BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Optimal designs for phase II/III drug development programs including methods for discounting of phase II results

Stella Erdmann[1*], Marietta Kirchner[1], Heiko Götte[2] and Meinhard Kieser[1]

Abstract

Background: Go/no-go decisions after phase II and sample size chosen for phase III are usually based on phase II results (e.g., the treatment effect estimate of phase II). Due to the decision rule (only promising phase II results lead to phase III), treatment effect estimates from phase II that initiate a phase III trial commonly overestimate the true treatment effect. Underpowered phase III trials are the consequence. Optimistic findings may then not be reproduced, leading to the failure of potentially expensive drug development programs. For some disease areas these failure rates are described to be quite high: 62.5%.

Methods: We integrate the ideas of multiplicative and additive adjustment of treatment effect estimates after go decisions in a utility-based framework for optimizing drug development programs. The design of a phase II/III program, i.e., the "right amount of adjustment", the allocation of the resources to phase II and III in terms of sample size, and the rule applied to decide whether to stop or to proceed with phase III influences its success considerably. Given specific drug development program characteristics (e.g., fixed and variable per patient costs for phase II and III, probable gain in case of market launch), optimal designs with respect to the maximal expected utility can be identified by the proposed Bayesian-frequentist approach. The method will be illustrated by application to practical examples characteristic for oncological studies.

Results: In general, our results show that the program set-ups with adjusted treatment effect estimate used for phase III planning are superior to the "naïve" program set-ups with respect to the maximal expected utility. Therefore, we recommend considering an adjusted phase II treatment effect estimate for the phase III sample size calculation. However, there is no one-fits-all design.

Conclusion: Individual drug development planning for a specific program is necessary to find the optimal design. The optimal choice of the design parameters for a specific drug development program at hand can be found by our user friendly R Shiny application and package (both assessable open-source via [1]).

Keywords: Optimization, Drug development program, Bias adjustment, Assurance, Probability of success, Sample size, Software

* Correspondence: erdmann@imbi.uni-heidelberg.de
[1]Institute of Medical Biometry and Informatics, University of Heidelberg, Im Neuenheimer Feld 130.3, D-69120 Heidelberg, Germany
Full list of author information is available at the end of the article

Background

Exploratory studies are usually carried out to provide a basis for deciding whether or not to proceed with a confirmatory trial and, if necessary, to provide information for planning purposes. In drug development programs, this strong link between exploratory (e.g., phase II) and confirmatory (e.g., phase III) studies favors integrated planning. In particular, the costs of phase III studies have increased remarkably in recent years [2, 3], while failure rates are quite high (approx. 45%, see [4] and the reference mentioned therein). Therefore, the availability and application of quantitative methods for decision making, which should be data-driven and objective, is desirable [5].

Already over 30 years ago, Hughes and Pocock [6] pointed out that decision rules in clinical trials can lead to a bias in the point estimate of the treatment effect, so that the true underlying effect might be overestimated at the time of an early positive decision. Twenty four years and various attempts of authors to adjust for overestimation of the treatment effect (in group sequential designs) later (e.g., [7] and references mentioned therein), Zhang et al. [8] still criticize that the cause and effect of this phenomenon is generally not well-understood. Trying to illustrate the problem, they provide a graphical explanation for the occurrence of overestimation. They argue that random variability (i.e., random highs and lows) of the treatment effect estimate is always present, but stabilizes around the true treatment effect as the trial continues to its end. However, when implementing a decision rule the variability favors the random highs: in a phase II/III drug development program with a go/no-go decision rule, it is only proceeded with phase III when large treatment effects are observed, but stopped when small effects occur. This selective handling of random variability may lead to overestimation of the magnitude of the treatment effect after phase II.

Ellenberg et al. [9] as well as Nardini [10] emphasize that the aim of treatment effect estimation is not to decide whether or not one therapy is better than the other, but to describe the size of therapeutic effects. Thus, we are concerned with a problem of estimation, not a problem of testing. Nardini concludes that estimates arising after a decision rule “should [consequently] not be taken at face value as true estimates of the new treatment’s effect”. Ellenberg et al. point out that statistical methods to adjust for this “random-high bias” exist, but criticize that “they are not applied as often as they should be”. Recently, the U.S. Food & Drug Administration reported 22 case studies since 1999 in which promising phase II clinical trial results were not confirmed in phase III clinical testing [11]. Such experiences are not rare: for some disease areas, the failure rate for phase III trials is reported to be as high as 62.5% [12] and about 50% for approval [13]. Chuang-Stein and Kirby [14] give cause for serious concern, as the severity of this may multiply, considering that the bigger the estimated effect from, e.g., a proof of concept trial, the greater the temptation to invest heavily and conduct multiple studies in parallel. They advise to use the concept of “assurance” for quantification of success probabilities and, moreover, to apply an adjustment for the overestimation of the treatment effect (e.g., [15]) when planning the next phase of a drug development program.

In our framework, we follow the concept of “assurance” [16, 17], which had first been introduced by Spiegelhalter et al. in 1986 with the concept of Bayesian predictive power (compare also “average power”) [18, 19]. This methodology was used later in various contexts by O’Hagan et al. [16, 17] (“assurance”), Chuang-Stein [20], Chuang-Stein and Yang [21] (“average success probability”) and finally by Gasparini et al. and Saint-Hilary et al. (“predictive probability of success”) [22, 23]. The idea is to use a prior distribution for the true assumed treatment effect for trial planning. This is in contrast to the “frequentist world”, where a fixed value is assumed. The “assurance” is then the weighted (unconditional) probability of a successful trial for a given effect, the weighting resulting from the likelihood that the therapy will achieve this effect. Due to synthesizing Bayesian principles in the planning phase and frequentistic decision-making procedures in the analysis, the above-mentioned approaches are described in the literature as “mixed Bayesian-frequentist”.

Kirby et al. [15] and Wang et al. [24] attempt to reduce the impact of overestimation by discounting the phase II treatment effect estimate by applying a multiplicative or additive adjustment, respectively. However, their suggestions are not universally applicable, and are rather “rules of thumb”, e.g., Kirby et al. suggest to use a retention factor of 0.9 times the assumed ratio of the phase III effect to phase II effect.

De Martini [4, 25] reports that the phase II sample size should be almost as large as the ideal phase III sample size (at least 2/3 of the latter) in order to have a sufficiently good information basis for phase III planning. He criticizes that in practice this ratio is only 1/4 on average and that an increase in sponsorship gains from drug development through larger phase II studies has not yet been well investigated. Larger phase II sample sizes would reduce the level of overestimation but increase the estimated phase III sample size [26] and could retrospectively be regarded as an unnecessary high investment in case of a no-go decision. Therefore, an optimal balance is required.

In this article, we integrate the general concepts of using a multiplicative or additive adjustment method to correct for overestimation of the treatment effect in a framework of utility-based optimization of phase II/III development programs [27]. That is, we want to critically examine adjustment methods from an economic point of view. In addition to simultaneously optimizing the phase II go/no-go decision rule and the sample size, we also optimize over the adjustment parameter used for the phase II treatment effect estimation to find "the right level of adjustment" for the specific situation at hand. Our approach can build the bridge between the long existing gap of theory and practice: we provide a Bayesian-frequentist hybrid framework, in which methods proposed for addressing the problem of overestimation of the treatment effect after go decisions are included in the optimization of drug development programs.

In the second section of this paper, we will introduce the basic setting and notation, explain the adjustment methods and show how they are incorporated in our optimization framework. After introducing the utility function and explaining the optimization procedure, we present optimal designs for exemplary settings of drug development programs in Section 3. We finish with a discussion in Section 4 and a conclusion in Section 5.

Methods

Basic setting

The considered drug development program consists of one exploratory phase II and one confirmatory phase III trial. Both are randomized trials with two arms (each with 1:1 sample size allocation), performed independently, investigating the same time-to-event primary endpoint and the same population. The true treatment effect is given by the negative logarithm of the true hazard ratio ($\theta = -\log(HR)$), which is the ratio of the hazard functions of the treatment and the control group. In order to reflect the uncertainty in the true treatment effect, θ can be modelled by a prior distribution $f(\theta)$. In phase II, the total number of events is denoted by d_2 and the maximum likelihood estimate of θ is given by $\hat{\theta}_2$. We assume that the estimator $\hat{\theta}_2$ is asymptotically normally distributed with $\hat{\theta}_2 \mid \theta \sim N(\theta, 4/d_2)$ (Note that the notation used will not differentiate between the treatment effect estimator (i.e., rule applied to estimate the quantity of interest, which is a random variable) and the treatment effect estimate (i.e., particular realization, fixed value), but by context it will be clear which quantity is meant.). Furthermore, we require that only phase II trials with promising results lead to a phase III trial. This is quantified by a go/no-go criterion with a go-decision in case of $\hat{\theta}_2 \geq \kappa$, where κ is a predefined threshold value. In case of a go decision, the number of events for the phase III trial is calculated based on the observed treatment effect of phase II. If the confirmatory analysis in phase III reveals a significant result, program success is declared (compare Fig. 1).

