


Performance evaluation of a new prognostic-efficacy-combination design in the context of telemedical interventions

Mareen Pigorsch^{1*} , Martin Möckel², Stefan Gehrig³, Jan C. Wiemer⁴, Friedrich Koehler⁵ and Geraldine Rauch^{1,6}

¹Charité – Universitätsmedizin Berlin, Institute of Biometry and Clinical Epidemiology, Berlin, Germany; ²Charité – Universitätsmedizin Berlin, Department of Emergency Medicine, Campus Virchow-Klinikum and Campus Charité Mitte, Berlin, Germany; ³Independent researcher, Berlin, Germany; ⁴BRAHMS, Thermo Fisher Scientific, Hennigsdorf, Germany; ⁵Charité – Universitätsmedizin Berlin, Department of Cardiology and Angiology Campus Mitte, Centre for Cardiovascular Telemedicine, Berlin, Germany; and ⁶Technische Universität Berlin, Berlin, Germany

Abstract

Aims Telemedical interventions in heart failure patients intend to avoid unfavourable, indication-related events by an early, individualized care, which reacts to the current patients need. However, telemedical support is an expensive intervention, and usually only patients with high risk for unfavourable follow-up events will be able to profit from it. Möckel et al. therefore adapted a new design which we call ‘prognostic-efficacy-combination design’. This design allows to define a biomarker cut-off and to perform a randomized controlled trial (RCT) in a biomarker-selected population within a single study. However, so far, it has not been evaluated if this double use of the control group for biomarker cut-off definition and efficacy assessment within the RCT leads to a bias in treatment effect estimation. In this methodological research work, we therefore want to evaluate whether the ‘prognostic-efficacy-combination design’ leads to biased treatment effect estimates and also compare it to alternative designs. If there is a bias, we further want to analyse its magnitude under different parameter settings.

Methods We perform a systematic Monte Carlo simulation study to investigate among others potential bias, root mean square error and sensitivity, and specificity as well as the total treatment effect estimate in various realistic trial scenarios that mimic and vary the true data characteristics of the published TIM-HF2 Trial. In particular, we vary the event proportion, the sample size, the biomarker distribution, and the lower bound for the sensitivity.

Results The results show that indeed the proposed design leads to some bias in the effect estimators, indicating an overestimation of the effect. However, this bias is relatively small in most scenarios.

Conclusions The ‘prognostic-efficacy-combination design’ can generally be recommended for clinical applications due to its efficiency compared to two separate trials. We recommend a sufficiently large sample size depending on the trial scenario. Our simulation code can be adapted to explore suitable sample sizes for other settings.

Keywords Telemedicine; Biomarker-selected population; Simulation study; Innovative study design; Bias estimation

Received: 4 November 2021; Revised: 29 July 2022; Accepted: 15 August 2022

*Correspondence to: Mareen Pigorsch, M.Sc., Charité – Universitätsmedizin Berlin, Institute of Biometry and Clinical Epidemiology, Charitéplatz 1, 10117 Berlin, Germany. Phone: +49 (0)30 450 562 176. Email: mareen.pigorsch@charite.de

Introduction

Telemedical interventions provide a wide field of treatment support options, which has the potential to better address the patients’ individual needs and thereby improve long-term outcomes. For example, a telemedical intervention can involve technological devices that allow closer supervision by

the physician while the patient is at home. Due to this, early signs of upcoming medical events are easier detected and appropriate care can be immediately provided.

Heart failure is a chronic disorder, for which an additional telemedical support was generally shown to be beneficial.^{1,2} However, telemedical support is an expensive intervention, and heart failure is one of the most prevalent chronic

diseases.³ Therefore, an efficient allocation is desirable. It is evident that a telemedical intervention intends to avoid unfavourable, indication-related events by an early, individualized care. Patients that have a sufficiently stable state of disease will most likely not profit from telemedical intervention. The estimated effect of the telemedical intervention in a population that includes such low-risk patients could accordingly misrepresent the true effect in a more targeted, high-risk population. Therefore, it would be ideal if the subgroup of potential telemedical profitters—the actual population of interest—could be identified in advance by an appropriate biomarker. To follow this idea, Möckel *et al.*⁴ implemented a post hoc analysis for a clinical trial. During the clinical trial TIM-HF2,¹ patients randomly drawn from the full population including *potential profitters* and *non-profitters* were randomly assigned to control (without telemedical support) or intervention (with telemedical support). At the time of inclusion, it was unknown which patients might profit from telemedical support and which not. At the time of post hoc analysis, however, information on biomarkers could be used to determine the subgroup of *potential profitters* at baseline retrospectively. The population of interest was thus retrospectively redefined as the sub-population of the *potential profitters* from telemedical care. This was possible because, at the studies' baseline assessment, several biomarkers were measured, which were known to be associated with the probability of an unfavourable event. However, at baseline, there were no established cut-off values for these biomarkers, allowing a prospective definition of the population of interest. Thus, in principle, a prognostic biomarker trial in the full population would have been necessary before being able to use biomarker cut-offs as inclusion criteria, defining the population of interest. As a separate biomarker study was not feasible due to time and financial restrictions, Möckel *et al.*⁴ determined the cut-off values for the biomarkers based on the observed biomarker values and the outcome event status of the patients in the control arm to identify a high-risk group that ideally comprises all *potential profitters* of telemedical care. These cut-offs were subsequently applied as post hoc inclusion criteria in both study arms, intervention and control, to filter for high-risk patients, that is, to retrospectively approach the study population of interest.

The final efficacy analysis was performed by comparing control and intervention within this high-risk subgroup identified post hoc via biomarker criteria. Such a post hoc inclusion criteria can also be seen as an adaptive design element. This new study design thus covers (i) the derivation of a cut-off for a biomarker and (ii) a randomized controlled trial (RCT) for a biomarker-selected population within a single study. In what follows, we will call this approach by Möckel *et al.*⁴ 'prognostic-efficacy-combination design'. The design is appealing due to its efficiency and could potentially be applied in future trials. However, the study design can be interpreted

as an adaptive design, where the inclusion criteria are adapted after completed recruitment and observation of patients. Biomarker cut-off definition and efficacy analysis are both based on the same sample of patients from the control arm. Therefore, it is generally possible that the treatment effect is biased, as an adaptive conditional study design element often yields biased effect estimators.^{5,6} Analogously, data analysis strategies that engage in 'double dipping', that is, in the re-utilization of information from the same data set at different analytical steps, are also prone to bias.⁷

In this methodological research work, we therefore evaluate whether the study design of Möckel *et al.*⁴ leads to biased treatment effect estimates. If there is a bias, we further want to analyse its size and the parameters influencing the size of the bias for deriving recommendations and warnings for future applications. To evaluate the performance of the design, we set up a simulation study that mimics and varies the true data characteristics of the TIM-HF2 Trial.²

Methods

The new prognostic-efficacy-combination design

In a classical efficacy study designed as a two-armed RCT, every patient with an index condition is randomized to a control group or an intervention group. The efficacy is assessed by comparing the primary efficacy endpoint between these two groups. In a classical prognostic biomarker study, a biomarker (most often continuous) is tested for its ability to separate patients with a prognostic index condition from those without this index condition, and an optimal cut-off for the biomarker is deduced.

