**RESEARCH ARTICLE**                                                                    **Open Access**

# Including non-concurrent control patients in the analysis of platform trials: is it worth it?

Kim May Lee[1]* and James Wason[1,2]

## Abstract

**Background:** Platform trials allow adding new experimental treatments to an on-going trial. This feature is attractive to practitioners due to improved efficiency. Nevertheless, the operating characteristics of a trial that adds arms have not been well-studied. One controversy is whether just the concurrent control data (i.e. of patients who are recruited after a new arm is added) should be used in the analysis of the newly added treatment(s), or all control data (i.e. non-concurrent and concurrent).

**Methods:** We investigate the benefits and drawbacks of using non-concurrent control data within a two-stage setting. We perform simulation studies to explore the impact of a linear and a step trend on the inference of the trial. We compare several analysis approaches when one includes all the control data or only concurrent control data in the analysis of the newly added treatment.

**Results:** When there is a positive trend and all the control data are used, the marginal power of rejecting the corresponding hypothesis and the type one error rate can be higher than the nominal value. A model-based approach adjusting for a stage effect is equivalent to using concurrent control data; an adjustment with a linear term may not guarantee valid inference when there is a non-linear trend.

**Conclusions:** If strict error rate control is required then non-concurrent control data should not be used; otherwise it may be beneficial if the trend is sufficiently small. On the other hand, the root mean squared error of the estimated treatment effect can be improved through using non-concurrent control data.

**Keywords:** Adding arms, Concurrent/ non-concurrent control, Platform trials

## Background

The implementation of innovative trial designs for drug development programmes has increased in recent years [1]. A platform trial is one such innovative design that helps improve efficiency. It has a multi-arm multi-stage design where a single control group is utilized to evaluate multiple experimental treatments for a single disease. Apart from the feature of dropping arms, a platform trial allows adding arms when new experimental treatments become available. This adds efficiency as adding arms takes less time and has a lower cost than setting up new trials from scratch.

A platform trial can be implemented through the specification of a master protocol [2–4]. Cohen et al [5] review the statistical aspects on the adding arm adaptation and present some trials that have adopted a platform approach. For pairwise comparisons between the effectiveness of newly added treatments and the control treatment, there are different views on utilising the control data [5]. To be more specific, the control data of a platform trial can be separated into non-concurrent and

*Correspondence: kim.lee@mrc-bsu.cam.ac.uk
[1]MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus , CB2 0SR Cambridge, UK
Full list of author information is available at the end of the article

concurrent. The former (latter) corresponds to the control data of the patients who are recruited before (after) a new experimental treatment is added. This definition applies to each newly added arm independently; the control data can be concurrent to a newly added treatment, and be non-concurrent to another treatment that is added at a later time point.

To our knowledge, methodology and guidance for adding arms remain limited and the operating characteristics of platform trials have not been well-studied. The literature on other trial adaptations such as dropping arms and changing treatment allocation ratio are more comprehensive [6–11]. In particular, many have explored the presence of time trends when more complex randomization procedures are recommended/ explored [12–21]. A trend may be present in a trial when there is a learning curve among the study personnel; a shift in the population baseline characteristics; and/ or the effect of the control treatment may change due to other reasons (e.g. improvement of the standard of care/ practice) [7, 22].

In order to minimise the impact caused by the presence of a trend, most researchers advocate the use of concurrent control data when making inference about the newly added treatment(s) [23–25]. Others stipulate that non-concurrent control data may be used when making inference about the newly added arm by adjusting for possible trends [26, 27]. Here we investigate whether using non-concurrent control data is worthwhile in platform trials. Several analysis approaches are considered. Within a two-stage setting where a new treatment is added at the end of stage one [25, 28], we explore the impact of i) the timing of adding a new arm ii) the sample size of the new arm and iii) the magnitude of a linear or a step trend [15] on the inference about the newly added experimental treatment. We provide recommendations about under which conditions non-concurrent control data should be used.

## Method

Consider a two-stage multi-arm trial that initially has $K$ treatments and a control treatment with a total sample size of $N = N_1 + N_2$ where $N_s$ is stage $s$ sample size, $s = 1, 2$. Denote $k = 0$ for a control treatment, $k = 1, ..., K$ for the initial experimental treatments. Let $n_{ks}$ be the sample size of treatment $k$ in stage $s$ with $n_{k1} + n_{k2} = n_k$. At the end of stage one, in particular after $N_1 = \sum_{k=0}^{K} n_{k1}$ patients have been randomized to the initial arms in stage one, a new treatment (denoted by $K + 1$) is added to the on-going trial with $n_{K+1,2}$ patients. This increases the second stage sample size to $N_2 + n_{K+1,2}$ and the overall sample size of the trial to $N + n_{K+1,2}$.

Denote $n_{01}/n_0$ as the timing of adding a new arm. Small $n_{01}/n_0$ indicates the newly available treatment is added to the trial after a small number of patients have been randomized to the initial arms in stage one. We consider the case that all arms finish recruitment simultaneously (although as long as controls continue to be recruited while treatment $K + 1$ is allocated patients, results apply more generally). For $j = 1, ..., N + n_{K+1,2}$, denoting the sequence of patient enrolment to the trial, let:

$$X_j = \sum_{k=0}^{K+1} \mu_k \cdot I(T_j = k) + \tau(j) + \epsilon_j \tag{1}$$

be the response of treatment $k$ of subject $j$; $\mu_k$ be the true effect of treatment $k$; $\tau(j)$ be a trend that could be a function of the patient ordering $j$; $\epsilon_j$ be the random errors that are identically and independently normally distributed with mean 0 and variance $\sigma^2$; and $T_j$ be the allocated treatment of subject $j$, with $I(\cdot)$ is an indicator function.

Let $\bar{X}_{k,s}$ and $\bar{X}_k$ denote the stage $s$ sample mean response of treatment $k$ and the overall mean response of treatment $k$. When there is no trend in the trial, i.e. $\tau(j) = 0(j)$, the sample mean estimators $\bar{X}_{k,s} \overset{iid}{\sim} N(\mu_k, \sigma^2/n_{ks})$ and $\bar{X}_k \overset{iid}{\sim} N(\mu_k, \sigma^2/n_k)$.

For illustration purpose, we consider the following when exploring the impact of a fixed trend on the inference:

- linear trend:

$$\tau(j) = \lambda (j - 1)/(N + n_{K+1,2} - 1) \tag{2}$$

- stepwise trend:

$$\tau(j) = \lambda \cdot c_j \tag{3}$$

where $c_j$ is an indicator that subject $j$ is enrolled in stage two, and $\lambda$ is a positive fixed value. Both of these trends inflate the value of the outcome variable in a sequential manner; the responses of the patients who were enrolled later are larger than those who were enrolled earlier.

### Simple approach: *Z*-test

We first consider a simple method that is analogous to ignoring the possibility of a trend or stage effect: a standard $Z$-test for testing each hypothesis

$H_{0k} : \mu_k = \mu_0$     against     $H_{1k} : \mu_k > \mu_0$,

using the sample means. When there is no trend, the (pairwise) marginal power of rejecting a null hypothesis to detect a difference in treatment effect, $\delta_k = \mu_k - \mu_0$, is

$$\Phi \left[ \frac{\delta_k}{\sqrt{\sigma^2/n_k + \sigma^2/n_0}} - \Phi^{-1}(1 - \alpha) \right],$$

where $\Phi$ and $\Phi^{-1}$ are the cumulative density function and the inverse of the cumulative density function of a standard normal distribution. For the newly added treatment, the power of rejecting $H_{0,K+1}$ depends on whether all the control data (of size $n_0$) or only the stage two control data

(of size $n_{02}$) is used in the inference. Elm et al [28] describe a $t$-test for the case where $\sigma^2$ is unknown.

