Contemporary Clinical Trials 82 (2019) 53–59

Contents lists available at ScienceDirect

Contemporary Clinical Trials

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

Testing monotherapy and combination therapy in one trial with biomarker consideration

Linda Z. Sun*, S. Peter Kang, Cong Chen

MRL, Merck & Co., Inc., Kenilworth, NJ, USA

ARTICLE INFO

Keywords:
Biomarker stratified design
Multi-arm trials
Shared control
FWER
Log-rank statistic

ABSTRACT

It is a common scenario that an experimental oncology therapy, as a monotherapy, may be more effective than standard of care (SOC) in a biomarker positive population but less so or even inferior to SOC in biomarker negative population. At the same time, due to synergistic or additive effect, the combination of the two may be more effective than SOC alone in the all-comer population. The conventional development paradigm is to conduct two separate Phase III trials, one with the monotherapy versus SOC in the biomarker positive population, and the other with the combination therapy versus SOC in the all-comer population. In this manuscript, we propose a one-trial design that stratifies by biomarker status and randomizes biomarker positive patients into three arms (combination therapy, monotherapy, and SOC) and biomarker negative patients into two arms (combination therapy and SOC). There are two hypotheses in the proposed design and each addresses a different question. The family-wise type-I error rate (FWER) is smaller, due to shared control, than that of two separate trials. Therefore, no FWER adjustment is necessary in the proposed design and each hypothesis can be tested at the conventional 2.5% (one-sided) alpha level. The population for comparison between the combination therapy and SOC is skewed in the proposed design. A two-step log-rank statistic is proposed to account for the skewness. Power and sample size of the proposed design are evaluated in comparison with the two-trial paradigm. The proposed design is more efficient.

1. Introduction

We are in an era of oncology drug development with fast advancement in two-dimensions: increasing number of new agents to be tested in clinical trials and increasing number of biomarker-classified populations. These expanding two dimensions and the interaction of them bring hope, opportunities, but also challenges. We are called to design efficient clinical trials to save time and resource, and ultimately to bring forth live-saving medicine to patients in an expeditious manner [1–4]. For the challenge of increasing number of new treatments to be tested, multi-arm trial design which compares several new treatments with a common control arm is simpler, faster, and cheaper [5], and has drawn more and more attention. For example, efficient utilization of common control is emphasized in FDA draft guidance about master protocols to expedite oncology drug development [6]. For the other challenge of increasing number of biomarker-defined patient populations, conventional design of enrolling all-comer population may not be efficient or appropriate. Inclusion of the biomarker negative patients may dilute the treatment effect and reduce the power of the study [7]. Sometimes, an all-comer trial may not even be ethical, if it is expected that the treatment activity cannot measure up to SOC in biomarker negative population based on existing knowledge and data. On the other hand, the high screen failure rate in the enrichment design [8,9], which only enrolls biomarker positive patients, may dampen the enrollment enthusiasm for the trial. Therefore, balancing the appropriate populations (enriched vs. all-comer) is key for an efficient design.

It is a common scenario in oncology that a new targeted therapy or immunotherapy may be more effective than SOC in a biomarker positive population but less so or even inferior to SOC in the biomarker negative population. However, due to synergistic or additive effect between the experimental therapy and the SOC, combination of the two may be more effective than SOC alone in the all-comer population regardless of biomarker status. One example is pembrolizumab monotherapy and its combination with chemotherapy for the first-line treatment of metastatic Non-Small Cell Lung Cancer (NSCLC). Monotherapy pembrolizumab is demonstrated to be superior to SOC chemotherapy in biomarker enriched population, defined as PD-L1 expression tumor proportional score (TPS) ≥ 50%, in Keynote-024 [10]. Pembrolizumab in combination with SOC is demonstrated to be superior to SOC in all-comer NSCLC in Keynote-189 [11]. Keynote-024

* Corresponding author at: 4467 Philbrook Sq, San Diego, CA 92130, USA.
E-mail address: linda_sun@merck.com (L.Z. Sun).

https://doi.org/10.1016/j.cct.2019.06.002
Received 8 March 2019; Received in revised form 1 June 2019
Available online 12 June 2019

and Keynote-189 were conducted sequentially because the proof-of-concept of the combination therapy results were not available yet when Keynote-024 started. However, for many follow-on tumor indications, the monotherapy and combination therapy don't have to be tested sequentially since the proof-of-concept has established. With the development of pembrolizumab in NSCLC as a motivating example, in this manuscript, we propose an efficient design which tests monotherapy and combination therapy in one trial with biomarker consideration. By combining the two trials into one, it not only saves resource and addresses enrollment challenges, but also brings the benefit of contemporaneous results of the monotherapy and the combination therapy to patients, physicians, regulators, and policy makers.