In our new design, these two elements are combined: the new prognostic-efficacy-combination design is based on two randomized patient groups, of which one receives the intervention (I) and the other serves as a control (C). For the sake of simplicity, we assume balanced group allocation with sample sizes per group given by n . The primary endpoint X is a binary event indicator, where an event has a negative impact for the patient. In our application, the event of interest could, for example, be 'hospitalization due to cardiovascular causes'. The composite event defined in the reanalysis of TIM-HF2 by Möckel *et al.*,⁴ which we use to justify our simulation setting was defined as '≥30 days lost per year follow-up time due to unplanned cardiovascular hospital admission or all-cause death'. As this event endpoint is rather complicated to assess, we used a simpler definition within this paper. We assume that the primary endpoint is Bernoulli-distributed:

$$X_i^I \sim \text{Bernoulli}(p^I), X_i^C \sim \text{Bernoulli}(p^C), i = 1, \dots, n,$$

where p^I , p^C define the underlying event probabilities under intervention and control condition. We assume that there is a

continuous biomarker B , which is intended to separate the patients that would experience an event in the future in the control arm ($X_i^C = 1$) from those that would not ($X_i^C = 0$). The general idea of our design is that only certain patients that actually will experience an unfavourable event in the future can potentially profit from a new intervention that might help to reduce the risk of an event. We will therefore denote the group of patients that will experience an event in the control arm in the future ($X_i^C = 1$) as the *potential profitters* (PP), whereas the group of patients without an event ($X_i^C = 0$) are denoted as the *non-profters* (NP). Due to randomization, the control arm C and the intervention arm I both are samples of mixed populations of *potential profitters* and *non-profters*. Note that the definition of *potential profitters* and *non-profters* takes the perspective of the planning stage of the trial, that is, one does not know yet if a patient will truly profit or not. Evaluating the efficacy of the intervention in this mixed (or full) population will dilute the treatment effect, as the actual population of interest is constituted only of the *potential profitters*. The aim of our new prognostic-efficacy-combination design is thus to assess the efficacy of the new intervention by comparing the intervention and the control only in the sub-population of the *potential profitters*. If an optimal biomarker would be available (sensitivity and specificity of 1), these sub-populations could be perfectly specified. In applications however, a biomarker will have imperfect sensitivity and specificity below 1. **The prognostic part** of the new design is conducted exclusively in the control arm C . Based on the data of the control arm, an optimal cut-off for the continuous biomarker is determined, where optimality refers to predefined sensitivity and/or specificity boundaries regarding the relevant end-

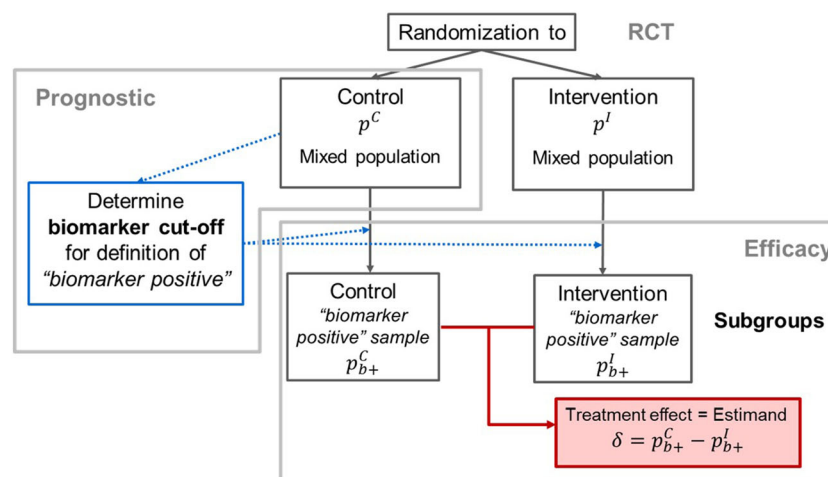
point. The choice of the boundaries can be adapted according to the disease-specific characteristics. By applying this biomarker cut-off in the intervention and in the control arm, the *potential profitters* sub-population in the intervention and in the control arm are approximated. As the biomarker is generally imperfect, the *biomarker positive* patients will not perfectly agree with the *potential profitters*. **The efficacy part** of the new design is then given by calculating the treatment effect comparing the *biomarker positive* patients in the intervention and the control arm. The treatment effect is expressed as an absolute proportion difference $\delta = p_{b+}^C - p_{b+}^I$, where the index $b+$ denotes the *biomarker positive* patients.

Note that the *potential profitters* are those patients which will experience an event in case it is not prevented by the new intervention. Therefore, by definition, the event proportion in the sub-population of *potential profitters* will equal 1 in the control arm ($p_{pp}^C = 1$) and might be lower in the intervention arm in case the intervention is effective ($p_{pp}^I \leq 1$). As the group of *potential profitters* is not perfectly identified by the biomarker, the group of *biomarker positive* patients will have lower event proportions ($p_{b+}^C \leq p_{pp}^C = 1$, $p_{b+}^I \leq p_{pp}^I \leq 1$). *Figure 1* schematically illustrates the new prognostic-efficacy-combination design, the involved (sub-)populations, and the design-specific parameters.

Reference approaches

The reference approach would be an independent external biomarker study to define the biomarker cut-off value and to use this cut-off for identifying the *biomarker positive* as a

Figure 1 Schematic illustration of the prognostic-efficacy-combination design.



prospective inclusion criteria within a subsequent, separate clinical trial. An alternative reference would be to split the control group into two parts: a training set for definition of the biomarker cut-off and a test set to be compared with the intervention group for the treatment estimation.

As a further theoretical comparator, it is also possible to take the true biomarker cut-off of the whole population for a given biomarker distribution. In practice, this exact distribution is unknown. The cut-off value of the reference design will approach the cut-off value obtained from the theoretical comparator with increasing study sample size.

Simulation and analysis

To evaluate if the prognostic-efficacy-combination design leads to biased treatment effect estimates, a simulation study⁸ was conducted using the statistical software R Version 4.1.1.⁹ Our simulation study can be subdivided into three steps. The first step is the data simulation step. In the second step, the analysis of the simulated study data was conducted using the prognostic-efficacy-combination design. In Step 3, the performance of the design was evaluated using different performance measures. These three steps are described in more detail in the following.

Step 1: Data simulation

We generated data for different scenarios, which are listed in the following. Each simulation scenario consists of $n_{sim} = 20\,000$ simulation runs. For illustration, we will describe the data generation process in detail for one simulation scenario, which closely resembles the estimated parameters and data of the TIM-HF2 trial. The sample size per arm in the basic scenario is $n = 750$. For the binary outcome, an event proportion of $p^c = 0.15$ was defined for the control arm, which implies that 15% of the population are *potential profitters* with an event proportion of 1 ($p_{PP}^c = 1$) and 85% are *non-profitters* with an event proportion of 0 ($p_{NP}^c = 0$). Further, we defined an event proportion of $p^i = 0.1$ for the intervention arm. Note that patients in the subgroup of *potential profitters* in the intervention arm will not *all* experience an event, as some events might be avoided by the intervention ($p_{PP}^i \leq 1$). The event proportion for this subgroup of *potential profitters* within the intervention arm is thus given by $p_{PP}^i =$

$$\frac{p^i}{p^c} = \frac{2}{3} \text{ and for the } \textit{non-profitters} \text{ by } p_{NP}^i = 0.$$

As the treatment effect to be estimated is $\delta = p_{b+}^c - p_{b+}^i$, the *biomarker positive* group needs to be identified to define the treatment effect. For a perfect biomarker, the *biomarker positives* correspond to the group of *potential profitters*, so in this case $\delta = p_{PP}^c - p_{PP}^i$. Therefore, for a perfect biomarker, the true treatment effect is $\delta = 1 - \frac{2}{3} = \frac{1}{3}$.

This can serve as a reference value even if the biomarker is not perfect.

