Statistics
in Medicine

Research Article

Received 17 September 2012, Accepted 9 May 2013 Published online 4 June 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5869

Bias in randomised factorial trials

Brennan C. Kahan[*†]

Factorial trials are an efficient method of assessing multiple treatments in a single trial, saving both time and resources. However, they rely on the assumption of no interaction between treatment arms. Ignoring the possibility of an interaction in the analysis can lead to bias and potentially misleading conclusions. Therefore, it is often recommended that the size of the interaction be assessed during analysis. This approach can be formalised as a two-stage analysis; if the interaction test is not significant, a factorial analysis (where all patients receiving treatment A are compared with all not receiving A, and similarly for treatment B) is performed. If the interaction is significant, the analysis reverts to that of a four-arm trial (where each treatment combination is regarded as a separate treatment arm). We show that estimated treatment effects from the two-stage analysis can be biased, even in the absence of a true interaction. This occurs because the interaction estimate is highly correlated with treatment effect estimates from a four-arm analysis. Simulations show that bias can be severe (over 100% in some cases), leading to inflated type I error rates. Therefore, the two-stage analysis should not be used in factorial trials. A preferable approach may be to design multi-arm trials (i.e. four separate treatment groups) instead. This approach leads to straightforward interpretation of results, is unbiased regardless of the presence of an interaction, and allows investigators to ensure adequate power by basing sample size requirements on a four-arm analysis. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: factorial trials; 2 × 2 factorial design; preliminary interaction test; two-stage analysis; randomised controlled trial

1. Introduction

Randomised controlled trials (RCTs) are the gold standard for assessing treatment efficacy; however, they can be both expensive and time consuming. It is not unusual for a trial to take 5–10 years from conception to results and cost millions of dollars or more. Factorial trials present an efficient and cheaper alternative to standard RCTs, as they allow investigators to test the individual effectiveness of two or more treatments in a single trial, without increasing the sample size. For a 2 × 2 factorial trial, this is carried out by randomising patients twice, once to treatment A or control A and then again to treatment B or control B. A factorial analysis can then be performed, where all patients who receive treatment A are compared with those who did not receive treatment A, and similarly for treatment B.

However, factorial trials rely strongly on the assumption that there is no interaction between treatments A and B, that is, the treatment effect of A does not depend on whether the patient also receives treatment B, and vice versa. If an interaction is present, then a factorial analysis will lead to biased results [1], and the appropriate method of analysis is that of a four-arm trial, where the three active arms (A alone, B alone, and A + B) are compared with the control group (neither A nor B).

Therefore, assessing the size of the interaction term during the analysis is often recommended [1, 2]. This can be formalised as a two-stage approach; a preliminary interaction test between treatments A and B is performed. If this interaction test is not significant at some pre-defined level (e.g. 5%), then a factorial analysis is performed. If the interaction test is significant, then a four-arm analysis is performed. However, recent reviews have shown that many factorial trials do not perform an initial interaction test [1, 2], and so large or even moderate interactions could give biased results and potentially misleading conclusions. Additionally, this type of two-stage analysis (where the method of analysis depends on an initial test of significance) has been shown to give biased estimates of treatment effect and to inflate the type I error rate in crossover trials with a preliminary test for carryover [3].

MRC Clinical Trials Unit, Aviation House, 125 Kingsway, London WC2B 6NH, U.K.
**Correspondence to: Brennan C. Kahan, MRC Clinical Trials Unit, Aviation House, 125 Kingsway, London WC2B 6NH, U.K.*
†E-mail: brk@ctu.mrc.ac.uk

Because randomised trials may be used as the basis for policy decisions, which may affect thousands or even millions of patients, it is essential that the method of analysis will give unbiased results with correct type I error rates. The aim of this paper is to explore the properties of a two-stage analysis in the context of factorial trials and to determine whether its use is appropriate in clinical trials. We consider only factorial trials where the primary aim is to assess the effectiveness of individual treatments, rather than assessing whether there is an interaction between treatment arms.

2. Motivating example: the Second Multicentre Intrapleural Sepsis Trial

The Second Multicentre Intrapleural Sepsis Trial (MIST2) was a 2×2 randomised factorial trial assessing whether either tPA or DNase were effective in reducing the size of patients' pleural effusions (measured as the area of the hemithorax occupied by effusion) in patients with a pleural infection [4]. Patients were randomised to one of the four treatment groups (tPA, DNase, tPA + DNase or double placebo) using minimisation. The primary outcome was the change from baseline to day 7 in the size of a patients' pleural effusion (the absolute difference was taken, rather than the relative difference). The trial recruited 210 patients (193 of whom were included in the primary analysis) and was not powered to detect an interaction (sample size calculations were based on a factorial analysis).