## Model-based approach: weighted linear regression

One way to account for the changes of a trial design is to consider a model-based approach. Elm et al [28] consider the following weighted linear regression model. Let $\bar{X}_s^{(k)} = \bar{X}_{k,s} - \bar{X}_{0,s}$ be the stage-wise difference in mean responses; fit a linear model to the samples of the stage-wise mean differences:

$$\bar{X}_s^{(k)} = \delta_k + \eta_s^{(k)}$$

where $\eta_s^{(k)}$ is normally distributed with mean zero and variance $\sigma^2/n_{ks} + \sigma^2/n_{0s}$. The covariance between mean responses, $cov(\bar{X}_1^{(l)}, \bar{X}_2^{(k)}) = 0 \; \forall l$ and $k = 1, ..., K$, as patient data are independent; whereas $cov(\bar{X}_s^{(l)}, \bar{X}_s^{(k)}) = \sigma^2/n_{0s}, \forall l \neq k$, where $l, k = 1, ..., K$, as the data of a shared control group is used in all the pairwise comparisons. For the newly added treatment, $cov(\bar{X}_2^{(K+1)}, \bar{X}_1^{(k)}) = 0$ and $cov(\bar{X}_2^{(K+1)}, \bar{X}_2^{(k)}) = \sigma^2/n_{02}$ for $k = 1, ..., K$, when only the concurrent control data is used.

Elm et al [28] have not considered using all the control data in estimating the difference in mean responses of treatment $K + 1$ and the control, i.e. defined as $\bar{X}_2^{(K+1)'} = \bar{X}_{K+1,2} - \bar{X}_0$. This estimate has a smaller variance than $\bar{X}_2^{(K+1)}$, and $cov(\bar{X}_2^{(K+1)'}, \bar{X}_s^{(k)}) = \sigma^2/n_0$ for $k = 1, ..., K$, and $s = 1, 2$, as $\bar{X}_0$ can be re-expressed as $\frac{n_{01}}{n_0}\bar{X}_{0.1} + \frac{n_{02}}{n_0}\bar{X}_{0.2}$.

Using weighted least squares estimation, an estimate of the vector of treatment effects relative to control, $\boldsymbol{\delta} = (\delta_1, ..., \delta_K, \delta_{K+1})^T$, can be obtained. The estimate, $\hat{\boldsymbol{\delta}}$, is unbiased and has a multivariate normal distribution with mean $\boldsymbol{\delta}$ when there is no trend. This joint distribution can be used to test the hypotheses, $\{H_{01}, ..., H_{0,K+1}\}$, in a similar way to the Dunnett test [29]. The marginal power of rejecting a hypothesis can be computed by considering the corresponding marginal distribution. Note that when $K > 0$, the marginal power is different to the power obtained from a $Z$-test (or a $t$-test) as the joint distribution accounts for the fact that a common control group is used in the analysis. Note also that $var(\hat{\delta}_{K+1})$ depends on the number of control patients.

We label these approaches as $WLS_{all}$ and $WLS_{s2}$ respectively, when all the control data and when only the concurrent control data are used in the estimation of $\delta_{K+1}$.

## Model-based approach: linear regression

In the context of an adaptive randomisation procedure, Villar et al. [19] considered fitting a linear regression model to adjust for the presence of a trend in their investigation where all arms start recruiting at the same time. We explore the following linear regression analyses for the trial that adds an arm at the end of stage one:

- $M_{a1}$: $X_j = \alpha + \sum_{k=0}^{K+1} \beta_k \cdot I(T_j = k) + \gamma \cdot j + \epsilon_j$
- $M_{a2}$: $X_j = \alpha + \sum_{k=0}^{K+1} \beta_k \cdot I(T_j = k) + \nu \cdot c_j + \epsilon_j$

where $\beta_0 = 0$ such that $\beta_k$ represents the difference in mean responses of treatment $k$ and the control treatment; and

- $M_{b1}$:
  $X_j = \alpha + \beta_0 \cdot I(T_j = 0) + \beta_{K+1}^{b1} \cdot I(T_j = K+1) + \gamma \cdot j + \epsilon_j$
- $M_{b2}$: $X_j = \alpha + \beta_0 \cdot I(T_j = 0) + \beta_{K+1}^{b2} \cdot I(T_j = K+1) + \nu \cdot c_j + \epsilon_j$

where $\beta_0 = 0$ such that $\beta_{K+1}^{b1}$ and $\beta_{K+1}^{b2}$ represent the difference in mean responses of the newly added treatment and the control treatment. Note that models $M_{a1}$ and $M_{a2}$ use all data to estimates all $\beta_1, ..., \beta_K, \beta_{K+1}$, whereas model $M_{b1}$ (and $M_{b2}$) uses the data of the control arm (both stages) and the newly added arm to estimate $\beta_{K+1}^{b1}$ (and $\beta_{K+1}^{b2}$) only. The parameter $\gamma$ in $M_{a1}$ and $M_{b1}$ represents the effect of continuous recruitment, whereas $\nu$ in $M_{a2}$ and $M_{b2}$ represents the effect of recruitment that happens after the new treatment has been added (Elm et al [28] interpret $\nu$ as a stage effect). It is worth mentioning that the amount of the control data used to estimate $\beta_{K+1}$ in $M_{a2}$ (and respectively $\beta_{K+1}^{b2}$ in $M_{b2}$) is similar to (same as) that of using only the second stage control data, e.g. $\bar{X}_{K+1,2} - \bar{X}_{0.2}$, since the newly added treatment was not present in stage one and the stage one control data would contribute to the estimation of the stage effect and that of $\beta_1, ..., \beta_K$ under such a model set-up. In other words, the stage one control data contributes very little to the estimation of the difference in mean responses of the newly added treatment and the control treatment when we adjust for $\nu \cdot c_j$ in the analysis.

We note that the inference from $M_{a1}$ without adjustment for the linear term in patient ordering is similar to that from $WLS_{all}$; and that from $M_{a2}$ is similar to $WLS_{s2}$. This is because the joint distribution of $\{\hat{\beta}_1, ..., \hat{\beta}_{K+1}\}$ is equivalent to the corresponding joint distribution of $\hat{\boldsymbol{\delta}}$. The subtle difference is the former approach has one more parameter (i.e. the intercept $\alpha$) than the weighted regression model, which reflects the mean response of the reference group assuming the model is true. On the other hand, testing $\beta_{K+1}^{b1}$ and $\beta_{K+1}^{b2}$ respectively are similar to considering an independent $Z$-test (or a $t$-test when assuming unknown variance) for the hypothesis of $H_{0,K+1}$.

## Metrics

We explore the above analysis approaches and consider

- bias of estimators;
- the type one error rate;
- the marginal power of rejecting a false null hypothesis;
- the root mean squared error,

$$rMSE = \sqrt{\text{bias of an estimate}^2 + \text{variance of an estimate}};$$

when including or excluding the stage one control data in the analysis of the newly added treatment. As we assume $\sigma^2$ is known, the presence of a trend would affect the accuracy of the estimate but not the variance of an estimate. We include the following metric to explore the reduction in the variance of the estimate:

$$BoS = \frac{V_0 - V_a}{V_0}$$

where *BoS* stands for borrowing of strength [30], $V_0$ and $V_a$ correspond to the variance of an estimate that is computed with only the concurrent control data and that with all the control data respectively. Note that *BoS* increases as $V_a$ decreases. This measure is similar to the *R*-squared that measures how close the data are to the fitted regression model. Here $BoS \approx 0$ indicates that the variability of the response data around its mean cannot be further explained by including non-concurrent control data; whereas $BoS \approx 1$ reflects that using all the control data results in a perfect estimate that has negligible variability. One may interpret BoS as an indicator of the benefit of including non-concurrent control data in the analysis of the newly added treatments. Small values show that including the non-concurrent control data provides little benefit to the estimation of the parameters in terms of precision gain. High value is desirable especially when the estimates of the variance are unbiased and are not affected by the presence of a trend.

