# Over-accrual in Bayesian adaptive trials with continuous futility stopping

Peer-reviewed author version

GARCIA BARRADO, Leandro & BURZYKOWSKI, Tomasz (2023) Over-accrual in Bayesian adaptive trials with continuous futility stopping. In: Clinical trials, 20 (3) , p. 252-260.

DOI: 10.1177/17407745231154685
Handle: http://hdl.handle.net/1942/39767

# Over-accrual in Bayesian adaptive trials with continuous futility stopping

Leandro Garcia Barrado[1,2] and Tomasz Burzykowski[1,2,3]

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

[2]International Drug Development Institute (IDDI), Belgium

[3]Department of Statistics and Medical Informatics, Medical University of Bialystok, Poland

**Corresponding author:**

Leandro Garcia Barrado, I-BioStat, Hasselt University, Agoralaan, B-3590 Diepenbeek, Belgium

Email: leandro.garciabarrado@uhasselt.be

**Abstract**

*Background* We explore frequentist operating characteristics of a Bayesian adaptive design that allows continuous early stopping for futility. In particular, we focus on the power versus sample size relationship when more patients are accrued than originally planned.

*Methods* We consider the case of a phase II single-arm study and a Bayesian phase II outcome-adaptive randomization design. For the former, analytical calculations are possible; for the latter, simulations are conducted.

*Results* Results for both cases show a decrease of power with an increasing sample size. It appears that this effect is due to the increasing cumulative probability of incorrectly stopping for futility.

*Conclusion* The increase in cumulative probability of incorrectly stopping for futility is related to the continuous nature of the early stopping, which increases the number of interim analyses with accrual. The issue can be addressed by, for instance, delaying the start of testing for futility, reducing the number of futility tests to be performed or by setting stricter criteria for concluding futility.

**Keywords**

Bayesian statistics, futility stopping, outcome-adaptive randomization

**Background**

Adaptive designs are becoming more often applied in randomized clinical trials (RCT)[1,2]. The U.S. Food and Drug Administration (FDA) defines an adaptive design as "a clinical trial design that allows for prospectively planned modifications to one or more aspects of the design based on accumulating data from subjects in the trial"[3].

Within the Bayesian framework, continuous monitoring and updating of available information is handled in an intuitive way. Thus, the framework is attractive for implementation of adaptive clinical trial designs. Bayesian adaptive designs that allow for early stopping have been proposed with high expectations and superior performance, as compared to their frequentist counterparts[4,5]. However, several issues with these designs have already been identified as well[6,7,8].

Bayesian adaptive designs with stopping rules are generally not concerned with the effect of monitoring treatment outcomes on the type-I error probability or power. However, for regulatory agencies such as the FDA, the frequentist properties of Bayesian designs are important[3,9]. Thus, properties of such designs should be investigated[10,11,12].

In Bayesian designs with futility stopping and a binary response, it is common to test the futility and efficacy of a particular treatment by applying a two-thresholds testing strategy, in which an (unacceptably) low threshold for the response probability is used to decide about efficacy and a (desirably) high threshold is applied to decide about futility[4,13]. The two thresholds are used in defining the hypotheses of efficacy and futility, respectively. Bayesian hypothesis-tests are then used during the trial to decide whether randomization to a particular treatment arm should be stopped. This stopping strategy was applied in the BATTLE trial[14,15], as well as more recently by Barry et al.[16] and Gu et al.[17].

A particular design is characterized by outcome-adaptive randomization and involves, next to the possibility of continuous stopping for futility, updating the randomization rates of enrolled subjects based on the results of the subjects already included in the trial. The design can be extended by

considering different strata of patients based on, for instance, biomarkers. Hereby, it becomes possible to assign patients within a particular stratum to the most promising treatment arm(s) during the course of the trial, while allowing to stop the trial early[14].

In this paper, we show that Bayesian adaptive designs with continuous futility stopping may lead to an undesirable relationship between power and sample size when accruing more patients than originally planned. We further investigate and confirm this issue by simulations considering the Bayesian outcome-adaptive randomization design with continuous futility stopping proposed by Barry et al.[16].

**Methods**

*Single-arm setting*

To set the scene, we consider a single-arm phase-II design with a single treatment and continuous Bayesian stopping for futility for a binary response. Assume that the number of responses $X_N$ among $N$ patients receiving the treatment is binomially-distributed with response-probability $\pi$. We want to test whether $\pi$ exceeds some unacceptable level $\pi_0$. Therefore, a Bayesian hypothesis test is considered based on the posterior distribution $P(\pi|x_N, N)$ of $\pi$ after observing $x_N$ responses. Hereto, consider the efficacy indicator $E$ defined as follows:

$$E = \begin{cases} 1 \; if \; P(\pi \geq \pi_0 | x_N, N) > \delta_E, \\ 0 \; otherwise, \end{cases} \qquad (1)$$

where $(1 - \delta_E)$ is a pre-defined size of a one-sided credible set. When $E = 1$, probability $\pi$ is deemed statistically significantly larger than $\pi_0$.

Moreover, to stop the trial early for futility, continuous monitoring of the posterior distribution $P(\pi|x_n, n)$ after every patient is applied during the trial. Let $n_0$ be the minimum number of patients required to be enrolled in the trial before allowing early stopping. Treatment is considered futile when $\pi$ is smaller than some desirable (target) response probability $\pi_1$. A Bayesian hypothesis test for futility can then be defined through the futility indicator $F_n$:

$$F_n = \begin{cases} 1 \; if \; P(\pi \geq \pi_1 | x_n, n) \leq \delta_F, \\ 0 \; otherwise. \end{cases} \qquad (2)$$

Denote by $(1 - \delta_F)$ the pre-defined size of a one-sided credible set. When $F_n = 1$, for $n = n_0, \dots, N$, the trial is stopped for futility after accrual of $n$ patients.