The statistical analysis plan called for an initial interaction test to be carried out at the 5% level. If the test was not significant, a factorial analysis would be performed (comparing all patients receiving tPA with all patients who did not, and all patients receiving DNase with all patients who did not). If the interaction test was significant, a four-arm analysis would be performed (tPA, DNase and tPA + DNase would each be compared with the double placebo group).

The initial interaction test was highly significant (effect size −14, 95% confidence interval (CI) −24 to −5, $p = 0.002$), and so the primary analysis was therefore a four-arm comparison. Neither tPA nor DNase were effective in reducing the area of the pleural effusion as compared with placebo (tPA 2, 95% CI −5 to 9, $p = 0.55$; DNase 4, 95% CI −2 to 11, $p = 0.14$). However, the combination treatment (tPA + DNase) was effective (effect size −8, 95% CI −13 to −2, $p = 0.005$). It should be noted that the mean size of pleural effusion at baseline was 43, so an absolute reduction between treatments of eight corresponds to a relative reduction of almost 20%.

If no interaction test had been conducted, and only a factorial analysis performed (as has been the case in previous reports of factorial trials [1, 2]), the results could have been highly misleading. Under a factorial analysis, tPA was found to significantly reduce the size of patients' pleural effusions compared with placebo (effect size −5, 95% CI −10 to −0.4, $p = 0.03$), and there was a nonsignificant trend towards a decrease in pleural effusions with DNase (effect size −3, 95% CI −8 to 2, $p = 0.28$). These results could have led to the widespread adoption of tPA as a treatment for pleural infection, meaning that patients not only would receive an ineffective treatment but would also be denied access to a potentially very effective treatment (tPA + DNase).

3. The two-stage analysis

Let $x_{a,b}$ denote the mean outcome for treatments $a, b (a = 0, 1, b = 0, 1$, where 0 indicates the patient did not receive that treatment, and 1 indicates they did). A general model can be written as follows:

$$y_i = \alpha + \beta_A + \beta_B + \beta_{\text{interaction}} + \varepsilon_i \quad (1)$$

where y_i is the outcome for patient i; β_A, β_B and $\beta_{\text{interaction}}$ denote the effects of treatments A and B, and their interaction; and ε_i is a random error term for patient i. The interaction term ($\beta_{\text{interaction}}$) is calculated as $(x_{00} + x_{11}) - (x_{01} + x_{10})$.

Under a factorial analysis, this model is simplified to the following:

$$y_i = \alpha + \beta_{A,\text{factorial}} + \beta_{B,\text{factorial}} + \varepsilon_i \quad (2)$$

where $\beta_{A,\text{factorial}}$ is calculated as $[(x_{10} - x_{00}) + (x_{11} - x_{01})]/2$ (assuming an equal sample size for each treatment combination), and similarly for $\beta_{B,\text{factorial}}$. A factorial analysis will be unbiased when treatments A and B do not interact. However, in the presence of a true interaction, the treatment effect will be biased by a factor of $\beta_{\text{interaction}}/2$. In the MIST2 trial, for example, the bias under a factorial analysis would be 7.2 (half the size of the interaction term), leading to a type I error rate of over 37.1%. Even

with a smaller interaction (e.g. 7.2), the type I error rate would still be 12.9% (derivations shown in the Appendix).

Therefore, a two-stage analysis is often adopted to guard against bias in the presence of an interaction. A preliminary test is performed on $\beta_{\text{interaction}}$ from model (1); if the test is not significant, a factorial analysis will be performed using model (2). If the interaction test is significant, the trial will be regarded as having four separate treatment arms, and the three active treatment groups (A, B, and A + B) will be compared with the control group (neither A nor B). Under a four-arm analysis, the model becomes the following:

$$y_i = \alpha + \beta_{\text{A,four-arm}} + \beta_{\text{B,four-arm}} + \beta_{\text{AB,four-arm}} + \varepsilon_i \tag{3}$$

where $\beta_{\text{A,four-arm}}$ is calculated as $x_{10} - x_{00}$, and similarly for $\beta_{\text{B,four-arm}}$ and $\beta_{\text{AB,four-arm}}$. It should be noted that this model is equivalent to model (1), although the parameterisations are different.

The idea behind the two-stage analysis is that it allows a factorial analysis to be performed when there is no interaction (leading to increased power) and a four-arm analysis when there is an interaction (giving unbiased treatment estimates).