In the next section, we consider $K = 1$ and a new arm is added at the end of stage one. Let the effect size of the difference in treatment effects be $\delta_1 = \delta_2 = 0.15$ under the alternative scenario, and $\delta_1 = \delta_2 = 0$ under the null

scenario, with $\sigma^2 = 1$, $n_{11} = n_{01} = N_1/2$, and $n_{12} = n_{02} = N_2/2$. We choose $n_0 = n_1 = 550$ to obtain 80% power and 5% type one error rate of rejecting $H_{01}$ using the *Z*-test.
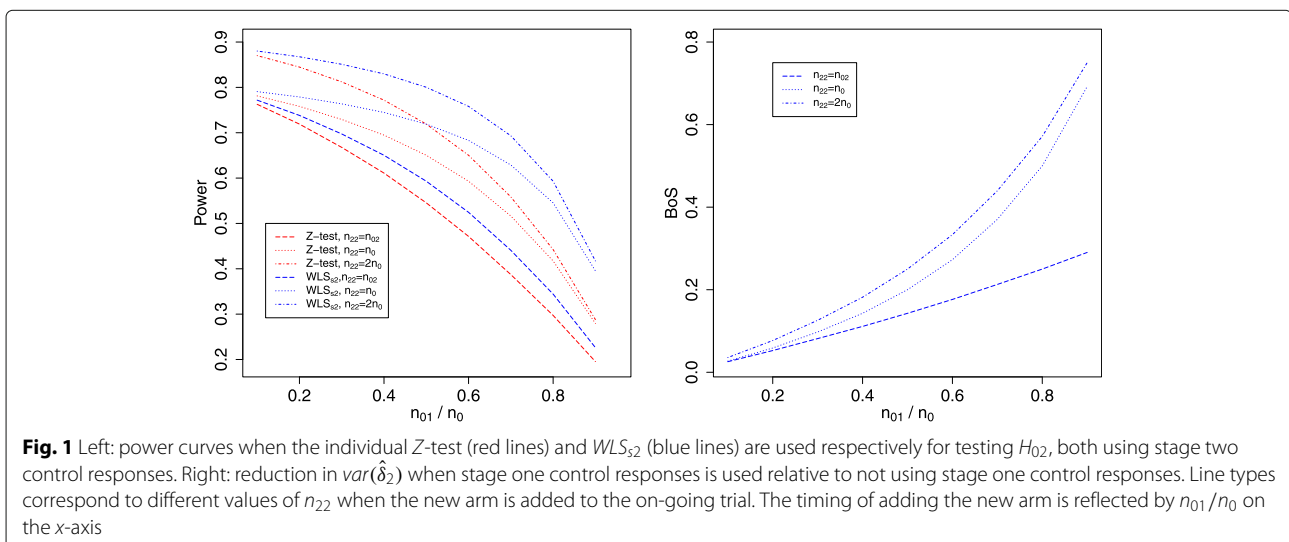
We conduct simulation studies to examine the impact of a trend with the timing of adding a new treatment, $n_{01}/n_0 = \{0.25, 0.5, 0.75\}$, and the sample size of the new arm, $n_{22} = \{n_{02}, n_0, 2n_0\}$. Following (1), we adjust the generated responses post randomization with $\lambda = \{0.02, 0.04, 0.06, 0.08\}$ in the linear trend (2) and the stepwise trend (3) respectively. These values indicate a trend of $\lambda \times 100\%$ of standard deviation. Each scenario is replicated 100 000 times. We illustrate $n_{22} = 2n_0$ as one of the options to explore the marginal benefit of having more patients (than necessary) in the newly added arm while keeping the size of the control arm the same as the initial plan.

## Results

### Analytical power and *BoS* when there is no trend

We now compare *BoS* and the power of rejecting $H_{02}$ when there is no trend and either all, or only concurrent control data are used in the inference.

We first demonstrate the differences between the independent *Z*-test (computed with only concurrent control data) and the model-based approach, $WLS_{all}$ and $WLS_{s2}$, in terms of power. For ease of presentation, we omit the comparisons to the presented linear regression models as they make adjustment for a trend (see the first row of plots in Figs. 2 and 3 for the corresponding marginal power). The left plot in Fig. 1 shows power curves when the individual *Z*-test (red lines) and $WLS_{s2}$, i.e. the approach with the bivariate normal distribution of $(\hat{\delta}_1, \hat{\delta}_2)$ (blue lines), are used respectively for testing $H_{02}$, both using only the concurrent control responses. The timing of adding the new



**Fig. 1** Left: power curves when the individual *Z*-test (red lines) and $WLS_{s2}$ (blue lines) are used respectively for testing $H_{02}$, both using stage two control responses. Right: reduction in $var(\hat{\delta}_2)$ when stage one control responses is used relative to not using stage one control responses. Line types correspond to different values of $n_{22}$ when the new arm is added to the on-going trial. The timing of adding the new arm is reflected by $n_{01}/n_0$ on the *x*-axis

arm to the on-going trial is reflected on the $x$-axis; small values indicate the arm is added after a small number of patients have been randomized to the initial treatment arms. Different lines correspond to having different $n_{22}$.

As expected, the power decreases with $n_{01}/n_0$ when excluding stage one control responses in the inference about the newly added treatment. Comparing the red lines to the blue lines in the left plot of Fig. 1, the power of the individual $Z$-test is lower than the marginal power obtained from $WLS_{s2}$ given the same $n_{22}$. When $n_{22} = n_{02}$, the difference between the power curves computed by the two approaches is less than 6% (shown by comparing the dashed lines). When $n_{22} \geq n_0$, the magnitudes of the differences between the power curves are larger especially for large $n_{01}/n_0$; e.g. compare the dotted-dashed lines at $n_{01}/n_0 = 0.8$, the marginal power obtained from $WLS_{s2}$ is 15% more than that computed from the individual $Z$-test. This finding highlights that considering the joint outcomes of the hypotheses through the joint distribution of the parameters is more efficient than using the individual $Z$-test when only the concurrent control data is used. This is because the former accounts for using a shared control group in the inference, whereas the $Z$-test uses the control data independently for each hypothesis test as if they were from several trials. Besides that, compare the (vertical) difference between the dotted-dashed line ($n_{22} = 2n_0$) and the dotted line ($n_{22} = n_0$), either red or blue, we see that the magnitude of the difference in the power decreases as $n_{01}/n_0$ increases. This indicates that the marginal benefit of having a larger sample size in one arm than necessary can be small, depending on when the new arm is added to the trial.

For $WLS_{all}$ with $n_{22} = n_{02}$, the power curve coincides with the red dotted line, i.e. the individual $Z$-test using only the concurrent control data and with $n_{22} = n_0$. This is not surprising as the same number of patients are used

in the inference, albeit the number of patients randomized to the newly added arm is different for the two comparators (i.e. $WLS_{all}$ considers $n_{02}$ responses of the new treatment and $n_0$ responses of the control treatment; the individual $Z$-test uses $n_0$ responses of the new treatment and $n_{02}$ responses of the stage two control treatment). For $WLS_{all}$ with $n_{22} = n_0$ and $n_{22} = 2n_0$, the marginal power of rejecting $H_{02}$ is 0.8 and 0.89 respectively across all $n_{01}/n_0$, which is close to the respective power obtained from $Z$-test using all the control data.

