Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Evaluating Futility of a Binary Clinical Endpoint using Early Read-Outs

Kelly Van Lancker[1] | An Vandebosch[2] | Stijn Vansteelandt[1,3] | Filip De Ridder[2]

[1]Department of Applied Mathematics, Computer Science and Statistics, Ghent, Belgium
[2]Janssen R&D, a division of Janssen Pharmaceutica NV, Beerse, Belgium
[3]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, United Kingdom

Correspondence
Kelly Van Lancker, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium. Email: kelly.vanlancker@ugent.be

Present Address
Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

Summary

Interim analyses are routinely used to monitor accumulating data in clinical trials. When the objective of the interim analysis is to stop the trial if the trial is deemed futile, it must ideally be conducted as early as possible. In trials where the clinical endpoint of interest is only observed after a long follow-up, many enrolled patients may therefore have no information on the primary endpoint available at the time of the interim analysis. To facilitate earlier decision-making, one may incorporate early response data that are predictive for the primary endpoint (e.g. , an assessment of the primary endpoint at an earlier time) in the interim analysis.

Most attention so far has been given to the development of interim test statistics that include such short-term endpoints, but not to decision procedures. Existing tests moreover perform poorly when the information is scarce, e.g. , due to rare events, when the cohort of patients with observed primary endpoint data is small, or when the short-term endpoint is a strong, but imperfect predictor. In view of this, we develop an interim decision procedure based on the conditional power approach which utilises the short- and long-term binary endpoints in a framework that is expected to provide reliable inferences, even when the primary endpoint is only available for a few patients, and has the added advantage that it allows the use of historical information. The operational characteristics of the proposed procedure are evaluated for the phase 3 clinical trial that motivated this approach, using simulation studies.

KEYWORDS:
Interim analysis, Biomarkers, Conditional Power, Futility

1 | INTRODUCTION

Interim analyses are routinely used in clinical trials to guide trial design modifications and early stopping. They are for instance used to decide whether the trial should be stopped for futility [1,2], in case the state of evidence at the time of the interim analysis leaves little hope that evidence of superiority would be found if the trial were to continue. The statistical rule to guide the decision of whether or not to stop the trial early for futility is often based on conditional power. [3,4,5,6] This monitoring approach quantifies the probability that the null hypothesis will be rejected at the end of the study with a given statistical test, given the primary endpoint data observed thus far. Computations of conditional power usually assume that the future primary endpoint data will either be generated under parameter values specified in the initial study design (i.e. , under the design assumption), or that they arise from the same distribution that generated the observed data collected so far, or that they will be generated under the null hypothesis. We will focus on the conditional power under the design assumption but our proposal will also apply for these alternative assumptions.

In trials where the clinical endpoint of interest is only observed after a long period of treatment-free follow-up, many enrolled patients may have no information on the primary endpoint available at the time of the interim analysis. Restricting the interim analysis to those patients with long-term information available, may then result in lack of information to support futility decisions. Delaying the time of the interim analysis to a time where a sufficient number of patients have their primary endpoint available, may be less ethical and increase the costs if the trial were deemed futile. In particular, it may rule out the possibility of stopping recruitment as all patients may then have been enrolled already. To support futility decisions, it then seems of interest to replace the long-term primary endpoint by a short-term surrogate, when available. This, however, is only justified under the strong assumption that tests for short- and long-term treatment differences are equivalent.[7]

For instance, the motivating phase 3 clinical trial was designed to evaluate the efficacy of a new experimental treatment for multidrug-resistant tuberculosis on top of the standard of care regimen (referred to as background regimen BR) as compared to placebo plus BR with regard to the proportion of subjects with a favorable treatment outcome (ClinicalTrials.gov, NCT00449644).[8] This requires confirmed culture conversion 60 weeks after randomization. Besides the clinical endpoint of interest, confirmed culture conversion (cure) was planned to be assessed 16 weeks after randomization (see Web Appendix A for an overview of the study design). An interim analysis to decide if the trial should be stopped for futility, would ideally employ the 16 week endpoint for patients for whom this, but not the 60 week endpoint is available. Basing the interim analysis only on the short-term endpoint is generally not satisfactory as it shifts the focus away from the primary long-term endpoint. Treatments may indeed be very similar with respect to their short-term effect, but differ with respect to their long-term effect, or vice versa. This may, for example, be the case when treatments differ in time to response. Several approaches in the literature therefore, instead, use information on short-term endpoints only to predict the primary long-term endpoint when it is missing. Marschner and Becker[9] introduced a Wald test based on the probability difference parameterization to address this problem for binary outcomes assessed at two time points during follow-up. Kunz et al[10] applied this estimator to a two-stage phase II oncology trial. Niewczas, König and Kunz[11] constructed decision-making rules based on the conditional power of the resulting test statistic assuming a Brownian motion structure. Sooriyarachchi et al[12] presented a score test, based on the log-odds ratio for success at the final time point, for incorporating patients with binary assessments taken at three fixed time points, with the third assessment time being the primary one. Whitehead et al[13] compared the performances of four methods for incorporating binary observations taken at two time points into interim analyses: the score and Wald approaches, each with the log-odds ratio and probability difference parameterizations. Their simulations have shown that all four approaches have good properties regarding the power and type 1 error in moderate to large sample sizes. Similar methods for continuous data have also been considered in the literature.[14,15] For example, Galbraith and Marschner[14] adapted and extended the methodology described in Marschner and Becker[9] to include continuous endpoints assessed at an arbitrary number of follow-up times. Hampson and Jennison[16] generalized this to a group sequential design for the situation where the primary endpoint is measured with delay.

In this article, we aim to improve performance in settings where the asymptotic theory, as used for the single known decision procedure for this problem,[11] may fail. In particular, when the number of patients with complete data at the interim assessment is small, or the considered event is rare, the variance estimator of Marschner and Becker[9] may perform poorly. A decision procedure based on this test statistic may then misrepresent the amount of information that is available at the interim assessment. Similar problems may occur when the short-term endpoint is a strong predictor of the primary endpoint, for then their degree of dependence may be difficult to assess well. We will overcome this problem by making use of a Bayesian procedure, which avoids asymptotic approximations, thereby giving rise to a decision procedure that is more widely applicable. This not only provides a more robust conditional power[17] but also brings the possibility of including historical data, which are often available from earlier phases in development. If, for example, accrual occurs very quickly, the availability of primary endpoint data will be limited until close to the end of the study. This makes it difficult to estimate the long-term response probability with a certain degree of precision early in the study. Using historical data to inform Bayesian priors may then help improve the certainty of decisions made during drug development and as a result reduce overall costs. The need to incorporate historical data was motivated by the tuberculosis example, introduced earlier in this section, where prior phase 2b trial data were available. Like Marschner and Becker[9], our focus throughout will be on binary short- and long-term endpoints.

1.1 | Motivating Context

The methods of this paper are likely to find greatest applicability in interim analyses of long-term binary endpoints where the treatment under study is intended to cause a favourable/unfavourable short-term response that then could be sustained or lost in the long-term. For example, in infectious diseases clinical trials, the clinical endpoint is often designed to evaluate some form of disease remission after a sufficiently long treatment-free follow-up period. Certain subjects may achieve response (virus-free) on treatment and continue to respond after treatment (and cure), while some relapse only after stopping the treatment. Others may not achieve response at all.

2 | INCORPORATION OF INTERMEDIATE DATA

2.1 | Setting and Definitions

The methodology is introduced for the case of a two-treatment randomized controlled trial in n patients where the primary long-term endpoint Y and the secondary short-term endpoint X on each subject are dichotomous observations. Assume that Y is evaluated τ_Y time after randomization, and that X is evaluated using the same criteria at time $\tau_X \leq \tau_Y$, both fixed and identical time points for all subjects. Define P_{Y_j} and P_{X_j} $(j \in \{0, 1\})$ as the probabilities of respectively a successful outcome at the end of trial (time point τ_Y) and a successful short-term read-out at time point τ_X, in the experimental $(j = 1)$ and control $(j = 0)$ arm. The primary hypothesis of interest, $H_0 : P_{Y_1} = P_{Y_0}$ or $\delta := P_{Y_1} - P_{Y_0} = 0$, will be tested against the one-sided alternative, $H_A : P_{Y_1} > P_{Y_0}$ or $\delta > 0$ at level α with power $1 - \beta$. To evaluate this hypothesis, a Z-statistic for the difference in proportions with pooled variance will be employed, although our proposal is readily applicable to any Z-test statistic of this hypothesis. Suppose now that an interim analysis of the primary long-term endpoint Y will be conducted at information fraction t_X for X and t_Y for Y, representing respectively the ratio of the number of patients with observed short-term endpoint data at the time of the interim analysis and the planned total sample size, and the ratio of the number of patients with observed long-term endpoint data at the time of the interim analysis and the planned total sample size. These indicate, starting from the beginning of the study, how far through the trial we are. Throughout, we will index an interim analysis by time t, with t fully determined by information fraction t_X and/or t_Y. In this paper, the main focus will be on the conditional power at a given information fraction t_X and/or t_Y assuming the distribution of the future primary endpoint data will be generated under the design assumed parameter values $P_{Y_1} = \pi_1$ and $P_{Y_0} = \pi_0$, with π_1 and π_0 the values that were used for powering the study.