However, the two-stage analysis is not perfect; factorial trials are generally underpowered to detect interactions [5]. Therefore, the two-stage method often leads to a factorial analysis, even in the presence of a moderate or even large interaction, leading to biased estimates and increased type I error rate.

A second issue with the two-stage analysis is that the estimate of the interaction term ($\beta_{\text{interaction}}$) is highly correlated ($\rho = -0.71$) with the treatment effect estimates for treatments A and B from a four-arm analysis ($\beta_{\text{A,four-arm}}$ and $\beta_{\text{B,four-arm}}$) (results shown in the Appendix). Therefore, when the interaction test is significant, the estimates from a four-arm analysis will be more extreme than they would be otherwise, indicating that the treatment effects from a four-arm analysis after a significant interaction will be biased.

The Appendix shows that when there is no true interaction, the expected bias for $\beta_{\text{A,four-arm}}$ and $\beta_{\text{B,four-arm}}$ after a significant interaction test (with a continuous outcome) is as follows:

$$\text{bias}\left(\hat{\beta}_{\text{four-arm}} | \text{abs}\left(\frac{\hat{\beta}_{\text{interaction}}}{\text{SE}(\beta_{\text{interaction}})}\right) \geqslant Z_{1-\alpha/2}\right) \geqslant \text{abs}\left(\frac{Z_{1-\alpha/2}\sigma}{\sqrt{n/4}}\right) \tag{4}$$

where n is the overall sample size, σ is the residual standard deviation, $Z_{1-\alpha/2}$ is the inverse of the cumulative normal distribution and α is the significance level for the interaction test.

We see from Equation (4) that by only using the four-arm analysis when the interaction term is significant, the expected treatment effect from a four-arm analysis will be biased. The size of this bias depends on the standard deviation, the sample size and the significance level for the interaction test. For the MIST2 trial ($n = 210$, SD = 16, 5% significance level), the bias would be >4.3 (approximately a 10% relative change). This bias can be either positive or negative, depending on the direction of the estimated interaction (because of the negative correlation between the estimate from the four-arm analysis and the interaction term, the bias from the four-arm analysis will be in the opposite direction as the estimated interaction).

It should be noted that this bias only applies to the main effects from a four-arm analysis (i.e. A vs control and B vs control); the treatment effect for A + B is uncorrelated with the interaction term and so will be unbiased when the interaction is significant. Additionally, the treatment effect estimates from a factorial analysis are uncorrelated with the estimate of the interaction and will be unbiased when the interaction test is not significant, and no true interaction exists.

These results indicate that the two-stage analysis can be biased even when the treatments do not interact. This is because a false-positive interaction will be found 5–10% of the time (depending on the alpha level), and when this occurs, the analysis from a four-arm trial will be biased. This bias could be in favour of either the treatment or control arm, depending on whether the estimated interaction is positive or negative, and will lead to inflated type I error rates.

This is demonstrated in Figure 1, which shows estimated treatment effects for $\beta_{\text{factorial}}$ or $\beta_{\text{four-arm}}$ after a false-positive interaction (i.e. when a significant interaction is found and the true interaction is 0). Figure 1A shows that a factorial analysis gives unbiased treatment effects; however, Figure 1B shows estimates from a four-arm analysis systematically deviate from the true treatment effect. The distribution of estimates from $\beta_{\text{four-arm}}$ is bimodal; this is because the direction of the bias depends on the direction of the estimated interaction term. Very few of the estimates are close to the true treatment effect. This highlights the reason the two-stage analysis leads to bias; after a false-positive interaction, unbiased estimates from a factorial analysis are replaced with biased estimates from a four-arm analysis.

