Biometrical Journal **xx** (xxx) xx, zzz–zzz / DOI: 10.1002/bimj.200100000

Using a Dose-Finding Benchmark to Quantify the Loss Incurred by Dichotomisation in Phase II Dose-Ranging Studies

Pavel Mozgunov[*,1], **Thomas Jaki** [1], and **Xavier Paoletti** [2 3]

[1] Medical and Pharmaceutical Statistics Research Unit, Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, UK

[2] Service de Biostatistique et dEpidemiologie & CESP OncoStat, INSERM, Institut Gustave Roussy, UVSQ, Villejuif, France

[3] Institute Curie, Paris, France

Received zzz, revised zzz, accepted zzz

While there is recognition that more informative clinical endpoints can support better decision-making in clinical trials, it remains a common practice to categorise endpoints originally measured on a continuous scale. The primary motivation for this categorisation (and most commonly dichotomisation) is the simplicity of the analysis. There is, however, a long argument that this simplicity can come at a high cost. Specifically, larger sample sizes are needed to achieve the same level of accuracy when using a dichotomised outcome instead of the original continuous endpoint. The degree of "loss of information" has been studied in the contexts of parallel-group designs and two-stage Phase II trials. Limited attention, however, has been given to the quantification of the associated losses in dose ranging trials. In this work, we propose an approach to estimate the associated losses in Phase II dose ranging trials that is free of the actual dose ranging design used and depends on the clinical setting only. The approach uses the notion of a non-parametric optimal benchmark for dose finding trials, an evaluation tool that facilitates the assessment of a dose finding design by providing an upper bound on its performance under a given scenario in terms of the probability of the target dose selection. After demonstrating how the benchmark can be applied to Phase II dose ranging trials, we use it to quantify the dichotomisation losses. Using parameters from real clinical trials in various therapeutic areas, it is found that the ratio of sample sizes needed to obtain the same precision using continuous and binary (dichotomized) endpoints varies between 70%-75% under the majority of scenarios but can drop to 50% in some cases.

Key words: Continuous Endpoint; Dichotomization; Dose Ranging Trials; Phase II; Non-parametric optimal benchmark.

1 Introduction

There is growing interest in endpoints that are more informative than binary endpoints. Indeed, there are many Phase II clinical trials in various therapeutic areas in which the clinical endpoint is measured on a continuous scale (see e.g. Verkindre et al., 2010; Karlson et al., 2016; Spertini et al., 2013). Nevertheless, it is a common practice to categorize them for the primary efficacy analysis. It has, however, been argued by many researchers that such simplicity can come at a high cost (Altman and Royston, 2006; Fedorov et al., 2009) as categorization almost always results in a loss of information (Kullback, 1997, Corollary 3.1). These losses inevitably result in larger sample sizes required to achieve a particular level of accuracy compared to when the original continuous measure is used. The degree of this "loss of information" in the context of clinical trials has been studied in different settings, e.g. parallel-group and two-stage Phase II designs (e.g. Senn, 2005; Wason and Seaman, 2013; Barnwell-Ménard et al., 2015; Lei et al., 2017). Senn (2005) has shown that the ratio of the sample sizes to achieve a particular level of precision is at least

*Corresponding author: e-mail: p.mozgunov@lancaster.ac.uk

64% in a parallel group clinical trials. In the setting of the two-stage adaptive trial, Wason et al. (2011) have shown that this ratio reached around 63%, and is around 50% in a case study. At the same time, in many therapeutic areas, the choice of dose of a drug to be tested in later phase trials is usually conducted in Phase II dose ranging clinical trial. Consequently, it is expected that dichotomising in dose finding trials will result in choosing the optimal dose with lower probability for a given sample size. However, limited attention has been dedicated to the quantification of the associated losses in the setting of dose finding trials.

To provide the answer to this question regardless of the design used and depending on the clinical context only, we propose to use an evaluation tool known as *non-parametric optimal benchmark* (or simply benchmark) originally proposed by O'Quigley et al. (2002). The benchmark was designed to provide a scenario-specific assessment of the accuracy of dose finding designs in terms of the proportions of correction selections (PCS) of the dose with the relevant target characteristics. For a design, the PCS is conventionally estimated by simulations in the scenarios chosen by researchers themselves. This can add subjectivity into the assessment as it might be more challenging to identify the target dose in some scenarios than in others. To overcome this, the benchmark provides the highest PCS a dose finding design can attain in the specific scenario. Comparing these upper bounds with the performance of a design can provide a more meaningful evaluation of a dose finding design operating characteristics. The benchmark can be used to evaluate any given design regardless of its nature, for example, adaptive or non-adaptive, as it accounts for the specifics of the scenario rather than the specifics of the design. We will show that the benchmark is not limited to design evaluations but can be also used to assess the consequences of *the trial assumptions*, and, specifically, the choice of the primary endpoint.

The original benchmark was proposed for studies with a binary endpoint only, and, therefore, its application was limited to Phase I oncology dose-escalation trials, in which binary endpoints prevail. A generalized benchmark for dose finding clinical trials with continuous responses was recently proposed by Mozgunov et al. (2018). While its application was demonstrated, again, in Phase I dose-escalation trials only, the generalization also opens the door for the benchmark application in the dose-ranging trials, in which endpoints measured on the continuous scale are more common. As the performance of Phase II dose ranging studies is also commonly studied by simulations, there is merit in the benchmark supporting a more informative evaluation of a dose ranging design. While various measures of a dose-finding design's performance can be of interest, in this work, we focus on the measure for which that benchmark is conventionally used, the PCS.

In this work, we firstly demonstrate how the generalized benchmark can be used to get a more meaningful evaluation of Phase II dose ranging designs given the specifics of these trials. Secondly, we use this result to quantify the losses associated with dichotomization of the continuous primary endpoint in dose ranging trials using the generalized benchmark. Effectively, the associated dichotomisation losses are found as the differences in the sample sizes required to attain particular PCS by the original benchmark for the binary (dichotomised) outcome and the generalised benchmark for the continuous outcomes. This allows to make the assessment free of the dose ranging design used. We will use several examples from various therapeutic areas, namely, chronic obstructive pulmonary disease, oncology, and cardiovascular disease, and study how various clinical trial parameters (dichotomisation threshold, variance of the responses, and location of the target dose) affect the dichotomisation losses. We will also study how trial parameters (variance of a response, dichotomisation threshold, and target probability) affect the associated losses.

