Bayesian biomarker-driven outcome-adaptive randomization with an imperfect biomarker assay

GARCIA BARRADO, Leandro & BURZYKOWSKI, Tomasz (2021) Bayesian biomarker-driven outcome-adaptive randomization with an imperfect biomarker assay. In: Clinical trials, 18 (2) , p. 137 -146.

DOI: 10.1177/1740774520964202
Handle: http://hdl.handle.net/1942/33822

Bayesian biomarker-driven outcome-adaptive randomization with an imperfect biomarker-assay.

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

[1]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

Objective: We investigate the impact of biomarker-assay's accuracy on the operating characteristics of a Bayesian biomarker-driven outcome-adaptive randomization (OAR) design.

Methods: In a simulation study, we assume a trial with two treatments, two biomarker-based strata, and a binary clinical outcome (response). Denote by P_{bt} the probability of response for treatment t $(t = 0 \text{ or } 1)$ in biomarker-stratum $(b = 0 \text{ or } 1)$. Four different scenarios in terms of true underlying response probabilities are considered: a null $(P_{00} = P_{01} = P_{10} = P_{11} = 0.25)$ and consistent $(P_{00} = P_{10} = 0.25, P_{01} = P_{11} = 0.5)$ treatment-effect scenario, as well as a quantitative $(P_{00} = P_{01} = P_{10} = 0.25,\ P_{11} = 0.5)$ and a qualitative $(P_{00} = P_{11} = 0.5,\ P_{01} = P_{10} = 0.25)$ stratum-treatment interaction. For each scenario, we compare the case of a perfect with the case of an imperfect biomarker-assay with sensitivity and specificity of 0.8 and 0.7, respectively. In addition, biomarker-positive prevalence values $P(B = 1) = 0.2$ and 0.5 are investigated.

Results: Results show that the use of an imperfect assay affects the operational characteristics of the Bayesian biomarker-based OAR design. In particular, the misclassification causes a substantial reduction in power accompanied by a considerable increase in the type-I-error probability. The magnitude of these effects depends on the sensitivity and specificity of the assay, as well as on the distribution of the biomarker in the patient population.

Conclusion: With an imperfect biomarker-assay, the decision to apply a biomarker-based outcome-adaptive randomization design may require careful reflection.

Keywords

Bayesian statistics, outcome-adaptive randomization, imperfect assay, biomarkers

Introduction

Bayesian outcome-adaptive randomization (OAR) designs for clinical trials are becoming popular.[1] While traditional designs for two-arm trials consider a fixed, e.g., 1:1 or 2:1 randomization ratio, OAR designs make use of the outcome information obtained for patients already included in the trial to continuously update the ratio. Since this generally results in more patients being assigned to the 'more promising' treatment, the adaptation is suggested to increase patient-specific benefits in clinical trials.[2,3]

Other extensions of the traditional randomized clinical trials are 'targeted'[4] and 'stratified'[5] designs. In 'targeted' designs, patients are pre-screened by using, e.g., biomarkers. Patients with a specific biomarker status are then selected and randomized to treatments which would be deemed the most promising for the status. Of course, this assumes that the mode of action of the treatment under investigation is well-known and that a biomarker assay is available. In case the assay misclassifies patients, the trial may suffer from a considerable loss of efficiency.[4] 'Stratified' biomarker designs randomize patients within each biomarker group to identify group-specific treatments. In case patients are misclassified in a 'stratified' biomarker trial, a considerable loss of power may be observed.[5]

By using OAR in 'stratified' designs, it is possible to assign patients within a particular biomarker stratum to the most promising treatment(s) during the course of the trial.[6] Such combined designs have also been proposed to test the efficacy of a novel targeted treatment while simultaneously identifying predictive markers for the treatment.[7]

Advantages of Bayesian biomarker-driven OAR designs, as compared to the fixed-randomization-ratio designs have been reported. Among others, a reduced total sample size or a decrease in the variation of the accrued sample size have been discussed.[6,7,8] However, several issues with OAR designs have also been identified, including potential bias due to time trends in the prognostic characteristics of the patient population,[9] statistical inefficiency due to imbalance in the number of patients assigned to

different treatment arms,[10] and a non-trivial probability of ending up with a substantially larger number of patients on the worse treatment arm.[11]

Given the demonstrated impact of assay accuracy on the results of 'targeted' and 'stratified' designs, one can ask whether similar effects apply to biomarker-based OAR designs. In this paper, we attempt to investigate this issue by using a simulation study.

Methods

We consider a phase-II trial design with Bayesian biomarker-based OAR as proposed by Barry et al.[8], reminiscent of the methodology developed in the influential BATTLE trial.[12] In the proposed design, patients are stratified into S mutually exclusive and exhaustive biomarker-based strata. The objective is to evaluate the efficacy of T treatments within each stratum by using a binary clinical outcome.

Hierarchical probit model

Probability of response is modelled by using a probit model.[6,8] Let y_{ist} denote the response of patient i in stratum s treated with treatment t. Assume a latent, normally-distributed random variable $z_{ist} \sim N(\mu_{st}, 1)$ and let $y_{ist} = 1$ if $z_{ist} > 0$. Then the probability of response $P_{st} \equiv P(y_{ist} = 1) = P(z_{ist} > 0) = \Phi(\mu_{st})$, where $\Phi()$ is the standard-normal cumulative-distribution function.

Subsequently, we define a hierarchical model by specifying that

$$P(y_{ist} = 1) = \Phi(\mu_{st}),$$

$$\mu_{st} \sim N(\omega_t, \sigma^2),$$

$$\omega_t \sim N(\alpha, \tau^2),$$

with hyper-parameters α, σ^2, and τ^2. The prior distribution for μ_{st} allows 'borrowing' information across strata within each treatment, with the extent of 'borrowing' controlled by parameter σ^2. The prior distribution for ω_t allows 'borrowing' information across treatments, with τ^2 controlling the extent of 'borrowing'.

