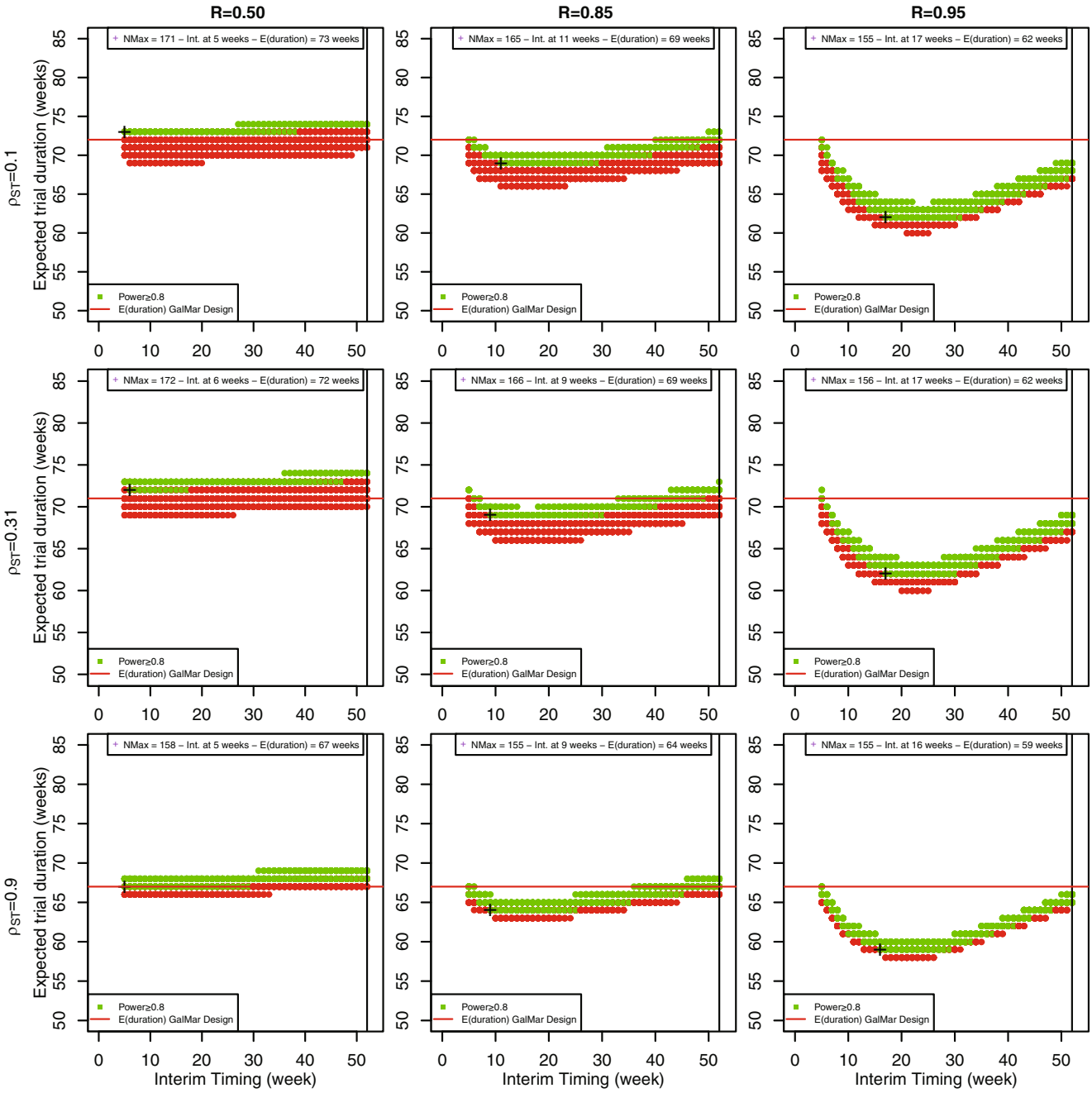

FIGURE 2 Expected trial duration for all combinations of $\rho_{ST} \in \{0.1,0.31,0.9\}$ (rows) and $R \in \{0.5,0.85,0.95\}$ (columns). Expected trial duration for different maximum sample sizes $n \in \{135,136,...,175\}$ is shown as a function of the timing (in weeks) of the first interim analysis, with the second interim analysis performed when 35% of the final measurements are available. Green points indicate trials with a power of at least 80% for $\beta_1 = 3$ and the type-I error probability of 0.05. Red solid horizontal line indicates the minimum expected trial duration for the Galbraight and Marschner design[3] for the particular value of ρ_{ST}. Black cross denotes the minimum expected trial duration at the earliest interim timing among trials with at least 80% power

trial the probability of stopping early for efficacy at week 21 (Interim 1) and week 59 (Interim 2) is equal to 32% and 31%, respectively (see also Appendix D of the Supplementary Materials).

In terms of the expected sample size, for $\rho = 0.9$ and $R = 0.95$, there is a gain of 1 patient per treatment arm as compared to both the Galbraith and Marschner[3] and the fixed-sample-size designs. In particular, the gain is obtained for a trial with a maximum sample size of $n = 151$ and the first interim analysis planned at week 13, for which the expected sample size is equal to 136 patients per treatment arm. For this trial, the probability of stopping early for efficacy at the first interim analysis, when only 78 patients per treatment have been accrued, is equal to 21% (see also Appendix D of the Supplementary Materials).