In Section 2, we recall the recently generalized benchmark (Mozgunov et al., 2018) in a setting of Phase II clinical trial. In Section 3, an example of the benchmark application to a Phase II clinical trial in chronic obstructive pulmonary disease is provided. In Section 4, the question of dichotomization in the setting of a cancer trial is considered. An evaluation when distributional assumption are violated is given in Section 5, and final remarks are provided in Section 6.

2 A Benchmark Dose Finding Trials with Continuous Endpoint

Consider a Phase II clinical trial with m doses, a total number of n patients and a continuous outcome, Y_{ij}, at dose d_j, $j = 1, \ldots, m$ for patient i, $i = 1, \ldots, n$ having cumulative distribution function (CDF) $F_j(y)$. The goal of the trial is to find the target dose (among the predefined set), which optimizes some decision criterion $T(\cdot)$. The target dose (TD), for example, can be the dose having the response closest to 50% of the maximum response (denoted by ED_{50}), or the dose having the probability of the response to be higher than a particular threshold closest to some target value.

Importantly, the benchmark is used at the planning stage of the trial only, when simulation studies are usually conducted. At this point, the CDFs $F_j(y)$ are assumed to be known to an investigator as they define the simulation scenario specified by an investigator and used to study the behavior of a design. While these are unknown in the actual trials, these are usually comprised of the set of the clinically feasible scenarios agreed with clinicians before the simulation study. Therefore, CDFs $F_j(y)$ are assumed to be known in the simulation, and consequently, by the benchmark. We describe the generalized benchmark below.

For a given sample size and CDFs $F_j(\cdot)$, the benchmark provides the probability that a particular dose is selected (according to the decision criterion T) by a procedure that assumes that outcomes of the patients are observed at each dose level and employes no functional relationship between dose and response. As any further (restrictive) model assumption are expected to decrease this probability, the probability produced by this selection procedure is a benchmark, the upper bound of the operating characteristics that could be attained. Formally, for the minimisation of the decision criterion, for each dose $d_{j'}$ the benchmark computes $\mathbb{P}\left[T\left(Y_{1j'}, Y_{2j'}, \ldots, Y_{nj'}\right) = \min_j T\left(Y_{1j}, Y_{2j}, \ldots, Y_{nj}\right)\right]$ where Y_{ij} has CDF $F_j(y)$. To compute these probabilities, a simulation-based approach was proposed by O'Quigley et al. (2002) and extended by Mozgunov et al. (2018).

The simulation-based evaluation of the benchmark employs the concept of the *complete information* that assumes that outcomes for each patient are known for each dose level. The patient's response to all doses can be generated in simulations based on $F_j(y)$ and using a patient's profile $u_i \sim \mathcal{U}(0, 1)$. The vector patient's of responses at each doses are known as the *complete information*. Mozgunov et al. (2018) have shown that the outcome for a patient with profile u_i at dose level d_j with corresponding CDF F_j is given by $y_{ij} = F_j^{-1}(u_i)$. Transforming profile u_i for each dose results in a vector of the complete information $(y_{i1}, \ldots, y_{im})$ for patient i. The procedure is repeated for all patients $i = 1, \ldots, n$, which results in the vector of responses for each dose level $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{nj})$, $j = 1, \ldots, m$. This means that for each dose level, there are n observations generated for n patients. These vectors are consequently used to compute the criterion $T(\cdot)$ for each dose. The dose for which the criterion is optimised is selected as the target dose. The procedure is repeated for S simulated trials. The number of repetitions to be used by the simulation-based implementation of the benchmark relates to the precision of the PCS estimate.

For a given simulation scenario with CDF $F_1, \ldots, F_j$, and decision criterion $T(\cdot)$, the benchmark for continuous outcome can be computed as follows:

1. Generate a sequence of patients' profiles $\{u_i\}_{i=1}^n$ from the Uniform distribution $\mathcal{U}(0, 1)$.
2. Transform u_i for dose level d_j using $y_{ij} = F_j^{-1}(u_i)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$ and store $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{nj})$.
3. Compute $T(\boldsymbol{y}_j)$ for all $j = 1, \ldots, m$, find dose J for which $T(\boldsymbol{y}_J)$ is optimised. Store the recommendation in the s^{th} simulation as $K_s = J$.
4. Repeat for $s = 1, \ldots, S$ simulated trials.
5. Use $\bar{K}^{(j)} = \sum_{s=1}^{S} \mathbb{I}\left(K_s = j\right)/S$ as the selection proportion of dose d_j for $j = 1, \ldots, m$.

An application of the benchmark as above for the assessment of dose ranging designs in the context of an actual clinical trial is given below.

3 An Evaluation of Phase II Dose Ranging Designs

Mielke and Dragalin (2017) studied several dose ranging designs in the context of Phase II clinical trial in chronic obstructive pulmonary disease (COPD) (Verkindre et al., 2010) with a total sample size of $n = 300$. The trial studied the lung function in terms of forced expiratory volume (FEV) within 1s, measured in litres (denoted by FEV1) of COPD patients to eight doses of a compound (and placebo)

$$d_j = 0, 12.5, 25, 37.5, 50, 62.5, 75, 87.5, 100 \text{ mg}.$$

The primary efficacy endpoint is the difference in FEV1 between a dose of the experimental treatment and placebo. The difference in FEV1 for patient i given dose d_j is denoted by y_{ij} and is assumed to have normal distribution $\mathcal{N}(\mu_j, 0.34^2)$, and the maximal difference (defined by clinicians) is assumed to be 0.15.

Mielke and Dragalin (2017) evaluated six non-adaptive and adaptive designs in this setting. We focus on one non-adaptive and one adaptive designs keeping the original notation:

- D_0: The design with fixed equal allocation to three active doses (25,50, and 100mg) and placebo
- D_5: The design assigns first half of the patients according to the compound D-optimal allocation which is computed using five candidate standardized models
  - M1: $d_j/(5 + d_j)$,
  - M2: $d_j/(15 + d_j)$,
  - M3: $d_j^3/(50 + d_j^3)$,
  - M4: $d_j - d_j^2/160$,
  - M5: $\exp(d_j/20) - 1$.

