Journal of Statistical Planning and Inference 199 (2019) 179–187

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Design of experiments for a confirmatory trial of precision medicine

Kim May Lee *, James Wason

MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, UK

ARTICLE INFO

Article history:
Received 8 December 2017
Received in revised form 7 June 2018
Accepted 16 June 2018
Available online 23 June 2018

Keywords:
Design of experiments
Regression model
Treatment randomization scheme
Weighted *L*-optimality

ABSTRACT

Precision medicine, aka stratified/personalized medicine, is becoming more pronounced in the medical field due to advancement in computational ability to learn about patient genomic backgrounds. A biomaker, i.e. a type of biological process indicator, is often used in precision medicine to classify patient population into several subgroups. The aim of precision medicine is to tailor treatment regimes for different patient subgroups who suffer from the same disease. A multi-arm design could be conducted to explore the effect of treatment regimes on different biomarker subgroups. However, if treatments work only on certain subgroups, which is often the case, enrolling all patient subgroups in a confirmatory trial would increase the burden of a study. Having observed a phase II trial, we propose a design framework for finding an optimal design that could be implemented in a phase III study or a confirmatory trial. We consider two elements in our approach: Bayesian data analysis of observed data, and design of experiments. The first tool selects subgroups and treatments to be enrolled in the future trial whereas the second tool provides an optimal treatment randomization scheme for each selected/enrolled subgroups. Considering two independent treatments and two independent biomarkers, we illustrate our approach using simulation studies. We demonstrate efficiency gain, i.e. high probability of recommending truly effective treatments in the right subgroup, of the optimal design found by our framework over a randomized controlled trial and a biomarker–treatment linked trial.

1. Introduction

A randomized controlled trial (RCT) has been the gold standard for testing a new intervention in medicine, especially in phase III confirmatory studies. Many treatments work differently in different patient subgroups, and in this case RCTs which enroll all patients are not necessarily the most efficient approach in phase III. Instead enriched designs that recruit patients likely to benefit have considerable advantages. There is a danger that enriching a phase III trial too much may lead to missing out on a patient subgroup that would have actually benefited.

This therefore motivates phase II trials of targeted agents investigating not only whether a drug works but in which patient subgroups it works in. Several 'biomarker-driven' trial designs have been proposed to allow investigation of multiple treatment arms in different patient subgroups (Buxton et al., 2014; Kaplan et al., 2013; Kaplan, 2015; Middleton et al., 2015). In the case where each treatment can be tested in each subgroup, the number of hypotheses to be tested in a trial can be very large. Some recent papers providing an overview of biomarker-driven trial designs include Antoniou et al. (2016), Renfro and Sargent (2017), Antoniou et al. (2017) and Parmar et al. (2017).

* Corresponding author.
E-mail address: kim.lee@mrc-bsu.cam.ac.uk (K.M. Lee).

https://doi.org/10.1016/j.jspi.2018.06.004

One important aspect of biomarker-driven trial designs that has not been well researched is how to use the information collected from a phase II trial assessing multiple treatments and biomarkers to design the most efficient phase III designs. In particular it would be very useful to have a framework which determines which treatments should be tested in phase III, and in which biomarker subgroups. There has been some work in the context of evaluating a single experimental treatment (Ondra et al., 2016), but to our knowledge none that investigates novel multi-arm phase II biomarker-driven trial designs.

Considering a regression model with first order interaction terms, we propose a tool to design a confirmatory trial based on the analysis of an observed phase II trial or a historical study. There are two elements in this tool: Bayesian data analysis on data of a phase II trial, and the application of design of experiments to finding an optimal design for future experiment. The focus of our tool is to find an efficient design that could reject false null hypotheses in the confirmatory trial with high power.

Bayesian data analysis is a flexible approach where the knowledge and confidence of clinicians can be incorporated into the framework via the specification of a prior distribution. When the sample size of the observed trial is small, we suggest bootstrapping the data for the Bayesian analysis, and conjecture a subset of hypotheses that would be tested in a confirmatory trial based on a posterior predictive distribution from the analysis. We then use the notion of design of experiments to find the optimal treatment randomization scheme for the future experiments based on these information. Design of experiments is an approach that provides guidance on data collection such that sufficient information could be collected for a future experiment. We consider a weighted version of L-optimal criterion that resemble the idea of Morgan and Wang (2010) where they consider weighted D-, A-, and E-optimal designs for a factorial model. Sverdlov and Rosenberger (2013) review methods on finding optimal allocation for multi-arm clinical trials, where the design depends on the unknown parameters of a factorial model. We note that the Bayesian data analysis in our framework is independent of the commonly used Bayesian optimal design framework, see for example Kathryn and Verdinelli (1995) for the review on Bayesian optimal design framework. Our framework can be generalized to finding a Bayesian optimal design for generalized linear and nonlinear models.

The structure of the paper is as follows. We present a statistical model and hypothesis testing procedure for the trial with biomarker setting in Section 2. We introduce our novel design approach in Section 3, and conduct simulation study to compare the performance of the proposed optimal designs with two commonly employed designs in Section 4. We discuss our work and provide some insights into future research topics in Section 5.

1.1. Motivating trial

As the motivation for the work that follows, we consider a phase II trial that, at the time of writing, is under consideration for funding. This trial will test two experimental targeted treatments (T1 and T2), against chemotherapy control, for high grade serous ovarian cancer. Two biomarkers are included (B1 and B2) with it being thought likely (but not definite) that T1 will work best in B1 positive patients and T2 in B2 positive patients. Patients can be positive for B1, B2, both or neither.

The endpoint used for efficacy is six month change in the level of circulating tumor DNA in the blood, which will be treated as normally distributed on the log scale. The objective of the phase II trial is to determine which of T1 and T2 should be tested in a larger phase III trial, and in which patient subgroups. The methodology in this paper will be used for helping to make this decision.