2. Method

2.1. Study design

The primary objectives of the proposed study are (1) to compare overall survival (OS) between monotherapy of experimental drug (E) and the standard of care (SOC) in biomarker positive (BMX+) patients; (2) to compare OS between the combination therapy E + SOC and SOC alone in all-comer population. Let H1 (monotherapy in BMX+) and H2 (combination therapy in all-comers) be the corresponding hypotheses of these two objectives. The study is considered to have met its objective if either null hypothesis H1 or H2 is rejected. In this randomized study, eligible patients will be stratified by biomarker status. Biomarker positive patients will be randomized in a 1:1:1 ratio to receive monotherapy E, combination therapy E + SOC, or SOC. Biomarker negative patients will be randomized in a 1:1 ratio to receive combination therapy E + SOC or SOC. Fig. 1 depicts the study design. Throughout this manuscript, we assume that the biomarker subpopulations are well-defined and pre-specified.

2.2. Analysis method

To test H1 (monotherapy vs. SOC in BMX+ population), we can use the regular log-rank test. The treatment effect (hazard ratio of E vs. SOC) can be estimated by regular Cox regression model.

To test H2 (combination therapy vs. SOC in all-comer population), the regular log-rank test by including patients assigned to the combination therapy arm and the SOC arm may not be appropriate, since these patients do not represent the true all-comer population. The prevalence of biomarker positive patients is lower in this population than the natural prevalence, because only two-thirds of the biomarker positive patients are assigned to combination therapy and SOC, while all biomarker negative patients are assigned to combination therapy and SOC in the proposed design.

To address this challenge, a two-step log-rank statistic is proposed.

- Step 1: Let W_1 be the log-rank statistic from the biomarker positive patients who are assigned to E + SOC or SOC, and W_2 from the biomarker negative patients.
- Step 2: Define

$$W = \frac{3}{2}W_1 + W_2$$

Because only $\frac{2}{3}$ of the biomarker positive patients from the true all-comer population are included in the combination therapy vs. SOC comparison, by giving $\frac{3}{2}$ weight to the biomarker positive statistic in W, it mimics the natural weighting of biomarker positive and negative in the true all-comer population. This weight is dictated by the randomization ratio upfront and there is no need to estimate it. Specific weight can be derived for non-equal randomization ratio designs.

This two-step log-rank statistic can also be considered as a modified stratified log-rank statistic. The stratified log-rank statistic is simply the sum of W_1 and W_2, while the proposed statistic gives different weighting to W_1 and W_2 to adjust the skewness of population caused by design. For testing H2, Chi-square test is applied to W and its variance, which can be estimated from estimates of W_1 and W_2 and their variances by noting that W_1 and W_2 are independent.

Similarly, a two-step approach can be used to estimate hazard ratio of E + SOC vs. SOC in all- comer population for the proposed design. Let θ be the hazard ratio, and $\beta = \log(\theta)$.

- Step 1: let $\hat{\beta}_1$ and $\hat{\beta}_2$ be the estimated treatment effect in biomarker positive patients and biomarker negative patients, respectively.
- Step 2: Define $\hat{\beta}$ to be the estimated treatment effect for all-comer population

$$\hat{\beta} = w_1\hat{\beta}_1 + w_2\hat{\beta}_2 \quad (1)$$

where $w_1 + w_2 = 1$. In the proposed design, because of the skewness of the population, we again propose to weigh each event in biomarker

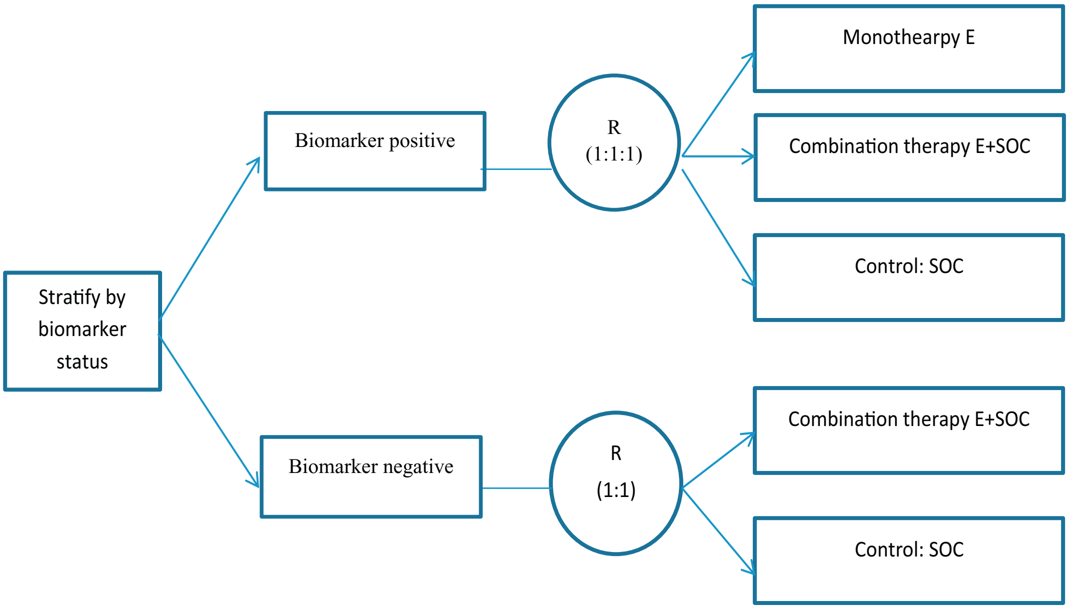

Fig. 1. Proposed study design.

positive group in this analysis by $\frac{3}{2}$ to recover the contribution of the biomarker positive and negative as the contribution in the true all-comer population. Let d_1 and d_2 be the number of events in the biomarker positive and negative patients respectively for estimating the combination therapy treatment effect in the proposed design,

then let.

