Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Improper analysis of trials randomised using stratified blocks or minimisation

Brennan C. Kahan[a]*and Tim P. Morris[a]

Many clinical trials restrict randomisation using stratified blocks or minimisation to balance prognostic factors across treatment groups. It is widely acknowledged in the statistical literature that the subsequent analysis should reflect the design of the study, and any stratification or minimisation variables should be adjusted for in the analysis. However, a review of recent general medical literature showed only 14 of 41 eligible studies reported adjusting their primary analysis for stratification or minimisation variables. We show that balancing treatment groups using stratification leads to correlation between the treatment groups. If this correlation is ignored and an unadjusted analysis is performed, standard errors for the treatment effect will be biased upwards, resulting in 95% confidence intervals that are too wide, type I error rates that are too low, and a reduction in power. Conversely, an adjusted analysis will give valid inference. The extent of this issue is explored using simulation for continuous, binary, and time-to-event outcomes where treatment is allocated using stratified block randomisation or minimisation.

Keywords: Clinical trials; Covariates; Adjustment; Minimisation; Stratification; Unadjusted analysis

1. Introduction

Well conducted randomised controlled trials are considered the gold standard for unbiased comparison of treatments as they ensure there are no systematic differences between treatment groups. The most basic method of allocating patients to a treatment is simple randomisation[1], where the probability of being assigned to either treatment is the same for all patients. However, this can lead to chance imbalances in important prognostic factors between treatment groups as well as the overall proportion receiving each treatment. A number of randomisation procedures have been developed that promote balance in key prognostic factors between treatment groups[2]. These include stratified block randomisation[3], stratified biased coin randomisation[4], stratified urn randomisation[5], optimum biased coin[6], and minimisation[7]. A recent article by Pond et al[8] showed that 85% of cancer clinical trials used at least one baseline variable in their randomisation, and Scott et al[9] showed that in 2001 45% of randomised trials in the Lancet or New England Journal of Medicine used either stratification or minimisation. These are the two most popular methods of balancing covariates in clinical trials, despite concerns surrounding both[10, 11], and will be the focus of this paper.

Stratified block randomisation involves grouping patients into strata defined by baseline characteristics, and performing block randomisation within each stratum. This ensures that the stratifying variables are balanced on completion of each block. For a small number of variables this method will work well, but tends to break down as the number grows larger. Four variables with two levels each leads to $2^4 = 16$ strata. If certain combinations of balancing variables are rare, some strata may not have even one completed block. Minimisation is a method that overcomes this problem. For each new patient entering a trial the baseline characteristics of patients in the two treatment groups are summarised and the patient is allocated to the treatment which would provide the best marginal balance in terms of prognostic factors. In practice the patient is usually assigned to the preferred treatment with probability π where $\pi > 0.5$. When neither treatment offers an advantage in terms of marginal balance, $\pi = 0.5$ as with simple randomisation.

[a]*MRC Clinical Trials Unit, Aviation House, 125 Kingsway, London, WC2B 6NH, U.K.*
[*]*Correspondence to: MRC Clinical Trials Unit, Aviation House, 125 Kingsway, London, WC2B 6NH, U.K. Email: brk@ctu.mrc.ac.uk*

Although stratified randomisation and minimisation are common in clinical trials, there remains some confusion over whether it is necessary to adjust for the stratification or minimisation variables in the analysis of the trial. A majority of the statistical literature suggests any baseline variable involved in randomisation should also be included in the analysis[9, 12, 13, 14, 15, 16, 17, 18] though some argue this is not always sensible[19]. Many textbooks and articles that describe stratification and minimisation do not mention any implications for the analysis[20, 21, 22, 23, 24, 25, 26]. It is important that this issue is clarified: an incorrect analysis could potentially result in a beneficial treatment being denied to patients or a treatment that is not beneficial being adopted.

A number of simulation studies have shown that ignoring the stratification or minimisation variables in the analysis may lead to invalid tests of significance[27, 28, 29, 30, 31]. Type I error rates are shown to be too low for continuous and time-to-event outcomes. Binary outcomes have not, to our knowledge, been studied using simulation.

Throughout, we will refer to variables which are used in the allocation scheme as 'balancing' variables, stratified block randomisation as 'stratification', and treatment allocation as 'randomisation' (since we only consider minimisation with some random element, as is usual in practice).