The assumed Gaussian priors are conjugate. Hence, the resulting conditional posteriors have a closed form. Consequently, as proposed by Barry et al.[8], Gibbs sampling can be used to estimate response probabilities P_{ist}.

Biomarker assay

In the remainder of the paper we assume availability of a binary biomarker. The true underlying biomarker-status of a patient is denoted by B, with possible values 0 and 1. The prevalence of biomarker-positive patients is $\theta \equiv P(B = 1)$. The biomarker-status of individual patients is established by using an assay A. Consequently, there are two assay-defined strata, indexed by $s = 0$ (assay-negative) and $s = 1$ (assay-positive). We allow the assay to be imperfect, i.e., to misclassify the biomarker-status of the patients. The accuracy of the assay is defined in terms of sensitivity

$$Se_A \equiv P(S = 1|B = 1)$$

and specificity

$$Sp_A \equiv P(S = 0|B = 0).$$

Trial design

Estimation of the hierarchical model by the Gibbs sampler requires that there is at least one patient in each stratum-by-treatment combination. Thus, initially, the fixed 1:1 randomization ratio is used during the accrual of the first n_0 patients. Afterwards, OAR is initiated. The randomization ratios within the strata are updated, following Barry et al.[8], by using the 'max-mapping' strategy. In particular, the randomization probability, r_{st}, is defined to be equal to the posterior probability that, in stratum s, treatment t is superior to all other treatments still under consideration in that stratum.

When a new patient is recruited in the OAR stage, the biomarker status of the patient is established by using the biomarker assay. Subsequently, the patient is randomized, using the current randomization ratios, to treatments available in the stratum to which the patient belongs. The response of the patient is observed and the updated data are used in a test for futility (Appendix A of the Supplementary Materials) which may result in irreversible suspension of one or more treatments in various strata. Subsequently, the randomization ratios are updated and used for the next patient to be accrued.

When the maximum number, N_{max} say, of accrued patients has been reached, the trial is terminated and the data are used to conduct a final test of futility. In case no significant futility result is obtained, a Bayesian test of the hypothesis of efficacy is performed (Appendix B of the Supplementary Materials). Note that the trial can also be stopped before reaching N_{max} 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

We consider a trial with two biomarker-based strata (biomarker-negative and -positive) and two treatments (control and experimental, say, indexed by 0 and 1, respectively). OAR is initiated after accruing $n_0 = 25$ patients. This choice ensures that, with a high probability, at least two patients are assigned to each stratum-treatment combination before starting OAR.

Four different scenarios are defined based on the value of the underlying true response probability P_{bt} of treatment t in biomarker-stratum b (see Table 1). In Scenario 1, the null setting, the response probability is unacceptably low for both treatments, i.e., $P_{bt} = \pi_0 = 0.25$. Scenario 2, the consistent treatment-effect setting, is characterised by the experimental treatment being efficacious in both strata, i.e., $P_{b1} = \pi_1 = 0.5$. In Scenario 3, there is a quantitative stratum-by-treatment interaction: both treatments are inefficacious in the biomarker-negative stratum, while the experimental treatment is efficacious in the biomarker-positive stratum. Finally, in Scenario 4, there is a qualitative stratum-by-treatment interaction: the control treatment is efficacious in the biomarker-negative stratum, while the experimental treatment is efficacious in the biomarker-positive stratum.

[Place Table 1 about here]

We also want to compare the case of a perfect biomarker-assay to the case of an imperfect assay. For the latter, the assay-accuracy is defined by $Se_A = 0.8$ and $Sp_A = 0.7$. These values are motivated by the diagnostic accuracy of the clinical diagnosis for Alzheimer's[13] or a PIK3CA mutation in breast cancer.[14] Additionally, we consider two biomarker-positive prevalence settings: $\theta = 0.5$ and 0.2.

Note that, depending on the assay accuracy, the actual prevalence of assay-positive patients and response probabilities may differ from the assumed true values. Table 2 presents 'actual' prevalences and response probabilities for each combination of the prevalence, assay-accuracy, and true response-probability scenario (see Table 1). It can be observed that, for the imperfect assay, the assay-positive prevalence $P(S = 1)$ is larger than true prevalence θ. This is due to the fact that the assay misclassifies the biomarker-status of some individuals. Misclassification has also an effect on the response probabilities for treatments. The effect is seen in Scenarios 3 and 4, in which an interaction between the biomarker-status and treatment-specific response probability is assumed (see Table 1). In particular, worth noting is the fact that, in scenario four and $\theta = 0.2$, misclassification in the assay-positive stratum changes the response probabilities so that the control treatment would appear as more efficacious than the experimental treatment, contrary to the assumed true values (see Table 1).

[Place Table 2 about here]

To examine the effect of misclassification on the operating characteristics of trials, 1000 trials were simulated for each of the ten settings indicated by the bold rectangles in Table 2. Maximum sample sizes N_{max} = 25, 50, 75, and 100 were considered to investigate the type-I-error probability and power to conclude efficacy for a treatment with the unacceptable and desired response probability, i.e., $P_{bt} = 0.25$ and 0.5, respectively.

As suggested by Barry et al.[8], parameters defining the prior distribution for μ_{jk} were chosen based on the goal of the analysis. Specifics of the prior distributions can be found in Appendix C of the Supplementary Materials.

The parameters for the tests for futility and efficacy were set as follows: $\pi_1 = 0.5, \delta_F = 0.01$, and $\pi_0 = 0.25, \delta_E = 0.9$, respectively. These values imply that treatment t is suspended indefinitely for futility in stratum s when the 99th percentile of the resulting futility-posterior of μ_{st} is smaller than or equal to 0.5. On the other hand, treatment t in stratum s is considered as efficacious when the 10th percentile

of the resulting efficacy-posterior of μ_{st} is larger than 0.25 (see Appendix A and B of the Supplementary Materials).