$w_1 = \frac{\frac{3}{2}d_1}{\frac{3}{2}d_1 + d_2}$ and $w_2 = 1 - w_1$.with w_1, w_2, mean and variance of $\widehat{\beta}_1$ and $\widehat{\beta}_2$, the mean and variance of $\widehat{\beta}$ can be estimated. So can be the hazard ratio $\theta = \exp(\beta)$.

3. Results

We compare the proposed design with the two-trial development paradigm in terms of sample size and the number of patients needed for screening when fixing the power through analytical illustration and hypothetical examples. The two-trial development paradigm is to conduct two separate Phase III clinical trials, one with the monotherapy versus SOC in the biomarker positive population (biomarker enriched trial) and the other with the combination therapy versus SOC in the all-comer population.

Multiplicity consideration associated with multi-arm Phase III trials is a relatively new area and regulatory guidance is yet to be developed [12]. Is it necessary to adjust for multiple comparison (experimental arms vs. control) in the multi-arm single trial when no such adjustment would be required had these comparisons been made in separate two-arm trials? Howard et al. [13] recently provides systematic recommendations and a decision diagram. If several experimental treatments share a control arm and thus are tested in a multi-arm trial to improve efficiency and the trial is focused on answering the efficacy question for each treatment separately, the hypotheses do not inform a single claim of effectiveness. In such case, from mathematical point of view, it is shown that the familywise error rate (i.e. at least one Type-I error rate) is always lower in multi-arm single trial sharing a common control than separate trials [14]. Therefore, no multiplicity adjustment is necessary to the multi-arm trial.

In the proposed design, hypotheses H1 and H2 are two separate research questions. They are tested in one trial to efficiently utilize (1) the common control and (2) patient resource since it includes both biomarker positive and negative patients. The outcome of one hypothesis does not necessarily imply the outcome of the other hypothesis. Therefore, multiplicity adjustment is not necessary in the proposed design [13,15]. That is, the alphas for H1 and H2 in the proposed design can be the same alphas for the two trials in the two-study development paradigm.

3.1. Sample size comparison through analytical illustration

Let X be a test statistic for a hypothesis, the general sample size formula is

$$(Z_{1-\alpha} + Z_{1-\beta})^2 = E(X)^2/var(X) \tag{2}$$

where α is the one-sided type I error, *1-β* is the power, *E(X)* and *var(X)* are the mean and variance of *X* for the target treatment effect, respectively. Since *var(X)* is a function of sample size, for given α, β, and target treatment effect, the sample size can be solved from Eq. (2).

In the proposed design, patients are randomized according to Fig. 1. In each of study of the two-trial design, patients are randomized in a 1:1 ratio to the experimental arm and control arm.

For the monotherapy vs. SOC comparison in biomarker positive population, both the proposed design and the two-trial design use log-rank as the test statistic, and thus the required sample size for testing H1 is the same for the two designs.

For the combination therapy vs. SOC comparison in all-comer population, we use log-rank test statistic in the two-study design. If *HR* is the target hazard ratio of combination therapy E + SOC vs. SOC, then according to Eq. (2) and the mean and variance of log-rank statistic, we have the following relationship

$$(Z_{1-\alpha} + Z_{1-\beta})^2 = \log(HR)^2 \cdot D/4 \tag{3}$$

where D is the number of events required for the all-comer trial to test combination therapy in the two-trial development paradigm.

In the proposed design, let D' be the total number of events in the whole trial, including the monotherapy arm, the combination therapy arm and the control SOC arm. Let p be the patient prevalence of biomarker positive. For ease of presentation, we assume that the biomarker is not prognostic and the proportion of events from biomarker positive patients out of all-comer patients is approximately the same as prevalence (p). (Sample size calculation under more general assumptions is provided in the Appendix). Therefore, there are approximately $2/3 \cdot p \cdot D'$ events in biomarker positive patients for the combination vs. SOC comparison, and $(1-p)D'$ events in biomarker negative patients. Then the total number of events to test H2 is approximately $(1 - 1/3 \cdot p)D'$.

As introduced in Section 2.2, W_1 and W_2 are log-rank statistics from the biomarker positive and negative patients, respectively, who are assigned to E + SOC or SOC in the proposed design. As $W = \frac{3}{2}W_1 + W_2$ is the test statistic for testing H2 in the proposed design, we have

$$E(W) = \log(HR) \bullet D'/4$$

$$Var(W) = \left(1 + \frac{1}{2}p\right) \bullet D'/4$$

According Eq. (2),

$$(Z_{1-\alpha} + Z_{1-\beta})^2 = \log(HR)^2 \cdot \left(\frac{D'}{4}\right) / \left(1 + \frac{1}{2}p\right) \tag{4}$$

From Eq. (3) and Eq. (4), we arrive at the following relationship between D' and D:

$$D' = \left(1 + \frac{1}{2}p\right)D \tag{5}$$

Therefore, the number of events needed to test H2 with the two-step test statistics W is

$$\left(1 - \frac{1}{3}p\right)D' = \left(1 + \frac{1}{6}p - \frac{1}{6}p^2\right)D > D \tag{6}$$

This shows that the number of events needed with the two-step log-rank test statistic W in the proposed design is more than what is needed in the all-comer study of the two-trial design by $\left(\frac{1}{6}p - \frac{1}{6}p^2\right)D$, because two-step log-rank test statistic has a larger variance (or less efficient as expected) than the regular log-rank statistic.