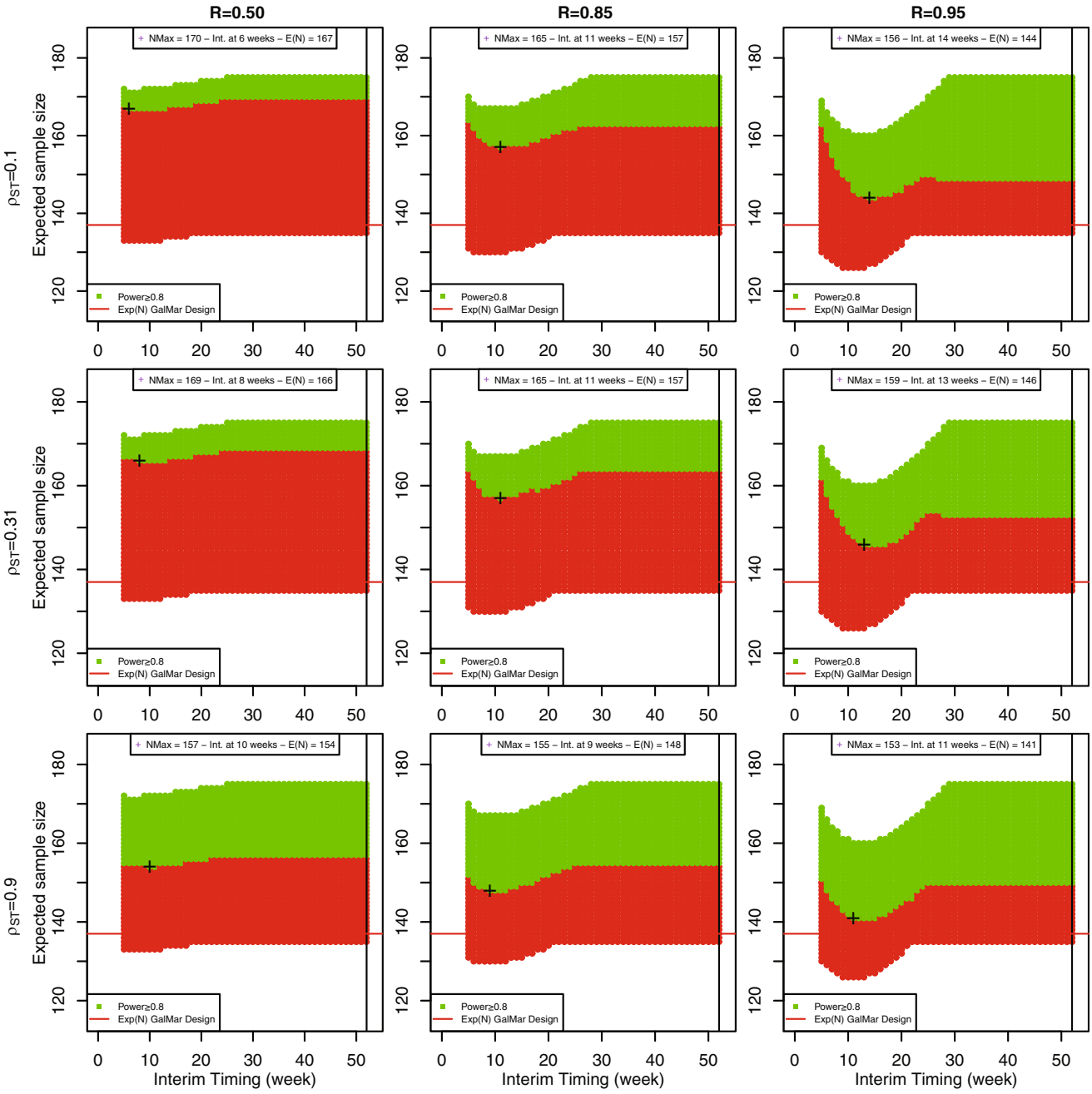

FIGURE 3 Expected sample size for all combinations of $\rho_{ST} \in \{0.1,0.31,0.9\}$ (rows) and $R \in \{0.5,0.85,0.95\}$ (columns). Expected sample size for different maximum sample sizes $n \in \{135,136,...,175\}$ is shown as a function of the timing (in weeks) of the first interim analysis, with the second interim analysis performed when 35% of the final measurements are available. Green points indicate trials with power of at least 80% for $\beta_1 = 3$ and the type-I error probability of 0.05. Red solid horizontal line indicates the minimum expected sample size for the Galbraith and Marschner design[3] (137 patients per treatment arm). Black cross denotes the minimum expected sample size at the earliest interim timing among trials with a power of at least 80%

4 | DISCUSSION

In the current manuscript, we have investigated the possibility of using an interim analysis based exclusively on an early outcome to expedite the evaluation of treatment's efficacy on a long-term continuous endpoint in a RCT. In general, the proposed design can be applied to any correlated continuous early outcome and continuous long-term endpoint. To fix ideas, we have considered the case of a trial with a longitudinally-measured continuous endpoint and

extended the approach considered by Galbraith and Marschner[3] by adding an interim analysis before any observations of the final (long-term) measurement have been collected.