Assuming a beta-distribution, $Beta(\alpha_p, \beta_p)$, as a prior for $\pi$ leads to the following posterior Beta distribution of $\pi$:

$$P(\pi|x_n, n) = Beta\big(\alpha_p + x_n, \beta_p + n - x_n\big).$$

By varying values of $\delta_E$ and $\delta_F$, one can vary the frequentist trial-operating characteristics. Denote by $P(E = 1|\pi = \pi_0)$ and $P(E = 1|\pi = \pi_1)$ the frequentist type-I error probability and power, respectively. Other operating characteristics of interest include the cumulative probability of correctly and falsely stopping the trial early, i.e., $\sum_{i=1}^{n} P(F_i = 1|\pi = \pi_0)$ and $\sum_{i=1}^{n} P(F_i = 1|\pi = \pi_1)$, respectively. Note that, in this case, direct calculations of the operating characteristics are feasible.

*Bayesian biomarker-driven outcome-adaptive randomization design*

To evaluate over-accrual in a more complex setting, we consider a phase-II trial setting with Bayesian biomarker-based outcome-adaptive randomization allowing to stop early for futility, as proposed by Barry et al.[16], hereafter referred to as *Barry design*. In the proposed design, patients are divided into $S + 1$ mutually exclusive and exhaustive biomarker-based strata. The objective of the trial (see Appendix A in Supplementary Materials) is to evaluate the efficacy of $T + 1$ treatments within each stratum by using a binary clinical outcome (response). A Bayesian futility hypothesis test is performed during the adaptive-randomization phase of the trial. At the end of the trial, after a final test of futility, a Bayesian hypothesis test is performed to obtain the conclusion regarding efficacy. Toward this aim, the futility and efficacy indicators in equations (1) and (2) are updated to account for all stratum-treatment combinations (see Appendix B in Supplementary Materials).

The trial begins with equal randomization within the considered strata. For every recruited patient, first, biomarker status is established. Subsequently, the patient is randomized to treatments available in the biomarker-stratum to which the patient belongs and the response of the patient is observed. When a predetermined number of patients, $n_0$, has been accrued, the updated data are used in a test of futility which may result in termination of accrual to one or more treatments in various strata. Subsequently, the randomization ratios are updated and used for the next patient. When the targeted number of patients has been reached, the trial is terminated and the data are used to conduct a final test of futility, and a test of efficacy. Because the futility and efficacy decisions are based on thresholds defined by $\pi_1$ and $\pi_0$, respectively, there is a probability of ending up with a set of observations

satisfying both criteria. One conservative way to avoid this issue is by favouring futility, i.e., including the final test of futility before testing for efficacy. Note that the trial can also stop before reaching the targeted sample size if all stratum-treatment groups have become closed for accrual based on the results of the futility test. In that case, no efficacy tests are conducted at the time of closing accrual to the last group.

*Simulation study*

To evaluate the effect of over-accrual, we consider direct calculations for the single-arm setting and conduct a simulation study for the Bayesian biomarker-based outcome-adaptive randomization design.

*Single-arm setting* We assume $\pi_0 = 0.25$, $\pi_1 = 0.5$, and $n_0 = 5$. Moreover, we set the prior-distribution parameters $\alpha_p$ and $\beta_p$ both equal to 2.5 for the test of futility, and both equal to 1 for the test of efficacy.

Note that, for the test for futility, the prior $Beta(2.5,2.5)$ is centred around 0.5, the target response probability, with 95% of its mass between 0.12 and 0.88. This results in a 'conservative' prior distribution that reduces the probability of stopping for futility when the sample size is small, i.e., early in the trial. The choice is motivated by the idea that stopping for futility should only be considered when there is enough information in the data that suggests an absence of effect.

For the test of efficacy, the prior parameter values result in a uniform (flat) prior for $\pi$. The choice is motivated by the idea that stopping for efficacy should be guided as much as possible by the data.

Finally, we consider the null-hypothesis setting with $\pi = 0.25$ and the alternative-hypothesis setting with $\pi = 0.5$.

To control the type-I error probability at 10% and obtain a power of at least 80% for a targeted sample size of 20 patients, we set $\delta_E = 0.94$ and $\delta_F = 0.095$ (see also Appendix C in Supplementary

Materials). To evaluate the effect of over-accrual, we consider a final sample size up to 100 patients and investigate the relationship between power and sample size.

*Barry design* In the simulations, two different settings are considered. The first setting corresponds to the *simulation study* in Barry et al[16]. It considers general outcome-adaptive randomization with one binary biomarker. In essence, this is equivalent to two phase-II trials conducted in two strata with stopping for futility. The second simulation setting corresponds to the *real-life example* presented by Barry et al.[16]. It considers a trial aimed at the development of new, or an appropriate application of existing, breast cancer therapies directed by biomarker information.

In the first simulation setting, the prevalence of biomarker-positive patients is set equal to $0.5$. In terms of underlying true response probabilities $\pi_{st}$ of treatment $t$ in stratum $s$, a quantitative stratum-by-treatment interaction is considered. No difference in response probabilities for treatments in the biomarker-negative stratum is assumed ($\pi_{00} = \pi_{01} = 0.25$), while the experimental treatment is more efficacious in the biomarker-positive stratum ($\pi_{11} = 0.5$, $\pi_{10} = 0.25$). This setting is equivalent to the 'single-marker' scenario considered by Barry et al.[16].