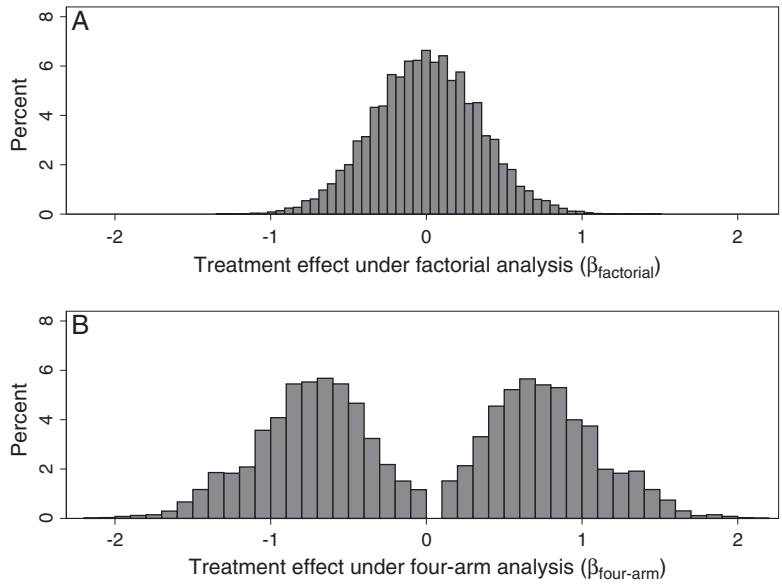

Figure 1. Estimated treatment effects (log(OR)) for factorial and four-arm analyses after a false-positive interaction (true log(OR) = 0). Data were generated using model (1) from the manuscript, with a sample size of 150 patients, and all treatment effects and the interaction term set to log(OR) = 0. A total of 1 000 000 simulations were performed, and a 5% significance level for the interaction test was used. Figures 1A and B show estimated log(OR) for treatment after a false-positive interaction using a factorial analysis and a four-arm analysis, respectively. Under a two-stage analysis, the unbiased treatment effects from a factorial analysis (Figure 1A) are replaced by biased estimates from a four-arm analysis (Figure 1B). The direction of the bias under a four-arm analysis depends on the direction of the estimated interaction term.

It can also be shown that even when a true interaction is detected, the estimated treatment effects from a four-arm analysis ($\beta_{A,\text{four-arm}}$ and $\beta_{B,\text{four-arm}}$) will still be biased (Appendix), indicating that the two-stage analysis is not protective against the presence of an interaction. It is interesting to note that, in the presence of a true interaction, if the interaction is not detected, the bias in the estimated treatment effect will be in the same direction as the interaction; however, if the interaction is detected, the bias in the estimated treatment effect will be in the opposite direction.

4. Simulation study

We performed a simulation study to assess the effects of a two-stage analysis on bias and the type I error rate in a variety of scenarios. We performed two sets of simulations, one using continuous outcomes and the other using binary outcomes.

For all simulations, patients were randomised to one of four treatment groups (treatment A only, treatment B only, treatment A + treatment B, or neither treatment) using permuted block randomisation with a block size of four (as this is the most commonly used block size [6]). We used 5000 replications for each scenario. We assessed the bias and type I error rate for treatment A only (as results for treatment B would be the same).

We performed a preliminary interaction test at the pre-specified level (information on significance levels can be found in the following text). If the interaction was significant, a four-arm analysis was performed ($\beta_{A,\text{four-arm}}$ from model (3)). If the interaction was not significant, a factorial analysis was performed ($\beta_{A,\text{factorial}}$ from model (2)). A significance level of 5% was used to test treatment effects.

4.1. Continuous outcomes

Simulations were based on the MIST2 dataset. The sample size was set to 210 patients, and outcomes were generated from model (1). ε was generated from a normal distribution with mean 0 and standard deviation 16. Both β_A and β_B were set to 0.

We investigated four different interaction sizes: 14 (as this was approximately the size of the interaction seen in the MIST2), 7 (half the size of the first interaction), 3.5 (a quarter the size of the first interaction) and no interaction. A significance level of 5% was used for the preliminary interaction test.

4.2. Binary outcomes

Latent outcomes were generated from model (1), where y_i represents the log odds for patient i; β_A, β_B and $\beta_{interaction}$ represent the treatment effects from treatments A and B, and their interaction (on a log scale); and ε_i is a random error term from a logistic distribution. Patient outcomes were set to 1 if $y_i > 0$ and 0 if $y_i < 0$. Both β_A and β_B were set to 0.

The following parameters were varied: (1) baseline event rates of 25% and 50% were used; (2) for each baseline event rate, both a small and a large sample size were used (see following text for more details); (3) significance levels of 5%, 10% and 20% were used for the preliminary interaction test; and (4) we used four different interaction sizes; OR for interaction of 5.0, 1.50, 1.13 and no interaction (see following text for more details).

Small and large sample sizes were calculated on the basis of 80% power to detect risk ratios of 0.80 and 0.50, respectively, under a factorial analysis. For a baseline event rate of 25%, this resulted in sample sizes of 350 and 2500 total patients, and for a baseline event rate of 50%, this resulted in sample sizes of 150 and 850 (sample sizes were rounded up).