To simulate the biomarker distribution B for the *potential profitters* and the *non-profitters*, we assumed a log-normal distribution with the following means and standard deviations:

$$B_{PP} \sim \text{Lognormal}(\mu_{PP} = 4.0, \sigma_{PP} = 0.5)$$

$$B_{NP} \sim \text{Lognormal}(\mu_{NP} = 3.0, \sigma_{NP} = 0.5)$$

This implies that on average, the *non-profitters* have a smaller biomarker value than the *potential profitters*; the standard deviation is assumed to be the same for both sub-populations. This results in a mixed biomarker distribution in the whole population B :

$$B \sim 0.15 \cdot B_{PP} + 0.85 \cdot B_{NP}$$

To determine the optimal biomarker cut-off from the simulated study data, we used a receiver operating characteristic (ROC) curve. In our application, we fix a lower bound for the empirical sensitivity as $sen = 0.95 \leq \widehat{sen}$. We then choose the biomarker cut-off that maximizes the empirical specificity under this condition. In doing so, we accept that the resulting specificity can be rather small, which is acceptable when the emphasis is on safety, that is, if false positives (e.g. *non-profitters* receiving telemedical care) are less clinically relevant than false negatives (e.g. *potential profitters* not receiving telemedical care).

To study the effect of varying data-generating processes and sensitivity requirements, we investigated different simulation scenarios where several parameters were varied from the basic scenario. The scenario with the parameters given above defines the basic scenario (underlined).

i Scenarios Bio1–Bio9: Different biomarker distributions

$$\mu_{PP} = \{3.3, 3.5, 3.8, \underline{4.0}, 4.3, 4.5, 4.8, 5.0, 6.0\}$$

ii Scenarios Ep1–Ep7: Different event proportions for the control arm

$$p^c = \{\underline{0.15}, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6\}$$

iii Scenarios Sam1–Sam8: Different sample sizes

$$n = \{150, 200, 250, 300, 400, \underline{750}, 1500, 5000\}$$

iv Scenarios Sen1–Sen6: Different lower sensitivity bounds to determine biomarker cut-off

$$sen = \{0.95, 0.9, 0.8, 0.7, 0.5, 0.3\}$$

Step 2: Analysis using the prognostic-efficacy-combination design

The estimation of this biomarker cut-off was done in three different ways: (i) in the control group of the current study, which refers to the new prognostic-efficacy-combination design (new design), (ii) in a separate sample from the population serving as external biomarker study (reference design) with the same sample size as the control group $n = 750$, and (iii) single random splitting of the control group of the current study into a training set for estimation of the biomarker cut-off and a test set to be compared with the intervention group for treatment estimation (splitting design). For the splitting design, the control group is divided into two groups of equal size, resulting in sample sizes $n_{training} = n_{test} = \frac{n}{2}$. As an additional theoretical comparator, the true biomarker cut-off (True) was analytically calculated as the (1–0.95) quantile of the biomarker distribution of the *potential profitters*.

The treatment effect (estimand) $\delta = p_{b+}^C - p_{b+}^I$ is then given using the true biomarker cut-off, and the corresponding estimator is $\hat{\delta}^k = \hat{p}_{b+}^C - \hat{p}_{b+}^I$, using method k , which is either the estimated biomarker cut-off from the prognostic-efficacy-combination design, the external biomarker study or the splitting design.

Step 3: Performance evaluation

To evaluate the performance of the prognostic-efficacy-combination design, we calculated the bias of the treatment effect δ as follows:

$$bias^k = \frac{1}{n_{sim}} \sum_{j=1}^{n_{sim}} (\hat{\delta}_j^k - \delta_j)$$

The bias is the difference between the estimated treatment effect and the true treatment effect using the analytically derived true cut-off, averaged over the n_{sim} simulation runs.

In addition, we calculated the root mean squared error (RMSE) of the treatment effect δ as follows:

$$RMSE^k = \sqrt{\frac{1}{n_{sim}} \sum_{j=1}^{n_{sim}} (\hat{\delta}_j^k - \delta_j)^2}$$

The smaller the RMSE, the less variable is the estimation of the treatment effect.

In addition to these performance measures, we report the estimands and corresponding estimates as averages over the $n_{sim} = 20\,000$ simulation runs:

- i estimands p_{b+}^C and p_{b+}^I and estimates \hat{p}_{b+}^C and \hat{p}_{b+}^I of the event proportions for *biomarker positives* in control

and intervention group and the corresponding standard deviations;

- ii estimand δ and estimates $\hat{\delta}^k$ of the treatment effect, for both cut-off derivation methods;
- iii estimands sen^C and sen^I and estimates \widehat{sen}^C and \widehat{sen}^I of the specificity resulting from the biomarker cut-off;
- iv estimands spe^C and spe^I and estimates \widehat{spe}^C and \widehat{spe}^I of the sensitivity resulting from the biomarker cut-off;
- v true biomarker cut-off and estimated biomarker cut-offs;
- vi estimand n_{b+}^C and estimates \hat{n}_{b+}^C of the size of the identified *biomarker positive* subgroup within the control arm.

Results

Tables 1–4 show the results of the simulation study for the different scenarios specified above: variations of biomarker distribution (Table 1, scenarios Bio1–Bio9), variations of event proportion p^C (Table 2, scenarios Ep1–Ep7), variations of sample size (Table 3, scenarios Sam1–Sam8), and variations of target sensitivity (Table 4, scenarios Sen1–Sen6). In each table, the row ‘True’ contains the estimands, whereas the rows ‘Reference’, ‘Split’, and ‘New’ contain resulting estimates from the external biomarker study design, the sample-splitting design, and the new prognostic-efficacy-combination design, respectively.

In general, the results in Tables 1–4 show that there is little bias for all examined parameter combinations. All three estimation methods result in some bias, where the bias in the new design is typically more pronounced than in the other two designs. The bias in the new design is always positive, indicating an overestimation of the treatment effect. For some extreme scenarios, the reference design and the splitting design result in underestimation. Regarding the empirical sensitivities in control and intervention, it can be observed that with the new design, the estimated sensitivity in the control is slightly higher than the theoretical value, whereas estimates from the reference or splitting design are slightly smaller than the theoretical value. For the intervention arm, empirical sensitivities are more similar between the designs than in the control arm. This can be explained with the fact that only in the new design, the biomarker cut-off is defined based on all outcomes in the control arm, and therefore, the sensitivity in the control arm is always bounded from below by 0.95. This also explains the fact that for the new design, biomarker cut-offs tend to be smaller than for the reference (Table 1). Comparing the new design with the splitting design, one can generally conclude that the new design leads to more bias, whereas the splitting design leads to higher RMSE and higher standard deviations of the event proportions for the biomarker positive and the estimated cut-off.

Table 1 Results for varying the mean-log μ_{pp} of biomarker distribution of potential profitters with constant $\sigma = 0.5$