The outcome-adaptive randomization stage is started after initial accrual of $n_0 = 25$ patients to ensure that, with high probability, at least two patients are assigned to each stratum-treatment combination before starting outcome-adaptive randomization and allowing the trial to stop early for futility. Following Barry et al.[16], we consider $\delta_F = 0.025$ and $\delta_E$=0.9. This results in a targeted sample size of $100$ patients that allows reaching a power of at least 0.8, while controlling the type-I-error probability at 0.1. To evaluate the effect of accruing more patients than initially planned, final sample sizes of $N \in \{25,50,75,100,125,150\}$ are considered.

The second simulation setting is based on the idea of designing a randomized phase II trial to evaluate a PI3K inhibitor therapy in advanced breast cancer patients. In the trial, four biomarker strata and two treatments (experimental and control) are considered, what leads to eight distinct stratum-treatment combinations (see Appendix D in Supplementary Materials).

In terms of the trial design characteristics, fixed 1:1 randomization within each stratum is considered until at least one patient is enrolled to each stratum-treatment combination. In this case, $n_0$ is random. The Bayesian hypothesis tests are defined by setting $\delta_F = 0.01$ and $\delta_E = 0.9$. As indicated in Barry et al.[16], this ensures a power of at least 0.9 while controlling the type-I error probability at 0.1 when a targeted sample size of $N = 168$ patients is considered. To evaluate the effect of over-accrual, final sample sizes of $N \in \{50,168,250,300,600\}$ are considered.

In both simulation settings, the considered target and unacceptable response rates, $\pi_1$ and $\pi_0$, are set at 0.5 and 0.25, respectively. Moreover, the parameters defining the prior distribution for $\pi_{st}$ are chosen based on the goal of the analysis (futility or efficacy), but independently of the simulation setting (see also Appendix E in Supplementary Materials).

For each final sample size, 1000 trials are simulated. In each simulation, the Gibbs sampler code developed by Barry et al.[16] was used. Based on the Raftery & Lewis diagnostic[18], 15,000 posterior samples were retained after a 15 iteration burn-in to achieve convergence for estimation of the required quantiles. Sufficiency of the number of burn-in iterations and non-informativeness of the initialisation was confirmed by inspection of randomly selected trace-plots and additional simulations (results not shown) with starting values sampled from U(-10,10). Computation time for one simulated trial with futility stopping ($N = 150$) was equal to about 4 hours on a 64-bit, 2.6 GHz, 8GB RAM machine using R 3.4.2 (x64)[19]. The R scripts can be found in Appendix F in Supplementary Materials.

**Results**

*Single-arm setting*

The results from the single-arm phase-II setting with continuous Bayesian stopping for futility are summarized in Figure 1. Panel a of Figure 1 shows the probability to conclude efficacy as a function of the considered final sample size. In the panel, results for both type-I error probability, based on the

null-hypothesis setting ($\pi = 0.25$), and power, based on the alternative-hypothesis setting ($\pi = 0.5$), are shown.

[Include Figure 1 about here]

From panel a of Figure 1 it can be seen that, for $\delta_E = 0.94$ and $\delta_F = 0.095$, a final sample size of 20 patients would ensure the type-I error probability to remain below 10% and power about 80%. Hence, we assume the threshold values to be calibrated for the targeted sample size of 20 patients. Over-accruing beyond 20 patients, however, leads to a counterintuitive decrease in power. For example, accruing up to 100 patients would decrease the power from 80% to about 70%.

An explanation of the decrease of power due to over-accrual is offered in panel b of Figure 1. The plot shows the cumulative probability of stopping the trial early for futility. For a final sample size of 20 patients, the probability to correctly stop the trial under the null-hypothesis setting is about 90%. On the other hand, the probability to incorrectly stop the trial under the alternative-hypothesis setting is about 19%. However, the cumulative probability of stopping early for futility is a strictly increasing function of sample size. Thus, over-accruing subjects beyond 20 patients increases the probability to values larger than 20%. This implies that the achievable power at the end of the trial falls below 80%.

*Barry design*

For the Barry design, we focus on the (simulation-based) proportion of trials which end with a statistically significant efficacy conclusion. For the efficacious biomarker-treatment combinations, this proportion is an estimate of power. For inefficacious combinations, the proportion estimates the type-I error probability. We also report the proportion of trials in which a particular stratum-treatment arm was stopped due to futility before or at the final sample size. For the efficacious stratum-treatment combinations, this proportion estimates the cumulative probability of incorrectly stopping for futility.

*Simulation-study setting* Panel a of Figure 2 shows that, for a trial with a final sample size of about 100 patients, the power to correctly conclude efficacy in the efficacious stratum-treatment combination

(green curve) is about 85%. On the other hand, for the inefficacious stratum-treatment combinations (red curves), the type-I error probability is at most 10%. Therefore, as noted by Barry et al.[16], for the required power and type-I-error-probability control objectives, the proposed threshold values imply a targeted sample size of 100 patients. However, accruing 150 patients decreases the power to about 80%, while decreasing the type-I error probability to about 2.5%.

[Include Figure 2 about here]

Panel b of Figure 2 shows that, for the inefficacious stratum-treatment combinations, the probability of stopping for futility is higher as compared to the efficacious combinations. For instance, in the biomarker-positive (S=1) stratum, the probability of stopping the inefficacious control treatment (T=0) before or at accruing exactly 100 patients is about 55%. The probability to incorrectly stop the trial early for an efficacious treatment-stratum combination increases with sample size. Over-accruing beyond 100 patients increases the probability to just below 20% at 150 patients, when 150 patients are accrued. As a result, and similarly to the situation observed for the single-arm design, the maximally achievable power decreases to around 80%.