2.2 | Conceptual Proposal

In the absence of early read-outs the conditional power calculations are based on only two cohorts of patients.[3,4,18] A first cohort includes all patients for whom the primary endpoint Y has been observed. The second cohort consists of all patients who have not yet been followed through to the long-term follow-up time τ_Y. Their future primary endpoint data are then assumed to follow a Bernoulli distribution with probability determined by the design assumptions. We will refer to this method, where one endpoint is used, as the standard method.

(Include Figure 1 about here)

When early response data are available, we can distinguish three cohorts of patients (Figure 1) at information fraction t: a first cohort of all subjects with both endpoints observed, a second cohort of all subjects with only the short-term endpoint X observed and a third cohort of all subjects without available data on X and Y. Our main motivation is to develop a method that extends the standard method by using all the available data and reduces to it when X is independent of Y.

The future primary endpoint data for patients in cohort 3 are assumed to follow a Bernoulli distribution with probability determined by the design assumptions in each treatment arm. Our aim is then to use short-term endpoint data X to improve the efficiency of the interim analysis on the long-term endpoint Y. This will be done by prediction of the unobserved primary endpoint Y based on early response data X for the subjects in cohort 2. Accordingly, the future primary endpoint data for early responders and early non-responders in each treatment arm of cohort 2 are assumed to follow a Bernoulli distribution conditional on the early response data with probability $P(Y = 1|X = 1, A = j)$ for early responders and $P(Y = 1|X = 0, A = j)$ for early non-responders in the experimental arm $(A = 1)$ and the control arm $(A = 0)$. It is tempting to extract these conditional probabilities directly from the available historical and cohort 1 data. However, this has the drawback that these probabilities then do not reduce to the design assumptions π_1 and π_0 when X is independent of Y, and that the approach hence does not extend the standard method. We therefore propose to parameterise the joint distribution of the repeated binary data in terms of P_{Y_j} and the conditional probabilities $P(X = 1|Y = 1, A = j)$ and $P(X = 1|Y = 0, A = j)$, denoted by respectively $P_{X_j|Y_j=1}$ and $P_{X_j|Y_j=0}$, in treatment arm $j \in \{0, 1\}$. This variation independent parameterisation allows to incorporate the design assumption via P_{Y_j}, and to extract $P_{X_j|Y_j=1}$ and $P_{X_j|Y_j=0}$ from all information available in cohort 1 at the interim analysis. In doing so, we are assuming that the data observed in cohort 1 are representative with regard to $P_{X_j|Y_j=1}$ and $P_{X_j|Y_j=0}$ for cohort 2. Using Bayes' theorem, we finally obtain

$$P(Y = 1|X = 1, A = j) = \frac{P_{X_j|Y_j=1}P_{Y_j}}{P_{X_j|Y_j=1}P_{Y_j} + P_{X_j|Y_j=0}(1 - P_{Y_j})}$$

$$P(Y = 1|X = 0, A = j) = \frac{(1 - P_{X_j|Y_j=1})P_{Y_j}}{(1 - P_{X_j|Y_j=1})P_{Y_j} + (1 - P_{X_j|Y_j=0})(1 - P_{Y_j})}.$$

We further define $P^d_{Y_j|X_j=1}$ and $P^d_{Y_j|X_j=0}$ as these conditional probabilities evaluated under the design assumptions:

$$\begin{aligned} P^d_{Y_j|X_j=1} &= \frac{P^{(1)}_{X_j|Y_j=1}\pi_j}{P^{(1)}_{X_j|Y_j=1}\pi_j + P^{(1)}_{X_j|Y_j=0}(1 - \pi_j)} \\ P^d_{Y_j|X_j=0} &= \frac{\left(1 - P^{(1)}_{X_j|Y_j=1}\right)\pi_j}{\left(1 - P^{(1)}_{X_j|Y_j=1}\right)\pi_j + \left(1 - P^{(1)}_{X_j|Y_j=0}\right)(1 - \pi_j)}, \end{aligned} \tag{1}$$

where the superscripts (m) for $m = 1, 2, 3$ express that the corresponding probability will be estimated based on the data in cohort m. We can now verify that the proposed method reduces to the standard method when X and Y are independent. Indeed, in that case $P^{(1)}_{X_j|Y_j=1} = P^{(1)}_{X_j|Y_j=0} = P^{(1)}_{X_j}$ and thus $P^d_{Y_j|X_j=1} = \frac{P^{(1)}_{X_j|Y_j=1}\pi_j}{P^{(1)}_{X_j|Y_j=1}\pi_j + P^{(1)}_{X_j|Y_j=0}(1-\pi_j)} = \pi_j$; similarly $P^d_{Y_j|X_j=0} = \pi_j$. Furthermore, when $X = Y$ in cohort 1 so that $P^{(1)}_{X_j|Y_j=1} = 1$ and $P^{(1)}_{X_j|Y_j=0} = 0$, then $P^d_{Y_j|X_j=1} = \frac{P^{(1)}_{X_j|Y_j=1}\pi_j}{P^{(1)}_{X_j|Y_j=1}\pi_j + P^{(1)}_{X_j|Y_j=0}(1-\pi_j)} = 1$ and similarly, $P^d_{Y_j|X_j=0} = 0$. In that case, the proposal thus acknowledges that all information on Y is obtained for patients in cohort 2. More generally, Figure 2 shows the relation between the log odds ratio (for the association between X and Y) and the marginal outcome probabilities $P^{(2)}_{Y_j}$ in cohort 2. If the observed X distributions in cohort 1 and 2 are equal, then $P^{(2)}_{Y_j}$ varies between the observed value $P^{(1)}_{Y_j}$ and the design assumption (Figure 2). If the observed X distributions in cohort 2 deviates from the X distribution observed in cohort 1, $P^{(2)}_{Y_j}$ need not lie in between $P^{(1)}_{Y_j}$ and the design assumption.
(Include Figure 2 about here)

2.3 | Analytical Proposal

In large studies where a test statistic Z with normal approximation is appropriate, we can approximate the conditional power for the observed X and Y data, denoted by D_t, at any interim analysis time t. For a given value of $\boldsymbol{P}^d_{Y|X} = (P^d_{Y_1|X_1=1}, P^d_{Y_1|X_1=0}, P^d_{Y_0|X_0=1}, P^d_{Y_0|X_0=0})'$ and assuming that $P_{Y_1} = \pi_1$ and $P_{Y_0} = \pi_0$ for given values π_1 and π_0 (e.g., design assumptions), using analytical expressions along similar lines as in Lan and Wittes [18] gives conditional power

$$\begin{aligned} \mathrm{CP}_t(\pi_1, \pi_0|\boldsymbol{P}^d_{Y|X}) &\equiv P(\text{reject } H_0 \text{ at final analysis}|D_t, \boldsymbol{P}^d_{Y|X}, \pi_1, \pi_0) \\ &= P(Z_1 \geq z_{1-\alpha}|D_t, \boldsymbol{P}^d_{Y|X}, \pi_1, \pi_0) \\ &= 1 - \Phi\left(\frac{z_{1-\alpha} - \mathrm{E}_c(Z_1)}{\sqrt{\mathrm{Var}_c(Z_1)}}\right), \end{aligned} \tag{2}$$