Biomarker	Event proportions for biomarker positive (\pm standard deviation)				Treatment effect $\delta = p_{b+}^c - p_{b+}^+$									
	Estimate (True)	p_{b+}^c	p_{b+}^+	\hat{p}_{b+}	δ	Bias	RMSE	\widehat{spe}^c	\widehat{spe}^+	\widehat{sen}^c	\widehat{sen}^+	\widehat{n}_{b+}^c	\widehat{n}_{b+}^+	Cut - Off (\pm SD)
Bio1	3.0	3.3	0.164 (± 0.015)	0.109 (± 0.012)	0.055		0.148	0.148	0.950	0.950	0.950	650.0	650.0	11.912
	Reference		0.165 (± 0.015)	0.110 (± 0.013)	0.055	0.0002	0.0032	0.154	0.155	0.946	0.946	645.4	645.4	12.016 (± 1.185)
	Split		0.165 (± 0.022)	0.110 (± 0.013)	0.055	0.0001	0.0151	0.159	0.159	0.942	0.942	321.1	321.1	12.068 (± 1.694)
Bio2	3.0	3.5	0.167 (± 0.016)	0.110 (± 0.013)	0.057	0.0016	0.0039	0.155	0.154	0.954	0.946	646.3	646.3	11.998 (± 1.200)
	True		0.185 (± 0.016)	0.123 (± 0.014)	0.062	0.0005	0.0050	0.259	0.260	0.950	0.950	579.0	579.0	14.550
	Reference		0.186 (± 0.019)	0.124 (± 0.015)	0.062	0.0005	0.0050	0.266	0.266	0.946	0.946	574.2	574.2	14.683 (± 1.445)
Bio3	3.0	3.8	0.187 (± 0.027)	0.125 (± 0.017)	0.062	0.0006	0.0173	0.270	0.270	0.942	0.942	285.6	285.6	14.754 (± 2.063)
	Reference		0.188 (± 0.021)	0.124 (± 0.015)	0.064	0.0021	0.0059	0.266	0.265	0.954	0.946	575.3	575.3	14.661 (± 1.465)
	Split		0.245 (± 0.021)	0.163 (± 0.018)	0.082	0.0011	0.0103	0.482	0.482	0.950	0.950	437.1	437.1	19.640
Bio4	3.0	4.0	0.248 (± 0.032)	0.165 (± 0.024)	0.083	0.0011	0.0103	0.486	0.486	0.946	0.946	434.1	434.1	19.823 (± 1.949)
	Reference		0.250 (± 0.046)	0.167 (± 0.030)	0.083	0.0016	0.0244	0.487	0.487	0.942	0.942	216.6	216.6	19.924 (± 2.778)
	Split		0.250 (± 0.035)	0.165 (± 0.025)	0.085	0.0033	0.0115	0.485	0.485	0.954	0.946	435.4	435.4	19.794 (± 1.974)
Bio5	3.0	4.3	0.317 (± 0.025)	0.211 (± 0.022)	0.106	0.0016	0.0161	0.639	0.639	0.950	0.950	337.2	337.2	23.988
	Reference		0.322 (± 0.048)	0.215 (± 0.035)	0.108	0.0016	0.0161	0.639	0.640	0.946	0.946	336.3	336.3	24.210 (± 2.378)
	Split		0.325 (± 0.068)	0.217 (± 0.044)	0.108	0.0025	0.0328	0.637	0.637	0.942	0.942	168.6	168.6	24.331 (± 3.387)
Bio6	3.0	4.5	0.324 (± 0.052)	0.214 (± 0.035)	0.110	0.0043	0.0177	0.639	0.638	0.954	0.946	337.5	337.5	24.174 (± 2.409)
	Reference		0.497 (± 0.034)	0.331 (± 0.032)	0.166	0.0018	0.0269	0.830	0.830	0.950	0.950	215.1	215.1	32.381
	Split		0.503 (± 0.075)	0.335 (± 0.055)	0.168	0.0018	0.0269	0.828	0.828	0.946	0.946	216.3	216.3	32.665 (± 3.202)
Bio7	3.0	4.8	0.505 (± 0.080)	0.334 (± 0.056)	0.171	0.0048	0.0284	0.827	0.827	0.954	0.946	217.4	217.4	32.616 (± 3.244)
	Reference		0.657 (± 0.037)	0.437 (± 0.039)	0.219	0.0002	0.0314	0.912	0.912	0.950	0.950	162.8	162.8	39.550
	Split		0.658 (± 0.082)	0.438 (± 0.063)	0.220	0.0002	0.0314	0.909	0.909	0.946	0.946	164.4	164.4	39.862 (± 3.895)
Bio8	3.0	5.0	0.660 (± 0.087)	0.437 (± 0.063)	0.223	0.0034	0.0327	0.909	0.908	0.954	0.946	165.3	165.3	39.803 (± 3.947)
	Reference		0.869 (± 0.030)	0.579 (± 0.045)	0.290	0.0006	0.0266	0.975	0.975	0.950	0.950	123.0	123.0	53.387
	Split		0.863 (± 0.059)	0.575 (± 0.056)	0.288	-0.0020	0.0266	0.972	0.972	0.947	0.947	124.1	124.1	53.642 (± 5.194)
Bio9	3.0	6.0	0.855 (± 0.087)	0.570 (± 0.067)	0.285	-0.0047	0.0462	0.969	0.969	0.944	0.945	62.8	62.8	53.635 (± 7.246)
	Reference		0.865 (± 0.062)	0.574 (± 0.056)	0.291	0.0006	0.0274	0.973	0.972	0.954	0.947	124.7	124.7	53.562 (± 5.263)
	Split		0.948 (± 0.021)	0.631 (± 0.046)	0.316	-0.0020	0.0194	0.991	0.991	0.950	0.950	112.8	112.8	65.207
Bio10	3.0	6.0	0.942 (± 0.036)	0.628 (± 0.050)	0.314	-0.0046	0.0341	0.988	0.989	0.948	0.948	113.4	113.4	65.254 (± 6.277)
	Reference		0.934 (± 0.056)	0.622 (± 0.055)	0.312	-0.0046	0.0341	0.988	0.988	0.946	0.947	57.1	57.1	64.975 (± 8.677)
	Split		0.943 (± 0.037)	0.627 (± 0.050)	0.317	0.0001	0.0197	0.990	0.989	0.954	0.948	114.0	114.0	65.159 (± 9.360)
Bio11	3.0	6.0	1.000 (± 0.001)	0.666 (± 0.046)	0.334	-0.0001	0.0060	1.000	1.000	0.950	0.950	106.9	106.9	177.252
	Reference		1.000 (± 0.001)	0.666 (± 0.046)	0.334	-0.0001	0.0060	1.000	1.000	0.950	0.950	106.9	106.9	175.108 (± 17.367)
	Split		1.000 (± 0.002)	0.666 (± 0.046)	0.334	<0.0001	0.0071	1.000	1.000	0.952	0.952	53.5	53.5	171.656 (± 24.372)
Bio12	3.0	6.0	1.000 (± 0.001)	0.666 (± 0.046)	0.334	<0.0001	0.0061	1.000	1.000	0.954	0.954	107.4	107.4	174.810 (± 17.583)
	Reference		1.000 (± 0.001)	0.666 (± 0.046)	0.334	<0.0001	0.0061	1.000	1.000	0.954	0.954	107.4	107.4	174.810 (± 17.583)
	Split		1.000 (± 0.001)	0.666 (± 0.046)	0.334	<0.0001	0.0061	1.000	1.000	0.954	0.954	107.4	107.4	174.810 (± 17.583)

New, prognostic-efficacy-combination design; Reference, external biomarker study; Split, split of control into training and test; True, true biomarker cut-off.

μ_{NP} - μ_{PP} : mean-log of biomarker distribution in non-profitters and potential profitters, respectively.