*Real-life example setting* The results for the real-life example simulation setting show the same trends as observed for the single arm and simulation study setting. Although a power of 90% is reached with the targeted 168 patients, over-accrual eventually leads to a power lower than 90%. Also, over-accruing patients to about 600 patients increases the cumulative probability of incorrectly stopping the trial for futility to just above 10%. With, as a result, reduction of the maximally achievable power at the end of the trial (see also Appendix G in Supplementary Materials).

*Reducing the decrease in power* The trends observed in panels a of Figures 1 and 2 imply that, paradoxically, accruing more patients to a trial may lead to a reduction of power. To alleviate this problem, one should reduce the cumulative probability of incorrectly stopping the trial early at larger sample sizes. Towards this aim, one could investigate different choices of the prior-distribution parameters ($\alpha_p$ and $\beta_p$) or the stopping rule for futility ($\delta_F$). Changing the former is not evident if

there is little knowledge about the true value of $\pi$. A more feasible approach is to investigate different values of $\delta_F$ and select those that may lead to acceptable values of the cumulative probability of incorrectly stopping early, as well as early stopping of inefficacious treatments.

To illustrate the approach, Figure 3 shows operating characteristics of the single-arm design with $\delta_F = 0.01$. Panel a of Figure 3 shows a slight increase of the type-I error probability (red curve) compared to the case of $\delta_F = 0.095$ (see panel a of Figure 1). The probability of stopping for futility in case of an inefficacious treatment (red curve in panel b of Figure 3) is affected to a greater extent as it decreases from about 90% to 50% for a trial with a final sample size of 20 patients, as compared to panel b of Figure 1. Decreasing $\delta_F$ reduces the cumulative probability of incorrectly stopping for futility (green curve in panel b of Figure 3) and decreases its rate of increase with sample size. As a result, the maximally achievable power becomes closer to 100% and shows only a slightly decreasing trend for larger values of the final sample size (green curve in panel a of Figure 3).

[Include Figure 3 about here]

The cumulative probability of incorrectly stopping for futility also depends on the number of patients $n_0$ accrued before the first futility test and on the total number of times the futility test is performed before testing for efficacy. To illustrate the impact of these factors, operating characteristics of two different variations of the single-arm scenario, assuming $\delta_F = 0.095$, were calculated.

First, we increase $n_0$ from 5 to 15. Panel b of Figure 4 shows that the cumulative probability of incorrectly stopping for futility for a final sample size of 100 patients slightly decreases, as compared to panel b of Figure 1. Consequently, the decrease in power, shown in panel a of Figure 4, is also reduced, as compared to the decrease in panel a of Figure 1.

[Include Figure 4 about here]

Second, Figure 5 summarizes the results when the test for futility is only performed after every 10 patients. Compared to the setting when futility is tested after every patient (see Figure 1), the

cumulative probability of incorrectly stopping for futility decreases from around 0.35 (panel b of Figure 1) to below 0.3 (panel b of Figure 5) for the final sample size of 100 patients.

[Include Figure 5 about here]

**Discussion and conclusions**

The presented results for a single-arm phase-II design with continuous Bayesian early futility stopping and a Bayesian biomarker-driven outcome-adaptive randomization design allowing for early stopping for futility indicate a counterintuitive (from a "classical" fixed-sample-size trial-design point of view) decrease in power in case of over-accrual. The decrease of power is due to the increasing cumulative probability of incorrectly stopping for futility implied by the repeated futility testing. When additional patients are considered, their outcomes may lead to posterior distributions for which the futility test criterion will be satisfied. Hence, for any additional patient, there is at least some probability to stop the trial for futility.

The cumulative stopping probability depends on the assumed prior distributions and the considered hypothesis-test criterion $\delta_F$. By considering smaller values for the latter, the increase in the cumulative probability of stopping can be reduced. Unfortunately, this also reduces the probability of correctly stopping for futility of an inefficacious treatment.

Higher probabilities of stopping early at the beginning of the trial will propagate to the cumulative stopping probability until the end of the trial. Delaying the first test of futility until more patients have been accrued, as well as decreasing the total number of futility tests to be performed, helps to decrease the cumulative stopping probability at the end of the trial.

The sample size at which power will start decreasing depends on the combination of the cumulative probability of stopping early and the power to conclude efficacy conditional on reaching a particular sample size. For example, in the real-life example setting, accruing beyond the targeted sample size of 168 patients initially shows an increase in power (see Figure G.1 in Appendix G in Supplementary Materials). This occurs when the combination of the probability of reaching a particular final sample size and the power conditional on reaching that sample size, exceeds the targeted power. From Figure H.2 in Appendix H in Supplementary Materials, one can see that, after an initial increase in power, the

probability of reaching the sample size of interest decreases to the point that, when combined with the conditional power, the unconditional power starts to decrease.

The real-life example setting, considered in this manuscript, corresponded to the example analysed by Barry et al.[16]. In that example, Barry et al. additionally capped the maximum number of patients in each treatment arm to avoid oversampling. Moreover, a lag was considered for futility testing and updating the randomization ratios to account for the length of follow-up needed to observe responses. Although these measures were not introduced by Barry et al. for that purpose, they may alleviate the decreasing power issue. Capping the maximum sample size for each treatment ultimately limits the total sample size that could be over-accrued. On the other hand, introducing a lag in futility testing delays the start of testing and, hence, reduces the cumulative probability of incorrectly stopping for futility.