2. Background and notation

Let vector $x_i = (x_{i1}, \ldots, x_{iL})$ be a biomarker profile of patient i where $x_{il} = 1$ represents patient i is positive for biomarker l, and $x_{il} = 0$ otherwise, $l = 1, \ldots, L$; T_{ik} be the experimental treatment indicator where $T_{ik} = 1$ indicates that patient i receives treatment k. The response model for patient i is

$$y_i = \alpha + \sum_{k=1}^{K} T_{ik}\beta_k + \sum_{l=1}^{L} x_{il}\gamma_l + \sum_{k=1}^{K}\sum_{l=1}^{L} T_{ik}x_{il}\delta_{kl} + \epsilon_i,$$

where α is the placebo/control effect for a patient with a negative biomarker profile, i.e. $x_i = (0, \ldots, 0)$, β_k is the main effect of experimental treatment k, γ_l is the main effect of biomarker l, and δ_{kl} is the interaction between treatment k and biomarker l. A placebo/control treatment is indicated by $T_{ik} = 0, \forall k = 1, \ldots, K$. The residual errors, ϵ_i, are assumed to be identically and independently distributed, and that they are normally distributed with zero mean and a common variance σ^2, i.e. $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, \ldots, n$.

As an example, consider a trial where there are two experimental treatments and two biomarkers, i.e. $K = 2$ and $L = 2$, and that each patient receives only one treatment (either $T_{i1} = 1$ or $T_{i2} = 1$) or a placebo/control treatment, $T_{i1} = 0$ and $T_{i2} = 0$. The response model is

$$\begin{aligned} y_i &= \alpha + \beta_1 T_{i1} + \beta_2 T_{i2} + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \delta_{11} x_{i1} T_{i1} + \delta_{12} x_{i2} T_{i1} + \delta_{21} x_{i1} T_{i2} + \delta_{22} x_{i2} T_{i2} + \epsilon_i \\ &= f(x_{il}, T_{ik})\theta + \epsilon_i, \end{aligned} \tag{1}$$

where

$$f(x_{il}, T_{ik}) = (1,\ T_{i1},\ T_{i2},\ x_{i1},\ x_{i2},\ x_{i1}T_{i1},\ x_{i2}T_{i1},\ x_{i1}T_{i2},\ x_{i2}T_{i2})$$

Table 1
Biomarker–treatment combinations, model parameters for each combination, difference between the effect of a treatment and the placebo/control within the subgroups, and randomization scheme of some designs.

i'	x_{i1}	x_{i2}	T_{i1}	T_{i2}	Treatment effect	Difference to control	ξ_{rct}	ξ_{tlt}	ξ_L^*
1	0	0	0	0	α		1/3	1/3	0.46
2	0	0	1	0	$\alpha + \beta_1$	β_1	1/3	1/3	0.54
3	0	0	0	1	$\alpha + \beta_2$	β_2	1/3	1/3	0.00
4	0	1	0	0	$\alpha + \gamma_2$		1/3	1/2	0.46
5	0	1	1	0	$\alpha + \gamma_2 + \beta_1 + \delta_{12}$	$\beta_1 + \delta_{12}$	1/3	0	0.00
6	0	1	0	1	$\alpha + \gamma_2 + \beta_2 + \delta_{22}$	$\beta_2 + \delta_{22}$	1/3	1/2	0.54
7	1	0	0	0	$\alpha + \gamma_1$		1/3	1/2	0.46
8	1	0	1	0	$\alpha + \gamma_1 + \beta_1 + \delta_{11}$	$\beta_1 + \delta_{11}$	1/3	1/2	0.54
9	1	0	0	1	$\alpha + \gamma_1 + \beta_2 + \delta_{21}$	$\beta_2 + \delta_{21}$	1/3	0	0.00
10	1	1	0	0	$\alpha + \gamma_1 + \gamma_2$		1/3	1/3	0.47
11	1	1	1	0	$\alpha + \gamma_1 + \gamma_2 + \beta_1 + \delta_{11} + \delta_{12}$	$\beta_1 + \delta_{11} + \delta_{12}$	1/3	1/3	0.00
12	1	1	0	1	$\alpha + \gamma_1 + \gamma_2 + \beta_2 + \delta_{22} + \delta_{21}$	$\beta_2 + \delta_{22} + \delta_{21}$	1/3	1/3	0.53

ξ_{rct} = randomized controlled trial; ξ_{tlt} =biomarker–treatment linked trial; ξ_L^* =optimal design for the first illustration.

is a row vector of the design matrix, denoted by X, of the regression model, and

$$\theta = (\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}, \delta_{21}, \delta_{22})^T$$

can be estimated by least squares estimator, $\hat{\theta} = (X^TX)^{-1}X^Ty$, that has $cov(\hat{\theta}) = (X^TX)^{-1}\sigma^2$, where

$$\begin{aligned} X^TX &= \sum_{i=1}^{n} f^T(x_{il}, T_{ik}) f(x_{il}, T_{ik}) \\ &= n \sum_{i'=1}^{m} \frac{n_{i'}}{n} f^T(x_{i'}, T_{i'}) f(x_{i'}, T_{i'}) \\ &= n \sum_{i'=1}^{m} p_{i'} f^T(x_{i'}, T_{i'}) f(x_{i'}, T_{i'}), \end{aligned}$$

$\sum_{i'=1}^{m} n_{i'} = n$, $n_{i'}$ and $p_{i'} = \frac{n_{i'}}{n}$ correspond to the number and proportion of patients who have the same $f^T(x_{i'}, T_{i'})$, and m denotes the number of unique $f^T(x_{i'}, T_{i'})$.

Without loss of generality, negative values indicate that the treatment is effective on the patients. We can test the existence of treatment effect on patients with different biomarker profiles by conducting hypothesis tests where

$$\begin{aligned} \text{null hypothesis:} \quad & c_r'\theta = 0, \\ \text{alternative hypothesis:} \quad & c_r'\theta < 0, \end{aligned}$$

c_r is a vector that indicates the linear combination of regression parameters, i.e. the corresponding treatment effect that is different to the placebo/control, $r = 1, \ldots, R$ is an index of the hypotheses that are possibly be tested in the study. As an example, consider patients who have biomarker profile $(x_{i1}, x_{i2}) = (1, 0)$, the difference between the effect of $T_{i1} = 1$ and $T_{ik} = 0\ \forall k$ on this subgroup is $\beta_1 + \delta_{11}$. Table 1 shows the treatment effect and the difference to the control/placebo effect for each subgroup where the responses follow (1), Table 2 shows the possible values of c_r when model (1) is the analysis model.