On the other hand, the proposed design can reduce the exposure of biomarker positive patients to the control arm. The number of events saved is

$$\frac{1}{3}pD' = \left(\frac{1}{3}p + \frac{1}{6}p^2\right)D \tag{7}$$

The net reduction of events is

$$\left(\frac{1}{3}p + \frac{1}{6}p^2\right)D - \left(\frac{1}{6}p - \frac{1}{6}p^2\right)D = \left(\frac{1}{6} \bullet p + \frac{1}{3} \bullet p^2\right)D > 0 \tag{8}$$

Therefore, as expected, the proposed design (one-trial design) requires fewer events than the total number of events required from the two-trial design.

Further, according to Eq. (7), there are approximately $\frac{1}{3}pD'$ events from the biomarker positive patients on the control SOC arm which contribute to both H1 and H2. Let Z_+ be the standardized log-rank test statistics for H1 and D_+ be the number of events to test H1. Let Z_w be the standardized two-step log-rank test statistics for H2. Then the correlation between Z_w and Z_+ is

$$\rho = \frac{\frac{3}{2} \bullet \frac{1}{3} pD'}{\sqrt{\left(1 + \frac{1}{2}p\right)D'} \bullet \sqrt{D_{+}}} \tag{9}$$

Therefore, the alpha levels for H1 and H2 can be > 0.025 to control FWER at the same level as the two-trial design since Z_w and Z_+ follow a bivariate normal distribution with correlation ρ. This implies that, as needed, the sample size of the proposed design can be further reduced when incorporating the correlation of the two test statistics for H1 and H2.

3.2. Sample size comparison through hypothetical examples

Now we use a hypothetical example to compare the sample size and number of patients screened for the one-trial design (the proposed design) vs. the two-trial design. In this example, the prevalence of biomarker positive is 33%. The target treatment effect of monotherapy E vs. SOC is hazard ratio HR = 0.65 in the biomarker positive population. The target treatment effect of combination therapy E + SOC vs. SOC is HR = 0.7 in all-comer population. For each hypothesis, one-sided α is 0.025 and power is 90%.

For the two-trial design, the monotherapy trial (Trial 1) needs 227 events from biomarker positive patients, and the combination therapy trial (Trial 2) needs 330 events from all-comer patients. If we assume that 70% randomized patients have events by the time of final analysis, Trial 1 needs 326 biomarker positive patients randomized (approximately 652 biomarker negative patients would be screen failure), and Trial 2 needs 472 all-comer patients. In total, the sample size for the two-trial design is 798. Strictly speaking, the 652 biomarker negative patients screened for Trial 1 cannot participate in Trial 2 even if they meet all other inclusion/exclusion criterion (Table 1).

For simplicity, we do not consider the correlation of the test statistics for H1 and H2 in the one-trial design (proposed design). That is, the alpha level (one-sided) is 0.025 for each of H1 and H2. To test H1, 326 biomarker positive patients are needed, as in the two-trial design. According to Eq. (6) in Section 3.1, to test H2, the proposed design needs 342 events from 489 patients. From Eq. (7), 61 biomarker positive patients on the control arm contribute to both H1 and H2. Therefore, the total sample size of the proposed design is 326 + 489–61 = 754. In practice, Eq. (5) shows that the one-trial design will first randomize 550 all-comers according to Fig. 1, and then enroll additional 204 biomarker positive patients, randomized to monotherapy or SOC to have 90% power for H1. In terms of screening, for this enrichment part of the trial, the 408 biomarker negative patients who are screen failures cannot participate even if they meet all other inclusion/exclusion criterion.

The sample size comparison of this example is summarized in Table 1. The one-trial design uses fewer patients than the two-trial design (754 vs. 798). Besides, the one-trial design enrolls about 250 biomarker negative patients (652–408) who would not have been eligible for Trial 1 in the two-trial design even if they meet other non-biomarker-status related enrollment criterion.