The size of two odds ratios (ORs) for the interaction were chosen on the basis of a review by McAlister *et al.* [1] of 38 interaction estimates from 26 trials. We took the absolute values of the ORs and selected the 50th percentile (OR = 1.13) and the 90th percentile (OR = 1.50). The size of the final interaction was based on the MIST2 dataset; we used the interaction estimate for the outcome 'need for surgery at 3 months' (OR = 4.90; this was rounded up to 5.0).

5. Simulation study results

5.1. Continuous outcomes

Results can be found in Table I. Under no interaction, the two-stage analysis led to biased estimates of treatment effect when the interaction test was significant (bias ±5.2, depending on the direction of the estimated interaction). This led to a type I error rate of 7.0%.

The two-stage analysis had low power to detect small or moderate interactions (12% and 35%, respectively). Estimated treatment effects were biased both when the interaction was not significant (bias 1.8 and 3.5 for small and moderate interactions, respectively) and when the interaction was detected (−3.6 and −2.2, respectively), which led to type I error rates of 13% and 25%, respectively. This demonstrates that in the presence of a true interaction, the two-stage analysis gives biased estimates of treatment effect, regardless of whether the interaction is detected.

The two-stage analysis had high power to detect a large interaction (88%). However, estimates were still slightly biased when the interaction was detected (−0.4) and were extremely biased when the interaction was not detected (6.9), leading to a type I error rate of 13%.

The absolute bias encountered here (approximately 1.8–6.9) corresponds to relative reductions between 4% and 16%, which would be considered clinically relevant in many cases. It should be noted that even in scenarios with relatively small bias (e.g. 1.8), which may not on its own be

Table I. Simulation results for continuous outcomes.

Interaction size	% of times interaction significant	Type I error rate (%)	Estimated treatment effect (true effect = 0)	
			Interaction significant	Interaction not significant
0	5.5	7.0	±5.2*	0.1
3.5	12.1	13.3	−3.6	1.8
7	34.9	24.6	−2.2	3.5
14	88.4	13.3	−0.4	6.9

*Direction of bias depends on the direction of the estimated interaction term.
The α level for the preliminary interaction test was set at 5%.

considered clinically relevant, the bias has the effect of shifting the distribution of estimated treatment effects so that extreme results are more likely, leading to an inflation in the type I error rate.

5.2. *Binary outcomes*

Results can be found in Tables II–V. Under no interaction, the two-stage analysis led to highly biased results when an interaction was detected (range of ORs 1.19 to 2.08), leading to type I error rates between

Table II. Simulation results for binary outcomes with $n = 150$ with baseline event rate 50%.

α level for interaction test	% of times interaction significant	Type I error rate (%)	Estimated OR (true OR = 1)	
			Interaction significant	Interaction not significant
Interaction size 1.0				
0.05	5.5	7.0	0.45 or 2.08*	0.99
0.10	10.3	7.6	0.50 or 1.89*	0.99
0.20	22.4	8.0	0.56 or 1.74*	0.99
Interaction size 1.13				
0.05	5.9	7.1	0.49	1.06
0.10	11.0	7.8	0.53	1.06
0.20	22.9	8.2	0.58	1.06
Interaction size 1.50				
0.05	9.5	10.8	0.54	1.22
0.10	16.5	10.7	0.60	1.21
0.20	30.2	10.5	0.65	1.22
Interaction size 5.00				
0.05	62.6	24.2	0.81	2.09
0.10	74.6	17.5	0.86	2.11
0.20	84.8	11.8	0.91	2.08

*Direction of bias depends on the direction of the estimated interaction term.

Table III. Simulation results for binary outcomes with $n = 850$ with baseline event rate 50%.

α level for interaction test	% of times interaction significant	Type I error rate (%)	Estimated OR (true OR = 1)	
			Interaction significant	Interaction not significant
Interaction size 1.0				
0.05	5.2	6.7	0.73 or 1.38*	1.00
0.10	10.8	7.3	0.76 or 1.32*	1.00
0.20	21.3	7.5	0.78 or 1.27*	1.00
Interaction size 1.13				
0.05	7.9	8.9	0.76	1.06
0.10	14.4	8.9	0.80	1.06
0.20	25.7	9.0	0.82	1.06
Interaction size 1.50				
0.05	31.9	23.0	0.85	1.22
0.10	44.4	19.4	0.88	1.22
0.20	58.1	15.4	0.91	1.22
Interaction size 5.00				
0.05	100	4.9	1.00	NA
0.10	100	4.9	1.00	NA
0.20	100	4.9	1.00	NA

*Direction of bias depends on the direction of the estimated interaction term.