We reject the null hypothesis, $c_r'\theta = 0$, if the test statistic,

$$\frac{c_r'\hat{\theta}}{[c_r'[cov(\hat{\theta})]c_r]^{1/2}} < -Z_{\alpha'},$$

where $Z_{\alpha'}$ is the $(1 - \alpha')\%$ quantile of a standard normal distribution, with α' Type 1 error. It is a desired property that $cov(\hat{\theta})$ is small in statistical analysis, reflecting that the data provides good and sufficient information to understand the underlying random process. In terms of a hypothesis test, smaller values of $cov(\hat{\theta})$ lead to higher probability to reject false null hypotheses (i.e. when the true value $c_r'\theta < 0$). We note that the real parameter values, θ, are not known in practice but the covariance matrix of the least squares estimator is inversely proportional to the information matrix, X^TX, which does not depend on θ.

3. Design of confirmatory trial

To investigate the effectiveness of treatments on biomarker subgroups, the biomarker profiles need to be known in advance. Conventional designs such as a randomized controlled trial would fail to study the treatment effect on different subgroups when the information on biomarker profiles is unavailable. On the other hand, a biomarker–treatment linked

Table 2
The possibly tested hypotheses when model (1) is the analysis model. The rth column represents the vector c_r.

r	1	2	3	4	5	6	7	8
θ	c_r							
α	0	0	0	0	0	0	0	0
β_1	1	1	1	1	0	0	0	0
β_2	0	0	0	0	1	1	1	1
γ_1	0	0	0	0	0	0	0	0
γ_2	0	0	0	0	0	0	0	0
δ_{11}	0	1	0	1	0	0	0	0
δ_{12}	0	0	1	1	0	0	0	0
δ_{21}	0	0	0	0	0	1	0	1
δ_{22}	0	0	0	0	0	0	1	1

trial would not administer a linked-treatment to a subgroup with a negative biomarker, and administer all possible tested treatments to the subgroup who has all negative biomarkers. If treatments work only in biomarker positive subgroups, an enrichment design could be implemented where the subgroup with all negative biomarkers is excluded in the trial. All of these commonly used designs consider equal randomization probabilities to assign treatments within subgroups. The columns, ξ_{rct} and ξ_{tlt}, in Table 1 show the randomization schemes of a randomized controlled trial and of a treatment linked trial for a study with two independent biomarkers and two independent treatments.

Instead of using these conventional designs, we propose to design a phase III/confirmatory trial based on an analysis of a phase II/exploratory study. The idea of our design framework is as follows: analyze an observed data using a Bayesian framework to provide guidance on selecting a subset of R possibly tested hypotheses and subgroups of patients to enroll in the confirmatory trial; formulate an analysis model for the future experiment and find an optimal randomization scheme. We propose to formulate an analysis model parsimoniously at the design stage to save costs on enrollment and data collection for the future confirmatory trial. For a chosen model, we consider the notion of design of experiments whereby the optimal treatment allocation scheme is of interest. An optimal design framework chooses the setting of a confirmatory experiment through the design matrix X such that some functions of $cov(\hat{\theta})$ are minimized. The following sections illustrate the key ideas of our framework: investigate which hypotheses are more beneficial to focus on in the future experiment, and find an efficient design that has high probability of recommending truly effective treatments in the right subgroups.

3.1. Specification of relative importance of hypotheses

We now illustrate the selection of hypotheses and subgroups using the data of a previous experiment such as a phase II study. The idea is to consider an analysis approach that aims to provide insight into a predictive distribution of model parameters of the future confirmatory trial. We consider a Bayesian approach of data analysis to account for the unseen uncertainty when selecting a subset of hypotheses that will be tested in the confirmatory trial, reflecting the prior belief or confidence of an experimenter in terms of what might happen in the future experiment. This is achieved by specifying a prior distribution for the model parameters. When the sample size of an observed study is small, we propose to use simple bootstrap procedure on the data, and conduct the Bayesian analysis in each bootstrap replication to overcome the variability of using only one set of data to make selection of hypotheses. If only the summary statistics of the phase II trial are available, we recommend to replicate the phase II trial and explore the operating characteristics of the optimal designs for a future experiment, where each optimal design is obtained based on the analysis of a replication of the phase II trial.

To illustrate our framework, we use a conjugate prior in the following presentation. Markov chain Monte Carlo (MCMC) sampling could be used when the posterior distribution is intractable. We define the total sample size of the phase II trial as n_{II}, the vector of observed responses as y_o, design matrix as X_0, and regression parameters of the response model of phase II as θ_{II}. For an observed study with a small n_{II}, we first bootstrap y_o and X_0 with replacement and conduct Bayesian data analysis on each bootstrap sample. For example, using a normal-inverse-gamma prior, $NIG(\theta_0, V_0, a, b)$, for (θ_{II}, σ^2), we obtain a posterior distribution $NIG(\theta_m^*, V^*, a^*, b^*)$,

$$\theta_m^* = \left(V_0^{-1} + X_0^T X_0\right)^{-1}(V_0^{-1}\theta_0 + X^T y_o),$$
$$V^* = \left(V_0^{-1} + X_0^T X_0\right)^{-1},$$
$$a^* = a + \frac{n_{II}}{2},$$
$$b^* = b + \frac{1}{2}\left[\theta_0^T V_0^{-1}\theta_0 + y_o^T y_o - (\theta^*)^T (V^*)^{-1}\theta^*\right]$$

for each bootstrapped sample. The marginal distribution of θ_{II} follows a multivariate t-distribution, $MVSt_{2a^*}(\theta_m^*, \Sigma^*)$ with $\Sigma^* = (\frac{b^*}{a^*})V^*$ and $2a^*$ degree of freedom.