  After an interim analysis, patients were allocated according to the most efficient design (using the D-efficiency criterion) out of a set of the designs predefined by Mielke and Dragalin (2017).

One of the properties studied by Mielke and Dragalin (2017) was the ability of designs to identify the dose (among a predefined set) corresponding to the mean difference in FEV1 closest to 50% of the maximum difference (ED_{50}) and 90% of the maximum difference (ED_{90}). Formally, this corresponds to the following decision criterion

$$T(\boldsymbol{y}_j, \gamma_{ED_{XX}}) = \left| \frac{\sum_{i=1}^{n} y_{ij}}{n} - \gamma_{ED_{XX}} \right| \tag{1}$$

where $\gamma_{ED_{XX}}$ is the level of difference in FEV1 corresponding to $XX\%$ of the maximum difference.

Two scenarios of the relationship for μ_j as a function of d_j, Emax and Sigmoid Emax, were considered for simulations

$$\mu_j^{(1)} = 0.15\frac{d_j}{d_j + 10}, \quad \mu_j^{(2)} = 0.15\frac{d_j^4}{d_j^4 + 35^4},$$

respectively, where 0.15 corresponds to the assumed maximum difference. Given the predefined set of doses these models translate into the following vectors of means

$$\mu_j^{(1)} = (0.00, 0.08, 0.11, 0.12, 0.12, 0.13, 0.13, 0.13, 0.14)$$

and

$$\mu_j^{(2)} = (0.00, 0.00, 0.03, 0.09, 0.12, 0.14, 0.14, 0.15, 0.15),$$

respectively. We used the benchmark to evaluate the designs D_0 and D_5 under these two scenarios.

The benchmark does not employ any particular model but can be used to select the doses (among predefined set) having differences in FEV1 closest to ED_{50} and ED_{90}. Given the maximum difference of 0.15, 50% of it is equal to 0.075 and 90% of it is equal to 0.135. Each of these values was used as the target values $\gamma_{ED_{50}}$ and $\gamma_{ED_{90}}$ in the criterion (1) for the benchmark to select the dose closest to the ED_{50} and the ED_{90}, respectively. These target values are scenario-related as they were chosen given the specified maximum difference in FEV1. These target values, however, are known in the benchmark setting, as it uses the simulation scenarios themselves. Consequently, the dose corresponding to the minimum of the criterion (1) using $\gamma_{ED_{50}} = 0.075$ was selected as the dose closest to the ED_{50}, and the dose corresponding to the minimum of (2.1) using $\gamma_{ED_{90}} = 0.135$ was selected as the dose closest to the ED_{90}.

The benchmark does not assume any subset of doses from the prespecified range as, for example, design D_0 does. Therefore, the selection of doses assumed under design D_0 is a feature of this design, and the benchmark can assess how well the equal allocation to the selected doses performs in terms of the PCS.

When reporting the results, Mielke and Dragalin (2017) considered the selection of the doses with relative effects between 25% and 75% for the estimation of the dose closest to ED_{50}. Similarly, the selection of doses with relative effects between 85% and 95% were considered for the estimation of the dose closest to the ED_{90}. For consistency, we provide the same characteristics of the benchmark. Table 1 shows the operating characteristics of the design against the respective benchmark. The results for the designs are extracted from Table 14.4 of the original work and the benchmark is evaluated using $S = 40000$ trial replications.

Table 1 The proportion of the ED_{50} and ED_{90} selections under two scenarios by the designs D_0 and D_5 and by the corresponding benchmark

	Scenario 1 (E_{max})		Scenario 2 (Sigmoid E_{max})	
	ED_{50}	ED_{90}	ED_{50}	ED_{90}
D_0	0.35	0.31	0.23	0.16
D_5	0.52	0.36	0.25	0.16
Benchmark	0.92	0.65	0.73	0.22

Comparing the proportion of correct ED_{50} and ED_{90} selections for designs D_0 and D_5, the adaptive design D_5 was able to find the target doses with a higher probability under Scenario 1, but shows comparable performance in Scenario 2. At the same time, comparing the proportion of selections between scenarios, it might seem that it was more than twice as challenging for D_5 to find ED_{50} and ED_{90} in Scenario 2 than in Scenario 1. Considering the ratio of correct selections with respect to the benchmark, the statement that ED_{50} was more challenging to estimate in Scenario 2 still holds, but the difference now was lower: the ratio was 0.56 $(0.52/0.92)$ for Scenario 1 against 0.34 $(0.25/0.73)$ for Scenario 2. The design D_5 might face problems identifying the target doses under Sigmoid Emax dose-effect shape scenarios. Similarly, while the proportions suggested that it was more than twice as hard for D_5 to find ED_{90} in Scenario 2 as in Scenario 1, the ratios of selections were equal to 0.55 $(0.36/0.65)$ and 0.72 $(0.16/0.22)$, respectively. This showed that, in fact, the performance in Scenario 2 is better than in Scenario 1. Furthermore, the benchmark also revealed that it is equally difficult for D_5 to find ED_{50} and ED_{90} in Scenario 1 while the analysis without the benchmark suggests a noticeable difference.

Overall, the benchmark leads to the conclusion that both designs perform uniformly in both scenarios in terms of identifying the ED_{90}, but have noticeable problems in finding the ED_{50} under the Sigmoid Emax scenario. Alternative specifications of D_0 (e.g. other allocation proportions) and of D_5 (e.g. other candidate models) should be investigated. In this section, we have focused on the PCS as the designs' main characteristic under the evaluation. However, other metrics could be of interest to a researcher. We discuss the potential merit of the benchmark application for other measures of performance in Section 6.

4 Implication of the Dichotomization in Dose Ranging Trials

Above, it was shown how the benchmark can help with the evaluation of different designs for dose ranging studies with continuous endpoints. However, it is commonplace to dichotomize outcomes measured on a continuous scale. The dichotomization inevitably will lead to different sample sizes required to achieve the desired accuracy. We aim to quantify these differences using the generalized benchmark taking into account various assumptions of the trial. Essentially, we compare two benchmarks: (i) the original benchmark O'Quigley et al. (2002) applied to the dose ranging trial with a binary (dichotomized) endpoint, and (ii) the generalised benchmark (Mozgunov et al., 2018) applied to the dose ranging trial with a continuous endpoint. As both of these approaches provide upper bounds on the PCS in the corresponding settings, the difference in them can be used to obtain a lower bound for the losses of the information associated with the dichotomization, and consequently, assess the increase in the sample size required after the dichotomization to reach a given accuracy.