Of course, power always depends on the true response probability (see Appendix I in Supplementary Materials). Assuming, for the purposes of designing a trial, a larger-than-the-true value of this probability may aggravate the decrease in power due to over-accrual. On the other hand, using a smaller-than-the-true value may help in addressing the issue. Unfortunately, the true probability is never precisely known. This implies that the risk of power loss due to over-accrual may always be present when designing a Bayesian outcome-adaptive randomization trial.

As noted by the reviewers of this paper, the aforementioned procedures to reduce the impact of the decrease in power are viable if unintentional over-accrual is anticipated at the design stage of the trial. When over-accrual is purposely considered after reaching the target sample size, other procedures could be contemplated. This type of over-accrual could be foreseen, for example, based on post-hoc external information[20]. In such cases, the trial's test strategies and/or its thresholds could be adapted to obtain a particular power conditional on having reached the target sample size (see Appendix H in Supplementary Materials). In practice, adapting the test's thresholds would follow the same approach as at the design stage of the trial.

In all investigated settings, the impact of the continuous monitoring on power seemed limited, as considerable over-accrual was required before the decrease in power would become meaningful from a practical point of view. However, the impact depends on the futility stopping and efficacy criteria, as well as the underlying true response probabilities. As the latter are never precisely known, it is not prudent to dismiss a priori the probability of a substantial power decrease in any case. Moreover, in the context of a particular early-phase setting like, for instance, expansion cohort trials, substantial increases in sample size are not uncommon[20].

The over-accrual issue, considered in this paper, is different from the concept of "overrunning" introduced in the context of, for instance, group-sequential clinical trials[21]. Given the nature of the considered designs, the time between accrual and available data should be relatively short to inform the outcome-adaptive randomization. Therefore, the issue of taking a decision while other data are being collected is less relevant. The setting considered in our manuscript can be described as delaying the stop of accrual and performing the test for efficacy past the sample size fixed at the design of the trial.

In conclusion, we have shown that the choice of a hypothesis-test strategy in a Bayesian biomarker-based outcome-adaptive randomization trial with stopping for futility may result in a decrease in power when sample size exceeds the sample size for which the trial was powered at. The strength of dependence between the magnitude of the decrease and the amount of over-accrual will be setting-specific. This is an undesired effect that should be kept in mind when designing any Bayesian outcome-adaptive randomization design, as well as during the trial when considering accruing patients beyond the planned sample size.

**Acknowledgements**

The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Hercules Foundation and the Flemish Government – department EWI.

**References**


1. Kairalla JA, Coffey CS, Thomann MA, et al. Adaptive designs: a review of barriers and opportunities. *Trials* 2012; 13: 145. https://doi.org/10.1186/1745-6215-13-145.

2. Mistry P, Dunn JA, Marshall A. A literature review of applied adaptive design methodology within the field of oncology in randomised controlled trials and a proposed extension to the CONSORT guidelines. *BMC Med Res Methodol* 2017; 17:108. https://doi.org/10.1186/s12874-017-0393-6.

3. U.S. Food and Drug Administration. Adaptive design clinical trials for drugs and biologics guidance for industry. Issued in Nov., 2019. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry

4. Berry SM, Carlin BP, Lee JJ, et al. *Bayesian adaptive methods for clinical trials.* Boca Raton: CRC Press, 2011.

5. Meinzer C, Martin R, Suarez JI. Bayesian dose selection design for a binary outcome using restricted response adaptive randomization. *Trials* 2017; 18: 420. https://doi.org/10.1186/s13063-017-2004-6.

6. Korn EL, Freidlin B. Commentary on Hey and Kimmelman. *Clin Trials* 2015; 12: 122-124. https://doi.org/10.1177%2F1740774515569611.

7. Thall P, Fox P, Wathen J. Statistical controversies in clinical research: Scientific and ethical problems with adaptive randomization in comparative clinical trials. *Ann Oncol* 2015; 26: 1621-1628. https://doi.org/10.1093/annonc/mdv238.

8. Korn EL, Freidlin B. Adaptive clinical trials: Advantages and disadvantages of various adaptive design elements. *J Natl Cancer I* 2017; 109: 1-6. https://doi.org/10.1093/jnci/djx013.

9. U.S. Food and Drug Administration. Guidance for the use of Bayesian statistics in medical device trials. Issued on Feb. 5, 2010. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-use-bayesian-statistics-medical-device-clinical-trials

10. Emerson SS, Kittelson JM, Gillen DL. Frequentist evaluation of group sequential clinical trial designs. *Stat Med* 2007; 26: 5047-5080. https://doi.org/10.1002/sim.2901

11. Gsponer T, Gerber F, Bornkamp B, et al. A practical guide to Bayesian group sequential designs. *Pharm Stat* 2014; 13: 71-80. https://doi.org/10.1002/pst.1593

12. Ventz S, Trippa L. Bayesian Designs and the Control of Frequentist Characteristics: A Practical Solution. *Biometrics* 2015; 71: 218-226. https://doi.org/10.1111/biom.12226.

13. Thall P., Simon R. Practical Bayesian Guidelines for Phase IB Clinical Trials. *Biometrics* 1994; 50: 337-394. https://doi.org/10.2307/2533377.

14. Kim ES, Herbst RS, Wistuba II, et al. The BATTLE trial: Personalizing Therapy for Lung Cancer. *Cancer Discov* 2011; 1: 44-53. https://doi.org/10.1158/2159-8274.CD-10-0010.