This article investigates the effects of ignoring stratification or minimisation variables in the analysis. In Section 2 we show that an unadjusted analysis after stratification leads to standard errors for the treatment effect that are biased upwards. In Section 3 we use simulation to show that stratification leads to correlation between the treatment groups, and investigate what impact this has on coverage rates if ignored in the analysis. We then present a series of simulations based on real trial data to determine to what extent these issues are likely to arise in practice. Section 4 reviews a cross section of recently reported randomised controlled trials to see how often balancing variables are used in randomisation, and how these are dealt with in the analysis. Section 5 is a discussion, and makes recommendations for trials randomised using stratified blocks or minimisation.

2. Correlated outcomes

We initially consider the simple case of matched pairs to investigate how ignoring the matching will affect the analysis, before considering the issue under general stratification.

Consider a study with a continuous outcome in which each patient is matched with another patient on some baseline variables. In each matched pair one patient is assigned to the treatment group, the other to the control group. Given n matched pairs, this can be thought of as a stratified trial with n strata, 2 patients per stratum and a total of $2n$ patients. This is similar to a 2×2 crossover trial where each patient receives both treatments.

Let Y_{1j} and Y_{2j} denote patient outcomes from the j^{th} matched pair, assigned to treatments 1 and 2 respectively. Assume that $\text{Var}(Y_{1j}) = \text{Var}(Y_{2j}) = \sigma^2$, the correlation between patients in the j^{th} pair is ρ, and the correlation between patients with different j is 0. It follows that $\text{Cov}(Y_{1j}, Y_{2j}) = \rho\sigma^2$.

Let $\bar{Y}_1$ and $\bar{Y}_2$ denote the mean outcomes in treatment groups 1 and 2 respectively. The variance of the treatment difference, $\bar{Y}_1 - \bar{Y}_2$, can be found using the following formula:

$$\text{Var}(\bar{Y}_1 - \bar{Y}_2) = \text{Var}(\bar{Y}_1) + \text{Var}(\bar{Y}_2) - 2\text{Cov}(\bar{Y}_1, \bar{Y}_2) \tag{1}$$

This is simplified to

$$\text{Var}(\bar{Y}_1 - \bar{Y}_2) = \text{Var}(\bar{Y}_1) + \text{Var}(\bar{Y}_2)$$

for a two sample t test, where $\bar{Y}_1$ and $\bar{Y}_2$ are assumed to be independent. It is easy to show when patients are matched in a pair, as in the above example, this assumption is not true:

$$\begin{aligned} \text{Cov}(\bar{Y}_1, \bar{Y}_2) &= \text{Cov}\left(\frac{1}{n}\sum_{j=1}^{n} Y_{1j}, \frac{1}{n}\sum_{j=1}^{n} Y_{2j}\right) && (2) \\ &= \frac{1}{n^2}\sum_{j=1}^{n} \text{Cov}(Y_{1j}, Y_{2j}) && (3) \\ &= \frac{\rho\sigma^2}{n} && (4) \end{aligned}$$

The correlation between $\bar{Y}_1$ and $\bar{Y}_2$ is then ρ, which is the same as the correlation between matched pairs.

The variance of the treatment difference not accounting for this correlation between $\bar{Y}_1$ and $\bar{Y}_2$ is

$$\frac{2\sigma^2}{n}$$

However, given that $\bar{Y}_1$ and $\bar{Y}_2$ are correlated by design, the *true* variance, continuing from (1), is

$$\frac{2\sigma^2}{n} - \frac{2\rho\sigma^2}{n} = \frac{2\sigma^2}{n}(1-\rho).$$

We then see that by not accounting for the correlation between $\bar{Y}_1$ and $\bar{Y}_2$, the variance of the treatment difference will be biased upwards by a factor of $(1-\rho)^{-1}$. Parzen et al[32] showed similar results for trials that are balanced on centre. This issue arises because of the covariance between $\bar{Y}_1$ and $\bar{Y}_2$. This is similar under general stratification. By assuming a correlation of ρ for all patients within a stratum, $\mathrm{Cov}(\bar{Y}_1, \bar{Y}_2)$ cannot be equal to 0, because the covariance of any two patients in the same stratum is non-zero, and always positive. This means that under stratification in general, standard errors of treatment effect will be biased upwards (though not necessarily to the same extent as in the case of matched-pairs), leading to confidence intervals that are too wide and p-values that are too large.