Below, we used a clinical trial example and associated data, to demonstrate how the benchmark can be used to quantify the loss of information resulting from the dichotomization in dose ranging clinical trials, and how it can inform the decision on the choice of the endpoint. Specifically, we conducted a comprehensive study studying several aspects of the dichotomization and clinical settings, specifically, influence of the means of outcomes and position of the target dose; influence of the dichotomization threshold; influence of the standard deviation of outcomes, and influence of the target probability.

4.1 Dichotomisation in a Phase II cancer clinical trial

In phase II cancer trials for which the primary endpoint is the tumor-shrinking effect, the underlying measurement is on a continuous scale. Despite that, it is commonplace to use the RECIST (Eisenhauer et al., 2009) to dichotomize this measurement and form a binary endpoint to assess the treatment. In this section, we use the information from a single-arm Phase II cancer clinical trial considered by Wason et al. (2011) to construct a setting of a dose ranging cancer clinical trial, and to evaluate the consequence of the dichotomization.

Following Wason et al. (2011), we consider the percentage decrease in the sum of lesion diameters at some dose of a drug, d_j, as a normally distributed random variable, $Y_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$. Here, positive values represent shrinkages in the tumor size, and the negative values - the growth of the tumor. In practice, this variable is used to form a binary response. If some realization of Y_j, y_j, is greater than some threshold ψ, then the binary random variable Z_j, takes values $z_j = 1$, and $z_j = 0$, otherwise. Clearly, the distribution of a random variable Z_j is characterized by a probability, p_j, which can be computed as

$$p_j = \mathbb{P}\left(Z_j = 1\right) = \mathbb{P}\left(Y_j > \psi\right) = 1 - \Phi_j(\psi)$$

where Φ_j is the distribution function of a normal random variable with mean μ_j and variance σ_j^2. The binary random variable, clearly, depends on the choice of the threshold ψ, and the parameter of distributions. The threshold is chosen by the investigator before the trial. For example, the RECIST for partial response corresponds to $\psi = 30\%$. Further, choosing parameters of the distributions, μ_j and σ_j correspond to choosing simulation scenarios over which a dose ranging design is to be evaluated.

From the actual trial data used by Wason et al. (2011), the estimated value of $\sigma = 36.4$ was elicited. Therefore, we used this value in the simulated scenarios, to asses the consequence of dichotomization in this setting. The smallest interesting value of the probability of the response is $p = 0.30$, meaning that an investigator is interested in a dose corresponding to shrinkage of a tumor by at least ψ in 30% of patients. Then, the goal of the dose ranging study is to find the minimum effective dose (among the respecified set of doses) having this probability closest to the target $\gamma = 0.3$. Formally, this translates into the following decision criterion to be used by the benchmark if the binary response is used

$$T(\boldsymbol{y}_j, \gamma = 0.3) = \left| \frac{\sum_{i=1}^{n} z_{ij}}{n} - 0.3 \right|,$$

and the decision criterion

$$T(\boldsymbol{y}_j, \gamma = 0.3) = \left| \left(1 - \Phi_j(\psi, \mu_j = \frac{\sum_{i=1}^{n} y_{ij}}{n}, \sigma_j^2 = var(\boldsymbol{y}_j) \right) - 0.3 \right|$$

if the continuous response is used. Importantly, the crucial step is to make the decision criteria used by the two benchmarks (binary and continuous) comparable. For the first set of simulations, we used the threshold value of $\psi = 30$ as it is being a common choice in clinical practice.

We specified six simulation scenarios on the continuous scale and started from considering the case of equal standard deviations $\sigma = 36.4$ for each dose, and different means. Means of the distributions in all scenarios are chosen such that d_1 is the target dose in Scenario 1, d_2 in Scenario 2 and so on. The scenarios are given in Table 2. The loss of the information due to the dichotomization was measured in terms of the variation in the sample size required to achieve the same accuracy when using the binary endpoint compared to the continuous one. The accuracy of 80% is targeted, meaning that one would like to select the correct dose in at least 80% of trials. The results are based on 40,000 replications that took on average 3–150 seconds of computer time (depending on the setting).

4.2 Numerical Results

The proportions of each dose selection by the benchmark using binary and continuous responses is given in Table 2.

Table 2 Comparison of the Benchmarks (B) using binary vand continuous endpoints and $\sigma = 36.4$ for diffirent values of the sample sizes n in six scenarios. The selection of the target doses are in bold. Results are based on 40000 replications.

Benchmark	n	d_1	d_2	d_3	d_4	d_5	d_6	Ratio of n
Sc 1, Mean (μ_j)		10	20	30	40	50	60	
B – Continuous	61	**80.2**	19.6	0.2	0.0	0.0	0.0	
B – Binary	61	**76.9**	22.5	0.7	0.0	0.0	0.0	
B – Binary	78	**80.4**	19.3	0.3	0.0	0.0	0.0	78.2%
Sc 2, Mean (μ_j)		0	10	20	30	40	50	
B – Continuous	104	6.5	**80.1**	13.3	0.0	0.0	0.0	
B – Binary	104	10.0	**72.6**	17.4	0.1	0.0	0.0	
B – Binary	143	7.1	**80.2**	12.7	0.0	0.0	0.0	72.7%
Sc 3, Mean (μ_j)		-10	0	10	20	30	40	
B – Continuous	104	0.0	6.5	**80.1**	13.3	0.0	0.0	
B – Binary	104	0.0	9.9	**72.6**	17.4	0.1	0.0	
B – Binary	143	0.0	7.1	**80.2**	12.7	0.0	0.0	72.7%
Sc 4, Mean (μ_j)		-20	-10	0	10	20	30	
B – Continuous	104	0.0	0.0	6.5	**80.1**	13.3	0.0	
B – Binary	104	0.0	0.0	9.9	**72.6**	17.4	0.1	
B – Binary	143	0.0	0.0	7.1	**80.2**	12.7	0.0	72.7%
Sc 5, Mean (μ_j)		-30	-20	-10	0	10	20	
B – Continuous	104	0.0	0.0	0.0	6.5	**80.1**	13.3	
B – Binary	104	0.0	0.0	0.0	9.9	**72.6**	17.5	
B – Binary	143	0.0	0.0	0.0	7.1	**80.2**	12.7	72.7%
Sc 6, Mean (μ_j)		-40	-30	-20	-10	0	10	
B – Continuous	32	0.0	0.0	0.0	1.1	18.8	**80.1**	
B – Binary	32	0.0	0.0	0.1	3.0	22.9	**73.9**	
B – Binary	51	0.0	0.0	0.0	0.8	18.7	**80.4**	62.7%