Let w_r denote the relative importance of hypothesis r, $r = 1, \ldots, R$. We propose to generate large samples from the posterior distribution (use MCMC sampling for intractable posterior distribution) to compute

$$P_r = P(c_r'\theta < \tau_r)$$

where $c_r'\theta$ is approximately normally distributed with mean $c_r'\theta_m^*$, and variance $c_r'\Sigma^* c_r$, and τ_r could be the minimum uninteresting treatment difference threshold for hypothesis r. We then find $E(P_r)$ where the expectation is averaged across the number of bootstrap replications, and compute w_r, the relative importance of hypothesis r, by

$$w_r = \begin{cases} E(P_r) & \text{if } E(P_r) \geq \kappa, \text{ where } \kappa \text{ is a user-specified threshold;} \\ 0 & \text{otherwise.} \end{cases}$$

Hypothesis r is selected and the corresponding subgroup is enrolled into the confirmatory trial if $w_r > 0$. These information is then used to formulate a parsimonious linear regression model and the design criterion for finding an optimal randomization scheme for the future confirmatory trial. Note that $\sum_{r=1}^{R} w_r$ need not sum up to 1 in the design problem.

3.2. Design of experiments

We now describe the design framework for finding an optimal design. After formulating the analysis model parsimoniously for the future experiment such that the parameters of interest are estimable, we want to find an optimal design,

$$\xi^* = \begin{Bmatrix} (x_1, T_1) & \cdots & (x_m, T_m) \\ p_1 & \cdots & p_m \end{Bmatrix},$$

that minimizes

$$\sum_{r=1}^{R} w_r\, c_r'[cov(\hat{\theta})]c_r \propto \sum_{r=1}^{R} w_r\, c_r' \left(n \sum_{i'=1}^{m} p_{i'} f^T(x_{i'l}, T_{i'k}) f(x_{i'l}, T_{i'k}) \right)^{-1} c_r, \tag{2}$$

where $w_r \geq 0$ reflects the relative importance of the corresponding hypothesis r. This setup has several well-known optimality criteria as special cases: the design criterion is called L-optimality when $w_r = 1\ \forall r$ (page 111 in Luc Pronzato and Andrej Pazman (2013)); it is an A-optimality when the summation sums the individual variances of the model parameters and is a c-optimality when $R = 1$ and $w_1 = 1$. The incorporation of w_r into the standard L-optimality resembles the idea of Morgan and Wang (2010) where the authors consider weighted D-, A-, and E-optimal designs for a factorial model.

A design ξ is called a continuous design when $np_{i'}$ need not be a positive integer in the optimization search but $\sum_{i'=1}^{m} p_{i'} = 1, p_{i'} \geq 0$. Otherwise ξ is called an exact design where $np_{i'}$ is a positive integer. The notion of a continuous design facilitates the search of an optimal design over a design region. A rounding procedure can be applied to the continuous design for practical implementation, see for example Pukelsheim and Rieder (1992). The interpretation of $p_{i'}$ in the conventional design framework context is that it is the suggested proportion of subgroups who have the corresponding $(x_{i'}, T_{i'})$. In the trial with biomarker setting, it is difficult to enroll the patient subgroups in practice according to the exact proportions. Hence, we use $p_{i'}$ to compute the randomization probability for each selected subgroup instead. Note that no rounding procedure is needed in this case. For a given total sample size, we find $p_{i'}$ of ξ^* by minimizing (2), subject to $\sum_{i'=1}^{m} p_{i'} = 1$. The relative importance of the selected subset of hypotheses and the subgroups are reflected by the values of w_r, $r = 1, \ldots, R$, and biomarker–treatment combination $(x_{i'}, T_{i'})$, $i' = 1, \ldots, m$. To avoid confusion, we use the same index r to denote the hypotheses in the design framework even some of the hypotheses are not selected to be tested in the future experiment.

As an example, consider that we are interested in testing only the following null hypotheses:

1. $\beta_1 = 0$; enroll biomarker–treatment combinations, $(x_{i1}, x_{i2}, T_{i1}, T_{i2}) = (0, 0, 0, 0)$ and $(0, 0, 1, 0)$.
2. $\beta_1 + \delta_{11} = 0$; enroll $(x_{i1}, x_{i2}, T_{i1}, T_{i2}) = (1, 0, 0, 0)$ and $(1, 0, 1, 0)$.
3. $\beta_2 + \delta_{22} = 0$; enroll $(x_{i1}, x_{i2}, T_{i1}, T_{i2}) = (0, 1, 0, 0)$ and $(0, 1, 0, 1)$.
4. $\beta_2 + \delta_{22} + \delta_{21} = 0$; enroll $(x_{i1}, x_{i2}, T_{i1}, T_{i2}) = (1, 1, 0, 0)$ and $(1, 1, 0, 1)$.

We propose to employ

$$y_i = \alpha + \beta_1 T_{i1} + \beta_2' T_{i2} + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \delta_{11} x_{i1} T_{i1} + \delta_{21} x_{i1} T_{i2} + \epsilon_i,$$

where $\beta_2' = \beta_2 + \delta_{22}$ instead of model (1) for the analysis of the future trial. An optimal design problem is to minimize

$$w_1 cov(\hat{\beta}_1) + w_2 cov(\hat{\beta}_1 + \hat{\delta}_{11}) + w_7 cov(\hat{\beta}_2 + \hat{\delta}_{22}) + w_8 cov(\hat{\beta}_2 + \hat{\delta}_{22} + \hat{\delta}_{21}),$$

subject to $\sum_{i'=1}^{12} p_{i'} = 1, p_{i'} = 0$ for $i' = 3, 5, 9, 11$. We note that the parameter δ_{12} would not be estimable when the optimal design is used because none of the selected patients would have $x_{i2} T_{i1} = 1$. Besides that, we can only estimate the combined effect of $\beta_2 + \delta_{22}$ here as the enrolled patients would have $T_{i2} = 1$ only if $x_{i2} = 1$. Note that if the experimenter chooses $\kappa = 0$ and hence $w_r > 0\ \forall r$, model (1) would be used as if in the randomized controlled trial where all subgroups are enrolled into the trial.