It is well known that an unadjusted analysis is inappropriate for a matched-pairs study (where pair would be included in the analysis) and for crossover trials (where subject would be included in the analysis). Although the correlation within strata of a parallel group trial may be smaller than the within-subject correlation for a crossover trial, the considerations are the same. The extent of the problem depends largely on the within-stratum correlation (i.e. the intraclass correlation). If this is non-negligible then there will be bias in the estimate of $\mathrm{Var}(\bar{Y}_1 - \bar{Y}_2)$. The within-stratum correlation depends on the strength of the relationship between the stratification variables and the outcome. The stronger the relationship, the larger the within-stratum correlation. Additionally, as more balancing variables are used in the randomisation process, the within-stratum correlation will increase (assuming the balancing variables are associated with the outcome). In practice, this means we should always expect non-negligible within-stratum correlation in stratified trials, as variables should only be used in balancing if they are expected to be related to outcome.

Stratification will also lead to correlated treatment groups in the cases of binary or survival outcomes. However, this is complicated by the fact that the standard error of the treatment effect is expected to increase when covariates are fitted[33]. For the binary case, Robinson and Jewell[34] show that adjusting for a balanced covariate that is associated with the outcome will cause a proportionally larger increase in the treatment effect estimate than in its standard error, meaning

$$\frac{\hat{\beta}_{\mathrm{adj}}}{SE(\hat{\beta}_{\mathrm{adj}})} > \frac{\hat{\beta}_{\mathrm{unadj}}}{SE(\hat{\beta}_{\mathrm{unadj}})},$$

where $\hat{\beta}$ is the estimate of the log–odds-ratio and the subscript denotes whether this is adjusted or unadjusted. This implies an increase in power for adjusted analyses. It is important to investigate what effect an unadjusted analysis will have in practice, given that the standard error is expected to increase after adjustment.

3. Simulation studies

3.1. General considerations

Simulation is used to investigate the impact of an unadjusted analysis after stratification or minimisation. We initially show that the correlation between treatment groups is *introduced* by stratified randomisation, as shown in (2). Secondly, we examine coverage rates of 95% confidence intervals for the treatment difference as the effect of the stratification variable on outcome increases. Finally, we present simulations based on data from real trials to explore what impact unadjusted analyses might have on coverage and power in practice. All simulations were done using Stata 11.1.

3.2. Correlation between treatment groups

This simulation study investigates the correlation between treatment groups after both simple randomisation and stratification. Two hundred patients were simulated for each replication. Eight thousand replications were used to estimate the correlation between treatment groups. A random sample of 200 replications is shown in Figure 1. Two hundred replications were chosen to maintain clarity of the graph. A continuous response Y_i for the i^{th} patient was simulated as:

$$Y_i = \alpha + \beta X_{\mathrm{treat}} + \gamma \boldsymbol{X}_{\mathrm{strat}} + \varepsilon_i, \tag{5}$$

where X_{treat} is a binary indicator equal to one if the patient receives the treatment, β is the additive effect of treatment on outcome, $\boldsymbol{X}_{\mathrm{strat}}$ are the stratification variables (here only one binary stratification variable is used) and γ are the regression coefficients corresponding to $\boldsymbol{X}_{\mathrm{strat}}$. γ was assigned values of 3 and then 6 and $\varepsilon_i \sim N(0,1)$. These effects are extreme and unlikely to occur in practice but are used to illustrate the point that the correlation between treatment groups increases as γ

increases. It should be noted that the correlation between treatment groups is independent of the treatment effect, and will be the same for any value of β (here, we used $\beta = 0$).

Figure 1 shows the strong correlation between $\bar{Y}_1$ and $\bar{Y}_2$ (where the subscripts denote the treatment group) under stratified randomisation. Under simple randomisation however, $\bar{Y}_1$ and $\bar{Y}_2$ are uncorrelated. This demonstrates that the correlation is introduced by the process of stratification. Under simple randomisation treatment groups will be uncorrelated, and so an unadjusted analysis will give valid results. The correlation between $\bar{Y}_1$ and $\bar{Y}_2$ increases as the strength of the relationship between the stratification variable and outcome increases.

3.3. Coverage

This simulation study investigates the coverage of 95% confidence intervals for the treatment effect as the effect of the stratifying variable increased. The coverage rate was defined as the proportion of times the 95% confidence interval contained β. Eight thousand replications were used so that if the true coverage was 95% then 8000 simulations would estimate a confidence interval for coverage to within $\pm 0.5\%$.