The dichotomization of the primary endpoint led to higher sample sizes in all scenarios. Under scenarios 2-5 with the target dose being not on the bound of the dose set, the ratio of sample sizes equaled 72.7%. The trial with the binary endpoint would require 39 more patients to achieve the same level of accuracy in these scenarios. Note that all of these scenarios are evaluated as equally difficult by the benchmark.

Under Scenario 1, where the target dose is the lowest dose, the loss in sample size is smaller - the ratio of sample sizes is 78.2%. On the other hand, the ratio of sample sizes in Scenario 6 with the target dose being the highest dose was noticeably smaller - 62.7%. A possible explanation for a greater loss in the sample size under Scenario 6 might be the fact that the right tail of the distribution is of interest. Taking the whole distribution into account allows to exclude dose d_5 as a candidate to be the target dose, and as the dose d_6 is the last one, it remained the only candidate and was selected using the smaller sample size. At the same time, the discrimination using the binary endpoint was more challenging and resulted in 29 more patients required to attain the same proportion of correct selections.

Overall, the same qualitative pattern of a higher ratio (compared to non-boundary scenarios) if the target dose was the lowest dose, and a smaller ratio if the target dose was the highest dose has been found in many different scenarios with different spacing between means μ_j and different variances (not shown).

4.3 Influence of the trial parameters

In the simulations above, the fixed threshold value $\psi = 30$, the target probability, $\gamma = 0.3$, and the standard deviation $\sigma = 36.4$ were chosen. The first two values are selected by a clinician and it is of interest to check whether their choices affects the difference in the sample size. The value of σ were chosen based on previous studies, and for the given values of the means at each dose level, would define the probability of a response of at least ψ at each dose, and as a result, the location of the target dose. Therefore, we also investigate how the dose-response scenario (in terms of the underlying probabilities) affect the dichotomisation losses. Again, we investigated the difference in terms of the sample sizes required to achieve 80% proportion of correct selections.

4.3.1 Influence of threshold ψ

As in the motivating trial, we fix $\sigma = 36.4$. We consider one scenario characterised by the following mean values

$$\mu_1 = -20, \mu_2 = -10, \mu_3 = 0, \mu_4 = 10, \mu_5 = 20, \mu_6 = 27.5,$$

and six thresholds $\psi = 0, 10, 20, 30, 40, 50$. These values correspond to the dose $d_1, \dots, d_6$ being the target doses, respectively. Again, we compare the approach with binary and continuous endpoints by the difference in the sample sizes required to get 80% accuracy. The results are given in Table 3.

As the same spacing between means (and the same variance) was used between doses around the target, the results under scenarios 1-4 matched exactly the results above. Under Scenario 5, however, the difference in means for d_5 and d_6 was now smaller, which resulted in larger sample sizes for both continuous and binary benchmark. The ratio of sample sizes remained nearly the same as in the rest of the scenarios, 72.4%, meaning that under the considered scenarios, the threshold value ψ did not have an impact on the dichotomisation losses in the sample size. The difference, however, can be found under Scenario 6 for which the ratio of sample sizes dropped to 45.7% compared to 62.7% using $\psi = 30$ and the same position of the target dose. As the right tail of the distribution is of interest, a higher threshold made it harder to discriminate between doses d_5 and d_6 in the setting with the binary endpoint. While it holds that the absolute sample sizes are decreased for both benchmarks, the relative loss from the dichotomization was dramatic.

Table 3 Benchmarks (B) using binary and continuous endpoints and $\sigma = 36.4$ for diffirent values of the sample sizes n and dichotomisation threshold $\psi = 0, 10, 20, 30, 40, 50$. The selection of the target doses are in bold. Results are based on 40000 replications.

Dose	n	d_1	d_2	d_3	d_4	d_5	d_6	Ratio of n
Mean (μ_j)		-20	-10	0	10	20	27.5	
Scenario 1: $\psi = 0$								
B – Continuous	61	**80.2**	19.6	0.2	0.0	0.0	0.0	
B – Binary	61	**76.9**	22.5	0.7	0.0	0.0	0.0	
B – Binary	78	**80.4**	19.3	0.3	0.0	0.0	0.0	78.2%
Scenario 2: $\psi = 10$								
B – Continuous	104	6.5	**80.1**	13.3	0.0	0.0	0.0	
B – Binary	104	10.0	**72.6**	17.4	0.1	0.0	0.0	
B – Binary	143	7.1	**80.2**	12.7	0.0	0.0	0.0	72.7%
Scenario 3: $\psi = 20$								
B – Continuous	104	0.0	6.5	**80.1**	13.3	0.0	0.0	
B – Binary	104	0.0	9.9	**72.6**	17.4	0.1	0.0	
B – Binary	143	0.0	7.1	**80.2**	12.7	0.0	0.0	72.7%
Scenario 4: $\psi = 30$								
B – Continuous	104	0.0	0.0	6.5	**80.1**	13.3	0.0	
B – Binary	104	0.0	0.0	9.9	**72.6**	17.3	0.2	
B – Binary	143	0.0	0.0	7.1	**80.2**	12.7	0.0	72.7%
Scenario 5: $\psi = 40$								
B – Continuous	163	0.0	0.0	0.0	2.9	**80.2**	16.9	
B – Binary	163	0.0	0.0	0.0	5.7	**74.2**	20.1	
B – Binary	225	0.0	0.0	0.0	3.2	**80.8**	16.0	72.4%
Scenario 6: $\psi = 50$								
B – Continuous	21	0.0	0.0	0.1	2.9	16.7	**80.2**	
B – Binary	21	0.3	0.3	1.8	9.4	33.5	**54.8**	
B – Binary	46	0.0	0.0	0.0	1.2	18.3	**80.4**	45.7%