Table IV. Simulation results for binary outcomes with $n = 350$ with baseline event rate 25%.

α level for interaction test	% of times interaction significant	Type I error rate (%)	Estimated OR (true OR = 1)	
			Interaction significant	Interaction not significant
Interaction size 1.0				
0.05	5.1	6.4	0.56 or 1.80*	1.00
0.10	10.6	6.8	0.60 or 1.66*	1.00
0.20	20.7	7.2	0.65 or 1.54*	1.00
Interaction size 1.13				
0.05	6.0	7.0	0.58	1.07
0.10	11.7	7.3	0.63	1.06
0.20	21.4	7.7	0.67	1.07
Interaction size 1.50				
0.05	13.0	14.2	0.65	1.24
0.10	21.1	13.6	0.70	1.24
0.20	34.8	12.3	0.75	1.24
Interaction size 5.00				
0.05	91.8	11.0	0.96	2.44
0.10	95.7	7.5	0.98	2.44
0.20	98.0	5.8	0.99	2.43

*Direction of bias depends on the direction of the estimated interaction term.

Table V. Simulation results for binary outcomes with $n = 2500$ with baseline event rate 25%.

α level for interaction test	% of times interaction significant	Type I error rate (%)	Estimated OR (true OR = 1)	
			Interaction significant	Interaction not significant
Interaction size 1.0				
0.05	5.1	6.0	0.80 or 1.25*	1.00
0.10	9.9	6.7	0.82 or 1.23*	1.00
0.20	19.5	7.1	0.85 or 1.19*	1.00
Interaction size 1.13				
0.05	9.8	9.9	0.84	1.06
0.10	17.2	9.8	0.87	1.06
0.20	30.3	9.5	0.89	1.06
Interaction size 1.50				
0.05	61.3	28.5	0.94	1.24
0.10	72.7	20.7	0.95	1.24
0.20	82.9	14.6	0.97	1.25
Interaction size 5.00				
0.05	100	5.0	1.00	NA
0.10	100	5.0	1.00	NA
0.20	100	5.0	1.00	NA

*Direction of bias depends on the direction of the estimated interaction term.

6.0% and 8.0%. Increasing the significance level for the preliminary interaction test led to smaller bias but increased type I error rates. This is because although increasing the significance level for the interaction will reduce the bias in the four-arm analysis (as per Equation (4)), this results in the treatment effect estimates being biased more often than they would under a smaller significance level, leading to more opportunities for false positives. Likewise, increasing the sample size reduced bias slightly but had little effect on type I error rates.

The two-stage analysis generally had low power to detect small or moderate interactions. As expected, estimated ORs were biased regardless of whether the true interaction was detected or not; when it was not detected, ORs were biased upwards. However, when the interaction was detected, the bias was downwards. As the size of the interaction increased, the bias was reduced after a significant interaction test but increased after a nonsignificant test.

The bias and type I error rates were large in all scenarios regardless of the sample size, the baseline event rate, the size of the interaction or the significance level for the preliminary interaction test; the one exception was for a very large interaction (i.e. OR = 5.00) and a large sample size. In this scenario, the power to detect the interaction was 100%; therefore, the four-arm analysis was always used, and results were unbiased. However, even when the power to detect the interaction was high (e.g. 80–95%), results were still largely biased, leading to type I error rates between 7.5% and 14.6%.

6. Discussion

Randomised factorial trials present an efficient method of assessing multiple treatments in a single trial. However, they rely on the assumption of no interaction between treatment arms, which is rarely known in advance. Reviews have shown that many trials do not assess this assumption (or at least do not report assessing this assumption) [1, 2]. However, as demonstrated by the MIST2 trial, performing a factorial analysis without first assessing the size of the interaction can be highly misleading and could lead to an ineffective or even potentially harmful treatment being adopted or an effective treatment being missed. Even a small interaction can bias treatment effects enough to result in large inflations of the type I error rate.

It is therefore recommended to assess the size of the interaction between treatments [1, 2]. This approach can be formalised as the two-stage analysis, where the method of analysis depends on the results of a preliminary test for interaction; if the test is not significant, a factorial analysis is performed. If the test is significant, a four-arm analysis is performed.

Although attractive in principle, the two-stage analysis has the following drawbacks: (a) it can lead to biased estimates and inflated type I error rates in the presence of no interaction; (b) it is not protective when interactions exist, as it has low power to detect a true interaction [5] (and so will often lead to a biased factorial analysis); and (c) even if a true interaction is detected, the analysis from a four-arm trial will still be biased. Neither increasing the significance level for the interaction test nor powering the trial for an interaction is effective in eliminating bias (except in the extreme situation where a trial has 100% power to detect the interaction).