Based on the hypothetical example, we now explore the impact of biomarker prevalence on the sample size of the proposed design. We keep all other design parameters the same (e.g. target $HR_1 = 0.65$ for H1 and $HR_2 = 0.7$ for H2) but assume different prevalence (p) of the biomarker positive population. The proposed design will start with all-comer enrollment according to Fig. 1. D' is the total number of events from this part of the enrollment. D' is determined according to Eq. (5) and Eq. (3) so that the number of events contributing to test H2 provides the desired power. Among D' number of events, the number of events contributing to H1 is

$$\frac{2}{3}pD' = \frac{2}{3}p\left(1 + \frac{1}{2}p\right)(Z_{1-\alpha} + Z_{1-\beta})^2/\log(HR_2)^2 \cdot 4$$

In Section 3.1, D_+ is denoted as the number of events required to test H1. According to Eq. (3),

$$D_+ = (Z_{1-\alpha} + Z_{1-\beta})^2/\log(HR_1)^2 \cdot 4$$

Compare $\frac{2}{3}pD'$ with D_+: If $\frac{2}{3}pD' < D_+$, it means the number of events for H1 from the all-comer enrollment of the proposed design is not sufficient, and additional biomarker positive patients need to be enrolled. The total number of events (T) is the events from the all-comer enrollment plus the events from the additional biomarker positive enrollment,

$$T = \left(1 - \frac{2}{3}p\right)\left(1 + \frac{1}{2}p\right)\frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{\log(HR_2)^2} \cdot 4 + \frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{\log(HR_1)^2} \cdot 4 \tag{10}$$

If $\frac{2}{3}pD' \geq D_+$, it means the biomarker positive patients enrolled during the all-comer enrollment is more than the required number for testing H1, i.e., H1 is overpowered and the sample size is driven by H2. In this scenario, the total number of events of the proposed design is that from the all-comer enrollment:

$$T = \left(1 + \frac{1}{2}p\right)\frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{\log(HR_2)^2} \cdot 4 \tag{11}$$

Fig. 2 shows the sample size change with different prevalence of the biomarker positive population. The sample size is always smaller under the one-trial design than the two-trial design.

In general, the saving of sample size increases with the prevalence as more patients on the control arm can contribute to both H1 and H2. Eq. (10) shows that the sample size of the one-trial design decreases with p increases. However, when the prevalence is higher than a certain cut-point ($p > 75\%$ in this example), as explained earlier, the sample size of the one-trial design is driven by H2 and overpowered for H1. In such case, Eq. (11) shows that the sample size of the one-trial design increases as p increases.

We further explore the impact of treatment effect of the monotherapy and combination therapy on the sample size of the proposed design, which is also confounded with the biomarker prevalence. The treatment effect of the monotherapy is kept the same as in the hypothetical example (i.e. target $HR_1 = 0.65$ for the monotherapy vs. SOC in biomarker positive population). When $\frac{2}{3}pD' < D_+$,Eq. (8) and Eq. (3) show that the saving in number of events of the one-trial design from the two-trial design is

$$\left(\frac{1}{6} \bullet p + \frac{1}{3} \bullet p^2\right)\frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{\log(HR_2)^2} \cdot 4$$

Table 1
Comparison of number of patients randomized and screened of one-trial design and two-trial design.

	One-trial design (Proposed Design)	Two-trial design
Sample size for H1 (monotherapy vs. SOC)	326	326
Sample size for H2 (combination therapy vs. SOC	489	472
Sample size saved due to shared control arm for both H1 and H2	(−)61	0
Total sample size	**754**	**798**
Number of screened patients who cannot be enrolled solely because of being biomarker negative	408	652

Biomarker prevalence p = 33%, target HR monotherapy = 0.65, HR combination = 0.70, one-sided alpha = 0.025, power = 90%, 70% randomized patients have events by the time of final analysis.

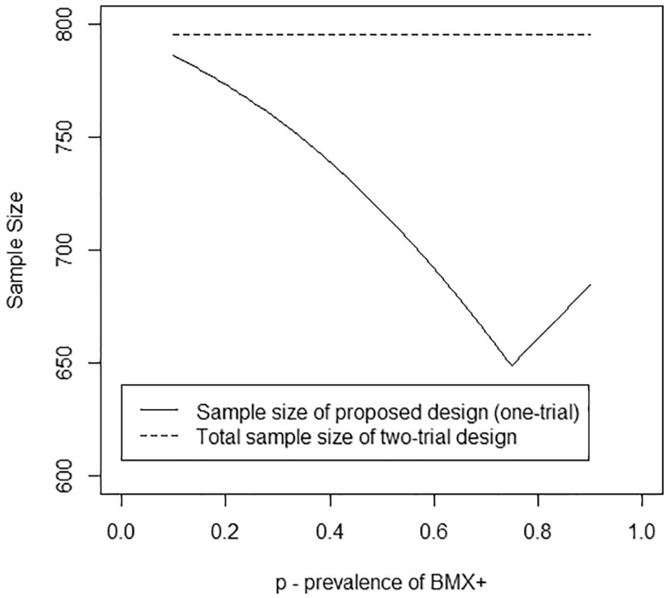

Fig. 2. Sample Size for Proposed Design with Different Prevalence of Biomarker Positive Population.