4.3.2 Influence of the dose-response scenario

In this part, we investigate the influence of the underlying probabilities of having a response of at least $\psi = 30$, that is $\mathbb{P}(Y_{ij} > 30)$, on the dichotomisation losses. Clearly, the probabilities at a given dose will depend on the combination of the mean and variance at this dose. Therefore, instead, of fixing one value (e.g. mean) and vary another (variance), we differentiate the scenarios in terms of the distances between probabilities at each dose. For example, for the standard deviation $\sigma = 5$, we considered the scenario with means $(24.82, 25.80, 26.63, 27.38, 28.08, 28.73)$ which correspond to the probabilities of being greater than 30 of

$$0.15, 0.20, 0.25, 0.30, 0.35, 0.40,$$

respectively. Then, we called it a scenario with the differences in probabilities equal 5%. The rest of scenarios are constructed similarly. This allowed a more comprehensive investigation under scenarios of varying difficulty. Specifically, we studied the difference of 2.5%, 5%, 10%, 15% and the standard deviations $\sigma = 5, 15, 35$. We fixed the position of the target for each value of σ to be d_4 in all scenarios. The results are given in Table 4

As expected, the ratio of sample sizes had extremely minor or no changes for different values of σ as the corresponding mean at each dose level was also varied to preserve the same distance between the probabilities of interest. For example, for the difference of 5%, the ratio of sample sizes is 69.3%–69.9% for different σ (the variations can be explained by the simulation error). This supports the point that the

Table 4 Sample sizes required to attain the PCS of 80% by the benchmark for continuous and binary endpoints under scenarios with different standard deviations and the dose-activity relationships. Results are based on 40000 replications.

Difference in probabilities	2.5%	5%	10%	15%
Scenario 1: $\sigma = 5$				
B – Continuous	1465	357	89	38
B – Binary	2166	515	122	50
Ratio	67.6%	69.3%	73.0%	76.0%
Scenario 2: $\sigma = 15$				
B – Continuous	1464	366	89	38
B – Binary	2158	525	122	50
Ratio	67.8%	69.7%	73.0%	76.0%
Scenario 3: $\sigma = 35$				
B – Continuous	1482	365	88	38
B – Binary	2166	522	120	50
Ratio	68.4%	69.9%	73.3%	76.0%

ratio of sample sizes (and sample sizes themselves) required to achieve particular PCS depends not on the variance itself but on the underlying probabilities of interest, i.e. combination of the variances and means. Importantly, despite both benchmarks taking the difficulty of the scenario into account (spacing between probabilities of interest), the results have shown that the dichotomization leads to higher relative losses as the distance between the true probabilities of $Y_j > \psi$ decreases. While sample sizes in both binary and continuous cases increased, the binary benchmark corresponds to faster growth. The sample size required by the continuous benchmark to achieved 80% of correct selection was nearly 2.3 times higher comparing scenarios with 15% and 10% differences, and approximately four times higher comparing scenarios with 10% and 5% differences. However, for the binary benchmark, the sample sizes increased by around 2.4 and 4.3, respectively. Therefore, the dichotomization leads to missing even more information when the distance between probabilities is relatively small: the ratio of sample sizes is around 76.0% for the distance of 15%, and drops by nearly 10% to 67.6% - 68.4% for the difference of 2.5%. This means that the continuous outcomes allow for a greater gain when a smaller difference in the underlying probabilities are of interest.

4.3.3 Influence of the target probability

In the above, we focused on the fixed target probability that the patient's outcome will be at least ψ. However, in various trial settings, different target probability can be of interest. Below, we study how the target probability affects the dichotomisation losses. We consider five scenarios with 10% difference between the underlying probabilities of having an outcome of at least $\psi = 30$, the same position of the target dose d_4, and the same standard deviation $\sigma = 36.4$ but with the target probabilities of $\gamma = 0.1, 0.3, 0.5, 0.7, 0.9$, respectively. The results are given in Table 5.

Table 5 Sample sizes required to attain the PCS of 80% by the benchmark for continuous and binary endpoints under scenarios with different target probabilities, γ. Results are based on 40000 replications.

Target probability, γ	10%	30%	50%	70%	90%
B – Continuous	28	88	104	89	28
B – Binary	46	120	147	121	42
Ratio	60.9%	73.3%	70.7%	73.6%	66.7%

The target probability in the trials does have an impact on the relative losses in the sample size if dichotomisation is employed. Specifically, the ratio of sample sizes varies between 60.9% and 73.6%. The smallest ratio (the largest losses in the sample size) can be found when targeting the probability of 10% and 90%, while a target value of 30% and 70% corresponds to smaller, but still considerable, losses. In terms of the absolute sample size gains, the largest sample sizes are required for both benchmarks when target the probability of 50% but the continuous benchmark would require 43 fewer patients to achieve 80% PCS.

Overall, the dichotomization of the continuous variable inevitably leads to higher required sample sizes. The maximum value of the ratio of sample sizes obtained in the simulation study was 78% under the scenario with the target dose being the first one. The least value of 45.7% was obtained for the highest threshold value. Under the majority of scenarios, the most common ratio of sample sizes was found to be 70%-75%. Interestingly, a similar ratio of sample sizes (72.4%) was found revisiting a recent Phase II cardiovascular clinical trial (Karlson et al., 2016) and using actual characteristics of the study (we refer the reader to Supplementary Materials for details). This makes a strong point why a continuous endpoint, if available, should be preferred over the binary one in cancer Phase II dose ranging trials.

5 Violation of the Normality Assumption

The underlying assumption in the comparisons provided above is that the responses are generated from a normal distribution, and that the continuous benchmark uses the normal distribution as well. However, if there is sufficient evidence that the normality assumption might be violated, then a non-normal distribution should be used. In this case, the benchmark as described in Section 2 will employ the assumed distribution as is construction in general terms of CDFs F_j of the outcomes given dose d_j allows for any distribution of outcome. Then, the general line of the comparison of the benchmark with non-normal outcomes and the dichotomised benchmark remains the same.