15. Zhou X, Liu S, Kim ES, et al. Bayesian adaptive design for targeted therapy development in lung cancer – a step toward personalized medicine. *Clin Trials* 2008; 5: 181-193. https://doi.org/10.1177/1740774508091815.

16. Barry WT, Perou CM, Marcom PD, et al. The use of Bayesian hierarchical models for adaptive randomization in biomarker-driven phase II studies. *J Biopharm Stat* 2015; 25: 66-88. https://doi.org/10.1080/10543406.2014.919933.

17. Gu X, Chen N, Wei C, et al. Bayesian two-stage biomarker-based adaptive design for targeted therapy development. *Stat Biosci* 2016; 8: 99-128. https://doi.org/10.1007/s12561-014-9124-2.

18. Raftery AE, Lewis SM. How many iterations in the Gibbs sampler? In Bayesian Statistics 4 (J. 0. Berger, J. M. Bernardo, A. P. Dawid and A. F. M. Smith, eds.) 763-773. Oxford Univ. Press, 1992.

19. R Core Team. R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. 2017; URL http://www.R-project.org.

20. Mehra R, Seiwert, TY, Gupta, S, et al. Efficacy and safety of pembrolizumab in recurrent/metastatic head and neck squamous cell carcinoma: pooled analyses after long-term follow-up in KEYNOTE-012. *Br J Cancer 2018,* 119**:** 153–159. https://doi.org/10.1038/s41416-018-0131-9

21. Whitehead J. Overrunning and underrunning in sequential clinical trials. *Control Clin Trials* 1992; 13: 106-121.

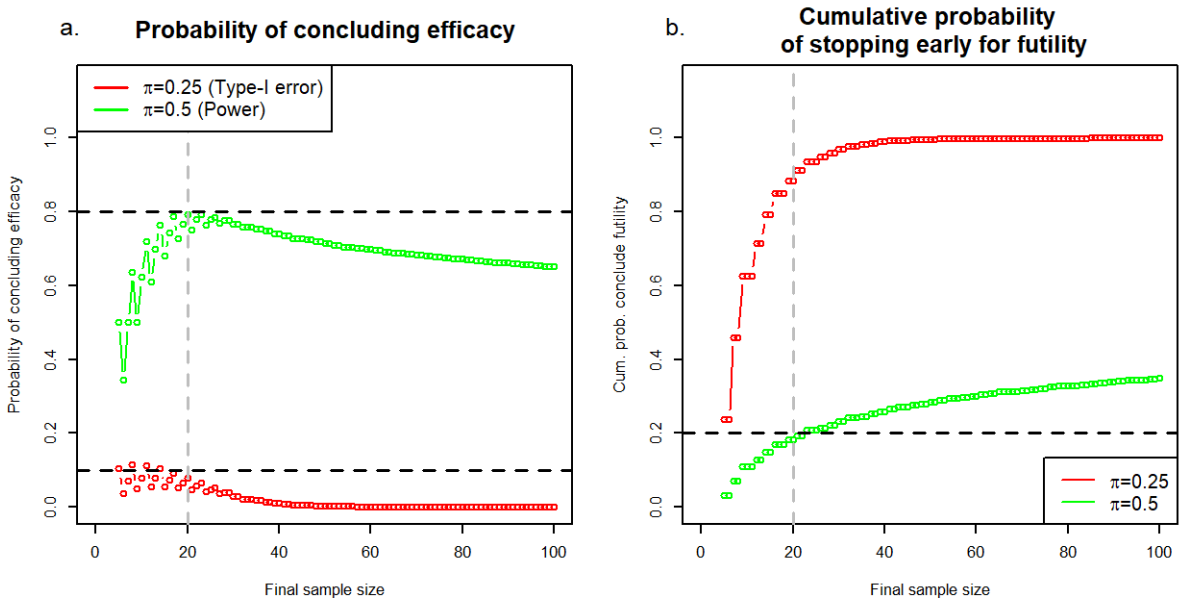


Figure 1: Operating characteristics for a single-arm trial with a single treatment designed for a targeted sample size of 20 patients (grey dashed line). a. Probability of correctly (power) and incorrectly (type-I error) concluding efficacy at the end of the trial as a function of the considered final sample size of the trial. b. Cumulative probability of incorrectly (green) and correctly (red) stopping for futility as a function of the considered final sample size. The alternative-hypothesis setting ($\pi = 0.5$) is denoted in green, the null-hypothesis setting ($\pi = 0.25$) is shown in red. Horizontal dashed lines indicate desired operating characteristics; type-I error probability of 0.1 and power of 0.8.