We now consider the *BoS* values that are computed with $var(\hat{\delta}_2)$ obtained from $WLS_{all}$ and $WLS_{s2}$ respectively. The right plot in Fig. 1 shows that when all the control data is included in the estimation, *BoS* increases with $n_{01}/n_0$. When $n_{22} = n_{02}$, *BoS* is close to 30% when the new treatment is added at $n_{01}/n_0 = 0.9$. When $n_{22}$ has the same number as $n_0$, we see that *BoS* becomes even larger. However, the marginal return of having large $n_{22}$ than necessary can be less notable: comparing the dotted-dashed line to the dotted line shows that the magnitudes of the differences between the *BoS* perhaps are less impressive compared to the likely cost of enrolling an extra $n_0$ patients for the new arm.

### Bias of the estimate when there is a trend

We now present the simulation findings of the inference about the newly added treatment when there is a trend.

Consider the estimated difference in mean responses of the newly added treatment and the control treatment. Table 1 shows the maximum absolute bias (and median absolute bias) in the parameter estimations when there is a trend. Each row corresponds to a trend with a value of $\lambda$ for all the scenarios with different combination of $n_{01}/n_0 = \{0.25, 0.5, 0.75\}$ and $n_{22} = \{n_{02}, n_0, 2n_0\}$. The values with an order of magnitude of -4 are set to zero, as we also obtain such a small number for the unbiased esti-

**Table 1** The maximum (median) absolute bias of the estimated difference in mean responses of the newly added treatment and the control arm when there is a trend. Values with -4 order of magnitude are set to zero

| $\lambda$ | $WLS_{all}$ | $WLS_{s2}$ | $M_{a1}$ | $M_{a2}$ | $M_{b1}$ | $M_{b2}$ |
|---|---|---|---|---|---|---|
| | | | Linear trend | | | |
| 2% | 0.007 (0.005) | 0 (0) | 0.001 (0) | 0 (0) | 0 (0) | 0 (0) |
| 4% | 0.015 (0.010) | 0 (0) | 0.001 (0) | 0 (0) | 0 (0) | 0 (0) |
| 6% | 0.022 (0.015) | 0 (0) | 0.001 (0) | 0 (0) | 0 (0) | 0 (0) |
| 8% | 0.030 (0.020) | 0 (0) | 0.001 (0) | 0 (0) | 0 (0) | 0 (0) |
| | | | Step trend | | | |
| 2% | 0.015 (0.010) | 0 (0) | 0.007 (0.004) | 0 (0) | 0.010 (0.005) | 0 (0) |
| 4% | 0.030 (0.020) | 0 (0) | 0.015 (0.008) | 0 (0) | 0.019 (0.010) | 0 (0) |
| 6% | 0.045 (0.030) | 0 (0) | 0.023 (0.012) | 0 (0) | 0.029 (0.015) | 0 (0) |
| 8% | 0.060 (0.040) | 0 (0) | 0.031 (0.016) | 0 (0) | 0.038 (0.019) | 0 (0) |

mators when there is no trend in our simulation, which is due to the Monte Carlo simulation error.

When there is a positive trend in the simulation, we find that the estimates obtained from $WLS_{s2}$, $M_{a2}$ and $M_{b2}$ respectively are unbiased. For those obtained from $WLS_{all}$, the magnitude of the bias is larger when there is a step trend than when there is a linear trend. This is because the presence of a positive trend inflates the estimate of the overall mean responses of the control treatment while the mean responses of the newly added treatment only have a trend that is similar to the mean responses of the second stage control treatment.

For those obtained from $M_{a1}$ and $M_{b1}$, the estimated parameter is almost unbiased when there is a linear trend but is biased when there is a step trend. This shows that adjustment with a linear term may not always yield accurate estimate especially when the structure of a trend is not linear.

**Marginal power in the presence of a trend**
The first row of plots in Figs. 2 and 3 show the marginal power of rejecting $H_{02}$ when there is no trend; other rows of plots correspond to different magnitudes of $\lambda$ when there is a linear and a step trend respectively. The $x$-axis of the plots indicates the value of $n_{22}$ and $n_{01}/n_0$. Each column of plots has a different range of $y$-axis due to having different values of $n_{22}$ when the new arm is added to the on-going trial. As expected, the marginal power obtained from $WLS_{s2}$ is similar to that from $M_{a2}$: the blue line (with $\times$) superimposes the red line (with $\triangle$).

We see that the marginal power obtained from $WLS_{all}$ is the highest and from $M_{b2}$ is the lowest among all the comparators. However, the marginal power obtained from $WLS_{all}$ is inflated when there is a positive trend. From our simulation, we find that relative difference in the inflated power to the non-inflated power is larger when $n_{22} = 550$ compared to when $n_{22} = 1100$. The reason could be when $n_{22}$ is small (relative to the fixed $n_{01}$), stage one control responses deviate more on average from the responses in stage two, which leads to a larger bias when using all the control data and hence increases the average number of rejected hypotheses.

On the other hand, the marginal power obtained from $M_{a1}$ and $M_{b1}$ respectively when there is a linear trend remains the same as when there is no trend (i.e. compare the black lines and cyan lines respectively across the row of plots in Fig. 2); but it is inflated when there is a step trend. The magnitude of the inflation increases with the values, $\lambda$, of the step trend (see Fig. 3).

We find that the marginal power obtained from $WLS_{s2}$, $M_{a2}$ and $M_{b2}$ respectively remains unchanged when there is a trend. This is not surprising as the impact of a positive trend in the mean responses of the newly added arm and the second stage control arm are of similar magni-

tude, which is being removed when the difference in mean responses is considered.

**Type one error rate in the presence of a trend**
Figures 4 and 5 show the type one error rate of rejecting $H_{02}$ in the presence of a linear trend and of a step trend respectively, when $\delta_1 = \delta_2 = 0$. A subtle difference in the type one error rate obtained from $WLS_{s2}$ and $M_{b2}$ respectively can be seen in the plots.

We observe that the presence of a step trend would lead to a larger inflation in the type one error rate obtained from $WLS_{all}$ than when there is a linear trend with the same $\lambda$, i.e. from comparing the green line (with $+$ sign) in Fig. 4 to those in Fig. 5. Besides, the magnitude of inflation is the largest when both the values of $n_{01}/n_0$ and $\lambda$ are the largest. Having larger $n_{22}$ also causes more inflation in the type one error rate given the same $n_{01}/n_0$. This may be because, given the same magnitude of the bias, large $n_{22}$ makes the denominator of the test statistics smaller, and hence increases the number of hypothesis rejections.

For the type one error rate obtained from $M_{a1}$ and $M_{b1}$, we see a similar finding to that of the marginal power. That is, when there is a linear trend, the type one error rate is maintained within the Monte Carlo simulation error, whereas it is inflated when there is a step trend (see the black line and cyan line respectively in Fig. 5). We find that the type one error rate obtained from $WLS_{s2}$, $M_{a2}$ and $M_{b2}$ respectively is maintained within the Monte Carlo simulation error when there is a trend.

**rMSE in the presence of a trend**
The plots of rMSE are presented in Additional file 1: Figures 1 and 2 in the supplement. When there is a positive trend, we find that $M_{b2}$ has the highest rMSE, which remains the same across all the considered values of $\lambda$. This is not surprising as the least amount of the data is used. The rMSE values obtained from $WLS_{s2}$ and $M_{a2}$ also remain unchanged when there is a trend.

When there is a linear trend, the rMSE obtained from $WLS_{all}$ increases with $\lambda$ while remaining smaller than that from $M_{a1}$ and $M_{b1}$ (which remain unchanged across the linear trends). However, when there is a step trend, the rMSE obtained from these approaches are inflated; the rMSE obtained from $WLS_{all}$ is higher than that from $M_{a1}$ when $\lambda \geq 4\%$. Nevertheless, across all the considered scenarios, the rMSE obtained from $M_{a1}$ is always smaller than the approaches that use less data, i.e. $M_{a2}$, $M_{b1}$, $M_{b2}$, and $WLS_{s2}$.