In general, the analysis model for a future experiment could be formulated by considering the model parameters of the selected hypotheses and the subgroup biomarker profiles. A linear combination of model parameters might be replaced by a new variable accordingly such that the information matrix is of full rank.

4. Illustration: a trial with two biomarkers and two treatments

In this section, we conduct simulation studies to illustrate the application of our framework to finding an optimal design for a confirmatory trial based on Bayesian analysis of a historical study, and make comparisons on the operating characteristics of a randomized controlled trial, ξ_{rct}, a biomarker–treatment linked trial, ξ_{tlt}, and optimal designs. Throughout the illustration, we consider the presence of two independent biomarkers with prevalence rate of 0.3 each, and two independent treatments that are linked to the biomarkers. In the first part of this illustration, we simulate a set of phase II data for finding an optimal design based on our framework, and use the bootstrap estimates as the true model parameters in the simulation of a confirmatory trial. We replicate a confirmatory trial using the randomization scheme of the designs to study the performance of different designs. In the second part of the illustration, we do not bootstrap a set of data but replicate the phase II trial according to a set of model parameters to explore the operating characteristics of different optimal designs, where each optimal design is found based on the analysis of a replicated phase II study. We compare the performance of these optimal designs with ξ_{rct} and ξ_{tlt} in the simulation of a confirmatory trial where the true model parameters are the same as those used in the replication of the phase II trial. The first illustration shows the role of $\kappa > 0$ on the optimal design when a set of phase II data is available; the latter illustration reflects the situation where only the summary statistics of a phase II trial are available, and shows the role of w_r on the optimal design when $\kappa = 0$.

Consider a phase II randomized controlled trial, ξ_{rct}, that studies the effect of two independent treatments, T_{i1} and T_{i2}, on patients where the information of two independent biomarkers, x_{i1} and x_{i2}, are available. For $n_{II} = 400$, we simulate biomarker profiles using binomial distributions with prevalence rate of 0.3 for each biomarker, treatment allocation with equal randomization probability, and a set of y_o according to model (1) with $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, $\sigma^2 = 1.15$ and $\theta_{II} = (-0.356, -0.213, 0.240, 1, 1, -0.182, 0.240, 0.527, -0.586)^T$. We note that model (1) has $r = 1, \ldots, 8$, possible hypotheses (see Table 2). We choose prior values $a = 0.0001$, $b = 0.0001$, $\theta_0 = (0, 0, 0, 0, 0, -0.1, 0, 0, -0.1)^T$, and V_0 a diagonal matrix with $diag(V_0) = (1, 1, 1, 1, 1, 2, 1, 1, 2)$. The values of θ_0 and V_0 are chosen to reflect the belief that the model parameters of a confirmatory experiment, apart from δ_{11} and δ_{22}, have no effect on the responses. The variances of δ_{11} and δ_{22} are relatively larger than the variances of other parameters show that there are more uncertainty in the belief that δ_{11} and δ_{22} may have negative effect on the responses. We consider bootstrapping the data 10 000 times in this illustration. For each bootstrap sample, we compute θ_m^* and Σ^*, and generate 10 000 samples from the posterior distribution, $MVSt_{2a^*}(\theta_m^*, \Sigma^*)$, to compute P_r by choosing $\tau_r = 0\ \forall r$. We then use all P_r from the bootstrap replications to compute $E(P_r)$ and w_r by choosing $\kappa = 0.5$. Based on the values of w_r, we formulate a model parsimoniously and find an optimal design by minimizing (2) using the function *fmincon* in Matlab. In this illustration, we obtained $E(P_1) = 0.761$, $E(P_2) = 0.683$, $E(P_3) = 0.326$, $E(P_4) = 0.367$, $E(P_5) = 0.097$, $E(P_6) = 0.032$, $E(P_7) = 0.883$ and $E(P_8) = 0.501$. Hence, we have $w_r = E(P_r)$ for $r = 1, 2, 7, 8$, and $w_r = 0$ for $r = 3, 4, 5, 6$, in this example. An optimal design denoted by ξ_L^* is obtained for this setting.

To study the operating characteristics of the design, we conduct simulation study with a larger sample size to reflect the practice of a confirmatory trial. We compare the performance of ξ_L^*, with a randomized controlled trial, ξ_{rct}, and a biomarker–treatment linked trial, ξ_{tlt}. These designs are different in terms of number of subgroups and treatment allocation scheme, see Table 1. We use the same prevalence rate, i.e. 0.3 for both independent biomarkers, and sample size of 1000 in the simulation of the confirmatory trial, whereby the treatment randomization scheme follows a design. In each simulation, we replicate the biomarker profiles 100 times; for each replication of biomarker profiles, we replicate treatment allocation according to the randomization scheme 100 times. The responses are generated according to model (1) with the expected model parameters from the bootstrap replications, $\theta = (-0.326, -0.155, 0.324, 0.794, 1.061, -0.014, 0.426, 0.290, -0.784)$, and $\sigma^2 = 1.15$. Using the simulated responses, we estimate the model parameters by least squares estimation, and compute the test statistics for hypothesis test with $\alpha' = 0.05$ in each replication. The number of times that a null hypothesis is rejected is averaged across the replications (both the replication of treatment allocation and biomarker profiles), giving the power of rejecting a null hypothesis if it is false, or a Type 1 error if it is true. In this illustration, we know $c_r'\theta < 0$, for $r = 1, 2, 7, 8$. We compute the expected number of correct rejection of false null hypotheses (ENCR) to make comparisons between the designs.