However, it is plausible that, at the planning stage, a normal distribution is assumed but outcomes come from a non-normal distribution (e.g. a symmetric distribution with heavier tails). It is then of interest to investigate if the presented results are robust to the normality assumption violation.

To study this, we use the same construction of the benchmark as above but amend the data generating algorithm. Let $\bar{V}_\nu$ be a random variable having a Student distribution of ν degrees of freedom (df). Then, the outcome of patient i given dose d_j is assumed to have non-standard Student's distribution of the form

$$V_{ij} = \mu_j + \sigma \bar{V}_\nu.$$

We consider five values of the degree of freedom: $\nu = \infty, 40, 20, 10, 7.5$, that correspond to various extent of heavy tails. Again, we consider the threshold of $\psi = 30$, $\sigma = 36.4$, and construct the scenarios such that the differences between the probabilities of outcome of at least ψ at neighbouring doses is 10%, and the target dose is d_4. The sample sizes required to achieve the PCS of 80% under various degrees of freedom are given in Table 6.

Table 6 Sample sizes required to attain the PCS of 80% by the benchmark for continuous (B–Continuous) and binary (B–Binary) endpoints under scenarios with outcomes generated from Student distribution with various degrees of freedom, $df = 40, 20, 10, 7.5$. Results are based on 40000 replications.

Degrees of Freedom, *df*	∞	40	20	10	7.5
B – Continuous	88	91	98	114	140
B – Binary	120	120	122	120	120
Ratio	73.3%	75.8%	80.3%	95.0%	116.7%

Under the scenario with the normality assumption satisfied, the ratio is the same as the ratio found above for the difference of 10% in the probabilities - 73.3%. As the degrees of freedom decrease, the sample size required by the continuous normal benchmark increases, while the sample size required by the binary benchmark remains nearly the same. Specifically, under the scenario with a slightly longer tails than for the normal distribution, the ratio of sample sizes is 75.8%, nearly 2.5% higher than under the normality assumption. For $\nu = 10$ that corresponds to a noticeably heavier tails and a wider range of values of outcomes (e.g. as high as 500), the ratio increases to as high as 95.0%. While the relative saving in the sample size become lower, the use of continuous outcomes can be provide gain the required number of patients. As degrees of freedom decrease further, for example, $\nu = 7.5$, the sample size for the continuous normal benchmark increases above the sample size of the binary benchmark resulting in the ratio above 100% - 116.7%. In this case, the use of continuous benchmark would not provide the gains in the sample size.

Finally, given that violation of normality assumption results in heavier tails of the outcomes distribution, we study how the target probability of the outcome being at least $\psi = 30$ affects the ratio of sample sizes (Table 7). We study various values of the target probabilities $\gamma = 0.1, 0.3, 0.5, 0.7, 0.9$ for two values of degrees of freedom, $\nu = 20, 10$.

Table 7 Sample sizes required to attain the PCS of 80% by the benchmark for continuous and binary endpoints under scenarios with different target probabilities, γ and degrees of freedom $df = 20, 10$. Results are based on 40000 replications.

Target probability, γ	10%	30%	50%	70%	90%
Degrees of Freedom $df = 20$					
B – Continuous	28	98	113	96	28
B – Binary	41	122	147	121	39
Ratio	68.3%	80.3%	76.9%	79.3%	71.8%
Degrees of Freedom $df = 10$					
B – Continuous	28	114	123	114	27
B – Binary	41	120	147	121	39
Ratio	68.3%	95.0%	83.7%	94.2%	69.2%

A similar pattern as in the normal case can be seen. The minimum ratio (the maximum losses in the sample sizes) is found when targeting the probabilities close to the bound of the unit interval, $\gamma = 0.1$ and $\gamma = 0.9$ - around 70% for both values of the degress of freedom, $\nu = 10, 20$. When targeting $\gamma = 0.30$ and $\gamma = 0.70$, the ratio are nearly the same for the fixed value of ν, nearly 80% for $\nu = 20$ and 95% for $\nu = 10$. The ratio is slightly decreased when targeting the probability of 50% - by nearly 3% under $\nu = 20$, and by 11% under $\nu = 10$ when compared to the case of $\gamma = 0.3$.

Overall, the benchmark based on the normality assumption is robust to slight or moderate violations of this assumption and can still provide substantional gains in the sample sizes under many various configurations compared to the binary benchmark. If the assumption of the normality of assumption is heavily violated, an appropriate distribution should be used when analysing the continuous outcomes. The benchmark can accommodate an arbitrary distribution of outcomes.

6 Discussion

In this work, the use of the dose finding benchmark to evaluate the choice of the trial design and primary endpoint was demonstrated in the setting of Phase II dose ranging clinical trials. It was found that the benchmark can provide insights for a more meaningful evaluation of one of the important characteristics of a dose ranging design, PSC, by accounting for the "difficulty" of simulation scenarios. The benchmark

was also found to be useful to assess the consequences of the dichotomization of the primary endpoint in dose ranging trials in terms of the loss of accuracy of the target dose identification as well as in the sample size increase to conserve the same level of accuracy. Under all considered scenarios, dichotomization leads to the inevitable increase in the sample size. The ratio of the sample sizes varied between 48% and 78%. Interestingly, putting these into the context of other clinical trials settings, the provided upper bound on the ratio found is below the respective bound for the parallel group trial and Simon's two-stage trials (Senn, 2005; Wason et al., 2011). This makes a strong point why the continuous endpoint, if clinically feasible, should be used in dose ranging studies. Importantly, estimating the impact on the sample sizes using the benchmark approaches allows illuminating all other factors, e.g. choice of the design for binary and continuous settings, as the benchmarks are based on the same information and on no parametric model.

It is worth mentioning that the main underlying concept of the benchmark used throughout this work, the complete information, is similar to the recently proposed generalized benchmark (Mozgunov et al., 2018): to obtain the upper bound of performance we assume that we know how each patient responds to each dose. One of the goals of this work is to demonstrate how versatile the benchmark is, and specifically, that it can be applied to other class of dose finding designs used in dose ranging trials that are usually different from methods used in Phase I dose-escalation studies. One of the important differences in the benchmark construction is the definition of the target dose employed by the benchmark applied in the paper. It was scenario-specific rather than a fixed value as many Phase I trials. Finally, it is worth mentioning that similar benchmark techniques can be used for the quantification of dichotomization costs in Phase I dose-escalation trials. However, we intentionally focus on the dose ranging setting throughout the work as the benchmark has not been applied to it before.