These findings highlight that even though the presence of a small trend would lead to a small bias in the estimation, the quality of the estimate computed using all the control data can be better than that computed using only the concurrent control data. This does not hold for larger trends.
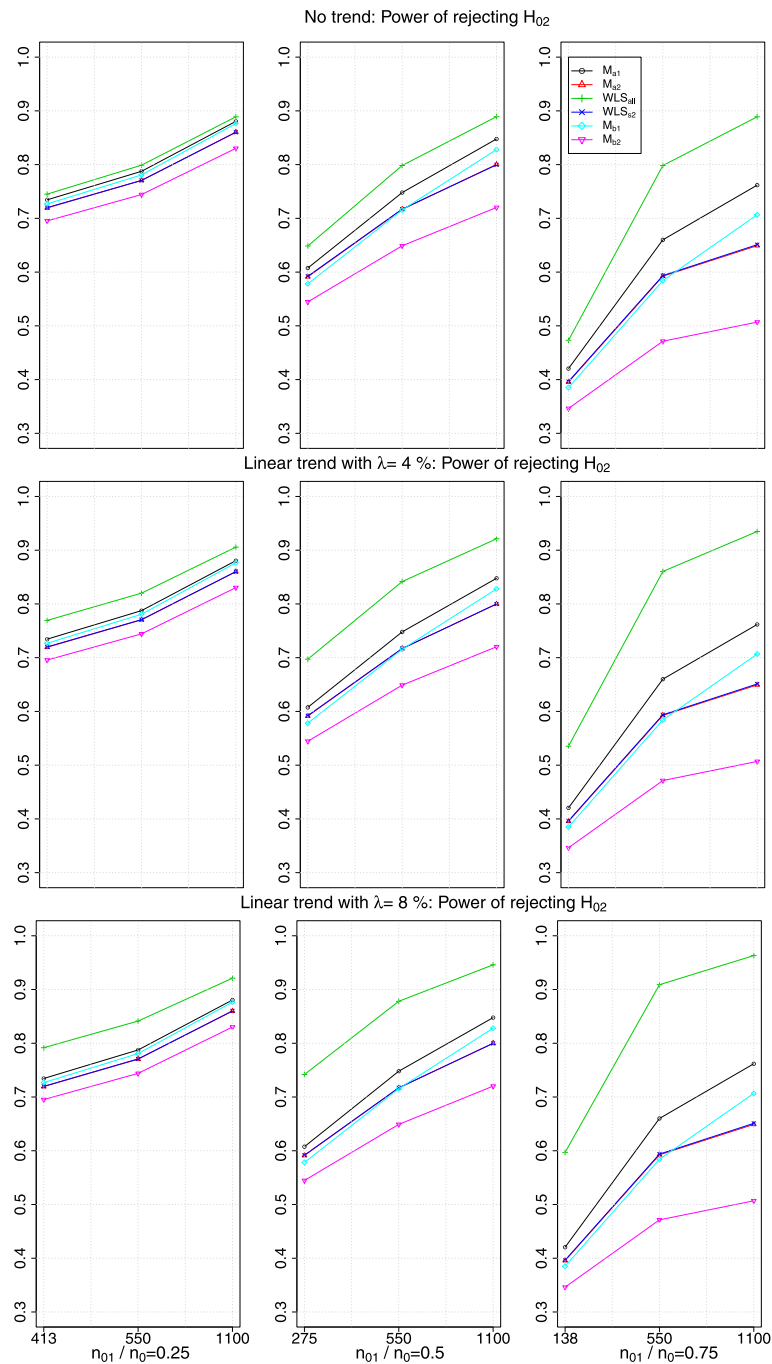
**Fig. 2** The power of rejecting $H_{02}$ in the presence of no trend or a linear drift with a magnitude of $\lambda > 0$ standard deviation (row-wise); *x*-axis indicates the value of $n_{22}$ and the timing of adding the new arm