Table 3 shows the effect sizes from different hypotheses and the operating characteristics of different designs in the simulation of a confirmatory trial. In this illustration, we find that the power of rejecting $c_7'\hat{\theta}$ is the largest whereas the power of rejecting $c_8'\hat{\theta}$ is the smallest; the Type 1 error of rejecting $c_5'\hat{\theta}$ and $c_6'\hat{\theta}$ are close to 0. This shows that the true values of hypotheses play a major part in hypothesis test where large magnitude of model parameters is easier to detect than others that are close to the bound chosen for a hypothesis test. The other reason that the power of rejecting $c_8'\hat{\theta}$ is so small is due to small subgroup sample size and the fact that the power of rejecting a null hypothesis with an interaction term is generally lower than that with only main effect parameters (Follmann, 2003). The expected sample size of this subgroup is $1000 * 0.3 * 0.3 = 90$, which is not sufficient for detecting the treatment effect that is very close to zero. Comparing the designs, the optimal design generally achieves larger power of rejecting the false null hypotheses. Consider the expected number of correct rejection of false null hypotheses (ENCR), we find that ξ_{rct} and ξ_{tlt} would lose about 11% and 7% efficiency over ξ_L^*. Looking at testing $c_1'\theta = 0$, we find that the power could be increased significantly if the optimal design is used in the experiment, i.e. 0.11 and 0.12 more than what ξ_{rct} and ξ_{tlt} could achieve. The randomization scheme of the optimal design for this corresponding subgroup is 0.46 for receiving placebo and 0.54 for receiving treatment 1, whereas the later designs are having 1/3 for receiving placebo, treatment 1 and treatment 2. This shows that enrolling a less informative biomarker–treatment combination could be waste of resources. In particular, $c_5'\theta = \beta_2$ is not significant in this illustration

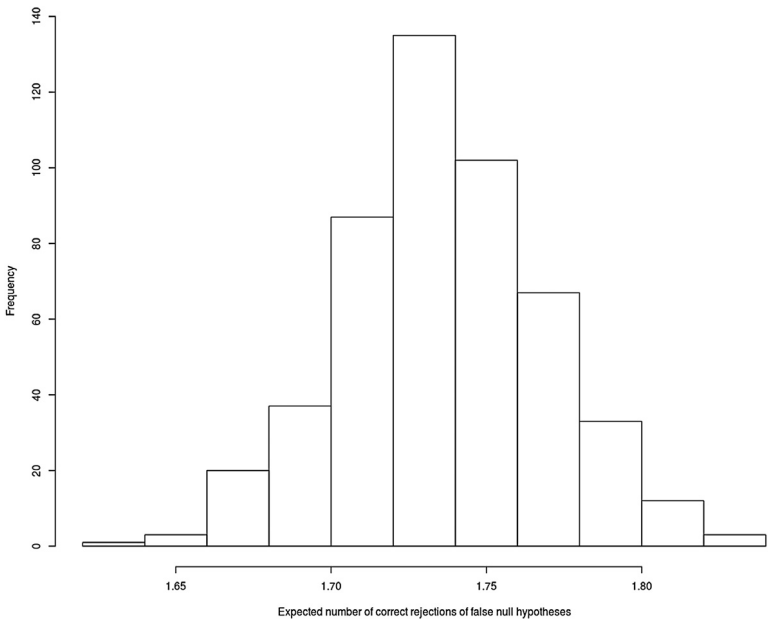

Fig. 1. Frequency of optimal designs that achieve the expected number of correct rejection of false null hypotheses (ENCR).

Table 3
Effect sizes from different hypotheses and probability of rejecting a null hypothesis in the simulation of different designs.

Effect sizes from different hypotheses in the simulation of a confirmatory trial									
	$c_1'\theta$	$c_2'\theta$	$c_3'\theta$	$c_4'\theta$	$c_5'\theta$	$c_6'\theta$	$c_7'\theta$	$c_8'\theta$	
	−0.155	−0.169	0.135	0.121	0.324	0.750	−0.460	−0.034	
Probability of rejecting a null hypothesis in the simulation of a confirmatory trial									
	$c_1'\hat{\theta}=0$	$c_2'\hat{\theta}=0$	$c_3'\hat{\theta}=0$	$c_4'\hat{\theta}=0$	$c_5'\hat{\theta}=0$	$c_6'\hat{\theta}=0$	$c_7'\hat{\theta}=0$	$c_8'\hat{\theta}=0$	ENCR
ξ_{rct}	0.43	0.31	0.01	0.02	0.00	0.00	0.90	0.09	1.74
ξ_{tlt}	0.41	0.37	0.02	0.02	0.00	0.00	0.96	0.08	1.81
ξ_L^*	0.54	0.37	NA	NA	NA	NA	0.96	0.08	1.95

ENCR= expected number of correct rejection of false null hypotheses.

and that having the biomarker–treatment combination $i' = 3$ (see Table 1) in the experiment is not beneficial. Conjecturing which hypothesis is significant and to be tested in future experiment is a crucial step in designing the trial.

To illustrate further, we consider the same trial setting but that only the summary statistics of a phase II trial are available. We consider $c_r'\theta = -0.2$ for $r = 1, 2, 7, 8$ and $c_r'\theta > 0$ for $r = 3, 4, 5, 6$, and $n_{II} = 300$. Instead of bootstrapping one data set (that gives different parameter estimates in each bootstrap replication), we replicate a phase II trial 500 times according to model (1) with an equal probability randomization scheme. These replications reflect the variability of the error term and the randomization scheme while $c_r'\theta$ remain the same across all the phase II replications. For each simulated phase II trial, we draw 10 000 samples from the posterior predictive distribution (i.e. $MVSt_{2a^*}(\theta_m^*, \Sigma^*)$ that has the same prior as the previous illustration) and compute $w_r = E(P_r)$ by setting $\kappa = 0$. An optimal design is then found based on the analysis of each phase II replication. Fig. 1 shows ENCR on the x-axis, that are achieved by the 500 optimal designs where each optimal design corresponds to a simulated phase II study. With total sample size of 1000, nominal Type 1 error of 5%, and true treatment difference of −0.2 for each hypothesis $r = 1, 2, 7, 8$, we find that the optimal designs are expected to reject at least 1.65 false null hypotheses in the simulation, whereas ξ_{rct} and ξ_{tlt} are giving 1.647 and 1.707 respectively.