Table 2 Results for varying event proportion p^c

Event proportions		Event proportions for biomarker positive (\pm standard deviation)										
Scenario	p^c	p^l	Estimand (True)		p_b^c		p_b^+		δ		Treatment effect $\delta = p_b^c - p_b^+$	
			Ref.	Split, New	\hat{p}_b^c	\hat{p}_b^+	$\hat{\delta}$	Bias	RMSE	\widehat{spe}^c	\widehat{spe}^l	\widehat{sen}^c
Ep1	0.15	0.1	True	0.317 (± 0.025)	0.211 (± 0.022)	0.106	0.639	0.639	0.950	0.950	337.2	23.988
			Reference	0.322 (± 0.048)	0.215 (± 0.035)	0.108	0.639	0.640	0.946	0.946	336.3	24.210 (± 2.378)
			Split	0.325 (± 0.068)	0.217 (± 0.044)	0.108	0.637	0.637	0.942	0.942	168.6	24.331 (± 3.387)
Ep2	0.2	0.1	New	0.324 (± 0.052)	0.214 (± 0.035)	0.110	0.639	0.638	0.954	0.946	337.5	24.174 (± 2.409)
			True	0.397 (± 0.026)	0.198 (± 0.021)	0.198	0.639	0.639	0.950	0.950	359.3	23.988
			Reference	0.399 (± 0.046)	0.200 (± 0.028)	0.200	0.637	0.638	0.947	0.948	359.9	24.090 (± 2.062)
Ep3	0.25	0.1	Split	0.403 (± 0.065)	0.201 (± 0.034)	0.202	0.637	0.637	0.945	0.944	179.7	24.227 (± 2.938)
			New	0.402 (± 0.050)	0.200 (± 0.028)	0.202	0.639	0.638	0.953	0.947	359.7	24.123 (± 2.080)
			True	0.467 (± 0.026)	0.187 (± 0.020)	0.280	0.639	0.639	0.950	0.950	381.4	23.988
Ep4	0.3	0.1	Reference	0.469 (± 0.044)	0.187 (± 0.025)	0.281	0.638	0.638	0.948	0.948	381.6	24.078 (± 1.850)
			Split	0.471 (± 0.062)	0.188 (± 0.029)	0.283	0.637	0.637	0.945	0.946	190.7	24.182 (± 2.629)
			New	0.471 (± 0.047)	0.187 (± 0.025)	0.284	0.639	0.638	0.953	0.948	381.8	24.091 (± 1.865)
Ep5	0.4	0.1	True	0.530 (± 0.025)	0.177 (± 0.019)	0.353	0.639	0.639	0.950	0.950	403.4	23.988
			Reference	0.531 (± 0.041)	0.177 (± 0.022)	0.354	0.638	0.638	0.948	0.948	403.2	24.072 (± 1.701)
			Split	0.533 (± 0.058)	0.178 (± 0.024)	0.356	0.638	0.638	0.946	0.946	201.5	24.167 (± 2.398)
Ep6	0.5	0.1	New	0.533 (± 0.044)	0.177 (± 0.022)	0.356	0.639	0.639	0.952	0.948	403.6	24.079 (± 1.697)
			True	0.637 (± 0.023)	0.159 (± 0.017)	0.477	0.639	0.639	0.950	0.950	447.5	23.988
			Reference	0.637 (± 0.034)	0.159 (± 0.018)	0.478	0.638	0.638	0.949	0.949	447.4	24.040 (± 1.455)
Ep7	0.6	0.1	Split	0.638 (± 0.049)	0.160 (± 0.020)	0.478	0.638	0.638	0.947	0.947	223.6	24.103 (± 2.050)
			New	0.639 (± 0.037)	0.159 (± 0.018)	0.479	0.639	0.639	0.952	0.949	447.7	24.052 (± 1.462)
			True	0.724 (± 0.020)	0.145 (± 0.016)	0.579	0.639	0.639	0.950	0.950	491.6	23.988
Ep8	0.7	0.1	Reference	0.724 (± 0.029)	0.145 (± 0.016)	0.579	0.638	0.638	0.949	0.949	491.7	24.007 (± 1.306)
			Split	0.725 (± 0.040)	0.145 (± 0.017)	0.580	0.637	0.637	0.948	0.948	245.8	24.035 (± 1.836)
			New	0.725 (± 0.030)	0.145 (± 0.016)	0.580	0.639	0.638	0.951	0.949	492.1	24.017 (± 1.305)
Ep9	0.8	0.1	True	0.798 (± 0.017)	0.133 (± 0.015)	0.665	0.639	0.639	0.950	0.950	535.8	23.988
			Reference	0.798 (± 0.023)	0.133 (± 0.015)	0.665	0.638	0.638	0.949	0.949	535.8	24.003 (± 1.188)
			Split	0.798 (± 0.032)	0.133 (± 0.015)	0.665	0.637	0.637	0.949	0.949	267.9	24.013 (± 1.696)
Ep10	0.9	0.1	New	0.799 (± 0.024)	0.133 (± 0.015)	0.666	0.639	0.638	0.951	0.949	536.2	24.006 (± 1.193)

Table 3 Results for varying sample size n

Scenario	n	Event proportions for biomarker positive (\pm standard deviation) Treatment effect $\delta = p_{b+}^c - p_{b+}^+$																						
		Sample size Estimand (True)	p_{b+}^c	\hat{p}_{b+}^c	δ	Bias	RMSE	\widehat{spe}^c	\widehat{sen}^c	\widehat{sen}^c	\widehat{sen}^c	\widehat{sen}^c	\widehat{sen}^c	\widehat{sen}^c	\widehat{sen}^c	\widehat{sen}^c	\widehat{sen}^c	\widehat{sen}^c	\widehat{sen}^c	Cut-off				
Sam1	150	True	0.317 (± 0.057)	0.212 (± 0.050)	0.105	0.0103	0.0490	0.6380	0.6390	0.9500	0.950	0.675	0.23	0.988										
		Reference	0.347 (± 0.110)	0.231 (± 0.082)	0.116	0.0128	0.0922	0.6500	0.6500	0.9270	0.927	0.655	0.25	0.338										
		Split	0.355 (± 0.152)	0.237 (± 0.101)	0.118	0.0247	0.0574	0.6380	0.6390	0.9180	0.916	0.653	0.25	0.667										
Sam2	200	New	0.361 (± 0.122)	0.231 (± 0.083)	0.130	0.0067	0.0382	0.6530	0.6500	0.9680	0.927	0.660	0.25	0.328										
		True	0.317 (± 0.049)	0.211 (± 0.043)	0.105	0.0047	0.0446	0.6380	0.6380	0.9500	0.950	0.638	0.23	0.988										
		Reference	0.337 (± 0.094)	0.225 (± 0.070)	0.112	0.0159	0.0382	0.6430	0.6430	0.9340	0.934	0.643	0.24	0.855										
Sam3	250	Split	0.331 (± 0.129)	0.221 (± 0.084)	0.110	0.0038	0.0325	0.6080	0.6080	0.9320	0.932	0.608	0.24	0.855										
		New	0.345 (± 0.097)	0.224 (± 0.070)	0.121	0.0072	0.0611	0.6430	0.6400	0.9660	0.935	0.643	0.24	0.769										
		True	0.317 (± 0.044)	0.211 (± 0.039)	0.106	0.0124	0.0367	0.6390	0.6390	0.9500	0.950	0.639	0.23	0.988										
Sam4	300	Reference	0.329 (± 0.082)	0.219 (± 0.062)	0.110	0.0047	0.0294	0.6350	0.6340	0.9390	0.939	0.635	0.24	0.434										
		Split	0.337 (± 0.118)	0.224 (± 0.077)	0.113	0.0026	0.0237	0.6260	0.6250	0.9310	0.931	0.626	0.24	0.661										
		New	0.337 (± 0.091)	0.219 (± 0.061)	0.118	0.0092	0.057	0.6360	0.6340	0.9650	0.939	0.636	0.24	0.410										
Sam5	400	True	0.317 (± 0.040)	0.212 (± 0.035)	0.106	0.0115	0.0332	0.6390	0.6390	0.9500	0.950	0.639	0.23	0.988										
		Reference	0.331 (± 0.076)	0.221 (± 0.056)	0.110	0.0081	0.0263	0.6430	0.6420	0.9390	0.939	0.643	0.24	0.596										
		Split	0.345 (± 0.109)	0.230 (± 0.072)	0.115	0.0026	0.0237	0.6490	0.6490	0.9280	0.927	0.649	0.25	0.265										
Sam6	750	New	0.338 (± 0.082)	0.221 (± 0.057)	0.117	0.0056	0.0489	0.6440	0.6420	0.9600	0.939	0.644	0.24	0.591										
		True	0.318 (± 0.035)	0.211 (± 0.030)	0.106	0.0016	0.0161	0.6390	0.6390	0.9500	0.950	0.639	0.23	0.988										
		Reference	0.326 (± 0.065)	0.217 (± 0.048)	0.109	0.0025	0.0237	0.6380	0.6380	0.9430	0.943	0.638	0.24	0.333										
Sam7	1500	Split	0.336 (± 0.092)	0.224 (± 0.061)	0.112	0.0025	0.0328	0.6420	0.6420	0.9350	0.935	0.642	0.24	0.793										
		New	0.332 (± 0.071)	0.217 (± 0.048)	0.115	0.0043	0.0177	0.6400	0.6390	0.9580	0.942	0.640	0.24	0.359										
		True	0.317 (± 0.025)	0.211 (± 0.022)	0.106	0.0008	0.0109	0.6390	0.6390	0.9500	0.950	0.639	0.23	0.988										
Sam8	5000	Reference	0.322 (± 0.048)	0.215 (± 0.035)	0.108	0.0015	0.0161	0.6390	0.6400	0.9460	0.946	0.639	0.24	0.210										
		Split	0.325 (± 0.068)	0.217 (± 0.044)	0.108	0.0015	0.0224	0.6370	0.6370	0.9420	0.942	0.637	0.24	0.331										
		New	0.324 (± 0.052)	0.214 (± 0.035)	0.110	0.0022	0.0115	0.6390	0.6380	0.9540	0.946	0.639	0.24	0.174										
Sam8	5000	True	0.317 (± 0.018)	0.211 (± 0.016)	0.106	0.0002	0.0056	0.6390	0.6390	0.9500	0.950	0.639	0.24	0.988										
		Reference	0.320 (± 0.034)	0.213 (± 0.025)	0.107	0.0006	0.0056	0.6390	0.6390	0.9480	0.948	0.639	0.24	0.094										
		Split	0.322 (± 0.048)	0.215 (± 0.031)	0.107	0.0006	0.0119	0.6390	0.6390	0.9460	0.946	0.639	0.24	0.187										
Sam8	5000	New	0.321 (± 0.036)	0.213 (± 0.025)	0.108	0.0006	0.0057	0.6390	0.6390	0.9520	0.948	0.639	0.24	0.090										
		True	0.317 (± 0.010)	0.211 (± 0.009)	0.106	0.0006	0.0056	0.6390	0.6390	0.9500	0.950	0.639	0.24	0.988										
		Reference	0.318 (± 0.018)	0.212 (± 0.014)	0.106	0.0006	0.0057	0.6390	0.6390	0.9490	0.949	0.639	0.24	0.050										
Sam8	5000	Split	0.318 (± 0.026)	0.212 (± 0.017)	0.106	0.0007	0.0057	0.6390	0.6390	0.9510	0.949	0.639	0.24	0.023										
		New	0.318 (± 0.020)	0.212 (± 0.013)	0.106	0.0007	0.0057	0.6390	0.6390	0.9510	0.949	0.639	0.24	0.023										