The benchmark, as used throughout this work, considered evaluation of dose ranging designs and quantification of dichotomisation losses in terms of the PCS. At the same time, other measures of performance can be also of interest. For example, one can be interested in the precision of the estimation of the target dose. We refer the reader to Supplementary Materials where we provide an example of the quantification of the dichotomisation losses in terms of a precision criterion. In this example, adding to the identification of the dose corresponding to a particular target probability, it is also required that the ratio of the upper and lower bounds of the corresponding confidence interval for the probability estimation to be below a particular value. For scenarios with various differences in the probabilities for different doses, we have found in this example that the ratio of sample sizes required to achieve a particular level of estimation precision on the target dose by the continuous and binary benchmark is around 70% under the majority of considered scenarios but can drop to as low as 15% when the variance of the outcomes is low compared to the difference in means. The maximum losses, therefore, are found to be higher than for the PCS criterion, and depend on σ.

This example above (as well as evaluation of PCS) concerns the case when the choice of the target dose is restricted to the set of candidate doses that are pre-specified in advance. Model-based dose finding designs, however, allow for interpolation between the pre-specified doses, and this aspect of dose finding designs' evaluation so far is overlooked by the considered implementation of the benchmark. Nevertheless, the benchmark might have a merit to support an evaluation of designs when interpolation is of interest. One of the core points of the benchmark's implementation is that it does not employ a parametric model on the dose-response relationship. Therefore, in term of the benchmark evaluation, once the compelete outcomes are generated for all patients at each dose level, one can fit a non-parametric regression and use it for the interpolation. The merit of this approach for the evaluation of other properties of Phase II dose ranging designs is subject to the future research.

Throughout the work, the scenarios with monotonically increasing dose-response relationships were considered only. The generalized benchmark used in this work can be applied in scenarios with non-monotonic shapes as it just requires the quantile transformation of patient profile u_i at each dose level. It would still provide an upper bound on the PCS. These bounds, however, might not be as tight as the bounds accounting for the uncertainty in the monotonic ordering, and might not be as useful as evaluation tools that can account for both uncertainties in the monotonic ordering and in the scenarios. Therefore, the

benchmark as considered can be of limited use in such settings. Currently, it remains unclear how such a benchmark can be constructed and is subject to our future research.

Supplementary Materials

Additional information including the example of the benchmark application and the dichotomisation costs evaluation in the setting of a cardiovascular disease may be found in Supplemental Materials.

Acknowledgements This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 633567. Xavier Paoletti is partially funded by the Institut National du Cancer (French NCI) grant SHS-2015 Optidose immuno project. This report is independent research arising in part from Prof Jaki's Senior Research Fellowship (NIHR-SRF-2015-08-001) supported by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care (DHCS).

References

Altman, D. G. and Royston, P. (2006) The cost of dichotomising continuous variables. *Bmj*, **332**, 1080.

Barnwell-Ménard, J.-L., Li, Q. and Cohen, A. A. (2015) Effects of categorization method, regression type, and variable distribution on the inflation of type-i error rate when categorizing a confounding variable. *Statistics in medicine*, **34**, 936–949.

Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M. et al. (2009) New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer*, **45**, 228–247.

Fedorov, V., Mannino, F. and Zhang, R. (2009) Consequences of dichotomization. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, **8**, 50–61.

Karlson, B. W., Wiklund, O., Palmer, M. K., Nicholls, S. J., Lundman, P. and Barter, P. J. (2016) Variability of low-density lipoprotein cholesterol response with different doses of atorvastatin, rosuvastatin, and simvastatin: results from voyager. *European Heart Journal–Cardiovascular Pharmacotherapy*, **2**, 212–217.

Kullback, S. (1997) *Information theory and statistics*. Courier Corporation.

Lei, Y., Carlson, S., Yelland, L. N., Makrides, M., Gibson, R. and Gajewski, B. J. (2017) Comparison of dichotomized and distributional approaches in rare event clinical trial design: a fixed bayesian design. *Journal of applied statistics*, **44**, 1466–1478.

Mielke, T. and Dragalin, V. (2017) Two-stage designs in dose finding. In *Handbook of Methods for Designing, Monitoring, and Analyzing Dose-Finding Trials* (eds. J. O'Quigley, A. Iasonos and B. Bornkamp), chap. 14, 247–265. CRC Press, Taylor and Francis Group.

Mozgunov, P., Jaki, T. and Paoletti, X. (2018) A benchmark for dose finding studies with continuous outcomes. *Biostatistics*, doi.org/10.1093/biostatistics/kxy045.

O'Quigley, J., Paoletti, X. and Maccario, J. (2002) Non-parametric optimal design in dose finding studies. *Biostatistics*, **3**, 51–56.

Senn, S. (2005) Dichotomania: an obsessive compulsive disorder that is badly affecting the quality of analysis of pharmaceutical trials. *Proceedings of the International Statistical Institute, 55th Session, Sydney*.

Spertini, F., Audran, R., Lurati, F., Ofori-Anyinam, O., Zysset, F., Vandepapelière, P., Moris, P., Demoitié, M.-A., Mettens, P., Vinals, C. et al. (2013) The candidate tuberculosis vaccine mtb72f/as02 in ppd positive adults: a randomized controlled phase i/ii study. *Tuberculosis*, **93**, 179–188.

Verkindre, C., Fukuchi, Y., Flémale, A., Takeda, A., Overend, T., Prasad, N. and Dolker, M. (2010) Sustained 24-h efficacy of nva237, a once-daily long-acting muscarinic antagonist, in copd patients. *Respiratory Medicine*, **104**, 1482–1489.

Wason, J. M., Mander, A. P. and Eisen, T. G. (2011) Reducing sample sizes in two-stage phase ii cancer trials by using continuous tumour shrinkage end-points. *European Journal of Cancer*, **47**, 983–989.

Wason, J. M. and Seaman, S. R. (2013) Using continuous data on tumour measurements to improve inference in phase ii cancer studies. *Statistics in medicine*, **32**, 4639–4650.