with $z_{1-\alpha}$ the $(1 - \alpha)$-quantile of the standard normal distribution, α the significance level and $\mathrm{E}_c(\cdot)$ and $\mathrm{Var}_c(\cdot)$ the conditional expectation and variance of the test statistic, given the observed data at the interim evaluation D_t, the parameter $\boldsymbol{P}^d_{Y|X}$ and given values π_1 and π_0 for the parameters of interest. To define $\mathrm{E}_c(\cdot)$ and $\mathrm{Var}_c(\cdot)$, let $m_{k,j}$ be the number of subjects in cohort k of treatment group j ($k = 1, 2, 3$; $j = 0, 1$), $z^{(1)}$ the test statistic in cohort 1 and n_1 and n_0 respectively the planned total sample sizes in the experimental and placebo arm. Using a test statistic Z_1 with normal approximation, the asymptotic mean $\mathrm{E}_c(\cdot)$ and variance $\mathrm{Var}_c(\cdot)$ can be expressed as (see Web Appendix B)

$$\begin{aligned}
\mathrm{E}_c(Z_1) &= z^{(1)}\sqrt{\frac{m_{1,0}}{n_0}} + \frac{\pi_1^*(\boldsymbol{P}^d_{Y|X}) - \pi_0^*(\boldsymbol{P}^d_{Y|X})}{\sigma\sqrt{\frac{1}{r}+1}}\frac{m_{2,0}}{\sqrt{n_0}} + \frac{\pi_1 - \pi_0}{\sigma\sqrt{\frac{1}{r}+1}}\frac{m_{3,0}}{\sqrt{n_0}} \\
\mathrm{Var}_c(Z_1) &= \frac{\sigma_p^2}{\sigma^2}\frac{m_{2,0}}{n_0} + \frac{m_{3,0}}{n_0},
\end{aligned}$$

where $r = \frac{n_1}{n_0}$ and

$$\begin{aligned}
\pi_1^*(\boldsymbol{P}^d_{Y|X}) &= P^{(2)}_{X_1} P^d_{Y_1|X_1=1} + \left(1 - P^{(2)}_{X_1}\right) P^d_{Y_1|X_1=0}, \\
\pi_0^*(\boldsymbol{P}^d_{Y|X}) &= P^{(2)}_{X_0} P^d_{Y_0|X_0=1} + \left(1 - P^{(2)}_{X_0}\right) P^d_{Y_0|X_0=0},
\end{aligned}$$

$$\begin{aligned}
\sigma_p^2 &= \frac{1}{(\frac{1}{r}+1)}\left[\frac{1}{r}\left\{P^{(2)}_{X_1} P^d_{Y_1|X_1=1}(1 - P^d_{Y_1|X_1=1}) + \left(1 - P^{(2)}_{X_1}\right) P^d_{Y_1|X_1=0}(1 - P^d_{Y_1|X_1=0})\right\}\right. \\
&\quad \left. + \left\{P^{(2)}_{X_0} P^d_{Y_0|X_0=1}(1 - P^d_{Y_0|X_0=1}) + \left(1 - P^{(2)}_{X_0}\right) P^d_{Y_0|X_0=0}(1 - P^d_{Y_0|X_0=0})\right\}\right], \\
\sigma &= \sqrt{\bar{p}(1-\bar{p})}, \text{ where } \bar{p} = \frac{n_1\pi_1 + n_0\pi_0}{n_1 + n_0}.
\end{aligned}$$

Note that the three terms of $\mathrm{E}_c(Z_1)$ correspond with the three different cohorts, while $\mathrm{Var}_c(Z_1)$ only includes terms for cohort 2 and 3. The fractions $\sqrt{\frac{m_{j,0}}{n_0}}$ $(j = 1, 2, 3)$ and $\frac{m_{j,0}}{n_0}$ $(j = 2, 3)$ represent how much information in the asymptotic expectation and variance, respectively, is explained by the different cohorts. This approximation uses $r = \frac{n_1}{n_0} = \frac{m_{1,1}}{m_{1,0}} = \frac{m_{2,1}}{m_{2,0}} = \frac{m_{3,1}}{m_{3,0}}$, which improves with increasing sample size. The above procedure is not readily feasible, however, because the conditional probabilities $\boldsymbol{P}^d_{Y|X}$ are unknown. Data from earlier studies (e.g., phase 2b data in TB study) may provide - albeit with uncertainty - an initial estimate. The observed data at the interim analysis allow to further improve these estimates, but acknowledging the resulting uncertainty can be difficult, e.g., when in a small cohort 1 we find $X = Y$ for all subjects. We will therefore approximate the joint sampling distribution of $\boldsymbol{P}^d_{Y|X}$ by the joint posterior distribution, which is ideally specified based on historical data. A detailed description of typical distributions is given in Section 2.3.1. We can then eliminate $\boldsymbol{P}^d_{Y|X}$ from the conditional power (2) via integration

$$\begin{aligned}
\mathrm{CP}_t(\pi_1, \pi_0) &= \int \mathrm{P}(\text{reject } \mathrm{H}_0 \text{ at final analysis} | D_t, \boldsymbol{P}^d_{Y|X}, \pi_1, \pi_0) f(\boldsymbol{P}^d_{Y|X} | D_t, \pi_1, \pi_0, H) \mathrm{d}\boldsymbol{P}^d_{Y|X} \\
&= \int \mathrm{CP}_t(\pi_1, \pi_0 | \boldsymbol{P}^d_{Y|X}) f(\boldsymbol{P}^d_{Y|X} | D_t, \pi_1, \pi_0, H) \mathrm{d}\boldsymbol{P}^d_{Y|X} \\
&= \mathrm{E}\left[\mathrm{CP}_t(\pi_1, \pi_0 | \boldsymbol{P}^d_{Y|X}) | D_t, \pi_1, \pi_0, H\right],
\end{aligned}$$

with $f(\boldsymbol{P}^d_{Y|X} | D_t, \pi_1, \pi_0, H)$ the (posterior) probability density of $\boldsymbol{P}^d_{Y|X}$ given the observed data D_t at the interim analysis, assumptions π_1 and π_0 about the parameters of interest and the prior information H which includes the historical data. We refer to this quantity as the expected conditional power. The (posterior) probability distribution of $\boldsymbol{P}^d_{Y|X}$ allows us to repeatedly sample a sufficient number of vectors $\boldsymbol{P}^d_{Y|X} = (P^d_{Y_1|X_1=1}, P^d_{Y_1|X_1=0}, P^d_{Y_0|X_0=1}, P^d_{Y_0|X_0=0})'$ from the (posterior) probability density $f(\boldsymbol{P}^d_{Y|X} | D_t, \pi_1, \pi_0, H)$, calculate the conditional power for each sampled vector of $\boldsymbol{P}^d_{Y|X}$ values via formula (2) and subsequently take the average over all these conditional power values. If the expected conditional power is below a certain, predefined cut off value, the trial will be stopped and futility recommended. Guidelines for determining this cut-off are given in Section 2.4.

For non-normal tests (e.g., Pearson chi-square test) or when the number of patients in cohort 1 is small (i.e., less than 5 expected events under the null hypothesis in one of the arms), we can determine the conditional power for a given value of $\boldsymbol{P}^d_{Y|X}$ via Monte Carlo simulations instead. As before, we then sample a sufficient number of replicates for $\boldsymbol{P}^d_{Y|X}$ from its (posterior) probability distribution. For each value of $\boldsymbol{P}^d_{Y|X}$, we simulate the unobserved primary endpoint for each of the patients in cohort 2 and 3. For the patients in cohort 2, this is done by sampling the values of the primary endpoint from a Bernoulli distribution with probability $P^d_{Y_j|X_j=1}$ or $P^d_{Y_j|X_j=0}$ $(j = 0, 1)$, depending on whether the subject has achieved the early response or not. These distributions are specific for each randomized arm. For patients in cohort 3, we sample the values of the primary endpoint from a Bernoulli distribution with probability π_1 and π_0 in respectively the experimental and control arm. Based on the observed and simulated primary endpoint data, we then evaluate the test. The conditional power is then defined as the fraction of trials in which the primary hypothesis is rejected at the prespecified significance level. When there are less than 5 expected events under the null hypothesis in one of the arms, then this is best based on Fisher's exact test or the test statistic after adding pseudo data, 1 success and 1 failure, to each sample group.[19]

2.3.1 | Posterior Distribution of $P^d_{Y|X}$

Assuming prior independence between $(P^d_{Y_1|X_1=1}, P^d_{Y_1|X_1=0})$ and $(P^d_{Y_0|X_0=1}, P^d_{Y_0|X_0=0})$, we can write $f(\boldsymbol{P}^d_{Y|X}|D_t, \pi_1, \pi_0, H)$ as the product of the posteriors $f(P^d_{Y_1|X_1=1}, P^d_{Y_1|X_1=0}|D_t, \pi_1, \pi_0, H)$ and $f(P^d_{Y_0|X_0=1}, P^d_{Y_0|X_0=0}|D_t, \pi_1, \pi_0, H)$. Since $P^d_{Y_j|X_j=1}$ and $P^d_{Y_j|X_j=0}$ are determined by P_{Y_j}, $P_{X_j|Y_j=1}$ and $P_{X_j|Y_j=0}$, a joint distribution for $P^d_{Y_j|X_j=1}$ and $P^d_{Y_j|X_j=0}$ is implied by a joint distribution for the latter two parameters and the assumed values π_1 and π_0 for P_{Y_j} in arm j. Assuming a priori independence between these three parameters, we can choose the prior distributions for the latter two separately.