New, prognostic-efficacy-combination design; Reference, external biomarker study; Split, split of control into training and test; True, true biomarker cut-off.

Table 4 Results for varying the lower bound for sensitivity *sen*

Scenario	<i>sen</i>	Event proportions for biomarker positive (\pm standard deviation)				Treatment effect $\delta = p_{b+}^c - p_{b+}^+$			
		p_{b+}^c	p_{b+}^+	\hat{p}_{b+}^c	\hat{p}_{b+}^+	δ	Bias	RMSE	Cut - Off ($\pm SD$)
Sen1	0.95	True	0.317 (± 0.025)	0.211 (± 0.022)	0.106	0.0016	0.0161	0.6390.6390.9500.950337.2.23.988	
		Reference	0.322 (± 0.048)	0.215 (± 0.035)	0.108	0.0025	0.0328	0.6390.6400.9460.946336.3.24.210 (± 2.378)	
		Split	0.325 (± 0.068)	0.217 (± 0.044)	0.108	0.0043	0.0177	0.6390.6370.9420.942168.624.331 (± 3.387)	
Sen2	0.9	True	0.324 (± 0.052)	0.214 (± 0.035)	0.110	0.0012	0.0185	0.6390.6380.9540.946337.524.174 (± 2.409)	
		Reference	0.402 (± 0.031)	0.268 (± 0.028)	0.134	0.0020	0.0379	0.7640.7640.9000.900251.928.767	
		Split	0.406 (± 0.053)	0.270 (± 0.040)	0.136	0.0041	0.0196	0.7630.7630.8960.896252.028.947 (± 2.310)	
Sen3	0.8	True	0.409 (± 0.075)	0.272 (± 0.049)	0.136	0.0004	0.0221	0.7610.7610.8920.893126.329.078 (± 3.295)	
		Reference	0.408 (± 0.059)	0.270 (± 0.041)	0.138	0.0039	0.0228	0.7620.7620.9040.897253.228.898 (± 2.330)	
		Split	0.534 (± 0.039)	0.355 (± 0.037)	0.178	0.0004	0.0221	0.8770.8770.8000.800168.635.844	
Sen4	0.7	True	0.535 (± 0.058)	0.356 (± 0.047)	0.179	0.0004	0.0221	0.8750.8750.7970.797169.235.980 (± 2.423)	
		Reference	0.536 (± 0.082)	0.357 (± 0.055)	0.179	0.0004	0.0449	0.8730.8730.7940.79585.136.059 (± 3.424)	
		Split	0.538 (± 0.066)	0.356 (± 0.047)	0.182	0.0039	0.0228	0.8750.8750.8040.797169.835.951 (± 2.419)	
Sen5	0.5	True	0.638 (± 0.043)	0.425 (± 0.045)	0.213	-0.0002	0.0249	0.9300.9300.6990.700123.342.005	
		Reference	0.637 (± 0.059)	0.424 (± 0.052)	0.213	-0.0013	0.0508	0.9280.9280.6990.699124.342.052 (± 2.620)	
		Split	0.636 (± 0.084)	0.424 (± 0.059)	0.212	0.0028	0.0253	0.9260.9260.6980.69962.742.074 (± 3.703)	
Sen6	0.3	True	0.640 (± 0.068)	0.424 (± 0.053)	0.216	0.0005	0.0302	0.9290.9280.7040.700124.842.017 (± 2.613)	
		Reference	0.795 (± 0.048)	0.529 (± 0.060)	0.266	-0.0011	0.059	0.9770.9770.4990.50070.754.598	
		Split	0.793 (± 0.057)	0.528 (± 0.063)	0.265	0.0021	0.0309	0.9760.9760.5000.50071.254.677 (± 3.229)	
New	prognostic- efficacy-combination design;	True	0.791 (± 0.081)	0.527 (± 0.067)	0.264	-0.0012	0.0396	0.9760.9760.5010.50136.054.678 (± 4.577)	
		Reference	0.796 (± 0.064)	0.528 (± 0.064)	0.268	-0.0024	0.0682	0.9770.9760.5020.50171.454.619 (± 3.226)	
		Split	0.902 (± 0.049)	0.600 (± 0.081)	0.301	0.0006	0.0392	0.9940.9940.3000.30037.470.966	
New	external biomarker study;	True	0.899 (± 0.053)	0.599 (± 0.082)	0.300	0.0012	0.0396	0.9940.9940.3040.30538.270.757 (± 4.358)	
		Reference	0.896 (± 0.076)	0.597 (± 0.083)	0.299	-0.0024	0.0682	0.9930.9930.3100.31019.670.476 (± 6.155)	
		Split	0.900 (± 0.056)	0.598 (± 0.082)	0.302	0.0006	0.0392	0.9940.9940.3040.30538.170.702 (± 4.371)	

New, prognostic-*efficacy-combination design*; Reference, external biomarker study; Split, split of control into training and test; True, true biomarker cut-off.