It indicates that the saving increases as the target treatment effect of the combination therapy decreases (or HR_2 increases). As expected, when the target treatment effect of the combination therapy decreases, the study requires a bigger all-comer sample size for H2. As a result, more patients on the control arm can be shared by H1 and H2 in the proposed design, leading to more savings. However, when $\frac{2}{3}pD' \geq D_+$, Eq. (5) and Eq. (3) show that the percentage of saving is

$$1-\left(1+\frac{1}{2}p\right)\frac{\log(HR_1)^2}{\log(HR_2)^2+\log(HR_1)^2}$$

In this scenario, the saving decreases as the target HR_2 of the combination therapy increases. Fig. 3 shows the percentage sample size savings under different target HRs of the combination therapy vs. SOC in all-comer population. When the biomarker positive prevalence is 33% and the HR_2 is between 0.7 and 0.8, which is the range of treatment effect of interest, the proposed design saves more sample size as the target treatment effect of the combination therapy decreases. When the biomarker positive prevalence is 50%, the saving of the proposed design first increases and then decreases as the target HR_2 of the combination therapy increases.

In summary, the use of a common control arm in the proposed design for testing H1 and H2 can save number of patients randomized. The extent of the saving depends on the biomarker prevalence and the target treatment effects for monotherapy and combination therapy. The proposed design saves the most when the events contributing to test H1 from the all-comer enrollment is the same as the number of events required to test H1 in Trial 1 of the two-trial design. In addition to savings in number of patients randomized, there is no waste of patients screened during the all-comer enrollment period of the proposed design since the biomarker negative patients can still be randomized in the study, while in the two-trial design the screen failures due to biomarker status for the biomarker positive trial (Trial 1) cannot be randomized to Trial 2.

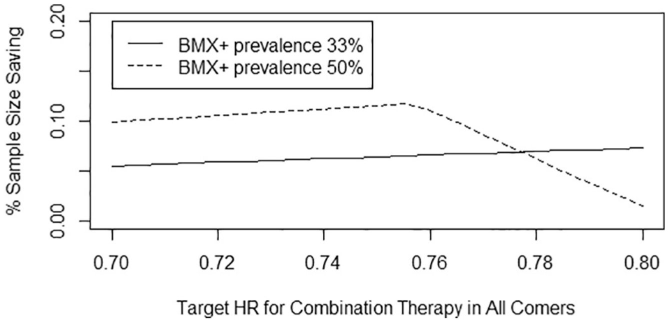

Fig. 3. Sample Size Saving (%) for Proposed Design Compared to the Two-Trial Design for Different Target Treatment Effect (HR) of the Combination Therapy.

4. Discussion

When monotherapy and combination therapy of an experimental drug are both of interest and the monotherapy may be effective only in biomarker enriched population, we propose a multi-arm design with stratification of biomarker to test the monotherapy and combination therapy in their respective target populations in a single trial. Compared to the conventional two-Phase-III-trial design, the proposed design not only saves time, number of patients randomized and screened, but also address some of the potential logistic challenges with the two-trial design, which sometimes may be serious enough to make the trial results uninterpretable. For example, in the two-trial paradigm, patients screened in the enrichment study for testing monotherapy cannot participate in the trial if they are in the biomarker negative group. If the all-comer study for testing the combination therapy is screening at the same time, these biomarker negative patients who screened for the monotherapy trial cannot be automatically considered for the combination therapy trials either. Otherwise the screening for the combination therapy trial would not be true all-comer population. In addition, if these trials screen and enroll at the same time at same sites, it begs a logistic question of which trial to participate for a biomarker positive patient. Investigators may suggest a biomarker positive patient to participate in the monotherapy trial or the combination trial based on some criteria outside of the inclusion/exclusion criteria of the protocols. For example, since the monotherapy is supposed to be more tolerable than the combination therapy, investigators may tend to enroll patients with ECOG performance status 1 and high disease burden to the monotherapy trial, while enrolling patients with better ECOG performance status and lower disease burden to the combination trial. This may cause difference in the biomarker positive population between the monotherapy trial and the combination therapy trial. If monotherapy trial doesn't demonstrate improvement from the standard of care, we cannot tell whether it's due to lack of efficacy or due to population selection bias.

For practical considerations of the proposed design, as alluded in the hypothetical example in Section 3.2, the biomarker positive patients randomized to the monotherapy and SOC arms during all-comer enrollment of the proposed design may be less than the sample size required to achieve the power for H1. In such case, the proposed design will stop enrollment of all-comers when the sample size required for H2 is met, and only to enroll biomarker positive patients until the required sample size for H1 is met. On the other hand, the biomarker positive patients randomized to the monotherapy and SOC arms during all-comer enrollment of the proposed design may be more than the sample size required to achieve the power of H1. In this case, the power for H1 in the proposed design may be higher than the biomarker enriched trial in the two-study-design. Regardless of which of the above two scenarios happens, in practice it's likely that the timing of the final analyses to test H1 and H2 are different. If the projected timing is close to each other, waiting until the later timing to conduct one final analysis for both H1 and H2 is a viable option. If the projected timing is far apart from each other, for example H2 is earlier than H1, one option is to conduct an interim analysis for H1 at the final analysis for H2. The implication of one hypothesis being positive earlier than the other should be considered in practice as well. If the combination therapy has results earlier than the monotherapy, and at that time the biomarker positive enrollment for H1 has not finished or just completed, it may bring challenge to the conduct of the trial after the combination therapy results. When designing the trial, if such scenario is expected to happen, it may be better to start the enrollment with biomarker positive patients only. After certain number of biomarker positive patients are enrolled, then open the enrollment to all-comers and randomize according to Fig. 1. When the proposed design is ongoing, monitoring and projecting number of events for H1 and H2 is a little more complicated than two-trial design due to the multi-arm feature. To execute the trial, the exact number of events and projection is usually unnecessary, so the study team can still monitor and project number of events approximately in a blinded fashion without having to set up an unblinded team to do so.