## Inference about the initial treatment

For the inference about the initial treatment, we find that the bias, type one error rate and power of rejecting $H_{01}$ are maintained at the same levels as those when there is no trend in the simulation. This finding is observed for all the model-based approaches. This may be due to using the restricted randomization procedure within each stage of the trial in the simulation, which ensures that not all $n_{ks}$ patients are being enrolled to a particular arm $k$ during the early or the late phase of the trial. As a result, the average impact of a trend on the stage-wise mean responses of the initial treatment and of the control treatment are similar, which are then being subtracted when the difference in the mean responses is considered. This observation is
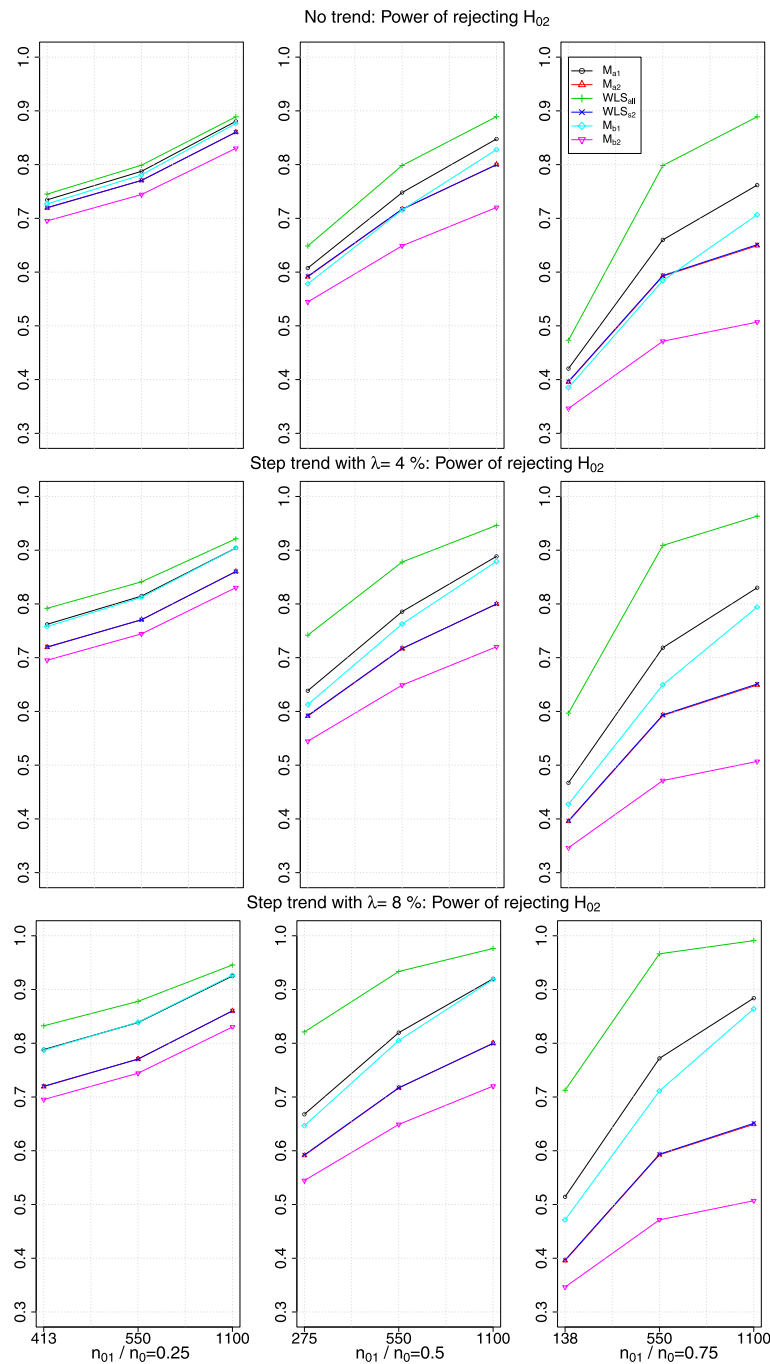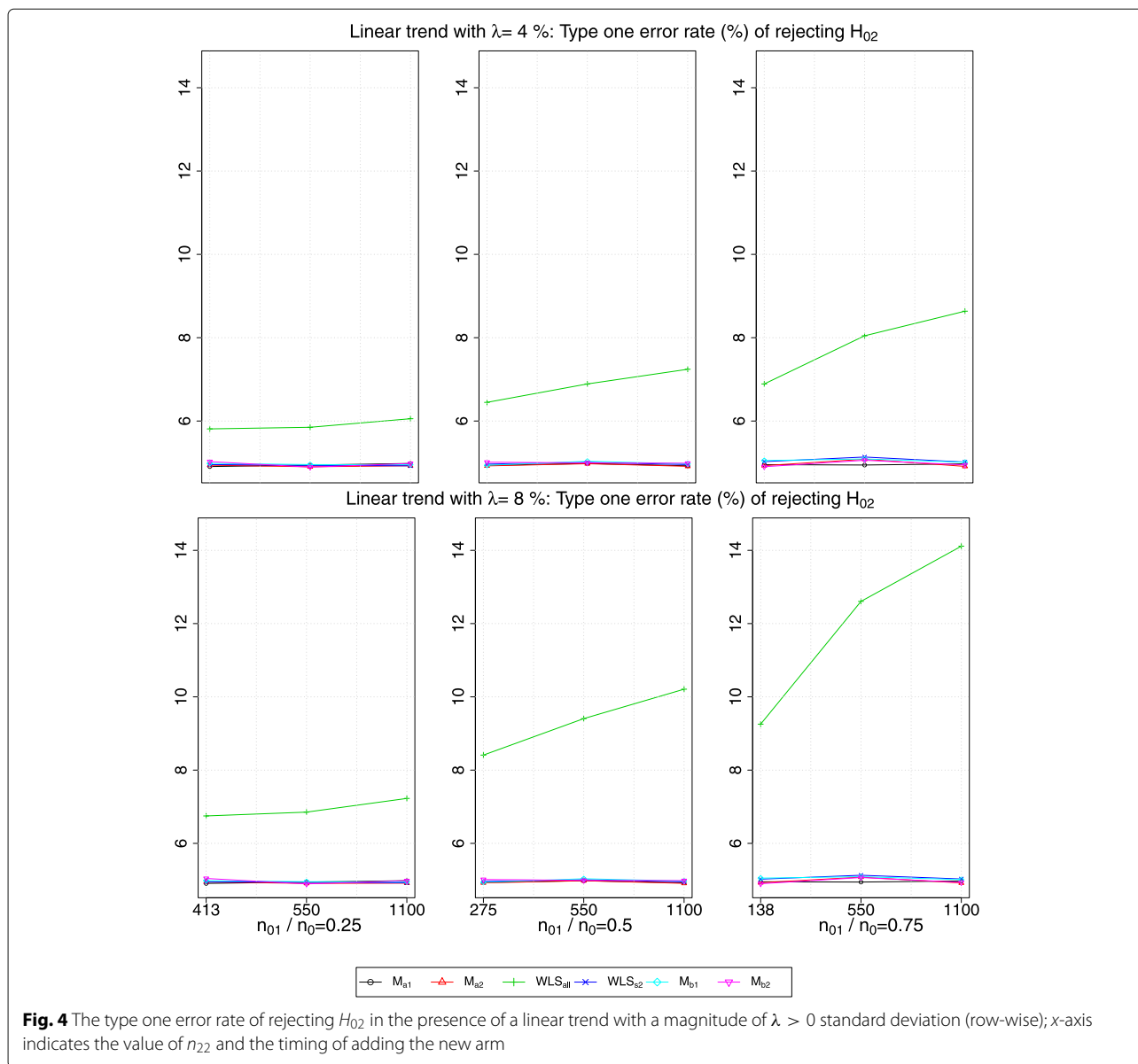
**Fig. 3** The power of rejecting $H_{02}$ in the presence of no trend or a step trend with a magnitude of $\lambda > 0$ standard deviation (row-wise); *x*-axis indicates the value of $n_{22}$ and the timing of adding the new arm

consistent with the finding of Ryeznik and Sverdlov [17] who compare different randomization procedures in the context of the standard ANOVA *F*-test.

## Discussion

When there is no trend, the advantages of including the non-concurrent control data are greater when the new arm is added at a later time point. Our simulation studies show that given the same magnitude of $\lambda > 0$, the presence of a step trend affects the validity of the inference more than that of a linear trend when all the control data are used in the analysis of the new arm. We also examine that using a linear regression model adjusting for a linear term may lead to spurious findings when there is a

**Fig. 4** The type one error rate of rejecting $H_{02}$ in the presence of a linear trend with a magnitude of $\lambda > 0$ standard deviation (row-wise); *x*-axis indicates the value of $n_{22}$ and the timing of adding the new arm
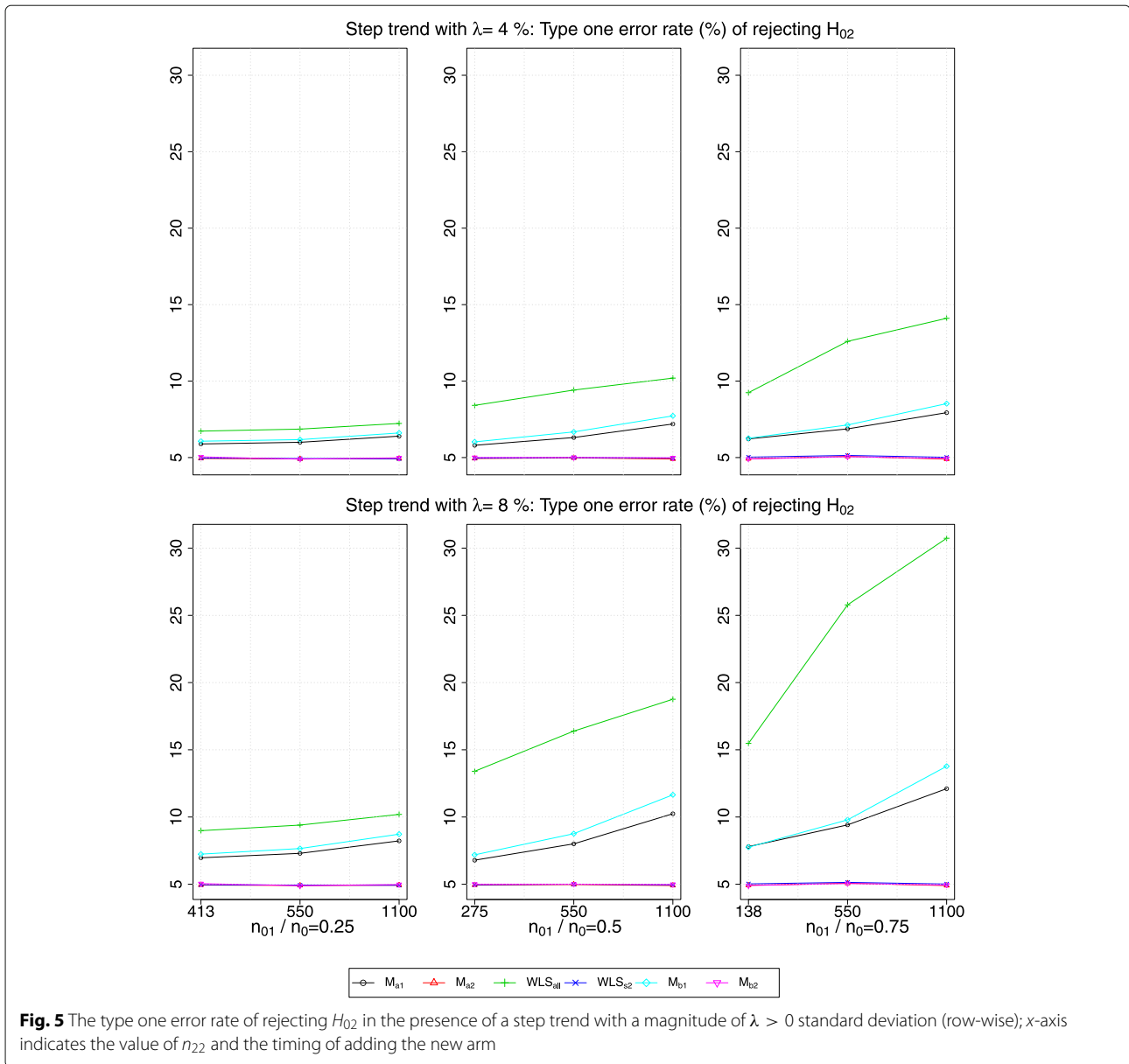
non-linear trend, though the rMSE of the corresponding estimate can be lower than other approaches that use only the concurrent control data. This represents a caveat for including non-concurrent control data in the analysis plan as it is impossible to know the structure of a trend at the planning stage of a trial. After data collection, one can only test the presence of a trend using some quality control techniques such as control charts and CUSUM analysis, see for example a review by Noyez [31]. The analogy between clinical trials and industrial processes is becoming clearer especially when the trial has a long enrolment period. [32–38]