We recommend non-informative Beta(0.5, 0.5) priors on $P_{X_j|Y_j=1}$ and $P_{X_j|Y_j=0}$ corresponding with a 2x2 table cross-classified according to X and Y with value 0.5 in each cell. When historical data are representative for the current trial with respect to the conditional probabilities $P_{X_j|Y_j=1}$ and $P_{X_j|Y_j=0}$, we can use these data to update these non-informative prior distributions as follows. The informative Beta prior distribution for $P_{X_j|Y_j=1}$ has parameters $0.5 + x$ and $0.5 + m - x$ if x out of the m patients with $Y = 1$ in treatment arm j of the historical dataset are early responders. Likewise, we can update the non-informative Beta(0.5, 0.5) prior distributions for $P_{X_j|Y_j=0}$ to informative priors with parameters $0.5 + y$ and $0.5 + s - y$ if y out of the s patients with $Y = 0$ in treatment arm j of the historical dataset are early responders. In practice, one needs to decide how much weight to give to the prior patients, given that these patients come from another study and thus should not be given larger weights than the cohort 1 patients. For more information on how to use prior belief of clinicians and historical data to elicit prior distributions, we refer the reader to Spiegelhalter, Freedman and Parmar[20]. If the historical data are not representative, non-informative priors may be more appropriate.

The likelihoods for $P_{X_j|Y_j=1}$ and $P_{X_j|Y_j=0}$ are based on the cohort 1 data and allow us to update the priors as before but now using the cohort 1 data instead of the historical data. The posterior distributions allow to repeatedly sample replicates for $P_{X_j|Y_j=1}$ and $P_{X_j|Y_j=0}$ independently. Assuming $P_{Y_j} = \pi_j$, we can derive values for $(P^d_{Y_j|X_j=1}, P^d_{Y_j|X_j=0})$ via the analytical expressions (1) in both treatment arms.

2.4 | Cut-Off for the Conditional Power

A cut-off on the conditional power values specifies the value below which we stop the trial for futility. Its choice is important when designing the decision rule; to avoid an increase in type 2 error while having sufficient 'power' of a correct futility decision.[21] If the cut-off point is chosen too low (high), the trial will be stopped too rarely (frequently). It is therefore important to select a cut-off value that limits the impact on the power of the trial to detect true superiority. The type II error probability can be partitioned as the probability of stopping for futility (and of a type II error) at the pre-specified interim analysis plus the probability of continuation and non-significance at the final analysis.[4]

One approach for choosing an appropriate cut-off is to maximize power loss as a result of futility stopping for the protocol planned superiority scenario(s). In particular, we infer a corresponding cut-off by running a sufficient number of simulations under the different scenarios. For a given cut-off and scenario, the probability to stop for futility is calculated as the fraction of trials in which the decision was to stop. We determine for each scenario the maximum cut-off for which the total power falls just below the pre-specified minimal total power for that scenario. The final cut-off is the minimum over all these cut-offs.

3 | SIMULATION STUDY

We carried out different simulation studies to compare the performance of the proposed method to the existing methods and to evaluate its operational characteristics under a variety of true futility and superiority scenarios. The cut-off criterion, as described in section 2.4, is used to investigate and compare the properties of the different methods.

3.1 | Simulation Settings

In the motivating phase 3 clinical trial, superiority of the experimental arm over the control arm is claimed if the proportion of subjects with favorable treatment outcome 60 weeks after randomization is significantly higher compared to the control arm at a one-sided significance level of 2.5%. Assuming a favorable treatment outcome rate of 60% and 73% in the control and experimental arm respectively, 275 patients in both arms are required to attain 90% power with a Z-statistic for the difference in proportions with pooled variance.[22]

Endpoint data are generated from a repeated binary data model under futility/superiority scenarios for the proportion of subjects achieving a favorable outcome Y, varying assumptions on the proportion of subjects achieving X and varying assumptions (strong, assumed and weak) on the log odds ratio $\log OR_j$ of Y for early responders versus early non-responders in treatment arm $j \in \{0, 1\}$. Particularly, the short-term endpoint data X are sampled from a Bernoulli distribution with probability P_{X_1} and P_{X_0} in respectively the experimental and control arm. Accordingly, the future primary endpoint data for early responders and early non-responders in each treatment arm are sampled from a Bernoulli distribution with probabilities conditional on the early response data. These conditional probabilities are determined by the marginal probabilities for X and Y as well as the log odds ratios. Table 1 provides an overview of the selected population models. Based on the observed phase 2b data, three correlation scenarios 'weak', 'assumed' and 'strong' were generated, whereby the correlation was weaker, the same and stronger respectively than the correlation assumed in the prior distribution. In addition to these correlations, a perfect predictor X of Y (SUP.1 and FUT.1) and a very strong predictor X of Y (FUT.3) are considered. We furthermore assume that at the interim analysis 10% of the patients have complete observations on the long-term outcome such that $t_Y = 0.10$ and 40% of short-term observations are available such that $t_X = 0.40$. This corresponds with an average of 54 patients in cohort 1, 166 in cohort 2 and 330 in cohort 3. These targeted proportions are recommended as the minimum values to obtain a reasonable probability to stop for true futility and a negligible probability to incorrectly stop under superiority in this example.[23] They are based on projected recruitment at that time.

(Include Table 1 about here)

First, we compare the performance of the proposed method to the standard method and the method introduced by Niewczas, König and Kunz[11] (referred to as method NKK). In adition, we investigated different strengths of association between both endpoints and the influence of the information fraction on the cut-off value and the 'power' of a correct futility decision.

Every scenario was run 100, 000 times. For each simulated trial the conditional power was calculated based on the analytical approximations given in Section 2.3 for 2500 posterior samples for $\boldsymbol{P}^d_{Y|X}$. Prior distributions were selected based on the real phase 2b data. Details are given in Web Appendix D. For each scenario and decision criterion, the probability to stop for futility was calculated as the fraction of trials in which the decision was to stop. The R-code is available in the R package FutilityStopping.

3.2 | Simulation Results

3.2.1 | Selecting a criterion for decision making

For method comparison a cut-off was set in such a way that a maximum reduction of 1% in power relative to an analysis without futility decisions is allowed under the most plausible superiority scenarios SUP.2 and SUP.5 (See Table 1) and a limited power loss (< 5%) under the less expected and even misleading scenarios (e.g. , SUP.3, whereby at interim the proportion of subjects achieving X is not different between the randomized groups). The probability to stop for futility and the overall power in function of the cut-off points are shown in Web Appendix G. Similar patterns in opposite directions are seen. The overall power decreases with increasing cut-off points, but stays around the design power of 0.90 up to cut-off points of around 0.51 under the different superiority scenarios, except under SUP.3. Since we also want a reasonable probability to stop early in case of true futility and the probability to stop for futility is increasing with increasing cut-off points, a stopping criterion of 0.51 for the expected conditional power appears to perform best in terms of a limited power inflation (< 1% under SUP.2 and SUP.5) and a reasonable probability of stopping for true futility. The probability to stop under a true futility scenario is 44% under FUT.1, 44.5% under FUT.3 and varies from approximately 7.5% to 9% for the scenarios under the weak log odds ratio, from approximately 24% to 34% for the scenarios under the assumed log odds ratio and from approximately 25% to 44% for the scenarios under the strong log odds ratio (Table 2 and Web Appendix G). The probabilities to stop under futility scenario FUT.5, whereby at interim a difference in proportion of subjects achieving a favorable treatment outcome X in favor of the experimental treatment is observed, under the weak, assumed and strong log odds ratio are 5%, 18% and 19% respectively.