Practical implementation

The hierarchical model was fitted by using the Gibbs sampler code developed by Barry et al.[8], updated to allow for imperfect-assay results (Appendix D of the Supplementary Materials). Based on the Raftery & Lewis diagnostic, [15] 15,000 posterior samples were retained after a 15 iteration burn-in, to achieve convergence for estimation of the required quantiles. No thinning was applied. The Gibbs sampler was run and results were analysed in R 3.4.2 (x64).[16] Computation time for one simulated OAR trial with $N_{max} = 100$ was equal to about 3 hours on a 64-bit, 2.6 GHz, 8GB RAM machine.

Results

The results of the simulation study are summarized in terms of four operational characteristics. First, the average number of accrued patients across the different stratum-treatment combinations is evaluated. Second, the proportion of statistically significant efficacy-test results, in function of N_{max}, is investigated. Finally, we inspect the effect of an imperfect assay on the average proportion of patients receiving an efficacious treatment, as well as the average proportion of patients having a positive response, in a trial with N_{max}=100.

Scenarios 1 and 2

The detailed results of Scenarios 1 and 2 are summarized in Appendix E of the Supplementary Materials. For the perfect-assay setting in Scenario 1, there is no clear preference for any treatment in neither of the strata and the probability of concluding efficacy, i.e., the type-I-error probability, is smaller than 0.05 for all combinations. For Scenario 2, considering a perfect-assay leads to a clear preference for the experimental treatment in both strata; for trials with $N_{Max} = 100$, the power is approximately 0.9 and the type-I-error probability is less than 0.1.

In both scenarios, treatment-specific response probabilities are constant across strata (Table 2). Thus, no effect of using the imperfect assay is to be expected. However, the results show a slight over-representation of assay-positive patients in the trial for the imperfect assay (see Table 2). For $N_{max} < 100$, this over-representation leads, in the assay-positive stratum, to an increase of power and a decrease of the type-I-error probability.

Regarding patient-specific outcomes in trials with N_{max}=100, in Scenario 1, both treatments are inefficacious, so none of the patients can be deemed as having received an efficacious treatment. Under Scenario 2, irrespectively of assay accuracy, about 76% of the patients receive an efficacious treatment. In addition, the proportion of responders is about 24% and 44% for Scenario 1 and Scenario 2, respectively, irrespectively of assay accuracy. It is worth noting that the value of 44% exceeds the response probability of $(0.25 + 0.5)/2 = 0.375$ expected under the fixed 1:1 randomization ratio.

This is the result of OAR assigning more patients to the 'most promising' (experimental) treatment arm in both strata.

Scenario 3

The results of Scenario 3 with $\theta = 0.5$ are shown in Figure 1 and Table 3. Panels **a** and **b** of Figure 1 present the average number of patients included in each stratum-treatment group in function of N_{max}. Panel **a** of Figure 1 shows that, after $n_0 = 25$ patients in the perfect-assay case, the accrual of patients to the experimental treatment in the assay-positive stratum enjoys the highest rate. Consequently, the accrual to the control treatment is substantially lower. This is the result of OAR putting more patients on the 'most promising' treatment in this stratum. Somewhat surprisingly, more patients are also randomized to the experimental treatment in the assay-negative stratum. This is due to the 'borrowing' of information between strata when testing for futility. Inspection of the prior distribution for this test reveals that the intra-treatment (across strata) correlation of μ_{st} is equal to 0.5 (see Appendix C of the Supplementary Materials).

Panel **b** of Figure 1 shows that the use of the imperfect assay reduces the number of assay-positive patients randomized to the experimental treatment. On the other hand, in the assay-negative stratum, the number of patients randomized to the experimental treatment increases. This is due to the effect of misclassification on the response probabilities, as described before and shown in Table 2. The effect is reflected in the randomization ratios and, consequently, the sample sizes for the experimental treatment.

[Place Figure 1 about here]

Panels **c** and **d** of Figure 1 present the estimated probability of concluding efficacy in function of N_{max}. For N_{max}=100, the power for the experimental treatment in the assay-positive stratum is substantially reduced from about 0.86 for the perfect biomarker-assay (see panel **c** of Figure 1 and Table 3) to about 0.65 for the imperfect assay (see panel **d** of Figure 1 and Table 3). Also, the estimate of the type-I-error probability for the experimental treatment in the assay-negative stratum doubles to about 0.2. The

decrease of power is due to the reduced response probability for assay-positive patients assigned to the experimental treatment (see Table 2). On the other hand, misclassification causes an increase in the response probability to the experimental treatment of assay-negative patients (see Table 2) and inflation of the type-I-error probability. There is no effect of the use of the imperfect assay on the estimated type-I-error probabilities for the control treatment, because the response probability for the control treatment is the same in both strata.

In terms of patient-specific outcomes, the average proportion of patients receiving an efficacious treatment is reduced from 0.6 (empirical standard error SE=0.17) for the perfect assay to 0.41 (SE=0.12) for the imperfect assay. Moreover, the number of positive-response patients is reduced from 0.39 (SE=0.07) to 0.34 (SE=0.07).

Scenario 4

The results of Scenario 4 with $\theta = 0.5$ are shown in Figure 2 and Table 3. Panels **a** and **b** of Figure 2 indicate that the accrual of patients to the efficacious stratum-treatment combinations is higher than for the inefficacious ones. For the perfect assay, the accrual rate is essentially the same for both efficacious combinations; a similar conclusion can be drawn for both inefficacious combinations. For the imperfect assay, the situation remarkably changes: the difference in the accrual rate between the efficacious and inefficacious treatments is reduced due to misclassification that alters the actual response probabilities. Moreover, the accrual rate of assay-positive patients is higher than that of assay-negative patients due to the increased assay-positive prevalence (Table 2).

[Place Figure 2 about here]

Panels **c** and **d** of Figure 2 present the estimated probability of concluding efficacy in function of N_{max}. For trials with a perfect biomarker-assay and $N_{max} = 100$, the estimated probability (power) for the two efficacious combinations (green color) is equal to about 0.90. The estimated type-I-error probability for the two inefficacious combinations (red color) is slightly below 0.1 (Table 3).