Due to the decision rule after phase II, the treatment effect estimate of phase II is biased. The bias is positive with $\kappa > 0$ as probability mass is shifted towards higher values:

$$\begin{aligned}
E\left[\hat{\theta}_2 | \hat{\theta}_2 \geq \kappa\right] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 1_{\{\hat{\theta}_2 \geq \kappa\}} \cdot \hat{\theta}_2 \cdot \frac{f\left(\hat{\theta}_2|\theta\right)}{P\left(\hat{\theta}_2 \geq \kappa|\theta\right)} d\hat{\theta}_2 \cdot f(\theta)d\theta \\
&+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 1_{\{\hat{\theta}_2 \geq \kappa\}} \cdot \hat{\theta}_2 \cdot 0\, d\hat{\theta}_2 \cdot f(\theta)d\theta \\
&> \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 1_{\{\hat{\theta}_2 \geq \kappa\}} \cdot \hat{\theta}_2 \cdot f\left(\hat{\theta}_2|\theta\right) d\hat{\theta}_2 \cdot f(\theta)d\theta \\
&+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 1_{\{\hat{\theta}_2 \geq \kappa\}} \cdot \hat{\theta}_2 \cdot f\left(\hat{\theta}_2|\theta\right) d\hat{\theta}_2 \cdot f(\theta)d\theta \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{\theta}_2 \cdot f\left(\hat{\theta}_2|\theta\right) \cdot f(\theta)d\hat{\theta}_2 d\theta = E\left[\hat{\theta}_2\right],
\end{aligned}$$

where here and in the following 1_A denotes the indicator function of event A and the density of the distribution of the respective argument is indicated by $f(.)$. The inequation holds as $\frac{1}{P(\hat{\theta}_2 \geq \kappa|\theta)} > 1$ and $\int_{-\infty}^{\infty} 1_{\{\hat{\theta}_2 \geq \kappa\}} \frac{f(\hat{\theta}_2|\theta)}{P(\hat{\theta}_2 \geq \kappa|\theta)} d\hat{\theta}_2 = 1$ and, therefore, the probability mass assigned to values less than κ in the unconditional expectation $E[\hat{\theta}_2]$ is distributed between values greater than κ in $E[\hat{\theta}_2|\hat{\theta}_2 \geq \kappa]$.

Note that the representation of the bias cannot be further simplified, neither by calculating $\mathrm{E}[\hat{\theta}_2] - \mathrm{E}[\hat{\theta}_2|\hat{\theta}_2 \geq \kappa]$ nor $\mathrm{E}[\hat{\theta}_2]/\mathrm{E}[\hat{\theta}_2|\hat{\theta}_2 \geq \kappa]$.

Therefore, in the following, multiplicative and additive adjustment methods for the treatment effect estimate obtained in phase II will be investigated. Afterwards, dependent on the respective adjustment method, launch criteria and approaches to calculate the number of events for phase III will be presented.

Additive and multiplicative adjustment methods

In this section, we introduce two methods (an additive and a multiplicative adjustment method) to adjust for the overestimation of the phase II treatment effect estimate. It should be mentioned that the terms "multiplicative" and "additive" relate to the specific type of scale and endpoint considered here.

Wang et al. [24] advise to apply an additive adjustment to the phase II treatment effect estimate if it is used for planning the sample size of phase III. They discuss using the lower limit of the one and two standard deviation confidence interval (CI) from the phase II trial (i.e., the lower limit of the CI for $\hat{\theta}_2$, corresponding to one or two standard deviations below the point estimate), respectively. We denote the significance level of the lower

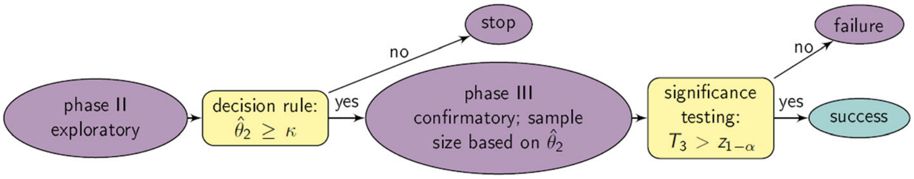

Fig. 1 Graphical illustration of basic phase II/III drug development program. The drug development program consists of one exploratory phase II trial, which is, in case of a go decision (i.e., treatment effect estimate of phase II $\hat{\theta}_2$ exceeds predefined threshold value $\kappa = -\log(HR_{go})$), followed by one confirmatory phase III trial, where the sample size planning is based on $\hat{\theta}_2$. The program is considered successful if phase III has a positive (significant) result (i.e., normalized log rank test statistic of phase III T_3 is above the $1-\alpha$ quantile of the standard normal distribution $z_{1-\alpha}$)

bound for the one-sided CI related to the phase II treatment effect estimate as $\alpha_{CI} \in [0.025, 0.5]$ and define the additive adjusted treatment effect estimate by $\hat{\theta}_2^{\alpha_{CI}} = \hat{\theta}_2 - z_{1-\alpha_{CI}} \cdot \sqrt{4/d_2}$, with $z_{1-\gamma} = \Phi^{-1}(1-\gamma)$, where $\Phi(.)$ denotes the distribution function of the standard normal distribution. Note that our version of the additive adjusted treatment effect estimate is a generalization of that of Wang et al., as they use the lower limit of the one and two standard deviation two-sided CI (i.e., in our notation $\alpha_{CI} = 0.32/2$ and $\alpha_{CI} = 0.05/2$) and we allow α_{CI} ranging from 0.025 to 0.5. For $\alpha_{CI} = 0.5$, the additive adjusted treatment effect estimate is not discounted as $\hat{\theta}_2 - z_{1-0.5} \cdot \sqrt{4/d_2} = \hat{\theta}_2$.

Kirby et al. [15] propose a multiplicative adjustment approach. They multiply the observed treatment effect estimate with a factor λ, which can be understood as a retention factor, that is, the fraction of the treatment effect retained. Integrated in our setting, we define $\hat{\theta}_2^{\lambda} = \lambda \cdot \hat{\theta}_2$, where the multiplicative adjustment parameter $\lambda \in [0.2, 1]$ can be viewed as the result of discounting the observed treatment effect of phase II by $1-\lambda$. Note that for $\lambda = 1$ the multiplicative adjusted treatment effect estimate is not discounted.

Go/no-go criteria, calculation of expected number of events for phase III and related program characteristics

When designing the phase II/III program, the observed treatment effect estimate of phase II plays a key role in two ways: 1. when making the go/no-go decision (selection s_1); 2. when calculating the phase III sample size (selection s_2; compare Fig. 1). At both instances, one has to decide whether or not to use an adjusted or unadjusted treatment effect estimate. To ease notation, the naïve (unadjusted) treatment effect estimate of phase II is denoted by $\hat{\theta}_2^{u} = \hat{\theta}_2$.

1.: If the treatment effect estimate $\hat{\theta}_2^{s_1}$, where $s_1 = \lambda$, α_{CI} or u (i.e., the multiplicatively adjusted, additively adjusted or unadjusted treatment effect estimate is selected for the decision rule), exceeds a predefined threshold value κ, it is decided to go to phase III and otherwise to stop the program. Hence, the expected probability to go to phase III can be determined by

$$p_{go}\left(\hat{\theta}_2^{s_1}\right) = \int_{-\infty}^{\infty} P\left(\hat{\theta}_2^{s_1} \geq \kappa | \theta\right) \cdot f(\theta) d\theta,$$

$s_1 = \lambda$, α_{CI} or u (compare Table A0 in the Additional file 1).

2.: In case of a go decision, the number of events for phase III is calculated based on the treatment effect estimate of phase II $\hat{\theta}_2^{s_2}$, $s_2 = \lambda$, α_{CI} or u, a desired power $1-\beta$, and a one-sided significance level α. For a balanced allocation ratio, it can be calculated by

$$D_3 = D_3\left(\hat{\theta}_2^{s_2}\right) = \frac{4 \cdot \left(z_{1-\alpha} + z_{1-\beta}\right)^2}{\left(\hat{\theta}_2^{s_2}\right)^2},$$

by assuming proportional hazards and asymptotic properties of the log-rank test statistic [28]. When going to phase III, the expected number of events (in phase II/III programs with decision rule $\hat{\theta}_2^{s_1} \geq \kappa$ and $\hat{\theta}_2^{s_2}$ used to calculate the number of events for phase III) can be determined by

$$\begin{aligned} d_3(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{s_2}) &= \mathrm{E}\left[D_3\left(\hat{\theta}_2^{s_2}\right) \cdot 1_{\{\hat{\theta}_2^{s_1} \geq \kappa\}}\right] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 1_{\{\hat{\theta}_2^{s_1} \geq \kappa\}} \cdot \frac{4 \cdot \left(z_{1-\alpha} + z_{1-\beta}\right)^2}{(\hat{\theta}_2^{s_2})^2} \\ &\cdot f(\hat{\theta}_2 | \theta) d\hat{\theta}_2 \cdot f(\theta) d\theta, \end{aligned}$$

(compare Table A0). The expectation of the estimate (of phase II) used for the sample size calculation can be calculated by

$$\begin{aligned} e_2 = e_2\left(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{s_2}\right) &= \mathrm{E}\left[\hat{\theta}_2^{s_2} | \hat{\theta}_2^{s_1} \geq \kappa\right] \\ &= \int_{-\infty}^{\infty} \frac{1}{\mathrm{P}\left(\hat{\theta}_2^{s_1} \geq \kappa | \theta\right)} \int_{-\infty}^{\infty} 1_{\{\hat{\theta}_2^{s_1} \geq \kappa\}} \\ &\cdot \hat{\theta}_2^{s_2} \cdot f(\hat{\theta}_2 | \theta) d\hat{\theta}_2 \cdot f(\theta) d\theta, \end{aligned}$$

for s_1, $s_2 = \lambda$, α_{CI} or u (compare Table A0) in order to calculate the bias $\mathrm{E}[\hat{\theta}_2^{s_2} | \hat{\theta}_2^{s_1} \geq \kappa] - \mathrm{E}[\hat{\theta}_2]$. As proposed by

De Martini [4, 25], the ratio of the number of events in phase II and III will also be calculated.