3.2.2 | Comparison With the Standard Conditional Power and Method NKK

In this section, we compare the expected conditional power with the standard approach and method NKK. The latter assumes a Brownion motion structure to calculate the conditional power based on a binary outcome measured at two time points. We refer the reader to Web Appendix E for further details.

In each trial, interim data were generated for 220 of the 550 patients ($t_X = 0.40$) and primary endpoint data for 54 of them ($t_Y = 0.10$). The cut-off value is determined using the same approach as for our proposal (see Section 3.2.1). Early stopping for futility is recommended when the conditional power falls below 0.73 when the standard approach is used, below 0.60 when method NKK is used or below 0.51 when the expected conditional power method is used. Note that the most appropriate cut-off value differs between the methods, since for each method a different cut-off value leads to a small to negligible probability of false early stopping in case of true superiority (maximum reduction of 1% in power under the most plausible superiority scenarios and a limited power loss (< 5%) under the less expected) and a reasonable probability of stopping for true futility. The probabilities to stop for futility under the different true futility scenarios are displayed in Table 2. Note that the superiority as well as the futility scenarios coincide when evaluating the standard conditional power since we are only evaluating the primary endpoint data which are simulated under the same distributions (e.g., binomial distributions with probabilities 0.60 and 0.73 for respectively the control and treatment arm under superiority).

(Include Table 2 about here)

It can be seen that the overall probability to stop for futility increases when using the proposal incorporating short-term endpoints compared to the standard approach, except at a weak log odds ratio where the results are similar. This is due to the different choice of the cut-off. In general, the proposal seems to perform better than method NKK, except when the log odds ratio is reasonably large (FUT.1 and strong FUT.3). The latter is due to the fact that method NKK ignores the uncertainty on the predicted values in cohort 2 when X happens to equal Y for all subjects in cohort 1. Acknowledging this uncertainty is important however. Consider for example an interim analysis where data for all patients are available; 50 in cohort 1 and 500 in cohort 2. Suppose that the data in cohort 1 shows that $X = Y$: 22 patients with $X = Y = 1$ and 3 patients with $X = Y = 0$ in the experimental arm, and 18 patients with $X = Y = 1$ and 7 with $X = Y = 0$ in the placebo arm. Suppose further that 160 of the 250 patients in the experimental arm of cohort 2 are early responders. If cohort 2 includes 177 early responders in the experimental arm, then the conditional power calculated by method NKK is 0, meaning that at the time of the interim analysis the probability to reject the null hypothesis at the end of the trial is 0; it equals 1 if cohort 2 includes 178 early responders in the experimental arm. In such settings, it is impossible to obtain a conditional power different from 0 or 1 and it may jump between these two extreme values when an early non-responder changes in an early responder or vice versa.

Similar simulations for $t_Y = 0.20$ and $t_X = 0.40$ show that the advantage of using the expected conditional power decreases with increasing cohort 1 sample size for a fixed interim sample size (see Web Appendix F). The difference in the amount of information between the proposed and standard method decreases since the amount of data in cohort 2 decreases. Compared with method NKK, this is due to the fact that the prior information becomes less important when the cohort 1 sample size increases.

3.3 | Operational Characteristics

3.3.1 | Strength of the Association

To further investigate the behaviour of the estimators, different strengths of association between X and Y were considered at $t_X = 0.40$ and $t_Y = 0.20$: independent predictor ($\log OR_1 = 0$ and $\log OR_0 = 0$), weak correlation ($\log OR_1 = 0.8$ and $\log OR_0 = 0.5$), weak/low correlation ($\log OR_1 = 1.3$ and $\log OR_0 = 0.8$) denoted by low3, weak/low correlation ($\log OR_1 = 2$ and $\log OR_0 = 1.5$) denoted by low2 , weak/low correlation ($\log OR_1 = 3$ and $\log OR_0 = 2$) denoted by low1, assumed correlation ($\log OR_1 = 4.1$ and $\log OR_0 = 2.3$), strong correlation ($\log OR_1 = 7.4$ and $\log OR_0 = 4.1$) and perfect predictor ($\log OR_1 = 100$ and $\log OR_0 = 100$). Identical probabilities of success for Y and X were chosen in both treatment arms: $P_{Y1} = P_{X1} = 0.73$ and $P_{Y0} = P_{X0} = 0.60$ for the superiority scenarios and $P_{Y1} = P_{X1} = P_{Y0} = P_{X0} = 0.60$ for the true futility scenarios.

(Include Figure 3 about here)

Figure 3 shows that the stronger the predictor, the higher the probability to stop under a true futility scenario without loss of power. For the superiority scenarios we would expect the graphs to be in the reverse order, but this does not seem to be the case. This is due to the fact that we rely more on the design assumptions as the log odds ratios become smaller (see Figure 2). If X and Y are independent, we completely rely on the design assumptions. The curve is therefore similar -but higher due to the extra variability induced by the posterior distribution- to the curve that we would get if we calculated the standard conditional power. Since in practice, the association between X and Y is unknown, it is recommended to assume a range of association structures between both endpoints (e.g., a perfect predictor, an X which is independent of Y, the log odds ratio as observed in historical data and the observed log odds ratio plus or minus 1.96 times its standard error) to protect against a large loss of power when determining the cut-off value.

3.3.2 | Varying Information Fractions

To investigate the influence of t_X on the cut-off value for the conditional power, we investigate the probability to stop for true futility under scenario FUT.2 (assumed log odds ratio) for the cut-off value that results in a power loss of at most 1% under scenario SUP.2 (assumed log odds ratio) for t_Y fixed at 0.10. The following information fractions for X were considered: $t_X = (0.20, 0.30, 0.40, 0.50, 0.60)$. The probability to stop for futility under FUT.2 is 20% when $t_X = 0.20$ (cut-off: 0.66), 28% when $t_X = 0.30$ (cut-off: 0.59), 35% when $t_X = 0.40$ (cut-off: 0.51), 38% when $t_X = 0.50$ (cut-off: 0.43) and 40% when $t_X = 0.60$ (cut-off: 0.35). The probability to stop for futility using the expected conditional power is shown in Figure 4. We see that the higher the information fraction for X, the lower the cut-off point can be chosen and the higher the probability to stop under a true futility scenario.

(Include Figure 4 about here)

To investigate the influence of t_Y on the cut-off value, we evaluate the probability to stop for true futility under scenario FUT.2 for the cut-off value that results in a power loss of at most 1% under scenario SUP.2 for t_X fixed at 0.40. We consider the following information fractions for Y: $t_Y = (0.05, 0.10, 0.15, 0.20)$. The probability to stop for futility under FUT.2 is 34% when $t_Y = 0.05$ (cut-off: 0.50), 35% when $t_Y = 0.10$ (cut-off: 0.51), 38% when $t_Y = 0.15$ (cut-off: 0.53) and 41% when $t_Y = 0.20$ (cut-off: 0.53). Thus, the higher the information fraction for Y, the higher the cut-off point that results in 1% power reduction and the higher the probability to stop under a true futility scenario. Otherwise, the availability of primary endpoint data Y (cohort 1 data) improves the performance of the futility assessment; especially under scenarios where the observed outcome (in terms of treatment difference) on the X data is different from the true treatment effect on Y, or lack thereof (see FUT.5 in Web Appendix G).

(Include Figure 5 about here)

4 | DISCUSSION

In trials where accrual occurs fast compared to the planned length of follow-up, few subjects will have primary endpoint data until close to the end of the study. This makes it difficult to accurately estimate the long-term response probability early in the study. We therefore proposed a method to incorporate short-term endpoint data (e.g. , an assessment of the primary endpoint at an earlier time) and prior information for decision-making in the interim analysis. Our proposal assumes that the prior information about the association between both endpoints and the observed association in cohort 1 are representative for cohort 2. This requires that the dependence between both endpoints does not change over time. Whether this assumption is biologically plausible, should be assessed a priori based on consultation with clinicians. In addition, we recommend comparing the baseline covariates of patients in cohort 1 and 2 to assess whether the observed data in cohort 1 may be representative of the future data. In future work we will propose methods that can weaken these assumptions by correcting for baseline covariates.