For the imperfect assay, the estimated power substantially decreases, as compared to the perfect-assay case. Moreover, there is a slight difference in favor of the control treatment in the assay-negative stratum. On the other hand, for the two inefficacious combinations, the estimated type-I-error probability substantially increases, with slightly higher values for the control treatment in the assay-positive stratum.

The reduction of power and increase of the type-I-error probability can again be attributed to the change of the response probabilities due to the imperfect nature of the biomarker assay (see Table 2). Additionally, the imbalance in the stratum-specific sample size, due to a higher accrual of assay-positive patients, increases power and decreases the type-I-error probability. As such, the imbalance mitigates the effects of misclassification.

Concerning the patient-specific outcomes, the average probability of patients receiving an efficacious treatment is reduced from 0.75 (SE=0.09) for the perfect assay to 0.57 (SE=0.08) when using the imperfect assay. Moreover, the use of the imperfect assay reduces the average proportion of positive-response patients from 0.44 (SE=0.06) to 0.39 (SE=0.07).

The results of Scenario 4 with $\theta = 0.2$ are shown in Figure 3 and Table 3. Panels **a** and **b** of Figure 3 show that, compared to $\theta = 0.5$ (panels **a** and **b** of Figure 2), more assay-negative subjects are enrolled in the trial and more subjects are being randomized to the treatment with the highest 'actual' response probability. This means that, in the imperfect-assay case, the inefficacious control treatment gets more patients in the assay-positive stratum (Table 2).

Panels **c** and **d** of Figure 3 present the estimated probability of concluding efficacy in function of N_{max}. In case of the perfect assay (panel **c** of Figure 3), assuming $\theta = 0.2$ results in a reduced power for the efficacious experimental treatment in the assay-positive stratum, as compared to $\theta = 0.5$. However, for $N_{max} = 100$, the power is still around 0.7 (see Table 3). The type-I-error probability is equal to around 0.1 in both strata. For the assay-negative stratum, no effect of the imperfect assay is observed neither for power nor for the type-I-error probability (panel **d** of Figure 3 and Table 3). For the assay-

positive stratum, however, the type-I-error probability to conclude the inefficacious control treatment as efficacious (0.56) is higher than the power to conclude the experimental treatment as efficacious (0.34). This result is due to the fact that, for the imperfect assay, the 'actual' response probability for the control treatment is higher than for the experimental treatment (see Table 2).

With respect to the patient-specific outcomes, the average probability of patients receiving an efficacious treatment is reduced from 0.77 (SE=0.09) for the perfect assay to 0.63 (SE=0.09) for the imperfect assay. Moreover, the use of the latter assay reduces the average proportion of positive-response patients from 0.44 (SE=0.06) to 0.41 (SE=0.06).

[Place Table 3 about here]

Discussion and conclusion

The presented results for the case of a perfect biomarker-assay are in line with current findings regarding Bayesian OAR. In particular, the results for our Scenario 3 (quantitative stratum-by-treatment interaction) and Scenario 4 (qualitative interaction) correspond to those reported by Barry et al.[8] In particular, Barry et al.[8] concluded a power of ≥ 0.8, with the type-I-error probability ≤ 0.1 for $N = 55$ and $N = 59$, for their single- and complementary-markers scenario, respectively. Panel **c** of Figures 1 and 2 show that, for our Scenarios 3 and 4, the same power and type-I-error probability are achieved at the same trial size.

It is worth noting that panel **d** of Figure 1 and 2 suggests a slight decrease of power when larger sample sizes are considered. Although discussion of this issue is beyond the scope of the current paper, this counterintuitive observation can be explained by the fact that sequentially testing for futility after every patient leads to an increasing type-II-error probability with an increasing sample size, restricting the maximally reachable power.

The use of an imperfect assay affects the operational characteristics of the Bayesian biomarker-driven OAR trial in several aspects.

First, assuming different sensitivity and specificity causes the assay prevalence to differ from the true underlying biomarker prevalence, as seen in Table 2. This, in turn, affects the power and the type-I-error probability. The effect is seen even if the response probability for a treatment is completely independent of biomarker-status, as in our Scenario 2.

Second, the assay misclassification alters the actual response probabilities for treatments within different strata (see Table 2). This is the case for the considered setting of two treatments, but also holds for the setting of multiple treatment arms within multiple strata. In this general setting, one can show that, in case of treatments with a different response probability, misclassification will always cause the smallest and largest response probability to increase and decrease, respectively (see Appendix F of the Supplementary Materials). The remaining response probabilities within the stratum may or may not increase or decrease, possibly leading to a different ordering of treatments based on their response probabilities. Therefore, in the general setting of more than one experimental treatment and one control treatment, power and the type-I-error probability may increase or decrease, depending on the resulting ordering of efficacious and inefficacious treatments within each stratum (Appendix F of the Supplementary Materials). In the setting of two treatments, the altered response probabilities lead to a decrease in power for the efficacious treatment and to inflation of the type-I-error probability for the inefficacious treatment, as compared to the perfect biomarker-assay case. Higher misclassification rates will affect response probabilities to a greater extent and cause larger effects on power and the type-I-error probability.

This implies that the effects observed for an imperfect assay in Scenarios 3 and 4 should be interpreted as a result of a combination of these two opposing effects, i.e., increased prevalence of positive-assay patients and change in the actual response probabilities. For both scenarios with $\theta = 0.5$, the power is drastically reduced from about 0.9 to at most 0.65 for trials accruing N_{max}=100 patients. On the other hand, the type-I-error probability doubles to 0.2 for the inefficacious treatments. To further investigate the effect of using an imperfect biomarker-assay, additional simulations under Scenario 4 were set up

(see Appendix G of the Supplementary Materials). These additional simulations confirm that for $\theta = 0.5$ and response probabilities of Scenario 4, interchanging the values of Se_A and Sp_A produces results that are symmetric to those shown in Figure 3 and Table 3.