Table 4 shows the probability of rejecting each null hypothesis $c_r'\theta = 0, r = 1, \ldots, 8$, achieved by ξ_{best}^*, ξ_{worst}^*, ξ_{rct} and ξ_{tlt}. The former two are optimal designs that have the highest and the lowest ENCR (in the simulation of a confirmatory trial) out of the 500 optimal designs, where each design corresponds to a replication of a phase II study. Across all designs that have different randomization schemes, we find that the Type 1 error of rejecting each true hypothesis is no greater than 0.03 when the nominal $\alpha' = 0.05$ is used in the simulation. Comparing the four false null hypotheses that have the same true value of −0.2, we find that the power of rejecting $c_1'\theta = 0$ is the highest and $c_8'\theta = 0$ is the lowest across all designs. This is not surprising as the subgroups have different sample sizes and that with prevalence rate of 0.3 for both biomarkers, the total sample size of $i' = 1, 2, 3$, is expected to be $1000 * (1 - 0.3) * (1 - 0.3) = 490$, and that of $i' = 10, 11, 12$, is $1000 * 0.3 * 0.3 = 90$. The last column of Table 4 shows that ξ_{rct} and ξ_{tlt} would lose about 11% and 7% efficiency in terms of

Table 4
Probability of rejecting a null hypothesis. We know $c_r'\theta < 0$, $r = 1, 2, 7, 8$, and $c_r'\theta > 0$, $r = 3, 4, 5, 6$. Each row corresponds to the performance of a design.

	$c_1'\hat{\theta} = 0$	$c_2'\hat{\theta} = 0$	$c_3'\hat{\theta} = 0$	$c_4'\hat{\theta} = 0$	$c_5'\hat{\theta} = 0$	$c_6'\hat{\theta} = 0$	$c_7'\hat{\theta} = 0$	$c_8'\hat{\theta} = 0$	ENCR
ξ^*_{best}	0.67	0.42	0.02	0.02	0.01	0.02	0.43	0.32	1.83
ξ^*_{worst}	0.61	0.39	0.01	0.02	0.01	0.02	0.35	0.28	1.64
ξ_{rct}	0.59	0.39	0.02	0.02	0.01	0.02	0.38	0.29	1.65
ξ_{tlt}	0.56	0.45	0.03	0.03	0.01	0.03	0.45	0.24	1.71

ENCR when compared with ξ^*_{best}. We find that ξ^*_{worst} achieved a similar ENRC when compared to ξ_{rct}, but not better than ξ_{tlt}. The latter finding is mainly due to the fact that ξ_{tlt} excluded $i' = 5,9$ (see Table 1) in the design whereas ξ^*_{worst} enrolled these biomarker–treatment combinations in the trial, when $i' = 5,9$ have limited contribution on testing false null hypotheses $r = 1, 2, 7, 8$. The former result might be due to the random error of that particular replication of a phase II trial. If we bootstrap this replicated data assuming that they are the observed phase II data, we would then proceed as in the previous illustration but with $\kappa = 0$, and could potentially obtain a better design than ξ^*_{worst} and ξ_{rct} for the future experiment.

5. Discussion

We have proposed a framework for designing confirmatory trials based on the analysis of a phase II study testing multiple experimental treatments in different biomarker subgroups. The design framework provides guidance on selecting biomarker–treatment combinations and a treatment randomization scheme.

When using a single regression model to analyze a multi-arm trial, the dimension of model parameters depends on the number of biomarker–treatment combinations. For example, in the presence of L binary biomarkers and K experimental treatments, there are $2^L \times K$ possible hypotheses that test treatment differences between treatments and a placebo. To find an efficient design, we propose a framework to select a subset of hypotheses out of the many possibly tested hypotheses based on Bayesian data analysis of a phase II/historical study, and use the notion of design of experiments to find the treatment randomization scheme for implementing a future experiment. We show that doing a traditional randomized trial enrolling all patient subgroups is not the most efficient approach when the treatments work only on certain subgroups.

We propose selecting the hypotheses to be tested in a confirmatory trial based on Bayesian data analysis of the phase II study. When the sample size of the observed study is small, we suggest to consider bootstrapping the available data to minimize the bias in estimation (due to small sample) that may cause bias in hypothesis selection. When only summary statistics are available, we suggest simulating the phase II study and exploring the operating characteristics of optimal designs before choosing one design for implementation. The above illustration shows that w_r, κ and τ play important roles in designing an experiment. In practice, κ and τ should be chosen carefully in discussion with clinicians based on their experiences. We note that our framework is flexible that each hypothesis may have different κ and τ. The specification of a prior distribution is another aspect that should reflect the knowledge of the clinicians on the likely magnitude of the treatment effect. When the posterior predictive distribution of model parameters is not in closed-form, we suggest to draw large samples using MCMC approaches, aiming to approximate the distribution of future model parameters by normal approximation.

Concerning the values of w_r, we find that small aberration on each w_r does not affect the optimal randomization scheme notably. The operating characteristics of a design depend on the true model parameters more than w_r. However, the true model parameters are never known in practice. Instead of enrolling subgroups according to the optimal proportion, p_i, $i = 1, \ldots, m$, as suggested by the classical design framework, we convert them into randomization probability as it facilitates the implementation of a trial that considers subgroup stratification. However, this approach would complicate the sample size calculation especially when the prevalence rates of biomarkers are at the extreme end of the range. We consider using a single model in the analysis such that subgroups with small sample sizes may borrow information from other subgroups that have larger sample sizes.

We have presented the framework for normally distributed responses. We note that our framework could be extended to nonlinear models where adjustments to the Bayesian data analysis, computation of w_r, and the design criterion would be required. When the model is nonlinear, the covariance matrix of the parameters may depend on the unknown parameters, leading to issues with finding an optimal design. For a chosen value of the model parameters, the classical optimal design framework could provide a locally optimal design. An alternative to this is to use a Bayesian optimal design framework whereby a prior distribution of the model parameters is incorporated into the design criterion for finding an optimal design. We note that the proposed framework is different to the commonly known Bayesian optimal design framework as we only use historical data to estimate w_r prior to finding an optimal design. Nevertheless, our framework could be extended to the nonlinear situation accordingly. For example, the observed phase II data may be used to construct the prior distribution for the unknown parameters in the Bayesian design framework, while the prior distribution in the Bayesian analysis of phase II data reflects the prior knowledge or confidence of the clinicians on the future experiment. These two different prior distributions could be the same or different. Other possible extensions of our framework could be the incorporation of missing data, see Lee et al. (2017a), Lee et al. (2017b), and incorporation of cost functions, see Cook and Fedorov (1995).