We generally found our proposal to perform better than the standard conditional power method and the conditional power method introduced by Niewczas, König and Kunz[11]. The larger amount of information, obtained by incorporating short-term endpoint data as well as historical data, resulted in a higher overall probability to stop for true futility.

The (minimal) values for t_X and t_Y used in the example are specific to the context. In practice, one should do a simulation study to determine the minimal number of patients needed in cohort 1 and 2 (or equivalently the proportion of patients with long-term and short-term endpoint information) to obtain a reasonable probability to stop for true futility and a negligible probability to incorrectly stop under the superiority scenarios. For the minimal t_X, we propose using the minimal proportion of primary endpoint data needed for the standard conditional power to obtain a reasonable probability to stop for true futility and a negligible probability to incorrectly stop under superiority as explained in Freidlin et al[23] (e.g. , 37% for a one-sided 0.025 level design with a power of 90%). Subsequently, we consider the minimal number/proportion of these patients needed in cohort 1 to obtain a negligible probability to stop under the superiority scenarios and a reasonable probability to correctly stop under the true futility scenarios.

In the motivating study, any missing observation was dealt with as an unfavourable outcome. For subjects in cohorts 2 and 3 who already discontinued from the trial it is then known that they will not meet the criteria of a favorable primary endpoint, even when they are not yet followed for the whole interval τ_Y. We recommend, however, that this information is excluded from the analysis to avoid that the analysis is dominated by drop-outs. Thus, for the subjects in cohort 2 only available data up to τ_X is included in the analysis, while for the subjects in cohort 3 no data is included. Although the proposed methodology has therefore been developed in the context of complete data, it still remains valid when data are missing completely at random. When data are missing at random, standard missing data methods (e.g., multiple imputation) can be used to impute the missing data. Conditional power computations are commonly evaluated at fixed values of the parameters P_{Y_1} and P_{Y_0}. It is more cautious however to average the conditional power function with respect to the current knowledge or opinion about P_{Y_1} and P_{Y_0}. The current knowledge about the underlying value of these parameters can be summarized using prior Beta distributions. The predictive power[24] can then be derived by averaging/integrating the conditional power over different values π_1 and π_0 for respectively P_{Y_1} and P_{Y_0}, each one weighted according to the current belief about its probability by means of a (posterior) distribution. In contrast to the conditional power approach, it produces an unconditional, predictive probability of rejecting the null hypothesis and it avoids having to assume specific values for P_{Y_1} and P_{Y_0}. It therefore also delivers a more robust power assessment.

This paper has focused on binary endpoints. An important generalization would be to enable continuous endpoints, whether censored or not.

Our proposal is limited to trial designs where only one specific short-term response is identified. In principle, when repeated measures at multiple intermediate time points are evaluated, including these repeated measures could provide further efficiency gains. In future work we will therefore develop a more generic proposal to enable an interim evaluation of the treatment effect based on a combination of biomarkers, patient characteristics and/or intermediate endpoints.

ACKNOWLEDGEMENTS

This project has received funding from VLAIO under the Baekeland grant agreement HBC.2017.0219.

Conflict of interest

The authors declare no potential conflict of interests.

SUPPORTING INFORMATION

Additional Supporting Information including source code to reproduce the results may be found in the Web Appendix.

DATA AVAILABILITY STATEMENT

The data are not publicly available due to confidentiality restrictions. The simulated data that support the findings of this study are available on request from the corresponding author.

References

1. Snapinn S, Chen MG, Jiang Q, Koutsoukos T. Assessment of futility in clinical trials. *Pharmaceutical Statistics* 2006; 5(4): 273-281. doi: 10.1002/pst.216

2. DeMets DL. Futility approaches to interim monitoring by data monitoring committees. *Clinical Trials* 2006; 3(6): 522-529. doi: 10.1177/1740774506073115

3. Halperin M, Lan KKG, Ware JH, Johnson NJ, DeMets DL. An aid to data monitoring in long-term clinical trials. *Controlled Clinical Trials* 1982; 3(4): 311 - 323. doi: http://dx.doi.org/10.1016/0197-2456(82)90022-8

4. Lachin JM. A review of methods for futility stopping based on conditional power. *Statistics in Medicine* 2005; 24(18): 2747–2764. doi: 10.1002/sim.2151

5. Proschan M, Lan K, Wittes J. *Statistical Monitoring of Clinical Trials: A Unified Approach*. Statistics for Biology and HealthSpringer New York . 2006.

6. Lan KG, Simon R, Halperin M. Stochastically curtailed tests in long-term clinical trials. *Communications in Statistics. Part C: Sequential Analysis* 1982; 1(3): 207-219. doi: 10.1080/07474948208836014

7. Prentice RL. Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine* 1989; 8(4): 431–440.

8. Diacon AH, Pym A, Grobusch MP, al e. Multidrug-Resistant Tuberculosis and Culture Conversion with Bedaquiline. *New England Journal of Medicine* 2014; 371(8): 723-732. doi: 10.1056/NEJMoa1313865

9. Marschner IC, Becker SL. Interim monitoring of clinical trials based on long-term binary endpoints. *Statistics in Medicine* 2001; 20(2): 177–192. doi: 10.1002/1097-0258(20010130)20:2<177::AID-SIM653>3.0.CO;2-K

10. Kunz CU, Wason JM, Kieser M. Two-stage phase II oncology designs using short-term endpoints for early stopping. *Statistical Methods in Medical Research* 2017; 26(4): 1671-1683.

11. Niewczas J, Kunz CU, König F. Interim analysis incorporating short- and long-term binary endpoints. *Biometrical Journal* 2019: 1–23. doi: 10.1002/bimj.201700281

12. Sooriyarachchi MR, Whitehead J, Whitehead A, Bolland K. The sequential analysis of repeated binary responses: a score test for the case of three time points. *Statistics in Medicine* 2006; 25(13): 2196–2214. doi: 10.1002/sim.2339

13. Whitehead A, Sooriyarachchi MR, Whitehead J, Bolland K. Incorporating intermediate binary responses into interim analyses of clinical trials: A comparison of four methods. *Statistics in Medicine* 2008; 27(10): 1646–1666. doi: 10.1002/sim.3046

14. Galbraith S, Marschner IC. Interim analysis of continuous long-term endpoints in clinical trials with longitudinal outcomes. *Statistics in Medicine* 2003; 22(11): 1787–1805. doi: 10.1002/sim.1311

15. Stallard N. A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Statistics in Medicine* 2010; 29(9): 959–971.

16. Hampson LV, Jennison C. Group sequential tests for delayed responses (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2013; 75(1): 3–54. doi: 10.1111/j.1467-9868.2012.01030.x

17. Rauch G, Kieser M. An expected power approach for the assessment of composite endpoints and their components. *Computational Statistics & Data Analysis* 2013; 60: 111 - 122. doi: https://doi.org/10.1016/j.csda.2012.11.001

18. Lan KKG, Wittes J. The B-Value: A Tool for Monitoring Data. *Biometrics* 1988; 44(2): 579-585.

19. Agresti A, Caffo B. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *American Statistician* 2000; 54(4): 280–288.

20. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian Approaches to Randomized Trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 1994; 157(3): 357–416.