Our results indicate that the use of an interim analysis based exclusively on an early outcome may lead to substantial gains, in terms of the expected trial duration and sample size, as compared to the situation in which interim analysis (analyses) have to include observations of the long-term endpoint. However, substantial gains are achievable only when the early outcome is a validated trial-level surrogate, that is, there exists a strong association between treatment effects on the early outcome and long-term endpoint.

Besides the trial-level correlation, other trial-specific factors may influence the potential gain of the proposed approach. For instance, the gain in the expected trial duration and sample size will increase with increasing probability of stopping for efficacy at the interim analysis based solely on the early-outcome data. The probability increases with decreasing values of d_T^2 and of the ratio d_T^2/d_S^2, because, in that case, the standard error of the test statistic used at the interim analysis decreases. The number of early-outcome observations available at the first interim analysis (n_S) affects the standard error as well. Note that n_S is a function of the difference in the follow-up time necessary to observe the early outcome and the long-term endpoint, and of the accrual rate. The longer the time difference, or the faster the accrual, the larger n_S and, thus, the larger potential gains, because the interim analysis based on early-outcome data may take place much earlier than any interim analysis that uses the long-term data. These additional factors entail that no specific trial-level correlation threshold values can be put forward since in particular settings different correlation values may suffice to render substantial gains.