This work has highlighted the impact of a trend when the variance of the observations is known. We emphasise

that not all methods assume known variance. We conjecture that the presented results would hold when an estimated variance is unbiased, i.e. the estimate is robust to the presence of a trend. Otherwise, the over- or underestimated variance would cause the test statistics to be biased and hence provide a spurious result [28]. We also note that when there is a negative trend, i.e. $\lambda < 0$, the type one error and the marginal power will be deflated, i.e. smaller than the nominal value. This is a direct consequence of a trend that causes responses to be smaller than the true values as the trial progresses.

All things considered, we would recommend that at the design stage of phase III trials, sample size calculations assume only concurrent control information will be

**Fig. 5** The type one error rate of rejecting $H_{02}$ in the presence of a step trend with a magnitude of $\lambda > 0$ standard deviation (row-wise); *x*-axis indicates the value of $n_{22}$ and the timing of adding the new arm

used. When the magnitude of a trend is small, we find including the non-concurrent control data in the analysis can improve the efficiency of estimating the parameter of interest. However there could be an inflation (or deflation) in the type one error rate and the marginal power of testing the corresponding hypothesis unless the positive (or negative) trend has a magnitude of less than 0.5% of the standard deviation (based on simulation results not presented here). This finding is similar to the finding of Kopp-Schneider et al [39] for the situation where external information is used in the inference of clinical trials: power gain is not possible when requiring a strict type one error rate control. When the recruitment to all arms do not finish simultaneously, the control data can be sepa-

rated into before and after adding an arm and those after the initial treatment arms finish recruitment. In this case, the presence of a trend across the three stages can be tested before utilising all the control data to increase the precision of the inference.

We acknowledge that an inflated power is not an issue if we are confident that the intervention is effective. However, using the trial result for other purposes may lead to negative consequences since an inflated power could mean that the estimated effect size is likely to be higher than it should. For example, doing a cost-effectiveness analysis using the estimated effect size may over-estimate the value of the intervention due to the presence of a positive trend. Future work could review real trials that add

arms and explore the presence of trends and their structure in the data. One can also investigate the potential of using non-concurrent control data for the situations where i) there are more than one treatment arm being added to the on-going trial; ii) there are more than two-stages; and iii) adaptive randomization procedures are considered.

## Conclusion

Platform trials can potentially speed up the drug development processes. The feature of continuous recruitment to the control arm may increase the precision of the inference about the newly added interventions in some situations; or reduce the required sample size for the evaluation of the newly added interventions at the cost of having a less stringent control of the error rates. In light of the presence of a potential trend, it is wise to compare the results of including and excluding non-concurrent control data in the analysis of the newly added interventions.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1186/s12874-020-01043-6.

---

**Additional file 1:** Figures 1 and 2 show the rMSE when there is a linear and a step trend respectively in the simulation.

---

**Abbreviations**
BoS: borrowing of strength; rMSE: root mean squared error

**Authors' contributions**
KL conceived and designed the research and drafted the manuscript. JW provided important intellectual input to interpretation of findings and manuscript writing. All authors critically appraised the final manuscript.

**Availability of data and materials**
Not applicable.

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare that they have no competing interests.

**Author details**
[1] MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus , CB2 0SR Cambridge, UK. [2] Population Health Sciences Institute, Newcastle University, Baddiley-Clark Building, Newcastle University Richardson Road, Newcastle upon Tyne, Newcastle upon Tyne, UK.