21. Schüler S, Kieser M, Rauch G. Choice of futility boundaries for group sequential designs with two endpoints. *BMC Medical Research Methodology* 2017; 17(1): 119. doi: 10.1186/s12874-017-0387-4

22. Rosner B. *Fundamentals of Biostatistics*. Cengage Learning . 2015.

23. Freidlin B, Korn EL, Gray R. A general inefficacy interim monitoring rule for randomized clinical trials. *Clinical Trials* 2010; 7(3): 197-208. PMID: 20423925doi: 10.1177/1740774510369019

24. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: Conditional or predictive power?. *Controlled Clinical Trials* 1986; 7(1): 8 - 17. doi: https://doi.org/10.1016/0197-2456(86)90003-6

GRAPHICAL ABSTRACT

Superiority	Y			X			Strong $\log OR$		Assumed $\log OR$		Weak $\log OR$		Note
Scenario	P_{Y0}	P_{Y1}	$P_{Y1}-P_{Y0}$	P_{X0}	P_{X1}	$P_{X1}-P_{X0}$	$\log OR_0$	$\log OR_1$	$\log OR_0$	$\log OR_1$	$\log OR_0$	$\log OR_1$	
SUP.1	0.60	0.73	0.13	0.60	0.73	0.13	-	-	100	100	-	-	$X = Y$
SUP.2	0.60	0.73	0.13	0.60	0.73	0.13	4.1	7.4	2.3	4.1	0.5	0.8	$P(X = 1) = P(Y = 1)$, but possibly $X \neq Y$
SUP.3	0.60	0.73	0.13	0.60	0.60	0	4.1	7.4	2.3	4.1	0.5	0.8	Futile X (False futility)
SUP.4	0.60	0.73	0.13	0.40	0.60	0.20	4.1	7.4	2.3	4.1	0.5	0.8	Larger effect in X (20%); $P(X = 1) < P(Y = 1)$ in control
SUP.5	0.60	0.73	0.13	0.60	0.80	0.20	4.1	7.4	2.3	4.1	0.5	0.8	Larger effect in X (20%); $P(X = 1) = P(Y = 1)$ in control
SUP.6	0.60	0.73	0.13	0.75	0.95	0.20	4.1	7.4	2.3	4.1	0.5	0.8	Larger effect in X (20%); $P(X = 1) > P(Y = 1)$ in control

Futility	Y			X			Strong $\log OR$		Assumed $\log OR$		Weak $\log OR$		Note
Scenario	P_{Y0}	P_{Y1}	$P_{Y1}-P_{Y0}$	P_{X0}	P_{X1}	$P_{X1}-P_{X0}$	$\log OR_0$	$\log OR_1$	$\log OR_0$	$\log OR_1$	$\log OR_0$	$\log OR_1$	
FUT.1	0.60	0.60	0	0.60	0.60	0	-	-	100	100	-	-	$X = Y$
FUT.2	0.60	0.60	0	0.60	0.60	0	4.1	7.4	2.3	4.1	0.5	0.8	$P(X = 1) = P(Y = 1)$, but possibly $X \neq Y$
FUT.3	0.60	0.60	0	0.60	0.60	0	-	-	7.4	10.7	-	-	$P(X = 1) = P(Y = 1)$, but possibly $X \neq Y$ (stronger association)
FUT.4	0.60	0.60	0	0.73	0.73	0	4.1	7.4	2.3	4.1	0.5	0.8	$P(X = 1) > P(Y = 1)$
FUT.4b	0.60	0.60	0	0.73	0.73	0	7.4	7.4	4.1	4.1	0.8	0.8	$P(X = 1) > P(Y = 1)$; same association in both arms
FUT.5	0.60	0.60	0	0.60	0.70	0.10	4.1	7.4	2.3	4.1	0.5	0.8	Effect of 10% in X (False superiority)
FUT.6	0.60	0.60	0	0.40	0.40	0	4.1	7.4	2.3	4.1	0.5	0.8	$P(X = 1) < P(Y = 1)$

TABLE 1 Parameters of the data generating models.

	Weak log odds ratio						
	FUT.1	FUT.2	FUT.3	FUT.4	FUT.4b	FUT.5	FUT.6
Proposal	-	8.45	-	7.50	8.12	5.34	9.37
NKK	-	2.26	-	2.30	2.27	2.19	2.22
Standard method	-	**12.32**	-	**12.32**	**12.32**	**12.32**	**12.32**

	Assumed log odds ratio						
	FUT.1	FUT.2	FUT.3	FUT.4	FUT.4b	FUT.5	FUT.6
Proposal	44.08	**34.24**	44.50	**24.15**	**25.63**	**17.93**	**26.07**
NKK	**63.46**	9.91	**53.53**	6.17	9.85	6.89	6.30
Standard method	12.32	12.32	12.32	12.32	12.32	12.32	12.32

	Strong log odds ratio						
	FUT.1	FUT.2	FUT.3	FUT.4	FUT.4b	FUT.5	FUT.6
Proposal	-	**44.39**	-	**28.85**	**29.26**	**19.42**	**25.12**
NKK	-	28.95	-	11.09	13.73	15.05	8.47
Standard method	-	12.32	-	12.32	12.32	12.32	12.32

TABLE 2 Probability to stop for futility under different true futility scenarios for the proposed method with the NKK method[11] and the standard conditional power for $t_X = 0.40$ and $t_Y = 0.10$: . The cut-off values were determined based on a maximum 1% power reduction under SUP.2 and SUP.5. Since FUT.1 is the futility scenario where $X = Y$, and thus $\log OR_1 = \log OR_0 = \infty$, the results are only displayed for the assumed log odds ratio. Similar for FUT.3 where we only have a (very) strong odds ratio.

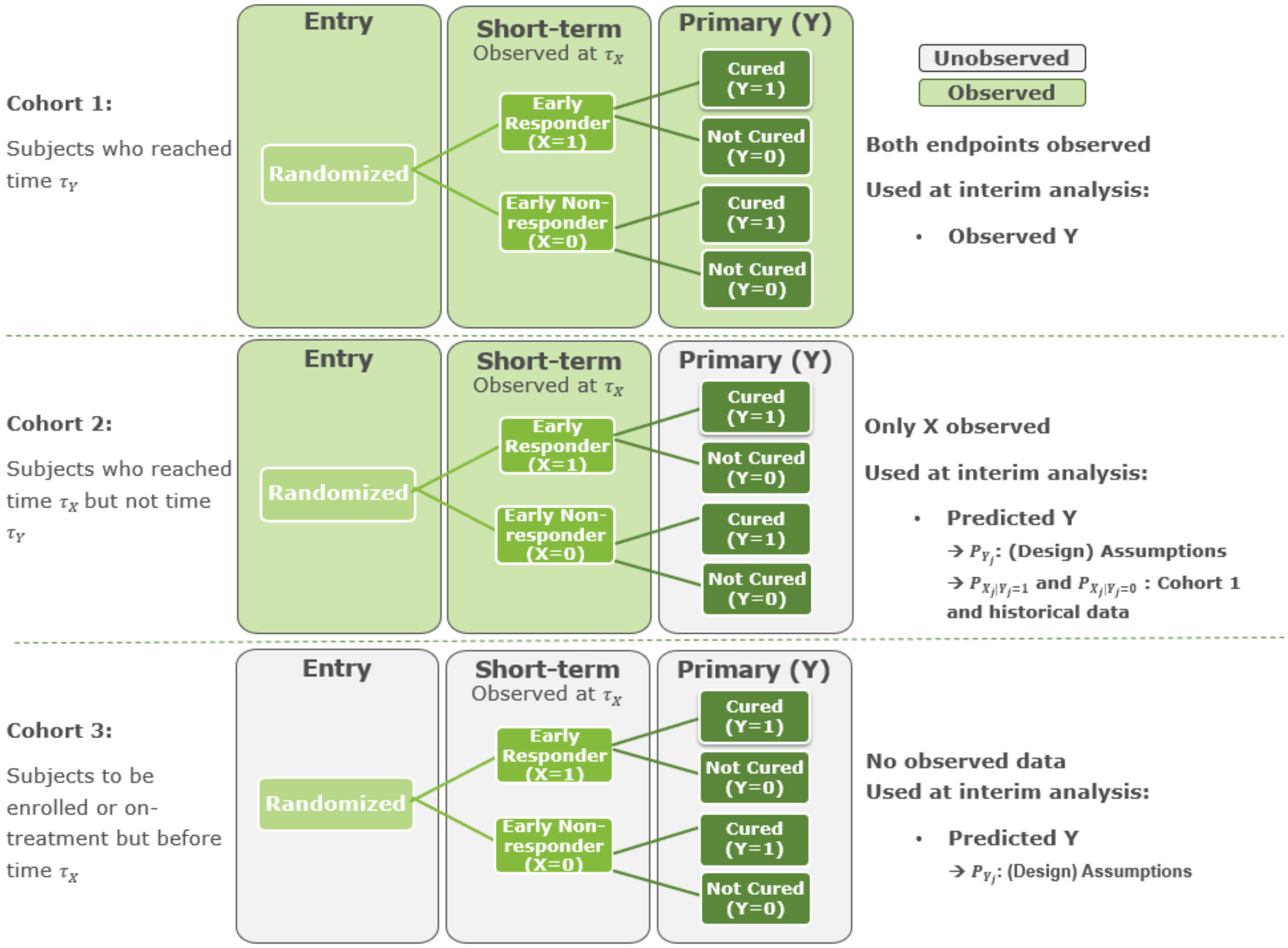

FIGURE 1 Available data at the interim analysis. Note that "cured" and "not cured" correspond with $Y = 1$ and $Y = 0$, respectively.