Furthermore, one can show that the operating characteristics of a trial using an imperfect assay coincide with those of a trial with a perfect assay run in a population characterised by true response probabilities and disease prevalence equal to the probabilities implied by the imperfect assay. Therefore, any advantages[6,7,8] and issues[9,10,11] applying to adaptive randomization, as compared to fixed- ratio randomization, are also relevant in case of an imperfect assay. For example, consider the case of the qualitative treatment effect interaction scenario (4) with true biomarker-positive prevalence of 0.5. Table 4 shows the results of a trial with a fixed one-to-one randomization with stopping for futility as described earlier. In this particular setting, both adaptive (Table 3) and fixed (Table 4) randomization are comparably affected by a decrease in power, indicated in bold, as well as an increase of the type-I-error probability. This is understandable because, irrespectively of the considered randomization paradigm, the use of an imperfect assay changes the assay-defined response probabilities, as indicated in Table 2. Moreover, for each assay setting, the power and type-I-error probability are very similar for the fixed-ratio and outcome-adaptive randomization (see Table H.1 in Appendix H of the Supplementary Materials).

[Place Table 4 about here]

In terms of the patient-specific characteristics, using an imperfect biomarker-assay substantially reduces, as compared to the perfect-assay case, the proportion of patients receiving the efficacious treatment and the proportion of patients having a positive response if there is an interaction between treatment and stratum, as in Scenario 3 and 4.

The assumption of immediate outcome assessment, used in the stimulation study, is a strong one. However, this assumption is not required to consider implementation of an OAR design. The impact of considering an imperfect biomarker-assay would not affect the difference between immediate or

lagged updating of the randomization ratios conditional on the affected response ratios and prevalence introduced by the imperfect assay. This point is illustrated in Appendix I of the Supplementary Materials regarding an imperfect assay simulation of the real real-life example from Barry et al.[8]

In conclusion, we have shown that the use of an imperfect assay affects the operational characteristics of the Bayesian biomarker-driven OAR design. Even in the simple settings considered in this manuscript, the effect may be substantial. The magnitude of the effect depends on the sensitivity and specificity of the assay, as well as on the distribution of the biomarker in the patient population. Therefore, the impact has to be evaluated on a case-by-case basis. Thus, with an imperfect biomarker-assay, the decision to apply a biomarker-based OAR design may require careful reflection.

Acknowledging, during the design simulations, that the considered assay may be imperfect, could potentially help in preventing organization of an underpowered trial. In particular, based on previous research or knowledge of the assay, one could try to consider a set of meaningful sensitivity and specificity combinations. If the impact on power would turn out to be minimal, one could still proceed with the design, though with caution. In case a large impact would be implied, one could consider powering the trial according to a worst-case scenario for the accuracy of the assay.

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. The authors are grateful to Marc Buyse for his comments related to various versions of this manuscript.

Declaration of conflicting interests

The Authors declare that there is no conflict of interest.

References

1. Biswas S, Liu DD, Lee JJ, et al. Bayesian clinical trials at the university of Texas md Anderson cancer center. *Clinical Trials.* 2009; 6: 205-216.
2. Berry DA. Adaptive clinical trials: the promise and the caution. *Journal of clinical oncology.* 2011; 29: 606-609.
3. Lee JJ, Chen N and Yin G. Worth adapting? Revisiting the usefulness of outcome-adaptive randomization. *Clinical Cancer Research.* 2012; 18: 4498-4507.
4. Simon R and Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research.* 2004; 10: 6759-6763.
5. Liu C, Liu A, Hu J, et al. Adjusting for misclassification in a stratified biomarker clinical trial. *Statistics in Medicine.* 2014; 15: 3100-3113.
6. Zhou X, Liu S, Kim ES, et al. Bayesian adaptive design for targeted therapy development in lung cancer – a step toward personalized medicine. *Clinical Trials.* 2008; 5: 181-193.
7. Gu X, Chen N, Wei C, et al. Bayesian two-stage biomarker-based adaptive design for targeted therapy development. *Statistics in biosciences.* 2016; 8: 99-128.
8. Barry WT, Perou CM, Marcom PK, et al. The use of Bayesian hierarchical models for adaptive randomization in biomarker-driven phase II studies. *Journal of Biopharmaceutical statistics.* 2015; 25: 66-88.
9. Korn EL and Freidlin B. Adaptive clinical trials: Advantages and disadvantageous of various adaptive design elements. *Journal of the national cancer institute.* 2017; 109: 1-6.
10. Korn EL and Freidlin B. Commentary on Hey and Kimmelman. *Clinical trials.* 2015; 12: 122-124.
11. Thall P, Fox P and Wathen J. Statistical controversies in clinical research: Scientific and ethical problems with adaptive randomization in comparative clinical trials. *Annals of oncology.* 2015; 26: 1621-1628.
12. Kim ES, Herbst RS, Wistuba II, et al. The BATTLE trial: Personalizing Therapy for Lung Cancer. *Cancer Discovery*. 2011; 1: 44-53.

13. Beach TG, Monsell SE, Philips LE, et al. Accuracy of the Clinical Diagnosis of Alzheimer Disease at National Institute of Aging Alzheimer's Disease Centers, 2005-2010. *Journal of Neuropathology & Experimental Neurology*. 2013, 71(4): 266-273.

14. Zhou Y, Wang C, Zhu H, et al. Diagnostic accuracy of PIK3CA mutation detection by circulating free DNA in breast cancer: A meta-analysis of diagnostic test accuracy. *PLoS one*. 2016; 11(6): e0158143.

15. Raftery AE and Lewis SM. How many iterations in the Gibbs sampler? *Washington univ seattle dept of statistics.* 1991; 763-773.

16. 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.

Tables

Table 1: Assumed true response probabilities (P_{bt}) in the considered simulation scenarios. Bold entries are efficacious biomarker-treatment combinations.