We therefore recommend the two-stage approach not be used in the analysis of factorial RCTs, as it can lead to biased results and inflated type I error rates, regardless of whether there is a true interaction.

This is very similar to the issue of carryover in crossover trials, where a two-stage analysis was the preferred analysis method. A preliminary test of carryover (equivalent to testing a treatment by period interaction) was performed; if significant, only data from the first period would be analysed. Otherwise, all data would be used in the analysis. However, in 1989, Freeman [3] showed that a two-stage analysis for crossover trials gave biased results and inflated the type I error rate, even if no carryover was present. As a result, it is recommended that crossover trials be designed to ensure no carryover (for example, by increasing the length of the washout period) and then performing an analysis based on both treatment periods, regardless of any tests for carryover [7].

However, it is unlikely that a similar approach for factorial trials (i.e. assuming no interaction and using a factorial analysis regardless of any interaction test) would gain widespread acceptance. There is no equivalent design feature such as a washout period with which to reduce the possibility of an interaction. It is possible that some trials may be designed with the intention of limiting the chance of an interaction, for example, by giving treatments at different time points or assessing treatments that work along different biological pathways. However, even in these scenarios, the possibility of an interaction still remains, and so an analysis that does not assess the interaction may give misleading results and cannot be recommended.

It will also be very difficult to assume no interaction on the basis of prior research. Even if previous data that assess the interaction are available, interaction estimates are so variable that it would be almost impossible to exclude at least a small or moderate interaction. Likewise, interactions cannot be ruled out on the basis of *in vitro* studies, as this may not translate to *in vivo* studies.

We therefore recommend that authors present both factorial and four-arm analyses to ensure the two analysis methods agree. If results do agree, then investigators can be confident in their conclusions. This

strategy is not ideal though, as estimates from the two analysis methods will generally be moderately different because of random variation. In particular, interpreting results when the two analysis methods disagree will be difficult. Additionally, the requirement for both analyses to give consistent results will lead to a large reduction in power, as most trials will have been powered on a factorial analysis but will also be assessed using a four-arm analysis, which uses only half the patients.

A preferable alternative may be to design trials as multi-arm (i.e. four separate treatment groups). This design has several advantages over that of a factorial trial. First, it ensures an unbiased comparison between treatment arms regardless of any interaction. Second, it allows investigators to ensure adequate power by basing sample size calculations on a four-arm comparison rather than a factorial analysis. Third, this approach leads to straightforward interpretation of the results, without the need to compare results from two different analysis methods.

Appendix A

A.1. Increased type I error rate under factorial analysis with an interaction

Let bias_{std} denote the standardised bias for an estimated treatment effect (that is, the bias divided by the standard error of the estimate). For a significance level of 5%, the type I error (α) is calculated as follows:

$$\alpha = 1 - (\Phi(1.96 - \text{bias}_{\text{std}}) - \Phi(-1.96 - \text{bias}_{\text{std}}))$$

A.2. Correlation between interaction and treatment effects

We first show then that the interaction estimate and the factorial analysis estimates are uncorrelated. Assume the variance for all patients (conditional on treatment) is σ^2.

Then, $V(X_{a,b}) = \frac{\sigma^2}{n/4}$ for all a, b (where n is the overall sample size). Then,

$$\begin{aligned}\text{Cov}\left(\beta_{\text{interaction}}, \beta_{\text{A,factorial}}\right) &= \text{Cov}\left((x_{00} + x_{11}) - (x_{01} + x_{10}), \frac{1}{2}\left[(x_{10} - x_{00}) + (x_{11} - x_{01})\right]\right) \\ &= \frac{1}{2}\left[-V\left(x_{00}\right) - V\left(x_{10}\right) + V\left(x_{11}\right) + V\left(x_{01}\right)\right] = 0\end{aligned}$$

Similar results can be shown for $\beta_{\text{B,factorial}}$. We now show that the interaction estimate and four-arm analysis estimates *are* correlated:

$$\begin{aligned}\text{Cov}(\beta_{\text{interaction}}, \beta_{\text{A,four-arm}}) &= \text{Cov}\left((x_{00} + x_{11}) - (x_{01} + x_{10}), (x_{10} - x_{00})\right) = -V(x_{00}) - V(x_{10}) \\ &= -\frac{2\sigma^2}{n/4}\end{aligned}$$

The correlation therefore is as follows:

$$\text{Corr}(\beta_{\text{interaction}}, \beta_{\text{A,four-arm}}) = \frac{-\frac{2\sigma^2}{n/4}}{\sqrt{\left(\frac{4\sigma^2}{n/4}\right)\left(\frac{2\sigma^2}{n/4}\right)}} = -\frac{1}{\sqrt{2}} = -0.71$$

Similar results can be shown for $\beta_{\text{B,four-arm}}$. A similar method can be used to show that $\beta_{\text{interaction}}$ and $\beta_{\text{AB,four-arm}}$ are uncorrelated.