The proposed design can easily be adapted to allow for stopping for futility. By considering the joint distribution of tests statistics (4), one can define an additional vector of critical values, controlling particular probabilities of stopping for futility. By subsequently specifying the critical values for efficacy stopping, conditional on stopping for futility, binding futility stopping can be defined. Results in Appendix E of the Supplementary Materials illustrate, using the considered ARMD-trial as an example, that an additional futility stopping can be introduced at the cost of only a small decrease in power.

One such situation that is important in practice is the use of an early outcome to stop the clinical development of an experimental treatment that shows no or little benefit on that outcome, especially when observation of the long-term clinical endpoint would take years to confirm early unpromising results. This approach is extensively used, albeit with unconfirmed surrogates, to screen out inactive treatments in oncology. For example, trials with patients with operable breast cancer select treatments for further development based on pathological complete response, an early outcome available at the time of surgery. Only treatments that show enough promise on the rate of pathological complete response qualify for inclusion in long-term clinical trials for confirmation of the treatment effects on event-free survival or overall survival, despite absence of evidence that pathological complete response is a valid surrogate for these long-term endpoints.[13] Friede et al[6] have studied this situation in the context of multiple sclerosis, when several experimental treatments become available for parallel screening. These authors proposed a testing procedure that preserves the type-I error. Further work is required to study the operating characteristics of their approach as a function of the strength of surrogacy of the early outcome for the long-term endpoint (trial-level R).

From a practical application point of view, planning a trial with an interim analysis based exclusively on the early-outcome data requires the results of a proper validation of the early outcome as a surrogate for the long-term endpoint, as described by Buyse et al.[8] In the context of longitudinally measured continuous outcomes, Pryseley et al[14] discuss methods that could be used to validate an earlier measurement as a surrogate for the final (long-term) measurement. Note that the results of the validation exercise have to point to a good trial-level surrogacy of the early outcome. This implies a limitation of the feasibility of the approach, because, in practice, the validation exercise would provide the estimates of the coefficients defining distribution (1). Adjusting for estimation error in those estimates would imply additional variability of the predicted treatment effect (3), which would, at least to some extent, reduce the potential gain in the expected trial duration and sample size.

The proposed design is evaluated in the setting where an interim analysis could be planned before any long-term endpoint data is collected. If, at the time of the interim analysis, any long-term endpoint information would be available, this should be considered in the analysis. Of course, in such a case, the individual-level correlation would be driving any potential gains in operating characteristics.[2,3,9]

It is also worth noting that validation of a surrogate endpoint by using the meta-analytic approach of Buyse et al[8] is not straightforward due to requirements related to availability and clinical relevance of historical randomized clinical-trial data. In particular, data are needed for patients from a population corresponding to the one considered for the planned trial. Also, the historical randomized trials should include interventions comparable (e.g., in terms of the mode

of action) to the one that will be investigated in the planned trial. Availability of such historical data for a particular case is not guaranteed. Without such data, implementation of the proposed design will not be feasible.

From a clinical point of view, an important consequence of stopping the trial at the interim analysis based exclusively on the early (surrogate) outcome may be the fact that clinical information could be lost due to not observing any data for the long-term endpoint. However, if such information is desired, one could consider continuing follow up of the patients that have been included in the trial until the interim analysis has taken place. It is worth noting that post-approval evaluation of clinical benefit is required, for instance, by FDA in case of accelerated approval of drugs.[15]

The current manuscript focusses on the continuous early-outcome and continuous long-term endpoint case. In that case, the derivation of the joint distribution of tests statistics (4) is relatively straightforward. For combinations of early-outcome and long-term endpoint of other types (binary, time-to-event, etc.), the derivation is more complex. In principle, however, it can be obtained by combining the results developed for group-sequential designs[16] and phase II/III designs[6] for standardized score-statistics.[17] The derivation is the focus of current research, the results of which will be reported elsewhere. In general, however, one can expect that the overall conclusions regarding the importance of trial-level correlation will also hold for pairs of early-outcome and long-term endpoint of types other than the continuous one.