		Biomarker-negative stratum (0)		Biomarker-positive stratum (1)	
		Control (P_{00})	Experimental (P_{01})	Control (P_{10})	Experimental (P_{11})
Scenario 1	**No treatment effect (Null)**	0.25	0.25	0.25	0.25
Scenario 2	**Consistent treatment effect**	0.25	**0.50**	0.25	**0.50**
Scenario 3	**Quantitative interaction**	0.25	0.25	0.25	**0.50**
Scenario 4	**Qualitative interaction**	**0.50**	0.25	0.25	**0.50**

Table 2: Biomarker-positive assay prevalence and 'actual' response probabilities for the perfect and imperfect biomarker assays. Bold entries are efficacious biomarker-treatment combinations.

Scenario	Population prevalence $P(B=1)$	Assay	Assay Prevalence $P(S=1)$	Assay-negative stratum (0)		Assay-positive stratum (1)	
				Ctrl (P_{00})	Exp (P_{01})	Ctrl (P_{10})	Exp (P_{11})
(1) Null	0.5	Perfect	0.5	0.25	0.25	0.25	0.25
		Imperfect	0.55	0.25	0.25	0.25	0.25
	0.2	Perfect	0.2	0.25	0.25	0.25	0.25
		Imperfect	0.40	0.25	0.25	0.25	0.25
(2) Consistent	0.5	Perfect	0.5	0.25	**0.50**	0.25	**0.50**
		Imperfect	0.55	0.25	**0.50**	0.25	**0.50**
	0.2	Perfect	0.2	0.25	**0.50**	0.25	**0.50**
		Imperfect	0.40	0.25	**0.50**	0.25	**0.50**
(3) Quantitative Interaction	0.5	Perfect	0.5	0.25	0.25	0.25	**0.50**
		Imperfect	0.55	0.25	0.31	0.25	**0.43**
	0.2	Perfect	0.2	0.25	0.25	0.25	**0.50**
		Imperfect	0.40	0.25	0.27	0.25	**0.35**
(4) Qualitative Interaction	0.5	Perfect	0.5	**0.50**	0.25	0.25	**0.50**
		Imperfect	0.55	**0.44**	0.31	0.32	**0.43**
	0.2	Perfect	0.2	**0.50**	0.25	0.25	**0.50**
		Imperfect	0.40	**0.48**	0.27	0.40	**0.35**

Table 3: Probability of obtaining a statistically significant efficacy test result for $N_{max} = 100$. Probabilities in normal font can be interpreted as the type-I-error probability, entries in bold as power.

Scenario	Population prevalence $P(B = 1)$	Assay	Assay-negative stratum (0)		Assay-positive stratum (1)	
			Ctrl (P_{00})	Exp (P_{01})	Ctrl (P_{10})	Exp (P_{11})
(1) Null	0.5	Perfect	0.02	0.02	0.02	0.02
		Imperfect	0.02	0.02	0.02	0.01
(2) Consistent	0.5	Perfect	0.07	**0.92**	0.08	**0.92**
		Imperfect	0.08	**0.92**	0.08	**0.92**
(3) Quantitative Interaction	0.5	Perfect	0.05	0.08	0.06	**0.86**
		Imperfect	0.05	0.20	0.06	**0.65**
(4) Qualitative Interaction	0.5	Perfect	**0.89**	0.08	0.08	**0.89**
		Imperfect	**0.73**	0.22	0.24	**0.68**
	0.2	Perfect	**0.87**	0.08	0.09	**0.71**
		Imperfect	**0.87**	0.11	0.56	**0.34**

Table 4: Probability of obtaining a statistically significant efficacy test result for $N_{max} = 100$ for a Bayesian fixed (one-to-one) randomization trial. Entries in normal font can be interpreted as the type-I-error probability, entries in bold as power.

Scenario	Population prevalence $P(B = 1)$	Assay	Assay-negative stratum (0)		Assay-positive stratum (1)	
			Ctrl (P_{00})	Exp (P_{01})	Ctrl (P_{10})	Exp (P_{11})
(4) Qualitative Interaction	0.5	Perfect	**0.86**	0.10	0.09	**0.87**
		Imperfect	**0.73**	0.24	0.27	**0.69**