The program is considered to be successful, if the one-sided null hypothesis $H_0: \theta \leq 0$ is rejected in favour of $H_1: \theta > 0$ at a one-sided significance level α. This is the case if $T_3 > z_{1-\alpha}$, where T_3 is the normalized log-rank test statistic in phase III, which is assumed to be asymptotically normally distributed, i.e., $T_3 = T_3 \mid \hat{\theta}_2, \theta \sim N(\theta/\sqrt{4/D_3}, 1)$. Note that significance testing is performed on phase III data only. Therefore, the expected probability of a successful program $PsP(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{s_2})$ (with decision rule $\hat{\theta}_2^{s_1} \geq \kappa$, and $\hat{\theta}_2^{s_2}$ used to calculate the number of events for phase III), which is defined as the expected probability of the joint event of going to phase III and achieving a significant result [25, 27], can be calculated by

$$PsP\left(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{s_2}\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 1_{\{\hat{\theta}_2^{s_1} \geq \kappa\}} \cdot \int_{\{z_{1-\alpha}\}}^{\infty} f\left(t_3|\hat{\theta}_2, \theta\right) dt_3 \cdot f\left(\hat{\theta}_2|\theta\right) d\hat{\theta}_2 \cdot f(\theta) d\theta,$$

where t_3 is a realization of $T_3 \mid \hat{\theta}_2, \theta$ (compare Table A0). One reviewer pointed out that this definition of a successful program records a false positive result (i.e. $T_3 > z_{1-\alpha}$ under H_0) as program success. We discuss this aspect in detail in Section A1 of Additional file 1. In reality, regulatory approval and with that a monetary gain, which is the core driver for our utility function, is achieved when a significant result is observed in phase III, acknowledging that there is a probability of α that it is a false positive decision. Thus, we keep the commonly used term "success" and *PsP* which should be regarded as probability of market access and not a probability of a correct decision.

Considered program set-ups

We investigate the impact of using adjusted treatment effect estimates (i.e., $\hat{\theta}_2^{\lambda}$ or $\hat{\theta}_2^{\alpha_{CI}}$) for the go/no-go decision and/or for the calculation of the number of events for phase III on the drug development program characteristics and compare the results to those where the unadjusted (naïve) treatment effect estimate $\hat{\theta}_2^{u}$ was used. Therefore, we investigate different program set-ups $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{s_2})$ which are defined by the selection of the treatment effect estimate used for the decision rule (selection s_1) and, in case of a go decision, by the choice of the treatment effect estimate used for the calculation of the number of events for phase III (selection s_2).

Table 1 gives an overview of the considered program set-ups. We compare the "unadjusted" set-up $(\hat{\theta}_2^{u}, \hat{\theta}_2^{u})$, where $\hat{\theta}_2^{u} = \hat{\theta}_2$ (i.e., s_1, $s_2 = u$), with two "multiplicatively adjusted" set-ups $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{\lambda})$ ($s_1 \in \{u, \lambda\}$, $s_2 = \lambda$), and two "additively adjusted" set-ups $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{\alpha_{CI}})$ ($s_1 \in \{u, \alpha_{CI}\}$, $s_2 = \alpha_{CI}$). Note that if $s_1 \neq u$, we define $s_2 = s_1$, which means that if an adjustment parameter is used for the decision rule, the same adjustment parameter is used for the calculation of the expected number of events for phase III (for reasons which will be given later).

Utility function

The aim is to optimize a phase II/III drug development program in terms of the adjustment parameters λ or α_{CI}, the number of events in phase II d_2, and the go/no-go decision threshold value κ. Therefore, we set up a utility function, which utilizes the difference between program costs and potential gains after successful market launch (compare Fig. 2 for a graphical illustration). For the costs, fixed (c_{02}, c_{03}) and variable per-patient (c_2, c_3) costs are included for the phase II and III trial, respectively. By dividing the number of events by the event rate ξ_i, the total number of patients can be calculated for the respective phase $i = 2, 3$. Obviously, only in case of a go decision the costs of the phase III trial apply. In case of program success, a benefit b is obtained, and we assume that the level of benefit depends on the observed treatment effect in the phase III trial as suggested by a report of the German Institute for Quality and Efficiency in Health Care [29]. As proposed by them, three effect size categories (small, medium and large) are used, whereby each category is defined by a threshold value (1, 0.95, 0.85) for the upper boundary of the 95% confidence interval for the *HR* (for details on the derivation of these threshold values, the interested reader may be referred to the "Anhang A"of [29]). The corresponding amount of benefit is denoted by b_1, b_2 and b_3, respectively. Based on this, costs $c(d_2, \kappa, s_2)$ and gain $g(d_2, \kappa, s_2)$ for a phase II/III program with program set-up $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{s_2})$ are given by

$$c(d_2, \kappa, s_2) = c_{02} + \frac{d_2}{\xi_2} \cdot c_2 + 1_{\{\hat{\theta}_2^{s_1} \geq \kappa\}} \cdot \left(c_{03} + \frac{D_3}{\xi_3} \cdot c_3\right)$$

$$g(d_2, \kappa, s_2) = 1_{\{\hat{\theta}_2^{s_1} \geq \kappa\}} \cdot \left(b_1 \cdot 1_{\{T_3 \in I_1\}} + b_2 \cdot 1_{\{T_3 \in I_2\}} + b_3 \cdot 1_{\{T_3 \in I_3\}}\right),$$

where $I_1 = (z_{1-\alpha}, -\log(0.95)/\sqrt{4/D_3} + z_{1-\alpha}]$, $I_2 = (-\log(0.95)/\sqrt{4/D_3} + z_{1-\alpha}, -\log(0.85)/\sqrt{4/D_3} + z_{1-\alpha}]$ and $I_3 = (-\log(0.85)/\sqrt{4/D_3} + z_{1-\alpha}, \infty)$ are transformations of the effect size intervals to intervals on the test statistic scale of T_3. Thus, the costs depend on the observed treatment effect in phase II and the gain depends on the observed treatment effect in phase II and III.

The utility is defined as the difference between costs and gain and expressed as a function of d_2 and κ over which it is simultaneously optimized. In the adjusted

Table 1 Overview of program set-ups $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{s_2})$

Program set-up $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{s_2})$	Adjustment of the estimate used for decision rule	Estimate used for decision rule	Adjustment of the estimate used for calculating the number of events for phase III	Estimate used for calculating the number of events for phase III
$S(\hat{\theta}_2^{u}, \hat{\theta}_2^{u})$ (unadjusted)	none ($s_1 = u$)	$\hat{\theta}_2^{u}$	none ($s_2 = u$)	$\hat{\theta}_2^{u}$
$S(\hat{\theta}_2^{u}, \hat{\theta}_2^{\lambda})$ (multiplicative)			multiplicative ($s_2 = \lambda$)	$\hat{\theta}_2^{\lambda}$
$S(\hat{\theta}_2^{u}, \hat{\theta}_2^{\alpha_{CI}})$ (additive)			additive ($s_2 = \alpha_{CI}$)	$\hat{\theta}_2^{\alpha_{CI}}$
$S(\hat{\theta}_2^{\lambda}, \hat{\theta}_2^{\lambda})$ (multiplicative)	multiplicative ($s_1 = \lambda$)	$\hat{\theta}_2^{\lambda}$	multiplicative ($s_2 = \lambda$)	$\hat{\theta}_2^{\lambda}$
$S(\hat{\theta}_2^{\alpha_{CI}}, \hat{\theta}_2^{\alpha_{CI}})$ (additive)	additive ($s_1 = \alpha_{CI}$)	$\hat{\theta}_2^{\alpha_{CI}}$	additive ($s_2 = \alpha_{CI}$)	$\hat{\theta}_2^{\alpha_{CI}}$

Program set-ups are defined by the estimate used for the go/no-go decision (selection s_1: "go if $\hat{\theta}_2^{s_1} \geq \kappa$ ") and by the calculation of the number of events for phase III (selection s_2: $D_3(\hat{\theta}_2^{s_2}), s_2 \in \{\lambda, \alpha_{CI}, u\}$, where $\hat{\theta}_2^{\lambda} = \hat{\theta}_2 \cdot \lambda$, $\hat{\theta}_2^{\alpha_{CI}} = \hat{\theta}_2 - z_{1-\alpha_{CI}} \cdot \sqrt{4/d_2}$, and $\hat{\theta}_2^{u} = \hat{\theta}_2$ are the multiplicatively adjusted, additively adjusted, and unadjusted treatment effect estimates of phase II).

program set-ups, the optimization is also over λ in the multiplicatively, and over α_{CI} in the additively adjusted set-ups, respectively. Thus, we define the utility for program set-up $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{s_2})$ by

$$u(d_2, \kappa, s_2) = g(d_2, \kappa, s_2) - c(d_2, \kappa, s_2),$$

where for the unadjusted program set-up $S(\hat{\theta}_2^{u}, \hat{\theta}_2^{u})$ $u(d_2, \kappa, s_2) = u(d_2, \kappa)$. To incorporate the development risk in terms of success probabilities, we consider the expected utility with respect to θ, $\hat{\theta}_2$ and T_3 $\mathrm{E}[u(d_2, \kappa, s_2)] = \mathrm{E}[g(d_2, \kappa, s_2)] - \mathrm{E}[c(d_2, \kappa, s_2)]$, where the expected costs and gain with respect to θ, $\hat{\theta}_2$ and T_3 are given by

$$E[c(d_2, \kappa, s_2)] = c_{02} + d_2/\xi_2 \cdot c_2 + c_{03} \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 1_{\{\hat{\theta}_2^{s_1} \geq \kappa\}} \cdot f\left(\hat{\theta}_2 | \theta\right) \cdot f(\theta) d\hat{\theta}_2 d\theta + c_3/\xi_3 \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 1_{\{\hat{\theta}_2^{s_1} \geq \kappa\}} \cdot D_3\left(\hat{\theta}_2^{s_2}\right) \cdot f\left(\hat{\theta}_2 | \theta\right) \cdot f(\theta) d\hat{\theta}_2 d\theta, E[g(d_2, \kappa, s_2)]$$
$$= \sum_{j=1}^{3} b_j \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 1_{\{\hat{\theta}_2^{s_1} \geq \kappa\}} \cdot 1_{\{T_3 \in I_j\}} \cdot f\left(t_3 | \hat{\theta}_2, \theta\right) \cdot f\left(\hat{\theta}_2 | \theta\right) \cdot f(\theta) dt_3 d\hat{\theta}_2 d\theta.$$