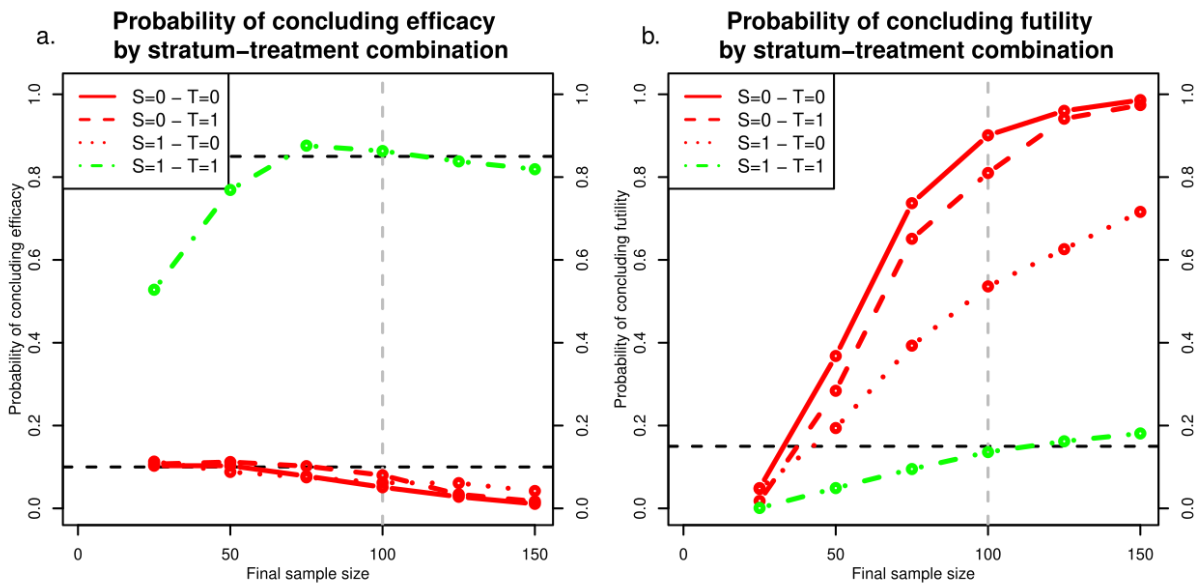


Figure 2: Operating characteristics of the simulation-study setting of the phase II Bayesian biomarker-based outcome-adaptive randomization design with a continuous futility stopping for a targeted sample size of 100 patients (grey dashed line). a. Proportion of trials concluding efficacy for each of the stratum-treatment combination as a function of the considered final sample size. b. Cumulative proportion of trials stopping early for futility for each stratum-treatment combination as a function of the considered final sample size. Efficacious stratum treatment combination ($\pi_{st} = 0.5$) marked in green, inefficacious combinations ($\pi_{st} = 0.25$) in red. S, stratum; T, treatment. Horizontal dashed lines indicate desired operating characteristics; type-I error probability of 0.1 and power of 0.85.

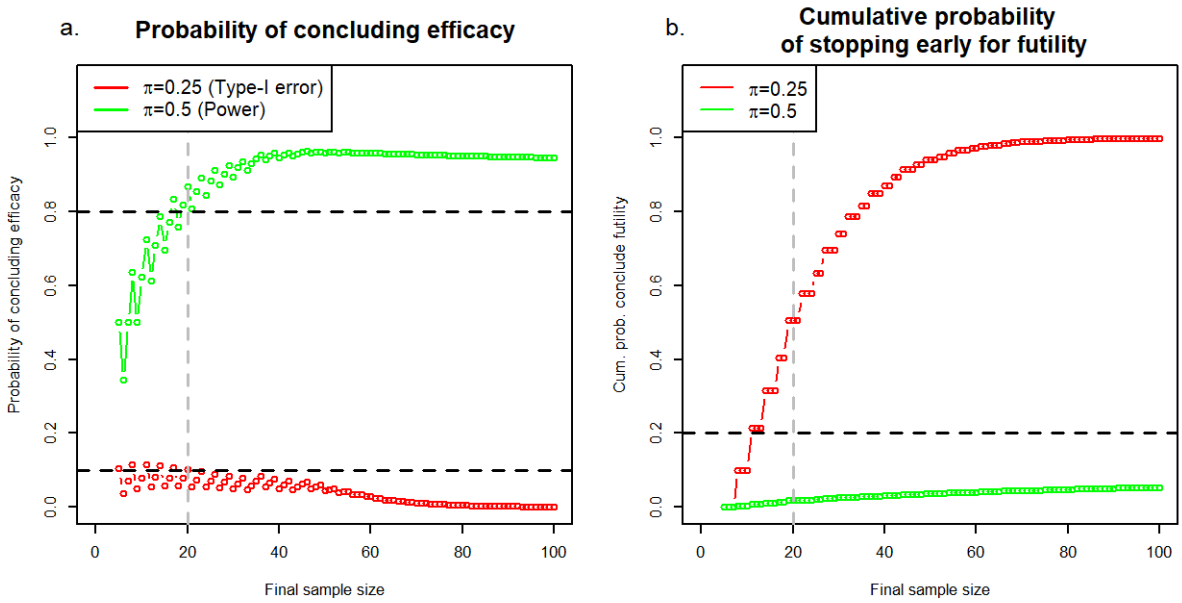


Figure 3: Operating characteristics for the single-arm setting with $\delta_F = 0.01$ for a targeted sample size of 20 patients (grey dashed line). a. Probability of correctly (power) and incorrectly (type-I error) concluding efficacy at the end of the trial as a function of the final sample size. b. Cumulative probability of incorrectly (green) and correctly (red) stopping for futility as a function of the final sample size. The alternative-hypothesis setting ($\pi = 0.5$) is denoted in green, the null-hypothesis setting ($\pi = 0.25$) is shown in red. Horizontal dashed lines indicate desired operating characteristics; type-I error probability of 0.1 and power of 0.8.