Figure Captions

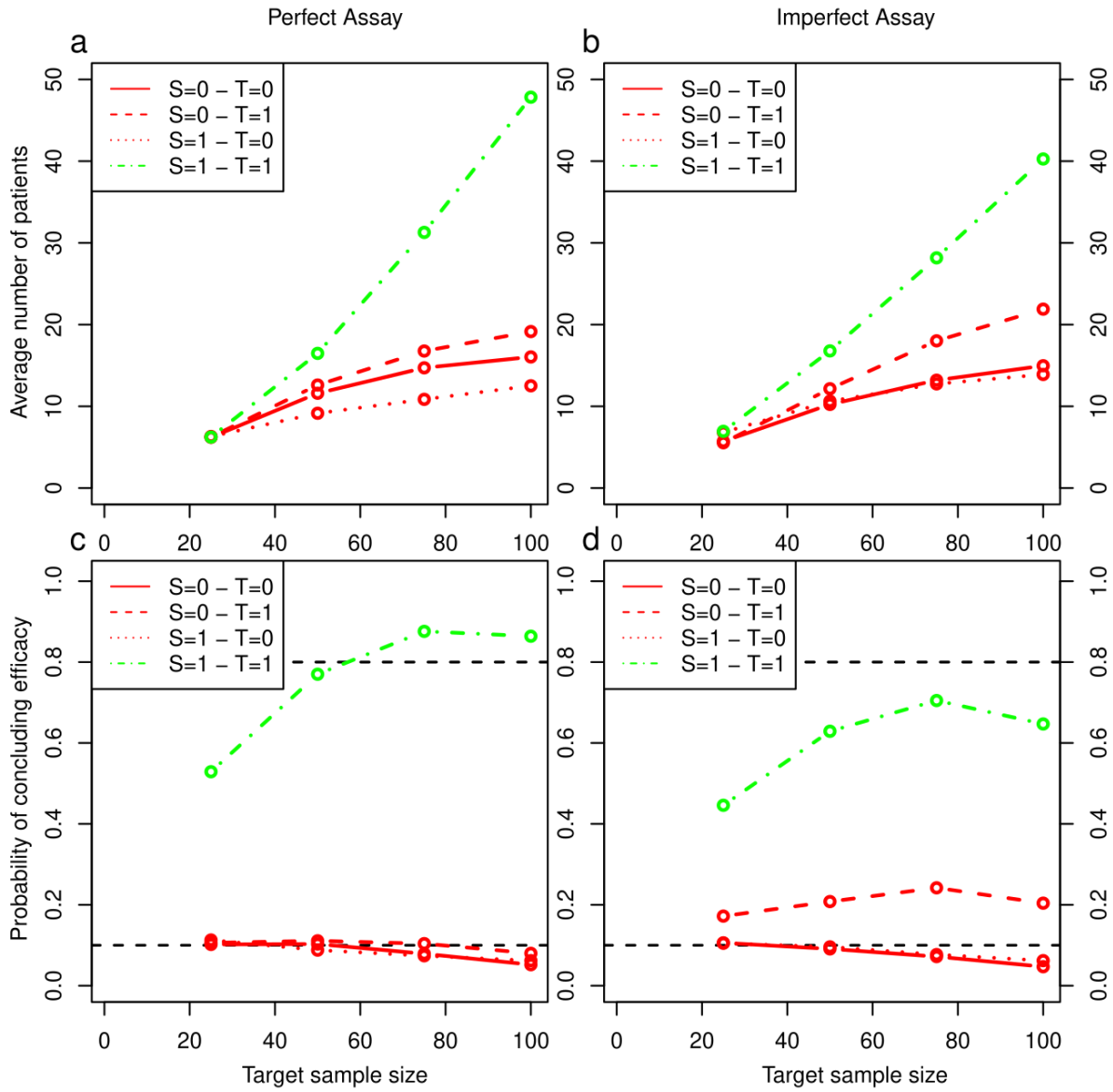

Figure 1. Average number of accrued patients (**a**-**b**) and average proportion of trials concluding efficacy by the accrued total sample size (**c**-**d**) for each stratum-treatment combination by the accrued total sample size over 1000 trials in Scenario 3 with $\theta = 0.5$. Results for the biomarker-negative ($S = 0$) stratum patients receiving the control ($T = 0$) and experimental ($T = 1$) treatment are indicated by the solid and dashed line, respectively. Results for the biomarker-positive ($S = 1$) patients receiving the control and experimental treatments are denoted by the dotted and dotted-dashed line, respectively. Green color marks the efficacious stratum-treatment combinations, red marks the inefficacious combinations.