In conclusion, we have shown that including an interim analysis based exclusively on an early outcome may lead to substantial gains in operational characteristics of a trial with a long-term endpoint of interest. However, a requirement is that the early outcome has to be a validated trial-level surrogate for the long-term endpoint. Further research is needed to exactly quantify the potential gains if, as it will be necessary in practice, estimates of the coefficients of the surrogate-validation model are used to implement the proposed design.

[Correction added on 24 December 2024, after first online publication: Last name of the author cited in the article has been corrected from Galbraight to Galbraith.]

ACKNOWLEDGMENTS

The research has been partly funded by the Service Public de Wallonie, under the research grant First Entreprice Docteur nr. 8162, jointly awarded to International Drug Development Institute (IDDI) SA and the Université Catholique de Louvain (UCLouvain).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

REFERENCES

1. Eisenstein EL, Collins R, Cracknell L, et al. Sensible approaches to reduce clinical trial costs. *Clin Trials*. 2008;5:75-84.
2. Marschner IC, Becker SL. Interim monitoring of clinical trials based on long-term binary endpoints. *Stat Med*. 2001;20:177-192.
3. Galbraith S, Marschner IC. Interim analysis of continuous long-term endpoints in clinical trials with longitudinal outcomes. *Stat Med*. 2003;22:1787-1805.
4. Royston P, Parmar MK, Qian W. Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Stat Med*. 2003;22:2239-2256.
5. Jenkins M, Stone A, Jennison C. An adaptive phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharm Stat*. 2011;10:347-356.
6. Friede T, Parsons N, Stallard N, et al. Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: an application in multiple sclerosis. *Stat Med*. 2011;30:1528-1540.
7. Chen X, Hartford A, Zhao J. A model-based approach for simulating adaptive clinical studies with surrogate endpoints used for interim decision-making. *Contemp Clin Trials Commun*. 2020;18:100562.
8. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*. 2000;1:49-67.
9. Li Y, Taylor JMG. Predicting treatment effects using biomarker data in a meta-analysis of clinical trials. *Statistics in Medicine*. 2010;29 (18):1875-1889. https://doi.org/10.1002/sim.3931
10. CATT Research Group, Martin DF, Maguire MG, et al. Ranibizumab and bevacizumab for neovascular age-related macular degeneration. *N Engl J Med*. 2011;364(20):1897-1908.
11. Burzykowski T, Buyse M. The correlation structure of longitudinal measurements of vision in patients with macular degeneration. *Pharm Stat*. 2011;10:115-121.

12. Branchoux S, Bellera C, Italiano A, Rustand D, Gaudin AF, Rondeau V. Immune-checkpoint inhibitors and candidate surrogate endpoints for overall survival across tumour types: a systematic literature review. *Crit Rev Oncol Hematol*. 2019;137:35-42.
13. Burzykowski T, Saad ED, Buyse M. Adoption of pathologic complete response as a surrogate end point in neoadjuvant trials in HER2-positive breast cancer still an open question. *JAMA Oncol*. 2017;3:416.
14. Pryseley A, Tilahun A, Alonso A, Molenberghs G. Using earlier measures in a longitudinal sequence as a potential surrogate for a later one. *Computational Statistics & Data Analysis*. 2010;54(5):1342-1354. https://doi.org/10.1016/j.csda.2009.11.024
15. Beaver JA, Howie LJ, Pelosof L, et al. A 25-year experience of US Food and Drug Administration accelerated approval of malignant hematology and oncology drugs and biologics: a review. *JAMA Oncol*. 2018;4:849-856.
16. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. New York, NY: CRC Press; 1999.
17. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. Chichester, UK: John Wiley & Sons; 1997.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Garcia Barrado L, Burzykowski T, Legrand C, Buyse M. Using an interim analysis based exclusively on an early outcome in a randomized clinical trial with a long-term clinical endpoint. *Pharmaceutical Statistics*. 2022;21(1):209-219. doi:10.1002/pst.2165

[Correction added on 25 December 2024, after first online publication: Last name of the author on reference 3 has been corrected from Galbraight to Galbraith.]