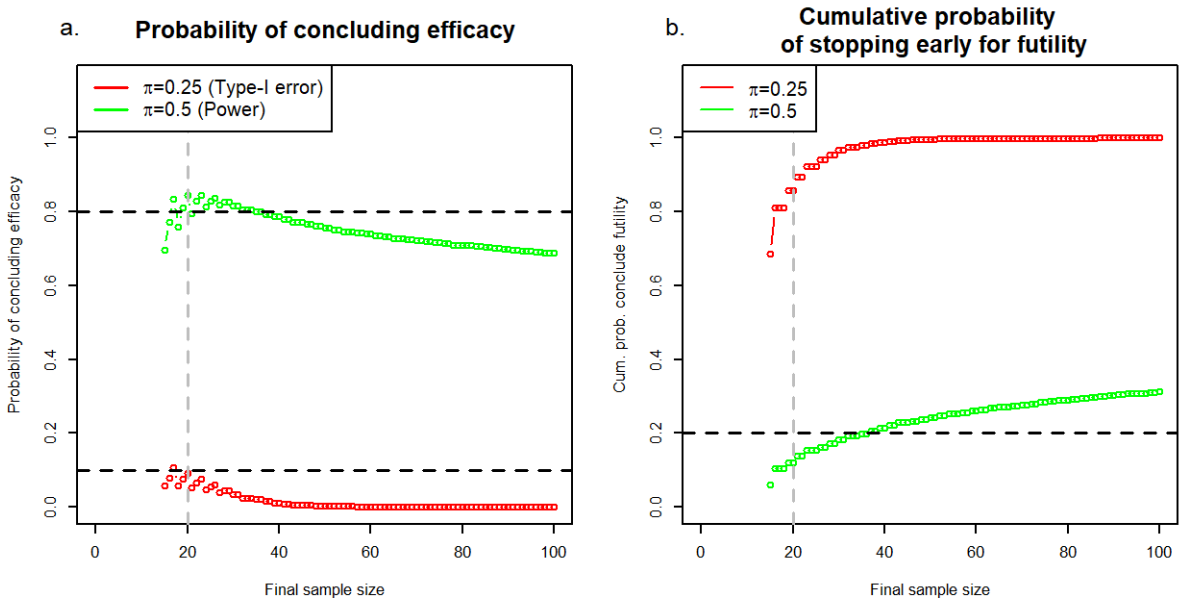


Figure 4: Operating characteristics for the single-arm setting for a targeted sample size of 20 patients (grey dashed line) with $n_0 = 15$. a. Probability of correctly (power) and incorrectly (type-I error) concluding efficacy at the end of the trial as a function of the final sample size. b. Cumulative probability of incorrectly (green) and correctly (red) stopping for futility as a function of the final sample size. The alternative-hypothesis setting ($\pi = 0.5$) is denoted in green, the null-hypothesis setting ($\pi = 0.25$) is shown in red. Horizontal dashed lines indicate desired operating characteristics; type-I error probability of 0.1 and power of 0.8.

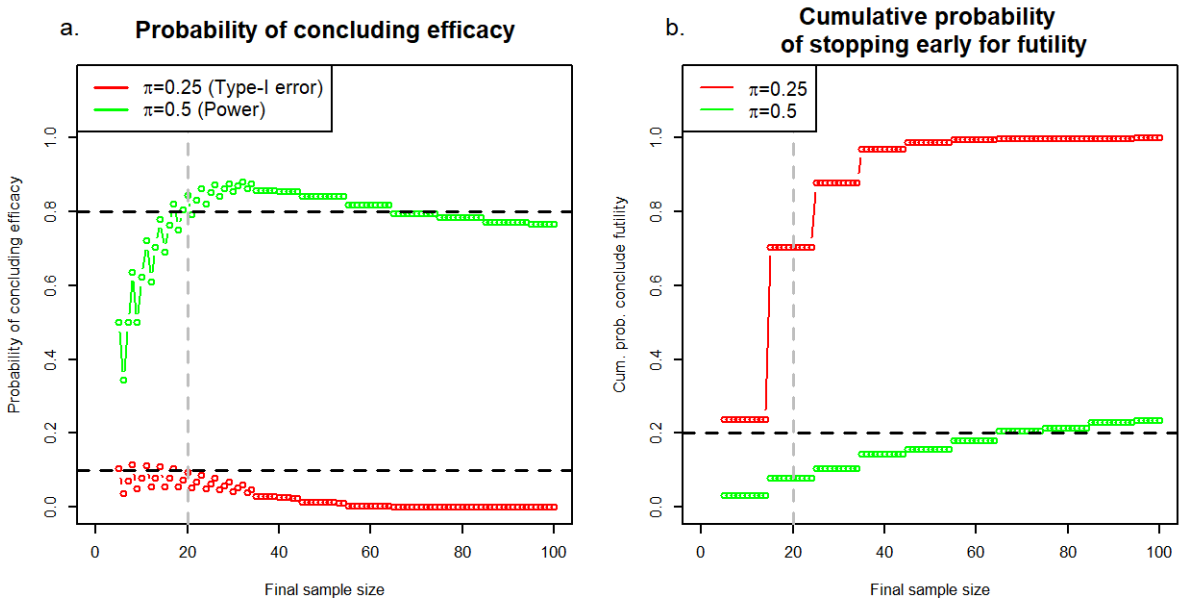


Figure 5: Operating characteristics for the single-arm setting for a targeted sample size of 20 patients (grey dashed line) with the futility hypothesis only tested after every 10 patients. a. Probability of correctly (power) and incorrectly (type-I error) concluding efficacy at the end of the trial as a function of the final sample size. b. Cumulative probability of incorrectly (green) and correctly (red) stopping for futility as a function of the final sample size. The alternative-hypothesis setting ($\pi = 0.5$) is denoted in green, the null-hypothesis setting ($\pi = 0.25$) is shown in red. Horizontal dashed lines indicate desired operating characteristics; type-I error probability of 0.1 and power of 0.8.