The aim is to find a design $\delta = (d_2, \kappa, s_2)$ that maximizes the expected utility $\mathrm{E}[u(d_2, \kappa, s_2)]$ for programs with program set-up $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{s_2})$. The optimization is carried out over d_2, κ, and λ in the multiplicatively or α_{CI} in

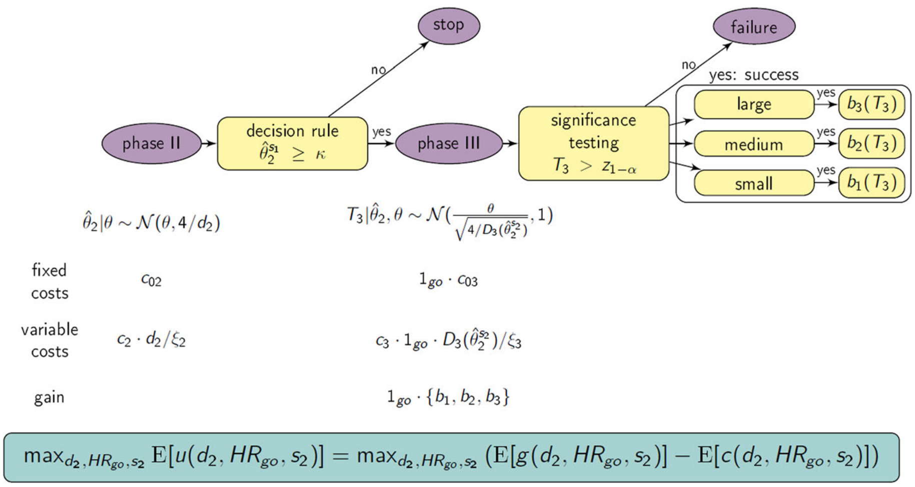
stop
failure
no
no
yes: success
phase II
decision rule
$\hat{\theta}_2^{s_1} \geq \kappa$
yes
phase III
significance testing
$T_3 > z_{1-\alpha}$
large
medium
small
yes
yes
yes
$b_3(T_3)$
$b_2(T_3)$
$b_1(T_3)$
$\hat{\theta}_2 | \theta \sim \mathcal{N}(\theta, 4/d_2)$
$T_3 | \hat{\theta}_2, \theta \sim \mathcal{N}(\frac{\theta}{\sqrt{4/D_3(\hat{\theta}_2^{s_2})}}, 1)$
fixed costs
c_{02}
$1_{go} \cdot c_{03}$
variable costs
$c_2 \cdot d_2/\xi_2$
$c_3 \cdot 1_{go} \cdot D_3(\hat{\theta}_2^{s_2})/\xi_3$
gain
$1_{go} \cdot \{b_1, b_2, b_3\}$
$\max_{d_2, HR_{go}, s_2} \mathrm{E}[u(d_2, HR_{go}, s_2)] = \max_{d_2, HR_{go}, s_2} (\mathrm{E}[g(d_2, HR_{go}, s_2)] - \mathrm{E}[c(d_2, HR_{go}, s_2)])$

Fig. 2 Graphical illustration of (adjusted) utility-based optimization. The treatment effect estimate of phase II may be adjusted for the decision rule (selection $s_1 \in \{u, s_2\}$ and/or for the calculation of the number of events for phase III (selection $s_2 \in \{\lambda, \alpha_{CI}, u\}$). The utility (including the costs and the gain) is optimized over the number of events for phase II d_2, the threshold value for the decision rule HR_{go}, and the adjustment parameter $s_2 = \lambda$ or α_{CI} (see Section 2.5 for details), ξ_i event rate in phase $i = 2, 3$, $b_j = b_j(T_3)$ benefit categories $j = 1, 2, 3$

the additively adjusted set-ups, respectively. The optimal design δ^* for each program set-up $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{s_2})$ is defined to be the design for which the expected utility is maximized, that is, $\mathrm{E}[u(\delta^*)] = \max_{\delta \in D} \mathrm{E}[u(\delta)]$, where $D = \{\delta = (d_2, \kappa, s_2)\}$ is the optimization set.

The optimization is solved by using numerical integration procedures written in the programming language R [30]. In order to facilitate the application of the approach, an user friendly R Shiny App (bias) and an R package (drugdevelopR including the R function optimal_bias) are provided open-source (both assessable via [1]).

Illustration of the framework by application to oncology trial example and practical extensions

In this paper, the parameters in the oncology trial example are chosen as in Kirchner et al. [27] to allow comparison of results. It should be noted that the example is primarily given to illustrate the framework and the chosen parameters should not be taken as face values. We tried to elicit the design parameters as realistic as possible to mimic an oncology drug development program by means of information from relevant literature and consultation with experts from the pharmaceutical industry in the field of oncology. However, it should be noted, that these parameters must be chosen carefully and specifically for each drug development scenario at hand.

The event rates for phase II and III are set to $\xi_i = 0.7$ for $i = 2, 3$. Therefore, the total sample size can be calculated by $d_i/0.7,\ i = 2, 3$. In practice, estimates on the event rates could be obtained by taking recruitment rates and duration as well as drop-out rates and treatment group specific hazards into account. However, using those parameters often leads to event rates around $\xi_i = 0.7$ as it is a compromise between data maturity and avoidance of long follow-up times if drop-out rates are higher than expected. If $\xi_i < 0.5$ the median event time might not be observed while if ξ_i is too high, the planned number of events might not be reached at all with substantial drop-out rates.

For phase III oncology trials, per-patient costs between 75,000 and 125,000 US *\$* are reported [31]. Therefore, per-patient costs for phase III of $\$10^5$ are considered and c_3 is set to 1 (in $\$10^5$). Furthermore, the per-patient costs for phase II are set to $c_2 = 0.75$ (in $\$10^5$). Due to, for example, additional biomarker measurements made in phase III, or because regulatory agencies may require more extensive data collection in phase III [32], higher per-patient costs in phase III compared with phase II are reasonable. In this example, the fixed costs for phase II and III are set to $c_{02} = 100$ and $c_{03} = 150$ (in $\$10^5$), respectively. To investigate different scenarios, the benefit parameters b_1, b_2 and b_3 are chosen to embody a low $(b_1, b_2, b_3)\ 1 : (1000, 2000, 3000),\ 2 : (1000, 2000, 4000),\ 3 : (1000, 3000, 4000)$ and a medium to large $(b_1, b_2, b_3) = 4 : (1000, 3000, 5000),\ 5 : (1000, 4000, 5000),\ 6 : (1000, 3000, 6000),\ 7 : (1000, 4000, 6000)$ over-all benefit (in $\$10^5$), where we assume a 5-year income period and profit margin of 0.2. Thus, seven different benefit scenarios (*bs* 1–7) will be considered. A mixture distribution consisting of the weighted sum of two normal distributions

$$\theta \sim w \cdot N(-\ log(0.69), (4/210)) + (1 - w) \cdot N(-\ log(0.88), (4/420)),$$

as proposed by Götte et al. [26] can be used to model the true treatment effect. The two normal distributions each depict a distribution for θ, whereby the means represent values of the assumed true treatment effect and the denominators of the associated variances can be viewed as "amount of certainty" about the treatment effect size in terms of numbers of events. The parameters of the distributions (i.e., means and variances) are elected such that a realistic range for the *HR* is covered (compare Fig. A2 in Additional file 1 and/or investigate the prior distribution with the help of our R shiny App prior [33]). The mean of the first of the two normal distributions characterizes a strong, the second one a moderate to low treatment effect, so that by ranging w from, e.g., 0.3 to 0.9 we can mirror pessimistic to more optimistic opinions about the true treatment effect. In practice, the choice of w can be guided by formal expert elicitation methods. Dallow et al. [34] presented an overview of such methods including elicitation of Gaussian mixture distributions. Note that the approach is general and allows for implementation of any alternative prior distribution. Again, elicitation methods (compare also, e.g., [35]) are a useful tool that may help (a group of) experts to quantify their opinions about the treatment effect as a probability distribution. Various software packages enable their practical application (compare, e.g., [36]).

In our framework it is also possible to account for, e.g., different population structures in phase II and phase III (due to different countries, centers, in-/exclusion criteria, ...) by assuming different distributions for the assumed true treatment effect in phase II and III (i.e., $\theta_2 \nsim \theta_3$), so that $\hat{\theta}_2 \mid \theta_2 \sim N(\theta_2, 4/d_2)$ and $T_3 \mid \hat{\theta}_2, \theta_2, \theta_3 \sim N(\theta_3/\sqrt{4/D_3}, 1)$. For ease of interpretation, all formulas and results presented in the main part are for the special case, where the true treatment effect is modelled by the same distribution for phase II and III (e.g., $\theta \sim \theta_2 \sim \theta_3$), and a brief investigation of this aspect can be found in Section A2 of Additional file 1.