## References

1. Bothwell LE, Avorn J, Khan NF, Kesselheim AS. Adaptive design clinical trials: a review of the literature and ClinicalTrials.gov. BMJ Open. 2018;8(2): 018320. https://doi.org/10.1136/bmjopen-2017-018320.
2. Woodcock J, LaVange LM. Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both. N Engl J Med. 2017;377(1):62–70. https://doi.org/10.1056/NEJMra1510062.
3. Hirakawa A, Asano J, Sato H, Teramukai S. Master protocol trials in oncology: Review and new trial designs,. Contemp Clin Trials Commun. 2018;12:1–8. https://doi.org/10.1016/j.conctc.2018.08.009.
4. Angus DC, Alexander BM, Berry S, Buxton M, Lewis R, Paoloni M, Webb SAR, Arnold S, Barker A, Berry DA, Bonten MJM, Brophy M, Butler C, Cloughesy TF, Derde LPG, Esserman LJ, Ferguson R, Fiore L, Gaffey SC, Gaziano JM, Giusti K, Goossens H, Heritier S, Hyman B, Krams M, Larholt K, LaVange LM, Lavori P, Lo AW, London AJ, Manax V, McArthur C, O'Neill G, Parmigiani G, Perlmutter J, Petzold EA, Ritchie C, Rowan KM, Seymour CW, Shapiro NI, Simeone DM, Smith B, Spellberg B, Stern AD, Trippa L, Trusheim M, Viele K, Wen PY, Woodcock J. Adaptive platform trials: definition, design, conduct and reporting considerations. Nat Rev Drug Discov. 20191–11. https://doi.org/10.1038/s41573-019-0034-3.
5. Cohen DR, Todd S, Gregory WM, Brown JM. Adding a treatment arm to an ongoing clinical trial: a review of methodology and practice. Trials. 2015;16(1):179. https://doi.org/10.1186/s13063-015-0697-y.
6. Karrison TG, Huo D, Chappell R. A group sequential, response-adaptive design for randomized clinical trials. Control Clin Trials. 2003;24(5):506–22. https://doi.org/10.1016/S0197-2456(03)00092-8.
7. Chow S-C, Chang M, Pong A. Statistical Consideration of Adaptive Methods in Clinical Development. J Biopharm Stat. 2005;15(4):575–91. https://doi.org/10.1081/BIP-200062277.
8. Feng H, Shao J, Chow S-C. Adaptive Group Sequential Test for Clinical Trials with Changing Patient Population. J Biopharm Stat. 2007;17(6): 1227–38. https://doi.org/10.1080/10543400701645512.
9. Mahajan R, Gupta K. Adaptive design clinical trials: Methodology, challenges and prospect,. Indian J Pharmacol. 2010;42(4):201–7. https://doi.org/10.4103/0253-7613.68417.
10. Pallmann P, Bedding AW, Choodari-Oskooei B, Dimairo M, Flight L, Hampson LV, Holmes J, Mander AP, Odondi L, Sydes MR, Villar SS, Wason JMS, Weir CJ, Wheeler GM, Yap C, Jaki T. Adaptive designs in clinical trials: why use them, and how to run and report them,. BMC Med. 2018;16(1):29. https://doi.org/10.1186/s12916-018-1017-7.
11. Altman DG. Avoiding bias in trials in which allocation ratio is varied. J R Soc Med. 2018;111(4):143–4. https://doi.org/10.1177/0141076818764320.
12. Coad DS. Sequential estimation with data-dependent allocation and time trends. Seq Anal. 1991;10(1-2):91–97. https://doi.org/10.1080/07474949108836227.
13. Hu F, Rosenberger WF, Zidek JV. Relevance weighted likelihood for dependent data. Metrika. 2000;51(3):223–43. https://doi.org/10.1007/s001840000051.
14. Rosenkranz GK. The impact of randomization on the analysis of clinical trials. Stat Med. 2011;30(30):3475–87. https://doi.org/10.1002/sim.4376.
15. Tamm M, Hilgers R-D. Chronological Bias in Randomized Clinical Trials Arising from Different Types of Unobserved Time Trends. Methods Inf Med. 2014;53(06):501–10. https://doi.org/10.3414/ME14-01-0048.
16. Hilgers R-D, Uschner D, Rosenberger WF, Heussen N. ERDO - a framework to select an appropriate randomization procedure for clinical trials. BMC Med Res Methodol. 2017;17(1):159. https://doi.org/10.1186/s12874-017-0428-z.
17. Ryeznik Y, Sverdlov O. A comparative study of restricted randomization procedures for multiarm trials with equal or unequal treatment allocation ratios. Stat Med. 2018;37(21):3056–77. https://doi.org/10.1002/sim.7817.
18. Lipsky AM, Greenland S. Confounding due to changing background risk in adaptively randomized trials,. Clin Trials. 2011;8(4):390–7. https://doi.org/10.1177/1740774511406950.
19. Villar SS, Bowden J, Wason J. Response-adaptive designs for binary responses: How to offer patient benefit while being robust to time trends?. Pharm Stat. 2018;17(2):182–97. https://doi.org/10.1002/pst.1845.
20. Jiang Y, Zhao W, Durkalski-Mauldin V. Time-trend impact on treatment estimation in two-arm clinical trials with a binary outcome and Bayesian

response adaptive randomization. J Biopharm Stat. 20191–20. https://doi.org/10.1080/10543406.2019.1607368.

21. Hilgers R-D, Manolov M, Heussen N, Rosenberger WF. Design and analysis of stratified clinical trials in the presence of bias. Stat Methods Med Res. 2019;096228021984614:. https://doi.org/10.1177/0962280219846146.

22. Greenland S. Interpreting time-related trends in effect estimates. J Chron Dis. 1987;40:17–24. https://doi.org/10.1016/S0021-9681(87)80005-X.

23. Wason J, Magirr D, Law M, Jaki T. Some recommendations for multi-arm multi-stage trials,. Stat Methods Med Res. 2016;25(2):716–27. https://doi.org/10.1177/0962280212465498.

24. Ventz S, Cellamare M, Parmigiani G, Trippa L. Adding experimental arms to platform clinical trials: randomization procedures and interim analyses. Biostatistics. 2018;19(2):199–215. https://doi.org/10.1093/biostatistics/kxx030.

25. Lee KM, Wason J, Stallard N. To add or not to add a new treatment arm to a multiarm study: A decision-theoretic framework. Stat Med. 2019;38(18):8194. https://doi.org/10.1002/sim.8194.

26. Saville BR, Berry SM. Efficiencies of platform clinical trials: A vision of the future. Clin Trials J Soc Clin Trials. 2016;13(3):358–66. https://doi.org/10.1177/1740774515626362.

27. Butler CC, Coenen S, Saville BR, Cook J, van der Velden A, Homes J, de Jong M, Little P, Goossens H, Beutels P, Ieven M, Francis N, Moons P, Bongard E, Verheij T. A trial like ALIC4E: why design a platform, response-adaptive, open, randomised controlled trial of antivirals for influenza-like illness?. ERJ Open Res. 2018;4(2):. https://doi.org/10.1183/23120541.00046-2018.

28. Elm JJ, Palesch YY, Koch GG, Hinson V, Ravina B, Zhao W. Flexible Analytical Methods for Adding a Treatment Arm Mid-Study to an Ongoing Clinical Trial. J Biopharm Stat. 2012;22(4):758–72. https://doi.org/10.1080/10543406.2010.528103.

29. Choodari-Oskooei B, Bratton DJ, Gannon MR, Meade AM, Sydes MR, Parmar MK. Adding new experimental arms to randomised clinical trials: impact on error rates. 2019. https://doi.org/1902.05336.

30. Jackson D, White IR, Price M, Copas J, Riley RD. Borrowing of strength and study weights in multivariate and network meta-analysis. Stat Methods Med Res. 2017;26(6):2853–68. https://doi.org/10.1177/0962280215611702.

31. Noyez L. Control charts, Cusum techniques and funnel plots. A review of methods for monitoring performance in healthcare. Interact Cardiovasc Thorac Surg. 2009;9(3):494–9. https://doi.org/10.1510/icvts.2009.204768.

32. Chang WR, McLean IP. CUSUM: A tool for early feedback about performance?,. BMC Med Res Methodol. 2006;6(1):8. https://doi.org/10.1186/1471-2288-6-8.

33. Sibanda T, Sibanda N. The CUSUM chart method as a tool for continuous monitoring of clinical outcomes using routinely collected data. BMC Med Res Methodol. 2007;7(1):46. https://doi.org/10.1186/1471-2288-7-46.

34. McLaren PJ, Hart KD, Dolan JP, Hunter JG. CUSUM analysis of mortality following esophagectomy to allow for identification and intervention of quality problems. J Clin Oncol. 2017;35(4_suppl):203–203. https://doi.org/10.1200/JCO.2017.35.4suppl.203.

35. Neuburger J, Walker K, Sherlaw-Johnson C, van der Meulen J, Cromwell DA. Comparison of control charts for monitoring clinical performance using binary data,. BMJ Qual Saf. 2017;26(11):919–28. https://doi.org/10.1136/bmjqs-2016-005526.

36. Fortea-Sanchis C, Martínez-Ramos D, Escrig-Sos J. CUSUM charts in the quality control of colon cancer lymph node analysis: a population-registry study. World J Surg Oncol. 2018;16(1):230. https://doi.org/10.1186/s12957-018-1533-0.

37. Redd D, Shao Y, Cheng Y, Zeng-Treitler Q. Detecting Secular Trends in Clinical Treatment through Temporal Analysis. J Med Syst. 2019;43(3):74. https://doi.org/10.1007/s10916-019-1173-0.

38. Fortea-Sanchis C, Escrig-Sos J. Quality Control Techniques in Surgery: Application of Cumulative Sum (CUSUM) Charts. Cir Esp (English Edition). 2019;97(2):65–70. https://doi.org/10.1016/j.cireng.2019.01.010.

39. Kopp-Schneider A, Calderazzo S, Wiesenfarth M. Power gains by using external information in clinical trials are typically not possible when requiring strict type I error control. Biom J. 2019;201800395:. https://doi.org/10.1002/bimj.201800395.

## Publisher's Note