For the one-trial design, multiplicity adjustment is not necessary since H1 and H2 do not inform a single claim of effectiveness [8]. However, since both the monotherapy and the combination therapy have the component of the experimental drug E, the effectiveness claims from H1 and H2 may be considered as somewhat related. If H1 and H2 are philosophically considered as informing a single claim of effectiveness (e.g. whether E is effective at all), the sample size comparison of the one-trial vs. two-trial design is the following: 1) If the single claim of effectiveness can be made when either H1 or H2 is positive, multiplicity adjustment for FWER may be necessary. Such adjustment (e.g. Bonferroni adjustment with one-sided 1.25% alpha for each hypothesis) applies to both one-trial design and two-trial design. Therefore, the sample size of the one-trial design is still smaller than the two-trial design under this scenario; 2) If the single claim of effectiveness needs both H1 and H2 to be positive, the chance of both H1 and H2 are falsely positive is inflated in the one-trial design due to the shared control group [8]. The false positive error rate for the two-trial design is 0.000625 if each trial's alpha is 0.025. For the one-trial design, in the hypothetical example, the correlation is about 0.2 according to Eq. (9). If each hypothesis in the one-trial design has 0.025 one-sided alpha, then the false positive error rate is 0.0016 according to the bi-variate normal distribution of Z_w and Z_+. To control the false positive error rate to be the same as that of the two-trial design (i.e. 0.000625), the alpha level for each hypothesis needs to be adjusted to 0.0142. Such adjustment will require a sample size of 863 for the one-trial design, which is larger than the total sample size of the two-trial design 798. Therefore, if effective claim needs both H1 and H2 to be positive, the two-trial design needs a smaller sample size than the proposed one-trial design. However, in oncology drug development, usually it does not require both H1 and H2 to be positive to claim effectiveness.

The proposed design can be extended easily to more complex Phase III trial design, e.g. group sequential design with interim analyses or different randomization ratios. Another potential extension of the proposed design is to add population adaptation in the design. The trial may start as an all-comer three-arm trial. Patients are stratified by biomarker status. There is a futility interim analysis for monotherapy in biomarker negative population. If the futility boundary is crossed, biomarker negative patients cannot be randomized to monotherapy arm anymore and will be randomized to combination therapy or control arm only. To test the combination therapy hypothesis H2, we divide the patients enrolled before and after the adaptation. For patients enrolled before adaptation, we use the regulator log-rank statistic. For patients enrolled after adaptation, we use the two-step log-rank statistic. Then we use the sum of these two statistics to test H2. Further, in real drug development scenarios, we may want to apply population adaptation to the combination therapy as well. The idea of the proposed design may be combined with the adaptive expansion of biomarker population design from Chen 2018 [10].

In addition, the idea of the two-step analysis method in this paper which adjusts for the biomarker skewness in a broader population may be applied in umbrella platform designs that consist of sub-studies to evaluate drugs targeting multiple biomarkers [6]. In such trials, one key design consideration is how patients with more than one biomarker of interest will be assigned to sub-studies. No matter which randomization method is used for patient allocation, the biomarker composition of the sub-studies is likely to be different from the natural prevalence. A similar two-step approach with weights determined by the pre-specified allocation rules may be considered for data analysis.

Cancer drug development is a fast-growing and complex area. The proposed design in this article provides a new idea to meet some of the challenges and the call to be more efficient in clinical trial design. While the proposed design is conceptually appealing, and the benefit is demonstrated in the hypothetical examples, practical issues need to be considered and worked out when implementing it in the real world.

Acknowledgement

The design idea was built on thoughts from Drs. Jon Cheng, Chris Gause, Joy Ge, Keaven Anderson from Merck & Co., Inc., Kenilworth, NJ USA, and Dr. Minori Koshiji from MacroGenics, Inc.

Appendix A. Appendix

A.1. Sample size calculation of the proposed design

In Section 3.1, for ease of presentation, we assume that the biomarker is not prognostic and the proportion of events from biomarker positive patients out of all-comer patients is approximately the same as prevalence (p). This assumes that the event rate in biomarker positive and negative patients are the same. In this Appendix, we present the sample size calculation of the proposed design under more general assumptions.

Let p be the prevalence of biomarker positive populations, r_+ be the event rate at the time of analysis for biomarker positive patients who are randomized to the control arm and combination arm, and r_- be the event rate for biomarker negative patients. Let N' be the total number of patients in the proposed design, and N be the number of patients required for the all-comer trial to test combination therapy in the two-trial development paradigm.

In the proposed design, the number of events on the control arm and combination arm is $\frac{2}{3}pN'r_+$ for biomarker positive patients, $(1-p)N'r_-$ for biomarker negative patients, and total number of events is $N'\left(\frac{2}{3}pr_+ + (1-p)r_-\right)$. The total number of events in the combination therapy trial in the two-trial design is $D = N(pr_+ + (1-p)r_-)$.