Table 1 shows the results for nine scenarios Bio1–Bio9 for varied biomarker distributions. The scenarios correspond to biomarker distributions that are increasingly well separated (see Figure 2). Intuitively, the more the biomarker distributions are separated, the higher are the specificities given a fixed sensitivity, and the smaller are the biomarker positive subgroups. Simultaneously, the treatment effect for the subgroup of biomarker positive increases, indicating that the group of potential profitters can be better identified. In addition to the results in Table 1, Figure 3A shows the distribution of estimates resulting from the three cut-off estimation methods relative to zero bias, which is indicated by the red horizontal line. For the first scenario, most of the patients are defined as biomarker positive; therefore, the bias is small. The size of the bias and the RMSE increase for all three methods (Reference, Split, and New) with better separation between the biomarkers of PP and NP. This behaviour changes from scenario Bio5 to Bio6, from which on the bias decreases again. Scenario B9 has a perfect specificity of 1, and the corresponding average event proportion is 1 in the biomarker positive subgroup of the control group, whereas the treatment effect is on average 0.334 and thus equal to the treatment effect for the potential profitters.

Table 2 shows the results for seven scenarios Ep1–Ep7 for varied event proportions. The group size of biomarker positive and the treatment effect increase with increasing event proportion under control condition, which is expected, as the group of potential profitters increases with higher p^C . At the same time, the bias of the treatment effect decreases for all three cut-off derivation methods. The bias of the externally defined cut-off (Reference) and of the cut-off definition via splitting design (Split) decreases faster than the bias of the new design. The RMSE remains relatively constant, but is slightly smaller for the more extreme scenarios Ep1 and Ep7, which can also be seen in Figure 3B.

Table 3 shows the results for eight scenarios Sam1–Sam8 for varied sample sizes. As expected, increasing the sample size results in a smaller bias and RMSE. The distribution of estimates from all approaches is visualized in Figure 3C. The specificity remains constant, as the biomarker distribution does not change. Though, for sample sizes of 750 and smaller (Sam1–Sam6), the estimated specificity differs slightly from the true value, and this difference increases with decreasing sample size. This additionally shows the loss of accuracy in estimation using small sample sizes.

Figure 2 Examples for biomarker distribution in the population for scenarios Bio1 ($\mu_{PP} = 3.3$), Bio4 ($\mu_{PP} = 4.0$), Bio8 ($\mu_{PP} = 5.0$), and Bio9 ($\mu_{PP} = 6.0$) with corresponding biomarker cut-off representing 95% sensitivity.

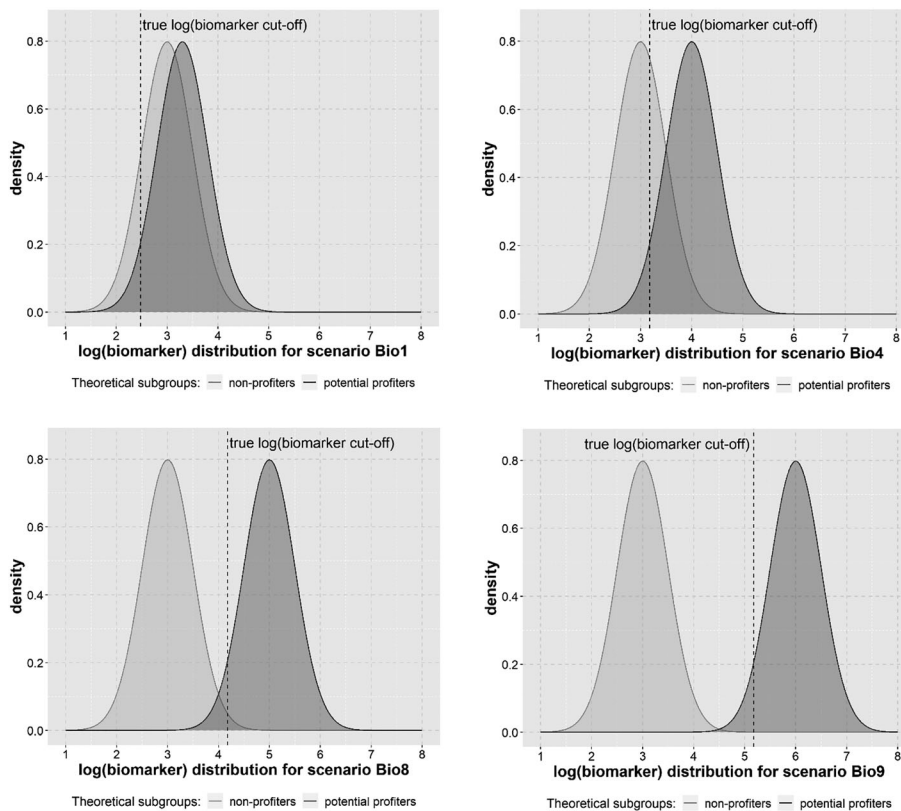


Figure 3 Boxplots for different scenarios; red line indicates targeted value.

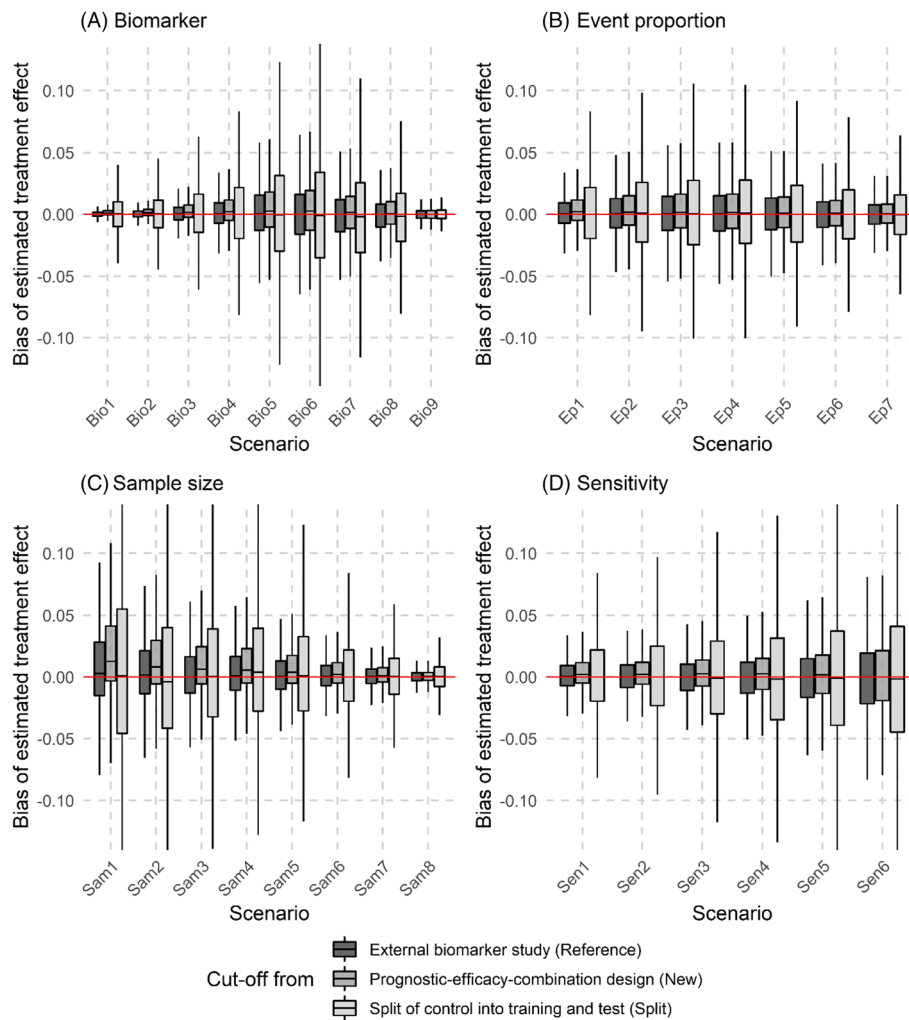


Table 4 and Figure 3D show the results for six scenarios Sen1–Sen6 for varying the lower bound for the sensitivity. As expected, specificity increases with decreasing sensitivity. At the same time, the size of the selected subgroup of *biomarker positive* decreases because the proportion of *PP* in this subgroup increases, and therefore, the treatment effect for this subgroup increases. The bias of the estimated treatment effect decreases, and the RMSE of the bias increases with smaller sensitivity for all the estimation methods.