One of the potential topics for future research is to account for the population drift or change in baseline measure of subgroups in the design framework. When data of control groups from several small trials that studied the same disease are available, having a design framework that considers this aspect could be beneficial for the clinical community. See for example Thall and Simon (1990) and Boonstra et al. (2016) who consider the design and information gain when historical control data is incorporated, and van Rosmalen et al. (2017) for the review of methods for incorporating historical control data. Besides that, we have not accounted for all the aspects of an analysis plan. For example, when a single statistical model is chosen to analyze all data, most of the hypotheses are not independent as the information from different subgroups is shared across themselves. Future work could focus on constructing optimal designs that account for issues such as multiplicity while optimizing the hypotheses selection as our design framework has done here. The issue that arose in subgroup sample size calculations may also be incorporated and addressed in the prospective design framework.

To conclude, we have proposed a novel design framework for designing a biomarker driven confirmatory trial based on the analysis of an observed experiment which provides an increased chance of determining which subgroups a targeted treatment genuinely works in.

Acknowledgment

This research has been funded by the Medical Research Council (grant codes MR/N028171/1 and MC_UP_1302/4).

References

Antoniou, M., Jorgensen, A.L., Kolamunnage-Dona, R., 2016. Biomarker-guided adaptive trial designs in phase II and phase III: A methodological review. PLoS One 11 (2), 1–30.

Antoniou, M., Kolamunnage-Dona, R., Jorgensen, A., 2017. Biomarker-guided non-adaptive trial designs in phase II and phase III: A methodological review. J. Personalized Med. 7 (1), 1.

Boonstra, P.S., Taylor, J.M., Mukherjee, B., 2016. Increasing efficiency for estimating treatment biomarker interactions with historical data. Stat. Methods Med. Res. 25 (6), 2959–2971.

Buxton, M.B., Natsuhara, K., DeMichele, A., Perlmutter, J., Hylton, N.M., Yee, D., van't Veer, L., Symmans, W.F., Hogarth, M., Lyandres, J., et al., 2014. Transforming the clinical trial process: The I-SPY 2 trial as a model for improving the efficiency of clinical trials and accelerating the drug-screening process. J. Clin. Oncol. 32.

Cook, D., Fedorov, V., 1995. Constrained optimization of experimental design. Statistics 26, 129–178.

Follmann, D., 2003. Subgroups and interactions. In: Advances in Clinical Trial Biostatistics. pp. 121–141.

Kaplan, R., 2015. The FOCUS4 design for biomarker stratified trials. Chin. Clin. Oncol. 4 (3).

Kaplan, R., Maughan, T., Crook, A., Fisher, D., Wilson, R., Brown, L., Parmar, M., 2013. Evaluating many treatments and biomarkers in oncology: a new design. J. Clin. Oncol. 31 (36), 4562–4568.

Kathryn, C., Verdinelli, I., 1995. Bayesian experimental design : A review. Statist. Sci. 10 (3), 273–304.

Lee, K.M., Biedermann, S., Mitra, R., 2017a. Optimal design for experiments with possibly incomplete observations. Statist. Sinica http://dx.doi.org/10.5705/ss.202015.0225.

Lee, K.M., Mitra, R., Biedermann, S., 2017b. Optimal design when outcome values are not missing at random. Statist. Sinica http://dx.doi.org/10.5705/ss.202016.0526.

Luc Pronzato, ., Andrej Pazman, ., 2013. Design of Experiments in Nonlinear Models. Springer, p. 416.

Middleton, G., Crack, L., Popat, S., Swanton, C., Hollingsworth, S., Buller, R., Walker, I., Carr, T., Wherton, D., Billingham, L., 2015. The National Lung Matrix trial: Translating the biology of stratification in advanced non-small-cell lung cancer. Ann. Oncol. 26 (12), 2464–2469.

Morgan, J.P., Wang, X., 2010. Weighted optimality in designed experimentation. J. Amer. Statist. Assoc. 105 (492), 1566–1580.

Ondra, T., Jobjörnsson, S., Beckman, R.A., Burman, C.-F., König, F., Stallard, N., Posch, M., 2016. Optimizing trial designs for targeted therapies. PLoS One 11 (9), e0163726.

Parmar, M.K., Cafferty, F.H., Sydes, M.R., Choodari-Oskooei, B., Langley, R.E., Brown, L., Phillips, P.P., Spears, M.R., Rowley, S., Kaplan, R., et al., 2017. Testing many treatments within a single protocol over 10 years at MRC Clinical Trials Unit at UCL: Multi-arm, multi-stage platform, umbrella and basket protocols. Clin. Trials 14, 451–461.

Pukelsheim, F., Rieder, S., 1992. Efficient rounding of approximate designs. Biometrika 79 (4), 763–770.

Renfro, L.A., Sargent, D.J., 2017. Statistical controversies in clinical research: Basket trials, umbrella trials, and other master protocols: A review and examples. Ann. Oncol. 28 (1), 34–43.

Sverdlov, O., Rosenberger, W.F., 2013. On recent advances in optimal allocation designs in clinical trials. J. Stat. Theory Pract. 7 (4), 753–773.

Thall, P.F., Simon, R., 1990. Incorporating historical control data in planning phase II clinical trials. Stat. Med. 9 (3), 215–228.

van Rosmalen, J., Dejardin, D., van Norden, Y., Löwenberg, B., Lesaffre, E., 2017. Including historical data in the analysis of clinical trials: Is it worth the effort?. Stat. Methods Med. Res..