In this example, we chose a wide range for κ (and d_2, as well as λ or α_{CI}, respectively) such that the optimization is not influenced by that choice. Therefore, the optimization set is $\mathrm{D} = \{\delta = (d_2, \kappa, s_2), d_2 \in \{50, 52, \ldots, 350\}, \kappa \in \{-\log(0.9),$

$-\log(0.89), \ldots, -\log(0.7)\}, s_2 = \lambda \in \{0.2, 0.225, \ldots, 1\}$ or $s_2 = \alpha_{CI} \in \{0.025, 0.075, \ldots, 0.5\}\}$. However, the lower bound of the decision rule set for κ can also be seen as representing a predefined clinically relevant effect size: phase III trials are then only conducted if the treatment effect observed in phase II is at least of that size. In Section A3 of Additional file 1, we present results of the procedure, where we chose $\min(\kappa) = -\log(0.8)$. Furthermore, it might be interesting to see how the optimal program design is influenced by the sponsor's real life budget constraint. Therefore, we also consider optimizing the drug development program with a constraint K on the expected costs of the program, i.e., $\mathrm{E}[c(d_2, \kappa, s_2)] \leq K$ (see Section A4 of Additional file 1 for more details). In pharmaceutical industry there are often discussions about skipping the phase II trial. For example, if competitors have already approved a drug with a similar mode of action one might see no need for further learning about the drug and go directly to a confirmatory trial. Our framework allows to systematically assess this aspect by setting $d_2 = 0$, $c_{02} = c_2 = 0$ and $p_{go} = 1$ (see Section A5 of Additional file 1 for more details). In addition, different definitions of the cost and benefit functions are possible. As mentioned above, the choice of three effect size categories (and therefore the benefit function) is based on a report of the German Institute for Quality and Efficiency in Health Care [29]. However, the presented framework could also be applied to an alternative set-up as, for example, the one proposed by Ding et al. [32]. Here, a proportional relationship between benefit and effect size is considered. In Section A6 of Additional file 1 we investigate this possibility in more detail.

Results

This section is organized as follows. It starts with general observations across all program set-ups $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{s_2})$. Then, we compare multiplicative $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{\lambda})$ vs. additive $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{\alpha_{CI}})$ vs. no adjustment $S(\hat{\theta}_2^{u}, \hat{\theta}_2^{u})$, where $s_1 = u$ or $s_1 = s_2$. The impact of adjusting the go/no go decision making, i.e., the differences between both multiplicative ($S(\hat{\theta}_2^{u}, \hat{\theta}_2^{\lambda})$ vs. $S(\hat{\theta}_2^{\lambda}, \hat{\theta}_2^{\lambda})$) and both additive adjustment methods ($S(\hat{\theta}_2^{u}, \hat{\theta}_2^{\alpha_{CI}})$ vs. $S(\hat{\theta}_2^{\alpha_{CI}}, \hat{\theta}_2^{\alpha_{CI}})$) are also presented. A discussion of the results is given in the next section.

The optimization results are presented in Table 2 (naïve setting, multiplicative adjustment), Table 3 (additive adjustment) and Figure 3, which show the optimal design parameters $\delta^* = (d_2^*, \kappa^*, s_2^*)$:

- optimal total number of events for phase II d_2^* (given by the optimal value of $d_2 \in D$),
- optimal go/no-go decision rule threshold value HR_{go}^* (given by the optimal value of $\kappa \in D$ in "HR-scale", i.e., $HR_{go}^* = \exp(-\kappa^*)$) and
- optimal adjustment parameter $s_2^* \in \{\lambda^*, a_{CI}^*\}$ (given by the optimal value of $s_2 \in D$) for the multiplicative and additive adjustment method, respectively,

with corresponding program characteristics for the optimal design:

- maximal expected utility $u^* = E[u(\delta^*)]$,
- expected number of events for phase III $d_3^* = d_3(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{s_2^*})$, where we chose a desired power of $1 - \beta = 0.9$ and a one-sided significance level $\alpha = 0.025$,
- total number of expected events in the program $d^* = d_3^* + d_2^*$,
- expected probability to go to phase III $p_{go}^* = p_{go}(\hat{\theta}_2^{s_1})$,
- expected probability of a successful program $sP^* = PsP(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{s_2^*})$ and
- expected estimate of phase II used for sample size calculation $\varepsilon_2^* = \exp(-e_2(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{s_2^*}))$ in "HR-scale",

for program set-ups $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{s_2})$, where $s_1 = u$ or $s_1 = s_2^* \in \{\lambda^*, a_{CI}^*\}$, benefit scenarios (*bs* 1-7) and weights for the prior distribution of the true underlying effect ($w = 0.3$, 0.6, 0.9), where $E[\hat{\theta}_2] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \hat{\theta}_2 \cdot f(\hat{\theta}_2|\theta) \cdot f(\theta) d\hat{\theta}_2 d\theta$.

Overall, larger assumed benefits (i.e., larger values for (b_1, b_2, b_3)) lead to more liberal optimal decision rules (i.e., larger values for HR_{go}^*) and higher investment in phase II (i.e., larger number of events for phase II d_2^*). This leads to a larger investment (in phase III), i.e., a higher expected probability to go to phase III p_{go}^* and a larger expected number of events in phase III d_3^*, respectively. This results in a larger expected probability of a successful program sP^* and thus in a larger maximal expected utility u^*.

In the multiplicatively adjusted program set-ups $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{\lambda})$, the maximal expected utility is always higher than the maximal expected utility in the additively adjusted program set-ups $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{\alpha_{CI}})$, which in turn is always higher than the maximal expected utility in the unadjusted program set-up $S(\hat{\theta}_2^{u}, \hat{\theta}_2^{u})$. It stands out that the investment in terms of numbers of events (i.e., d_2^*, d_3^*, d^*) tends to be higher in the adjusted program set-ups compared to the unadjusted program set-up, especially for scenarios with higher benefits and more optimistic prior. The expected probability to go to phase III p_{go}^* is notably lower in the adjusted program set-ups compared to the unadjusted program set-up, whereas the expected probability of a successful program sP^* is higher.

Dividing the optimal number of events in phase II by the expected number of events in phase III (i.e.,

Table 2 Optimal design parameters for unadjusted and multiplicatively adjusted program set-ups

	Unadjusted								Multiplicatively adjusted																	
	Program set-up $S(\hat{\theta}_2^u, \hat{\theta}_2^u)$								Program set-up $S(\hat{\theta}_2^u, \hat{\theta}_2^\lambda)$									Program set-up $S(\hat{\theta}_2^\lambda, \hat{\theta}_2^\lambda)$								
bs	HR_{go}^*	d_2^*	ε_2^*	d_3^*	d^*	p_{go}^*	sP^*	u^*	λ^*	HR_{go}^*	d_2^*	ε_2^*	d_3^*	d^*	p_{go}^*	sP^*	u^*	λ^*	HR_{go}^*	d_2^*	ε_2^*	d_3^*	d^*	p_{go}^*	sP^*	u^*
$\boldsymbol{w}$ **= .3**, i.e., $\mathbf{exp}(-\mathbf{E}[\hat{\boldsymbol{\theta}}_\mathbf{2}])$=**.82**																										
1	.80	82	.65	146	228	.46	.24	76	.750	.76	81	.70	170	251	.38	.25	99	.750	.81	84	.70	161	245	.37	.25	100
2	.82	109	.67	189	298	.49	.28	188	.700	.77	116	.73	222	338	.39	.29	235	.725	.83	112	.73	214	326	.40	.29	235
3	.83	133	.68	218	351	.51	.31	299	.750	.80	133	.74	275	408	.45	.33	343	.750	.84	133	.73	252	385	.43	.32	343
4	.84	144	.69	248	392	.53	.33	432	.700	.80	158	.75	320	478	.44	.35	509	.725	.85	161	.75	296	457	.44	.34	508
5	.85	161	.70	284	445	.55	.35	569	.700	.81	196	.76	366	562	.46	.38	690	.700	.86	182	.76	348	530	.45	.37	690
6	.85	172	.70	287	459	.55	.35	567	.750	.82	179	.75	357	536	.48	.38	640	.750	.86	175	.75	347	522	.48	.37	640
7	.86	193	.71	331	524	.57	.38	712	.700	.82	196	.77	413	609	.48	.40	828	.700	.87	200	.77	412	612	.48	.40	828
$\boldsymbol{w}$ **= .6**, i.e., $\mathbf{exp}(-\mathbf{E}[\hat{\boldsymbol{\theta}}_\mathbf{2}])$=**.76**																										
1	.82	133	.65	213	346	.61	.43	370	.775	.79	126	.71	265	391	.55	.45	412	.775	.83	140	.71	259	399	.55	.45	411
2	.84	147	.66	262	409	.65	.46	598	.725	.80	175	.73	348	523	.58	.50	696	.725	.85	168	.73	343	511	.58	.50	696
3	.85	182	.67	299	481	.68	.50	764	.775	.82	196	.73	374	570	.62	.53	849	.750	.86	189	.73	390	579	.62	.53	849
4	.86	196	.68	333	529	.70	.52	1012	.700	.82	210	.75	462	672	.62	.56	1172	.700	.87	217	.75	462	679	.62	.56	1172
5	.86	210	.68	336	546	.70	.52	1267	.675	.82	245	.76	505	750	.62	.57	1523	.675	.88	238	.76	542	780	.64	.58	1521
6	.86	217	.68	338	555	.70	.53	1200	.750	.83	235	.74	450	685	.64	.57	1342	.750	.87	238	.74	453	691	.65	.57	1343
7	.87	217	.69	374	591	.72	.54	1460	.700	.83	259	.76	521	780	.65	.59	1693	.700	.88	249	.76	535	784	.65	.59	1693
$\boldsymbol{w}$ **= .9**, i.e., $\mathbf{exp}(-\mathbf{E}[\hat{\boldsymbol{\theta}}_\mathbf{2}])$=**.71**																										
1	.84	154	.65	278	432	.78	.60	693	.800	.81	161	.70	346	507	.73	.64	753	.775	.85	158	.71	370	528	.73	.65	753
2	.86	182	.66	332	514	.82	.65	1039	.725	.82	203	.73	472	675	.76	.70	1193	.725	.86	210	.73	447	657	.75	.69	1193
3	.86	207	.66	338	545	.83	.66	1255	.750	.83	217	.73	480	697	.78	.72	1384	.750	.87	231	.73	486	717	.79	.73	1384
4	.87	221	.67	367	588	.84	.68	1623	.700	.83	252	.75	562	814	.79	.75	1871	.700	.88	245	.75	573	818	.79	.75	1871
5	.88	235	.67	399	634	.86	.70	1996	.650	.83	287	.76	661	948	.80	.77	2394	.675	.89	280	.76	665	945	.81	.78	2394
6	.88	245	.67	401	646	.86	.70	1855	.725	.84	277	.74	570	847	.81	.76	2072	.750	.88	266	.73	544	810	.81	.76	2072
7	.88	256	.67	402	658	.86	.70	2233	.700	.85	301	.75	664	965	.83	.79	2589	.700	.89	298	.75	647	945	.82	.78	2590