Discussion

To estimate the treatment effect of a telemedical intervention for a biomarker-selected population likely to profit from the intervention, one would traditionally first conduct a separate biomarker study to define an optimal biomarker cut-off.

In a second step, an RCT for the biomarker-identified target population can then be realized. This process including two separate studies is time consuming and costly. Therefore, the new prognostic-efficacy-combination design that combines the two studies in one single study is appealing. With the simulation study presented in here, we showed that the new design, however, causes some, albeit limited, upward bias in treatment effect estimation. To reduce bias, we recommend a sufficiently large sample size depending on the specific trial scenario. For the various settings investigated here, a sample size of $n = 750$ per group works quite well. Strikingly, our simulations show that a random sample-splitting approach with the sample size of $n = 750$, which entails the same cost and effort for the clinical investigators as the new design, would only minimally reduce bias, but substantially increase RMSE. As every simulation study can only provide a limited number of scenarios and settings, we recommend testing a desired configuration of parameters

before using the design. To do so, we provide the code of our simulation.⁸ Besides bias, simulations can also help to assess the reliability of inference (e.g. coverage of confidence intervals) under different settings.

It is important to keep in mind that throughout we assumed that a suitable biomarker and a desired sensitivity for the detection of the target population were specified in advance to data analysis, with only the optimal cut-off being unknown. As in any design, flexibly exploring multiple subgroup definitions by different biomarkers and cut-offs post hoc can result in substantial additional bias and is discouraged.¹⁰ In terms of further adaptive design elements, one could also think of including an interim analysis for defining the biomarker cut-off. If the biomarker cut-off is only applied to the patients recruited after interim analysis, this would drastically reduce the sample size and is basically a variation of the reference approach. If the cut-off is applied to all patients, including those enrolled before interim analysis, the approach would constitute a mixture of the reference design (for the patients recruited after interim analysis) and the new design (for the patients recruited before interim analysis).

Note that in our work, we focus on one predefined biomarker. Combining or selecting several biomarkers is in principle possible, but not part of our simulation study. The consequence of incorporating multiple biomarkers should, among other factors, depend on the additive value of their information (i.e. their correlation) and the specific algorithm to combine them for patient selection. Using two instead of one biomarker could make it more sensitive to find the subgroup of interest. As we investigated our design for different sensitivity scenarios, we would conclude that higher sensitivity leads to slightly higher biases. This is in line with the consideration that the inclusion of more variables into a predic-

tion model is prone to overfitting and therefore higher biases.

Consistent with our results, other authors in the social sciences have suggested that bias in experimental treatment effect estimation due to in-sample definition of prognostic subgroups could be addressed with single random sample splitting.¹¹ They suggest further refinements like repeated splitting and leave-one-out estimators to control the inefficiency of the single random splitting approach, which was also apparent in our simulations. Such extensions could prove valuable in our setting as well. Other extensions might cover different types of endpoints that are non-binary, for example, continuous or time-to-event endpoints.

Acknowledgements

Open Access funding enabled and organized by Projekt DEAL.

Conflict of interest

MP and GR: None declared. MM: Received honoraria for lectures from Roche Diagnostics, AstraZeneca, Bayer Vital, Daiichi-Sankyo, Boehringer Ingelheim, and BRAHMS Thermo Fisher Scientific. He serves as a consultant for BRAHMS Thermo Fisher Scientific and Bayer and has received research funding from Roche Diagnostics and Radiometer. SG: Former employee of BRAHMS Thermo Fisher Scientific. JCW: Employee of BRAHMS Thermo Fisher Scientific. FK: Received grants from German Federal Ministry of Education and Research and received honoraria from Novartis, Sanofi, Boston Scientific, Linde, and Medtronic.

References

- Koehler F, Koehler K, Deckwart O, Prescher S, Wegscheider K, Winkler S, Vettorazzi E, Polze A, Stangl K, Hartmann O, Marx A, Neuhaus P, Scherf M, Kirwan BA, Anker SD. Telemedical interventional management in heart failure II (TIM-HF2), a randomised, controlled trial investigating the impact of telemedicine on unplanned cardiovascular hospitalisations and mortality in heart failure patients: Study design and description of the intervention. *Eur J Heart Fail* 2018; **20**: 1485–1493.
- Koehler F, Koehler K, Deckwart O, Prescher S, Wegscheider K, Kirwan B-A, Winkler S, Vettorazzi E, Bruch L, Oeff M, Zugck C, Doerr G, Naegel H, Störk S, Butter C, Sechtem U, Angermann C, Gola G, Prondzinsky R, Edelmann F, Spethmann S, Schellong SM, Schulze PC, Bauersachs J, Wellge B, Schoebel C, Tajsic M, Dreger H, Anker SD, Stangl K. Efficacy of telemedical interventional management in patients with heart failure (TIM-HF2): A randomised, controlled, parallel-group, unmasked trial. *Lancet* 2018; **392**: 1047–1057.
- Chapel JM, Ritchey MD, Zhang D, Wang G. Prevalence and medical costs of chronic diseases among adult Medicaid beneficiaries. *Am J Prev Med* 2017; **53**: S143–S154.
- Möckel M, Koehler K, Anker SD, Vollert J, Moeller V, Koehler M, Gehrig S, Wiemer JC, von Haehling S, Koehler F. Biomarker guidance allows a more personalized allocation of patients for remote patient management in heart failure: Results from the TIM-HF2 trial. *Eur J Heart Fail* 2019; **21**: 1445–1458.
- Chang M, Chow S-C, Pong A. Adaptive design in clinical research: Issues, opportunities, and recommendations. *J Biopharm Stat* 2006; **16**: 299–309.
- Cerqueira FP, Jesus AMC, Cotrim MD. Adaptive design: A review of the technical, statistical, and regulatory aspects of implementation in a clinical trial. *Ther Innov Regul Sci* 2020; **54**: 246–258.
- Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI. Circular analysis in systems neuroscience: The dangers

- of double dipping. *Nat Neurosci* 2009; **12**: 535–540.
8. Pigorsch M, Gehrig S, Rauch G. Simulation code for performance evaluation of a new prognostic-efficacy-combination design. GitHub. 2021. <https://github.com/mareenp/prognostic-efficacy-combination.git> (28 July 2022).
 9. R Core Team. R: A language and environment for statistical computing. Vienna, Austria; 2020.
 10. Lipkovich I, Dmitrienko A, D'Agostino RB Sr. Tutorial in biostatistics: Data-driven subgroup identification and analysis in clinical trials. *Stat Med* 2017; **36**: 136–196.
 11. Abadie A, Chingos MM, West MR. Endogenous stratification in randomized experiments. *Rev Econ Stat* 2018; **100**: 567–580.