As $W = \frac{3}{2}W_1 + W_2$ is the test statistic for testing H2 in the proposed design, we have

$$E(W) = \log(HR)\bullet\frac{\left(\frac{3}{2}\bullet\frac{2}{3}pr_+ + (1-p)r_-\right)N'}{4} = \log(HR)\bullet(pr_+ + (1-p)r_-)N'/4$$

$$Var(W) = \left(\frac{9}{4}\bullet\frac{2}{3}pr_+ + (1-p)r_-\right)N'\frac{1}{4} = \left(\frac{3}{2}pr_+ + (1-p)r_-\right)N'/4$$

According to Eq. (2) and Eq. (3),

$$\frac{E(W)^2}{var(W)} = \log(HR)^2 \cdot D/4$$

We get,

$$N' = \frac{\frac{3}{2}pr_+ + (1-p)r_-}{(pr_+ + (1-p)r_-)}N$$

The sample size required to test H2 in the proposed design is $\left(1-\frac{1}{3}p\right)N'$. The sample size needed with the two-step log-rank test statistic W in the proposed design is more than what is needed in the all-comer study of the two-trial design by

$$\left(1-\frac{1}{3}p\right)N' - N = \frac{\frac{1}{2}p(1-p)r_+ - \frac{1}{3}p(1-p)r_-}{(pr_+ + (1-p)r_-)}N \tag{12}$$

On the other hand, the sample size saved with the proposed design due to shared control is

$$\frac{1}{3}pN' = \frac{\frac{1}{2}p^2r_+ + \frac{1}{3}p(1-p)r_-}{(pr_+ + (1-p)r_-)}N \tag{13}$$

Therefore, the net sample size (number of patients) saving of the proposed design is Eq. (13)–Eq. (12)

$$\frac{p\left(\left(p-\frac{1}{2}\right)r_+ + \frac{2}{3}(1-p)r_-\right)}{(pr_+ + (1-p)r_-)}N \tag{14}$$

It's easy to see that the conditions for Eq. (14) to be positive, i.e. the proposed design saves sample size are

$$p > \frac{1}{2}$$

Or

$$p < \frac{1}{2} \text{ and } \frac{r_-}{r_+} > \frac{\frac{1}{2}-p}{\frac{2}{3}(1-p)} = f(p)$$

That is, for $p > 50\%$, the proposed design always saves sample size compared to traditional design. For $p < 50\%$, if the event rate ratio of biomarker negative and positive patients is above the line *f*(*p*), which is always < 0.75, then the proposed design will save sample size. In practice, most likely the event rate ratio is above *f*(*p*), and thus the proposed design is more efficient than traditional design in terms of sample size.

References

[1] J. Woodcock, L.M. LaVange, Master protocols to study multiple therapies, multiple diseases, or both, N. Engl. J. Med. 377 (1) (2017) 62–70.

[2] D.A. Berry, The brave new world of clinical cancer research: adaptive biomarker-driven trials integrating clinical practice with clinical research, Mol. Oncol. 9 (2015) 951–959.

[3] A. Hirakawa, J. Asano, H. Sato, et al., Master protocol trials in oncology: review and new trial designs, Contemp. Clin. Trials Commun. 12 (2018) 1–8.

[4] L.A. Renfro, D.J. Sargeant, Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols, a review and examples, Ann. Oncol. 28 (2017) 34–43.

[5] M.K. Parmar, J. Carpenter, M.R. Sydes, More multiarm randomised trials of superiority are needed, Lancet 384 (2014) 283–284.

[6] F.D.A. draft guidance, Master Protocols: Efficient Clinical Trial Design Strategies to Expedite Development of Oncology Drugs and Biologics Guidance for Industry, (2018).

[7] C. Chen, X. Li, W. Li, R.A. Beckman, Adaptive expansion of biomarker populations in phase 3 clinical trials, Contemp. Clin. Trials 71 (2018) 181–185.

[8] FDA draft guidance, Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products, (2012).

[9] S.J. Wang, J. Hung, Adaptive enrichment with subpopulation selection at interim: methodologies, applications and design considerations, Contemp. Clin. Trials 36 (2013) 673–681.

[10] M. Reck, et al., Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer, N. Engl. J. Med. 375 (2016) 1823–1833.

[11] L. Gandhi, et al., Pembrolizumab plus chemotherapy in metastatic non-small-cell lung cancer, N. Engl. J. Med. 378 (2018) 2078–2092.

[12] J.M. Wason, L. Stecher, A.P. Mander, Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? Trials 15 (2014) 364.

[13] D.R. Howard, J.M. Brown, S. Todd, W.M. Gregory, Recommendations on multiple testing adjustment in multi-arm trials with a shared control group, Stat. Methods Med. Res. 27 (2018) 1513–1530.

[14] M. Proschan, D. Follman, Multiple comparisons with control in a single experiment versus separate experiments: why do we feel differently? Am. Stat. 49 (1995) 144.

[15] B. Freidlin, E.L. Korn, R. Gray, et al., Multi-arm clinical trials of new agents: some design considerations, Clin. Cancer Res. 14 (2008) 4368–4371.