Optimal design parameters λ^*, d_2^* and HR_{go}^*, corresponding value of maximal expected utility u^*, expected estimate used for sample size calculation ε_2^*, expected number of events in phase III when going to phase III d_3^*, expected total number of events of program d^*, expected probability to go to phase III p_{go}^*, and expected probability of a successful program sP^* for the optimal design, for $c_2 = 0.75$, $c_3 = 1$, $c_{02} = 100$, $c_{03} = 150$ in \$ 10^5, $\xi_2 = \xi_3 = 0.7$, $1 - \beta = 0.9$, $\alpha = 0.025$ (one sided), benefit scenarios *bs* 1–7, weights for the prior distribution $w = 0.3, 0.6, 0.9$, for the unadjusted program set-up $S(\hat{\theta}_2^u, \hat{\theta}_2^u)$ and multiplicatively adjusted program set-ups $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^\lambda)$, respectively

d_2^* / d_3^*), leads to values of 0.55–0.64, 0.55–0.64, 0.58–0.67, 0.43–0.54 and 0.42–0.54 in program set-up $S(\hat{\theta}_2^u, \hat{\theta}_2^u)$, $S(\hat{\theta}_2^u, \hat{\theta}_2^{\alpha_{CI}})$, $S(\hat{\theta}_2^{\alpha_{CI}}, \hat{\theta}_2^{\alpha_{CI}})$, $S(\hat{\theta}_2^u, \hat{\theta}_2^\lambda)$ and $S(\hat{\theta}_2^\lambda, \hat{\theta}_2^\lambda)$, respectively. Furthermore, it can be observed that the treatment effect estimate of phase II used for sample size calculation in the optimal design is overestimated in the unadjusted setting ($\varepsilon_2^* < \exp(-\mathrm{E}[\hat{\theta}_2])$ as indicated by the black circles and yellow line in Figure 3). This overestimation is lower in the adjusted settings and can even result in an underestimation (compare multiplicative settings for $w = 0.9$).

The operating characteristics for the optimal designs (e.g., u^*, sP^*) compared between the two multiplicatively and the two additively adjusted program set-ups do not vary (much) for each benefit scenario *bs* and choice of weight for the prior distribution *w*, respectively. However, there are differences in the optimal choice of the threshold value for the decision rule HR_{go}^*: in the program set-ups with adjusted phase II treatment effect estimate used for decision making ($S(\hat{\theta}_2^\lambda, \hat{\theta}_2^\lambda)$ and $S(\hat{\theta}_2^{\alpha_{CI}}, \hat{\theta}_2^{\alpha_{CI}})$), HR_{go}^* is always larger (by 0.04 to 0.06 and by 0.01 to 0.07, respectively) than in the program set-ups with unadjusted treatment effect used for decision making ($S(\hat{\theta}_2^u, \hat{\theta}_2^\lambda)$ and $S(\hat{\theta}_2^u, \hat{\theta}_2^{\alpha_{CI}})$).

Discussion

To find optimal drug development designs, the costs of the program (fixed/variable costs for phase II/III), the assumed benefit, and the development risk (i.e., the

Table 3 Optimal design parameters for additively adjusted program set-ups

Additively adjusted																		
Program set-up $S(\hat{\theta}_2^{u}, \hat{\theta}_2^{a_{CI}})$										Program set-up $S(\hat{\theta}_2^{a_{CI}}, \hat{\theta}_2^{a_{CI}})$								
bs	$a_{CI}{}^*$	HR_{go}^*	d_2^*	ε_2^*	d_3^*	d^*	p_{go}^*	sP^*	u^*	$a_{CI}{}^*$	HR_{go}^*	d_2^*	ε_2^*	d_3^*	d^*	p_{go}^*	sP^*	u^*
$\boldsymbol{w = .3}$, i.e., $\mathbf{exp}(-\mathbf{E}[\hat{\boldsymbol{\theta}}_2])=\mathbf{.82}$																		
1	.450	.78	88	.66	140	228	.42	.24	78	.450	.80	84	.65	138	222	.42	.23	78
2	.400	.79	113	.68	188	301	.43	.27	194	.400	.83	116	.69	192	308	.43	.28	194
3	.425	.81	140	.69	220	360	.47	.31	302	.425	.84	133	.69	229	362	.47	.31	302
4	.375	.81	155	.71	261	416	.46	.32	442	.400	.85	161	.71	261	422	.48	.33	442
5	.350	.81	189	.72	278	467	.46	.34	593	.350	.86	182	.72	289	471	.47	.34	593
6	.400	.83	190	.72	310	500	.50	.36	573	.425	.86	186	.71	311	497	.52	.36	573
7	.375	.83	204	.72	336	540	.50	.37	729	.375	.87	203	.73	346	549	.51	.37	729
$\boldsymbol{w = .6}$, i.e., $\mathbf{exp}(-\mathbf{E}[\hat{\boldsymbol{\theta}}_2])=\mathbf{.76}$																		
1	.450	.81	140	.66	226	366	.60	.43	372	.425	.83	130	.67	226	356	.58	.43	372
2	.350	.80	168	.69	278	446	.58	.46	614	.350	.85	172	.69	282	454	.58	.46	614
3	.425	.83	175	.68	304	479	.64	.49	772	.425	.85	193	.68	296	489	.64	.49	772
4	.350	.82	224	.70	341	565	.62	.51	1045	.325	.87	228	.71	366	594	.62	.52	1045
5	.250	.81	273	.72	406	679	.60	.53	1338	.275	.88	249	.72	411	660	.62	.53	1338
6	.375	.84	252	.70	395	647	.67	.54	1221	.375	.87	252	.70	376	628	.66	.54	1222
7	.300	.83	287	.72	437	724	.65	.56	1515	.300	.88	273	.72	419	692	.64	.55	1515
$\boldsymbol{w = .9}$, i.e., $\mathbf{exp}(-\mathbf{E}[\hat{\boldsymbol{\theta}}_2])=\mathbf{.71}$																		
1	.425	.82	168	.66	296	464	.75	.61	695	.450	.84	168	.66	284	452	.76	.61	695
2	.350	.82	203	.69	371	574	.76	.65	1068	.350	.86	210	.68	355	565	.76	.65	1068
3	.400	.84	224	.68	381	605	.80	.68	1272	.400	.87	228	.68	385	613	.80	.68	1272
4	.325	.83	252	.70	433	685	.79	.69	1681	.300	.88	280	.70	446	726	.79	.70	1682
5	.250	.82	308	.71	482	790	.78	.71	2122	.225	.89	315	.72	510	825	.78	.71	2122
6	.350	.84	287	.69	434	721	.81	.71	1898	.350	.88	294	.69	438	732	.82	.71	1898
7	.275	.83	315	.71	489	804	.80	.72	2333	.275	.89	326	.71	500	826	.81	.73	2334

Optimal design parameters $a_{CI}{}^*$, d_2^* and HR_{go}^*, corresponding value of expected utility u^*, expected estimate used for sample size calculation ε_2^*, expected number of events in phase III when going to phase III d_3^*, expected total number of events of program d^*, expected probability to go to phase III p_{go}^*, and expected probability of a successful program sP^* for the optimal design, for $c_2 = 0.75, c_3 = 1$, $c_{02} = 100, c_{03} = 150$ in \$ 10^5, $\xi_2 = \xi_3 = 0.7$, $1 - \beta = 0.9$, α = 0.025 (one sided), benefit scenarios *bs* 1–7, weights for the prior distribution $w = 0.3, 0.6, 0.9$ for the additively adjusted program set-ups $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{a_{CI}})$

expected probability of a successful program) are taken into account. By maximizing the expected utility with respect to the design parameters (adjustment parameter, number of events for phase II and threshold value for the go/no-go decision rule), optimal phase II/III drug development program designs can be found. Therefore, it enables quantitative reasoning for the design (i.e., the optimal "amount of adjustment", sample size and decision rule) for specific drug development programs at hand.

We investigated two adjustment methods (additive and multiplicative adjustment), several benefit scenarios (e.g., low, medium, large overall benefit), different distributions for the true treatment effect (with the same and different distributions in phase II and III), scenarios with a real life budget constraint, scenarios with a predefined clinically relevant effect, and scenarios where phase II could be skipped, hence presented a method for the implementation of a variety of possible oncology drug development program scenarios, and an opportunity for assessing associated changes of the optimal design parameters. Of course, the implementation of alternative (e.g., proportional relationship between benefit and effect size) or more complex planning situations and broader application to other research areas are possible by choosing relevant (e.g., cost and benefit) parameters appropriately [37–39]. As the framework has been shown to be very flexible, frequent scenarios in oncology drug development are adequately mapped with our approach. However, certain situations may be simplified. For example, in our framework the development program consists entirely of just one phase II trial and one phase III trial, which is, however, not unusual in oncology. For situations that two or more phase III trials are performed, the framework of optimal planning of