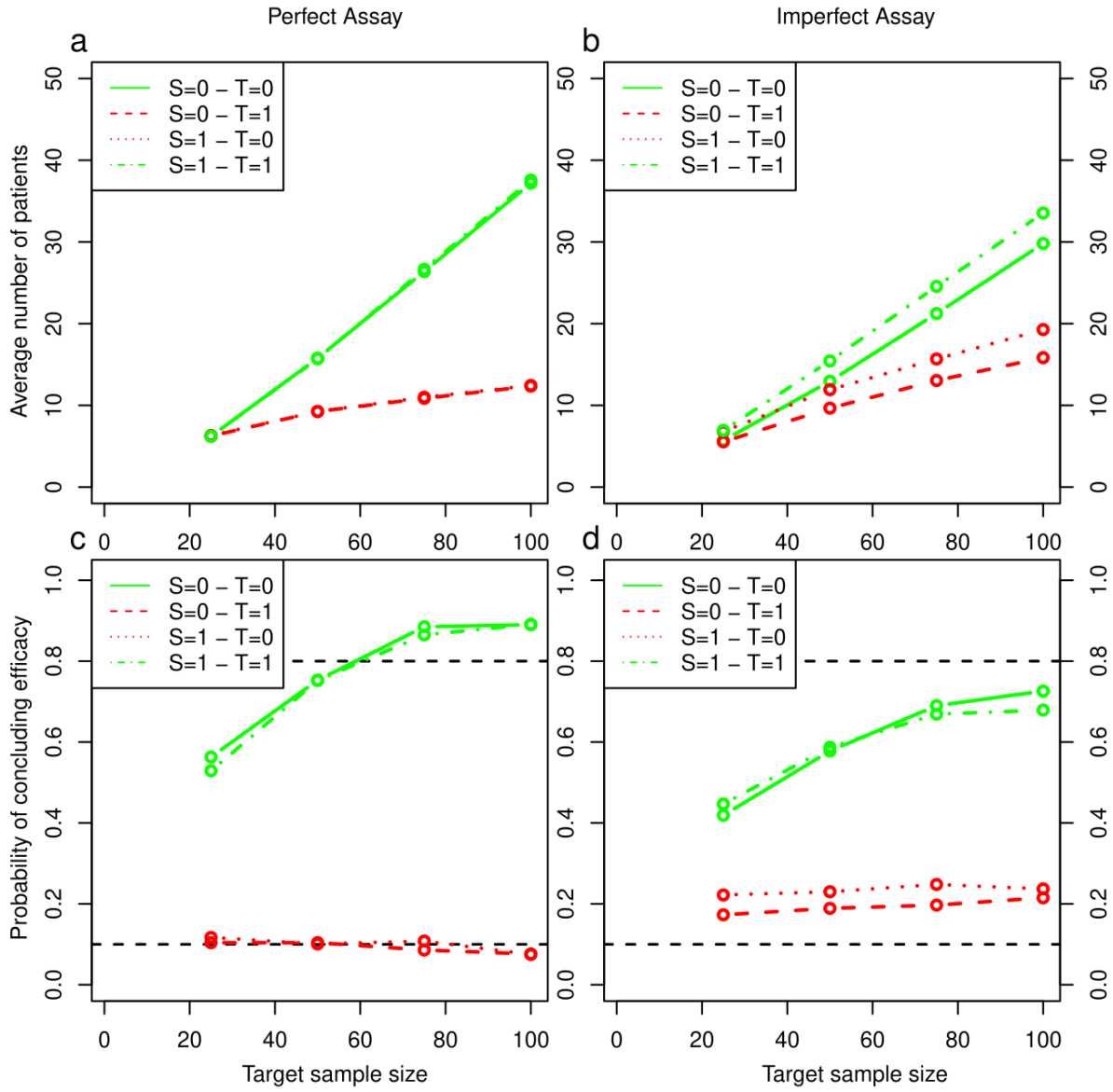

Figure 2. Average number of accrued patients (**a**-**b**) and average proportion of trials concluding efficacy by the accrued total sample size (**c**-**d**) for each stratum-treatment combination by the accrued total sample size over 1000 trials in Scenario 4 with $\theta = 0.5$. Results for the biomarker-negative ($S = 0$) stratum patients receiving the control ($T = 0$) and experimental ($T = 1$) treatment are indicated by the solid and dashed line, respectively. Results for the biomarker-positive ($S = 1$) patients receiving the control and experimental treatments are denoted by the dotted and dotted-dashed line, respectively. Green color marks the efficacious stratum-treatment combinations, red marks the inefficacious combinations.

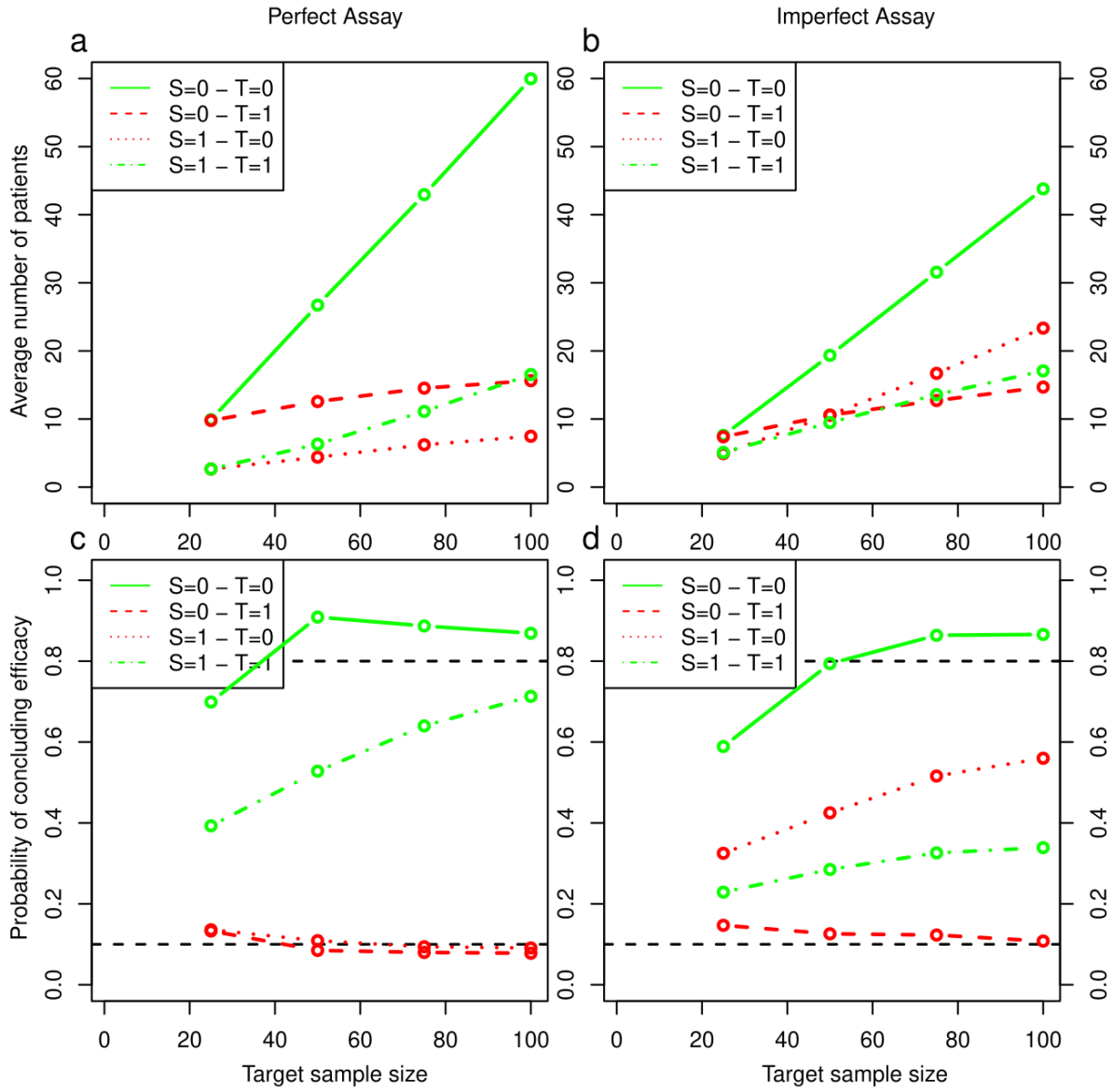

Figure 3. Average number of accrued patients (**a**-**b**) and average proportion of trials concluding efficacy by the accrued total sample size (**c**-**d**) for each stratum-treatment combination by the accrued total sample size over 1000 trials in Scenario 4 with $\theta = 0.2$. Results for the biomarker-negative ($S = 0$) stratum patients receiving the control ($T = 0$) and experimental ($T = 1$) treatment are indicated by the solid and dashed line, respectively. Results for the biomarker-positive ($S = 1$) patients receiving the control and experimental treatments are denoted by the dotted and dotted-dashed line, respectively. Green color marks the efficacious stratum-treatment combinations, red marks the inefficacious combinations.