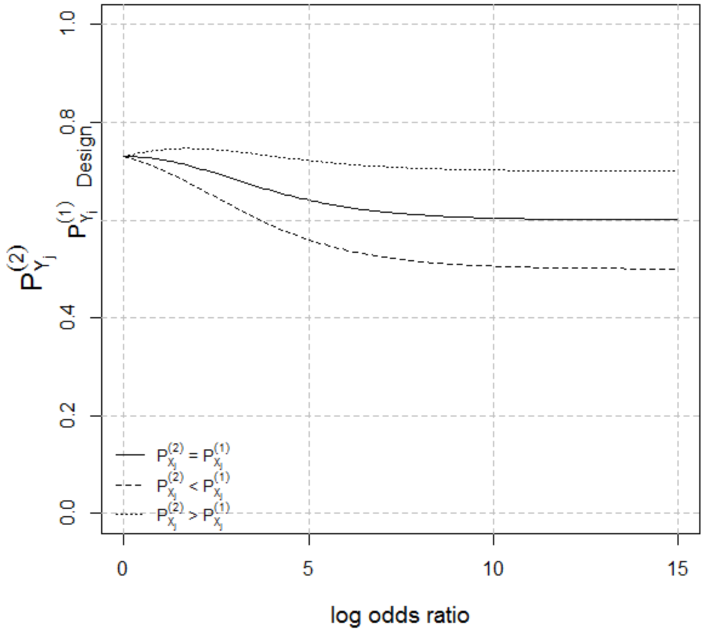

FIGURE 2 Change in predicted $P_{Y_j}^{(2)} = P_{Y_j|X_j=1}^{d} P_{X_j}^{(2)} + P_{Y_j|X_j=0}^{d} \left(1 - P_{X_j}^{(2)}\right)$ when $P_{Y_j}^{(1)}$ is smaller than the design assumption as a function of the strength of the association $\log\left(\frac{P(Y=1|X=1,A=j,C=1)P(Y=0|X=0,A=j,C=1)}{P(Y=0|X=1,A=j,C=1)P(Y=1|X=0,A=j,C=1)}\right)$, where $C = 1$ for subjects in cohort 1.

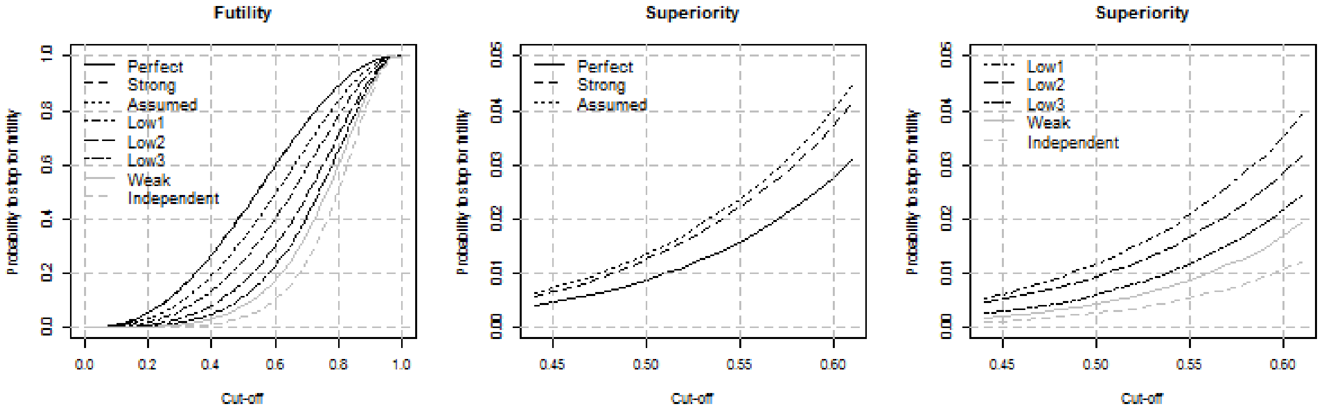

FIGURE 3 Probability to stop for futility under FUT.2 and SUP.2 for different strengths of the association between X and Y in function of cut-off points using the expected conditional power.

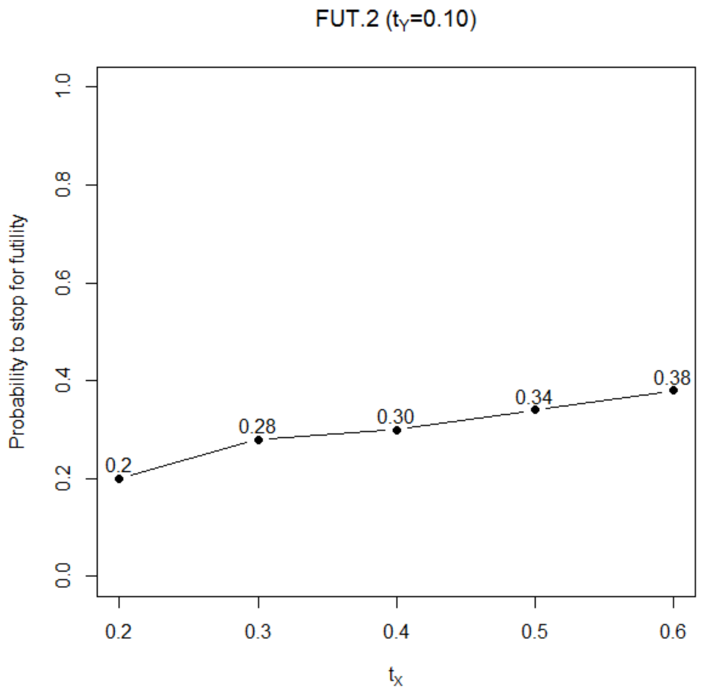

FIGURE 4 Probability to stop for futility for varying choices of t_X and fixed $t_Y = 0.10$ under scenario FUT.2 (assumed log odds ratio) for the cut-off value that results in a power loss of at most 1% under scenario SUP.2.

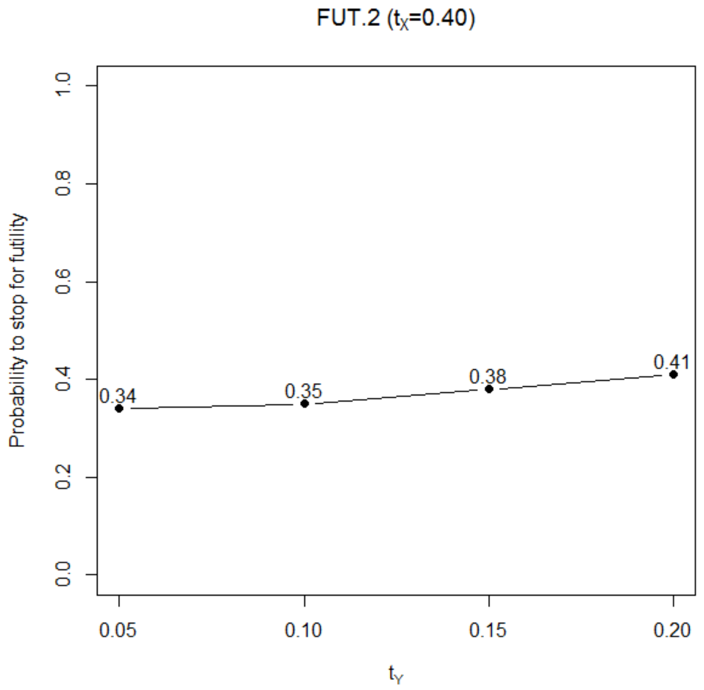

FIGURE 5 Probability to stop for futility for varying choices of t_Y and fixed $t_X = 0.40$ under scenario FUT.2 (assumed log odds ratio) for the cut-off value that results in a power loss of at most 1% under scenario SUP.2.