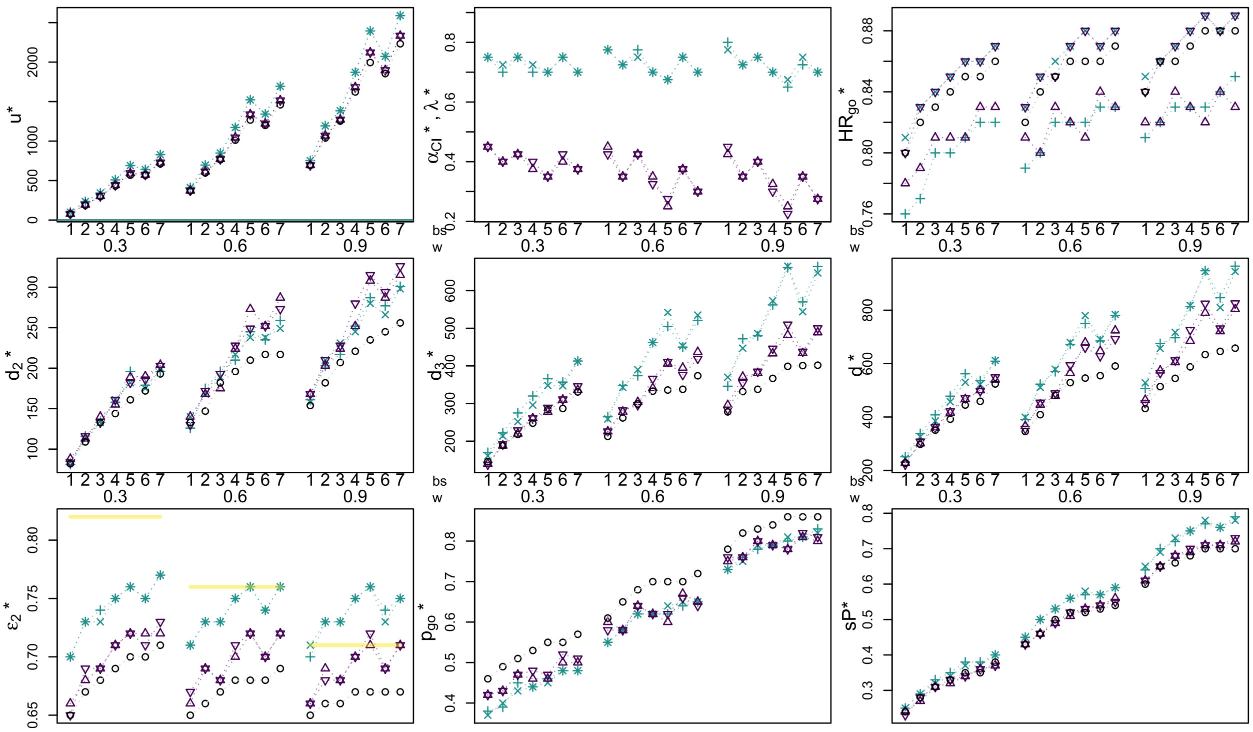

Fig. 3 Optimization results. Maximal expected utility u^*, corresponding optimal design parameters $\delta^* = (d_2^*, HR_{go}^*)$, $\delta^* = (d_2^*, HR_{go}^*, \lambda^*)$ or $\delta^* = (d_2^*, HR_{go}^*, \alpha_{CI}^*)$, expected probability to go to phase III p_{go}^*, expected probability of a successful program sP^*, expected estimate used for sample size calculation ε_2^*, expected number of events in phase III when going to phase III d_3^* and expected total number of events of program d^* in the optimal design, for $c_2 = 0.75$, $c_3 = 1$, $c_{02} = 100$, $c_{03} = 150$ in \$ 10^5, $\xi_2 = \xi_3 = 0.7$, $1 - \beta = 0.9$, $\alpha = 0.025$ (one sided), for program set-ups $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{s_2})$, s_1, $s_2 = \lambda$, α_{CI} or u (that is $S(\hat{\theta}_2^{u}, \hat{\theta}_2^{u})$: black circle; $S(\hat{\theta}_2^{u}, \hat{\theta}_2^{\lambda})$, $S(\hat{\theta}_2^{\lambda}, \hat{\theta}_2^{\lambda})$: green cross; $S(\hat{\theta}_2^{u}, \hat{\theta}_2^{\alpha_{CI}})$, $S(\hat{\theta}_2^{\alpha_{CI}}, \hat{\theta}_2^{\alpha_{CI}})$: violet triangle), benefit scenarios bs 1–7, and weights for the prior distribution $w = 0.3, 0.6, 0.9$, where the yellow line indicates $\exp(-\mathrm{E}[\hat{\theta}_2])$. Note that the symbols used to show the program characteristics of both multiplicatively and additively adjusted program set-ups, i.e., green crosses and violet triangles, appear as stars when plotted on top of each other

development programs was presented in a recent article by Preussler et al. [40]. Furthermore, we assumed the phase II trial to be two-armed. In the field of oncology dose investigations are often performed before and not as a part of phase II. However, in other indications dose-finding is performed in phase II. Methods for optimizing phase II/III programs with multi-armed phase II/III studies are presented in Preussler et al. [41]. Futility investigations in the phase III trial and/or considering a "seamless design" for the final analysis may be a worthwhile option, and it will be a topic of future research to investigate their impact on the optimal design. We assumed that the endpoint used in phase II and phase III is the same. We are currently exploring the situation that a surrogate (like progression-free or disease-free survival) is captured in phase II and overall survival is the primary endpoint in phase III. Another important point is that time-effects are not considered in this article. The program is unaccounted for the duration of development which is amongst others discussed in Preussler et al. [41]. That work presents in detail how to incorporate the impact of trial duration into the framework (compare Supplementary Material A2 [41]). However, when trying to incorporate "time" into the utility function, many aspects have to be considered. For example, one could take into account the "life cycle" of a drug as proposed by Patel & Ankolekar [42] who describe a typical life cycle by an early growth phase followed by a plateau, after which the sales decline as the patent expires. Furthermore, if there are several competitors investigating a similar drug then the company, who is the first to bring the drug to the market, usually gets the higher market share, i.e., higher gain. However, including these aspects requires competitor information and assumptions about their unknown future observed treatment effects. Any such assumptions are usually associated with very high uncertainty. Instead of trying to include too many (unknown) aspects into the utility function a rather simplified approach, as presented here, is advisable. If after observing phase II data further information about the potential of the drug, dose, target population or (time-dependent) benefits are available the probability of success (compare [43]) and the utility function could be updated to support go/no-go decisions as well as the design of the phase III trial.

In general, our results show that the adjusted program set-ups are superior to the unadjusted program set-up with respect to the maximal expected utility. This is associated with higher investments in terms of number of events and lower expected probabilities to go to phase III in the adjusted program set-ups compared to the unadjusted approach. Thus, in the adjusted program set-ups it is less often decided to go to phase III, but in case of a go decision, the investment in terms of sample size is higher. These aspects are particularly true for the multiplicatively adjusted program set-ups, which have also higher expected probabilities of a successful program compared to the additively adjusted and unadjusted program set-ups. Simply said, the money is spent more wisely when adjustment methods are used.

Values for the adjustment parameters that do not lead to an adjustment (i.e., $\alpha_{CI} = 0.5$ and $\lambda = 1$ in the additively and multiplicatively adjusted program set-ups, respectively) were included but never selected in the optimization. Thus, the results suggest that adjustment should always be considered, which is in line with Chuang-Stein and Kirby [14]. Furthermore, we see that in the unadjusted case there is an overestimation of the treatment effect after phase II, which is mitigated by the adjustments. In the multiplicative setting it is even shown that an overcorrection and thus an even larger investment in terms of sample size can be worthwhile with respect to the expected utility. Note that the focus is on maximal expected utility and the expected estimate of phase II is only a supporting variable, i.e., obtaining a "perfectly" unbiased estimator is not the goal in this application. With regard to the optimal number of events in phase II compared to phase III ($d_2^* \ / \ d_3^*$), it can be seen that with the framework in the unadjusted and additive case one ends up in the "desirable" (according to De Martini [4, 25]) range of 2/3 and also in the multiplicative case with lower $d_2^* \ / \ d_3^*$, one still exceeds the often used 1/4. However, it should be noted that the total optimal sample size is highest for the multiplicative case.

Both multiplicatively adjusted (i.e., $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{\lambda})$) and additively adjusted (i.e., $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{\alpha_{CI}})$) program set-ups do not differ in their maximal expected utility, whereas the program set-ups with adjusted estimate used for decision making (i.e., $S(\hat{\theta}_2^{\lambda}, \hat{\theta}_2^{\lambda})$ and $S(\hat{\theta}_2^{\alpha_{CI}}, \hat{\theta}_2^{\alpha_{CI}})$) have larger optimal threshold values for the decision rule than program set-ups where only the estimate used for calculating the expected number of events for phase III is adjusted (i.e., $S(\hat{\theta}_2^{u}, \hat{\theta}_2^{\lambda})$ and $S(\hat{\theta}_2^{u}, \hat{\theta}_2^{\alpha_{CI}})$). Considering only these two aspects, adjustment of the treatment effect estimate used for the decision rule may be omitted when also optimizing the threshold value for the decision rule: this only leads to larger values for HR_{go}^* (i.e., more liberal decision rules) which compensate the adjusted (more conservative) treatment effect estimates. For the same reason, program set-ups $S(\hat{\theta}_2^{\lambda}, \hat{\theta}_2^{u})$ and $S(\hat{\theta}_2^{\alpha_{CI}}, \hat{\theta}_2^{u})$ (i.e., multiplicative or additive adjustment used for the decision rule and no adjustment applied for the calculation of the number of events for phase III) are not considered. Furthermore, as adjustment of the treatment effect estimate used for the decision rule may be omitted when also optimizing over the threshold value for the decision rule, we did not consider program set-ups where different adjustment parameters used for the decision rule and the calculation of the expected number of events are optimized (in our notation $S(\hat{\theta}_2^{\lambda_1}, \hat{\theta}_2^{\lambda_2})$ and $S(\hat{\theta}_2^{\alpha_{CI1}}, \hat{\theta}_2^{\alpha_{CI2}})$).