A.3. Bias in four-arm analysis after a significant interaction

Assuming that $\beta_{\text{interaction}}$ and $\beta_{\text{four-arm}}$ follow a bivariate normal distribution with correlation as above, then the conditional expectation of $\beta_{\text{four-arm}}$ (for either treatment A or B) given that $\beta_{\text{interaction}}$ is significant is as follows:

$$\begin{aligned}&E\left(\hat{\beta}_{\text{four-arm}} | \text{abs}\left(\frac{\hat{\beta}_{\text{interaction}}}{\text{SE}(\beta_{\text{interaction}})}\right) \geqslant Z_{1-\alpha/2}\right) \geqslant \\ &E\left(\hat{\beta}_{\text{four-arm}}\right) + \frac{\text{SE}(\beta_{\text{four-arm}})}{\text{SE}(\beta_{\text{interaction}})}\rho\left(Z_{1-\alpha/2}\text{SE}(\beta_{\text{interaction}}) - E(\beta_{\text{interaction}})\right)\end{aligned}$$

where ρ denotes the correlation between the estimates of $\beta_{\text{four-arm}}$ and $\beta_{\text{interaction}}$.

When a true interaction is present, this expression is the following:

$$E\left(\hat{\beta}_{\text{four-arm}}|\text{abs}\left(\frac{\hat{\beta}_{\text{interaction}}}{\text{SE}(\beta_{\text{interaction}})}\right) \geqslant +Z_{1-\alpha/2}\right) \geqslant \beta_{\text{four-arm}} + \left(\frac{Z_{1-\alpha/2}\sigma}{\sqrt{n/4}} - \beta_{\text{interaction}}\right)$$

When there is no interaction, this expression reduces to the following:

$$E\left(\hat{\beta}_{\text{four-arm}}|\text{abs}\left(\frac{\hat{\beta}_{\text{interaction}}}{\text{SE}(\beta_{\text{interaction}})}\right) \geqslant Z_{1-\alpha/2}\right) \geqslant \beta_{\text{four-arm}} + \frac{Z_{1-\alpha/2}\sigma}{\sqrt{n/4}}$$

Acknowledgements

The author would like to thank Tim Morris and Dan Bratton for their helpful comments on the manuscript. The author would also like to thank two anonymous referees whose comments helped to improve the manuscript.

References

1. McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: a systematic review. *Japan Automobile Manufacturers Association* May 21 2003; **289**(19):2545–2553. DOI: 10.1001/jama.289.19.2545.
2. Montgomery AA, Astin MP, Peters TJ. Reporting of factorial trials of complex interventions in community settings: a systematic review. *Trials* 2011; **12**:179. DOI: 10.1186/1745-6215-12-179.
3. Freeman PR. The performance of the two-stage analysis of two-treatment, two-period crossover trials. *Statistics in Medicine* 1989; **8**(12):1421–1432.
4. Rahman NM, Maskell NA, West A, Teoh R, Arnold A, Mackinlay C, Peckham D, Davies CWH, Ali N, Kinnear W, Bentley A, Kahan BC, Wrightson JM, Davies HE, Hooper CE, Gary Lee YC, Hedley EL, Crosthwaite N, Choo L, Helm EJ, Gleeson FV, Nunn AJ, Davies RJO. Intrapleural use of tissue plasminogen activator and DNase in pleural infection. *The New England Journal of Medicine* Aug 11 2011; **365**(6):518–526. DOI: 10.1056/NEJMoa1012740.
5. Green S, Liu PY, O'Sullivan J. Factorial design considerations. *Journal of Clinical Oncology* Aug 15 2002; **20**(16): 3424–3430.
6. Kahan BC, Morris TP. Reporting and analysis of trials using stratified randomisation in leading medical journals: review and reanalysis. *British Medical Journal* 2012; **345**:e5840.
7. Senn S. Cross-over trials in Statistics in Medicine: the first '25' years. *Statistics in Medicine* Oct 30 2006; **25**(20): 3430–3442. DOI: 10.1002/sim.2706.