Conclusions

Based on our results, we highly recommend using (multiplicatively) adjusted phase II treatment effect estimates for calculation of the phase III number of events in a phase II/III drug development program with go/no-go decision rule (compare Chuang-Stein & Kirby [14], Kirby et al. [15] and De Martini [4, 25]). However, as our results also show that the optimal design parameters of each method depend on the cost and benefit parameters as well as on the applied prior distribution, no general rule exists. In contrast, the design parameters should be determined by applying our proposed optimization procedure for specific values of the parameters in the respective drug development program. Therefore, we provide an user friendly R Shiny App (bias) and an R package (drugdevelopR including the R function optimal_bias) open-source (both assessable via [1]).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12874-020-01093-w.

Additional file 1. In the Additional file 1, an overview of formulas in program set-ups $S(\hat{\theta}_2^{s_1}, \hat{\theta}_2^{s_2})$,$s_1$, $s_2 = \lambda$, α_{CI}, u (A0) and investigation of an alternative definition of program success is given (A1). Furthermore, more details and results of the application example when modelling different population structures in phase II and III (A2), when using a predefined minimal clinically relevant effect for phase III planning (A3), when using a budget constraint (A4), when skipping phase II (A5) and when using a linear function for modelling the gain (A6) are presented. The file Code.R includes the main function calls for generating the datasets and tables, using the `R` package `drugdevelopR`.

Abbreviations

α_{CI},λ: Adjustment parameter for additive and multiplicative adjustment method, respectively; *bs*: Benefit scenario; *CI*: Confidence interval; d_2,d_3,d: Total number of events for phase II, III and the program, respectively; *HR*: True assumed hazard ratio; κ: Threshold value for the go/no-go decision rule, $\kappa = -\log(HR_{go})$; s_1,s_2: Estimate used for go/no-go decision and calculation of number of events, respectively; θ: True assumed treatment effect, $\theta = -\log(HR)$

Acknowledgements
We thank the reviewer for their valuable comments, which improved the manuscript remarkably.

Authors' contributions
SE, MKir, HG, MKie developed the proposed method. SE wrote the manuscript and associated software code. All authors read and approved the final manuscript.

Authors' information
The first author of this manuscript (SE) changed her name from Stella Preussler to Stella Erdmann. Therefore, the (first) author of [1, 33, 40], [41] and this manuscript is the same.

Funding
We would like to thank the Deutsche Forschungsgemeinschaft (DFG) for supporting this research by the research grant KI 708/2–1 (financing a statistical position of the Institute of Medical Biometry of the University Hospital Heidelberg). Furthermore, we would like acknowledge the financial support by the DFG within the funding program "Open Access Publishing", by the Baden-Württemberg Ministry of Science, Research and the Arts and by Ruprecht-Karls-Universität Heidelberg (payment of the publishing costs). The role of the funding bodies was of financial manner only, therefore they took no direct part in the analysis or in writing the manuscript. Open access funding provided by Projekt DEAL.

Availability of data and materials
The datasets used can be generated with the help of the R package drugdevelopR and the code containing the respective function calls is provided in the additional files (see file Code.R).

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests. HG has no conflict of interest with the subject matter of this manuscript while being an employee of Merck Healthcare KGaA.

Author details
[1]Institute of Medical Biometry and Informatics, University of Heidelberg, Im Neuenheimer Feld 130.3, D-69120 Heidelberg, Germany. [2]Merck Healthcare KGaA, Frankfurter Str. 250, D-64293 Darmstadt, Germany.

Received: 21 February 2020 Accepted: 3 August 2020
Published online: 09 October 2020

References

1. Erdmann, S. drugdevelopR: bias. https://web.imbi.uni-heidelberg.de/bias/. Accessed 02 Jul 2020.
2. DiMasi JA, Hansen RW, Grabowski HC, Lasagna L. Research and development costs for new drugs by therapeutic category. Pharmaco Economics. 1995;7:152–69.
3. DiMasi JA, Feldman L, Seckler A, Wilson A. Trends in risks associated with new drug development: success rates for investigational drugs. Clinical Pharmacology & Therapeutics. 2010;87:272–7.
4. De Martini D. Empowering phase II clinical trials to reduce phase III failures. Pharm Stat. 2020;19:178–86.
5. Antonijevic Z. Optimization of Pharmaceutical R&D Programs and portfolios: design and investment strategy. Heidelberg: Springer; 2015.
6. Hughes MD, Pocock SJ. Stopping rules and estimation problems in clinical trials. Stat Med. 1988;7:1231–42.
7. Fan X, DeMets DL, Lan KG. Conditional bias of point estimates following a group sequential test. J Biopharm Stat. 2004;14:505–30.
8. Zhang JJ, Blumenthal G, He K, Tang S, Cortazar P, Sridhara R. Overestimation of the effect size in group sequential trials. Clin Cancer Res. 2012;18: 18,4872–6.
9. Ellenberg SS, DeMets DL, Fleming TR. Bias and trials stopped early for benefit. Jama. 2010;304:156–9.
10. Nardini C. Monitoring in clinical trials: benefit or bias? Theoretical Medicine and Bioethics. 2013;34:259–74.
11. US Food and Drug Administration. 22 case studies where phase 2 and phase 3 trials had divergent results. 2017. Available at http://go.nature.com/2mayug4. Accessed 02 Jul 2020.
12. Gan HK, You B, Pond GR, Chen EX. Assumptions of expected benefits in randomized phase III trials evaluating systemic treatments for cancer. J Natl Cancer Inst. 2012;104:590–8.
13. Arrowsmith J. Trial watch: phase II failures: 2008–2010. Nat Rev Drug Discov. 2011;10:328–9.
14. Chuang-Stein C, Kirby S. The shrinking or disappearing observed treatment effect. Pharm Stat. 2014;13:277–80.
15. Kirby S, Burke J, Chuang-Stein C, Sin C. Discounting phase 2 results when planning phase 3 clinical trials. Pharm Stat. 2012;11:373–85.
16. O'Hagan A, Stevens JW, Montmartin J. Bayesian cost-effectiveness analysis from clinical trial data. Stat Med. 2001;20:733–753.2005.
17. O'Hagan A, Stevens JW, Campbell MJ. Assurance in clinical trial design. Pharm Stat. 2005;4:187–201.
18. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: conditional or predictive power? Control Clin Trials. 1986;7:8–17.
19. Spiegelhalter DJ, Freedman LS. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. Statistics Med. 1986;5:1–13.
20. Chuang-Stein C. Sample size and the probability of a successful trial. Pharm Stat J Appl Stat Pharm Ind. 2006;5:305–9.
21. Chuang-Stein C, Yang R. A revisit of sample size decisions in confirmatory trials. Statistics in Biopharmaceutical Research. 2010;2:239–48.
22. Gasparini M, Di Scala L, Bretz F, Racine-Poon A. Some uses of predictive probability of success in clinical drug development. Epidemiology, biostatistics and. Public Health. 2013;10:1.
23. Saint-Hilary G, Barboux V, Pannaux M, Gasparini M, Robert V, Mastrantonio G. Predictive probability of success using surrogate endpoints. Stat Med. 2019;38:1753–74. https://doi.org/10.1002/sim.8060.
24. Wang SJ, Hung HM, O'Neill RT. Adapting the sample size planning of a phase III trial based on phase II data. Pharm Stat. 2006;5:85–97.
25. De Martini D. Adapting by calibration the sample size of a phase III trial on the basis of phase II data. Pharm Stat. 2011;10:89–95.
26. Götte H, Schüler A, Kirchner M, Kieser M. Sample size planning for phase II trials based on success probabilities for phase III. Pharm Stat. 2015;14:515–24.
27. Kirchner M, Kieser M, Götte H, Schüler A. Utility-based optimization of phase II/III programs. Stat Med. 2016;35:305–16.
28. Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika. 1981;68:316–9.
29. IQWiG. Allgemeine Methoden. Version 5.0, 10.07.2016, Technical Report.
30. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2019. Available at https://www.R-project.org/. Accessed 02 Jul 2020.
31. Steensma DP, Kantarjian HM. Impact of cancer research bureaucracy on innovation, costs, and patient care. J Clin Oncol. 2014;32:376–8.
32. Ding M, Rosner GL, Müller P. Bayesian optimal design for phase II screening trials. Biometrics. 2008;3:886–94.
33. Erdmann, S. drugdevelopR: prior. https://web.imbi.uni-heidelberg.de/prior/. Accessed 02 Jul 2020.
34. Dallow N, Best N, Montague TH. Better decision making in drug development through adoption of formal prior elicitation. Pharm Stat. 2018; 17:301–16.
35. O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, et al. Uncertain judgements: eliciting experts' probabilities. Chichester: Wiley; 2006.
36. Devilee JLA, Knol AB. Software to support expert elicitation: an exploratory study of existing software packages; 2011.
37. DiMasi JA, Grabowski HG, Vernon J. R&D costs and returns by therapeutic category. Drug Information J. 2004;38:211–23.
38. Adams CP, Brantner W. Spending on new drug development. Health Econ. 2010;19:130–41.
39. Morgan S, Grootendorst P, Lexchin J, Cunningham C, Greyson D. The cost of drug development: a systematic review. Health Policy. 2011;100:4–17.
40. Preussler S, Kieser M, Kirchner M. Optimal sample size allocation and go/no-go decision rules for phase II/III programs where several phase III trials are performed. Biom J. 2019;61(2):357–78.

41. Preussler S, Kirchner M, Götte H, Kieser M. Optimal designs for multi-arm phase II/III drug development programs. Statistics in Biopharmaceutical Res. 2019. https://doi.org/10.1080/19466315.2019.1702092.
42. Patel NR, Ankolekar S. A Bayesian approach for incorporating economic factors in sample size design for clinical trials of individual drugs and portfolios of drugs. Stat Med. 2007;26:4976–88.
43. Götte H, Kirchner M, Sailer MO, Kieser M. Simulation-based adjustment after exploratory biomarker subgroup selection in phase II. Stat Med. 2017;36:2378–90.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.