Data were again generated using model (5). Here, $\varepsilon_i \sim N(0,1)$ for all simulations. As in Section 3.2 the choice of β is unimportant as the coverages rates are independent of treatment effect in this scenario, and so the results will be the same for any value of β (again β was set to 0). X_{strat} was a binary stratification variable, and patients were assigned to each level of X_{strat} with probability 0.5. γ was set to values of 0 to 3 in increments of 0.2. Each simulated patient was randomised twice: once using simple randomisation and once using stratified randomisation with a block size of eight. A block size of eight was chosen so that it would be small enough to promote balance, but not so small as to be unrealistic. Both adjusted and unadjusted analyses were performed. Four scenarios were therefore considered for each simulated dataset:

1. Simple randomisation, unadjusted analysis
2. Simple randomisation, adjusted analysis
3. Stratified randomisation, unadjusted analysis
4. Stratified randomisation, adjusted analysis

For each replication a 95% confidence interval for the treatment effect was calculated. These confidence intervals were then used to calculate the coverage for each scenario.

Figure 2 shows results for $n = 250$ (similar results, not shown, were observed for $n = 100$, 500 and 1000). This shows that ignoring balancing variables in the analysis will lead to confidence intervals that do not have nominal coverage. At the extreme, the coverage levels approach 100%. Even when γ is small relative to the residual standard deviation the coverage is affected. In contrast, adjusting for balancing factors in the analysis or using simple randomisation always gave nominal coverage.

3.4. Simulations based on real trial data

Thus far we have only considered the impact of an unadjusted analysis in the case of stratified randomisation with a continuous outcome. In this section we examine the impact of an unadjusted analysis for continuous, binary, and time-to-event outcomes after both stratification and minimisation have been used. Simulations were carried out using parameters from real trial data. Five trials were used: FASTER[35], MIST2[36], RE01, GBSG, and PBC. Information on RE01, GBSG and PBC can be found in Royston and Sauerbrei[37]. Of these five trials, only FASTER and MIST2 used balancing variables in their randomisation. For the other trials, several prognostic factors were chosen as candidate balancing variables. Sections 3.4.1 to 3.4.5 give more information on each individual trial. As above, 8000 replications were used for each scenario.

Balancing variables were generated from a multivariate normal distribution. Variances and covariances were based on the original dataset so that the proportion of patients in each stratum was similar to the original study. Binary and ordinal data were then categorised based on cut-points specified to give the desired proportions. Continuous observations that fell outside the range of the original dataset were replaced with the minimum or maximum from the original dataset.

Sample sizes of 100, 200, 500, and 1000 were used, as well as the original sample size from each study. The treatment effect was chosen to give approximately 80% power when using the original sample size.

Stratification was used with random permuted blocks of size 8. Minimisation allocated treatments in turn and then calculated the marginal covariate imbalance for each. The preferred treatment was then chosen by a biased coin with $\pi = 0.8$. (This reflects practice since, with the exception of permuted blocks, deterministic methods are generally discouraged [12, 11].) For stratification and minimisation, continuous variables were dichotomised at their observed mean from the original dataset, but were included in the analysis as continuous variables. The GBSG and MIST2 trials were exceptions, as progesterone receptor status from GBSG was dichotomised at its observed median due to skewness, and baseline pleural effusion in MIST2 was dichotomised at the same value as used in the original trial.

Coverage, power, and standard errors were compared between adjusted and unadjusted analyses. For each analysis we defined a model-based standard error and an empirical standard error as follows:

$$SE_{\text{model}} = \frac{1}{r}\sum_{j=1}^{r} SE_j$$

and

$$SE_{\text{empirical}} = \sqrt{\frac{1}{r-1}\sum_{j=1}^{r}(\hat{\beta}_j - \beta)^2},$$

where r denotes the number of simulations, and SE_j denotes the standard error of the treatment effect for the j^{th} simulation. We define the % bias in model-based standard errors as

$$100\left(\frac{SE_{\text{model}} - SE_{\text{empirical}}}{SE_{\text{empirical}}}\right).$$

Values > 0 indicate model-based standard errors are biased upwards, while < 0 indicates they are biased downwards.

Continuous outcomes were again generated from model (5), as in sections 3.2 and 3.3. Binary outcomes were generated using a latent response model. Latent variables $Y_i^\star$ were generated and the observed responses Y_i were classified as 1 if $Y_i^\star > 0$ and 0 otherwise. The latent response was simulated from a model similar to (5) but where ε_i follows a logistic distribution with mean 0 and variance $\frac{\pi^2}{3}$. Here β and γ represent log-odds ratios. Because the expected treatment effect would be different for adjusted and unadjusted models (even when $\boldsymbol{X}_{\text{strat}}$ are balanced[33]), coverage of the treatment effect was compared when β was set to 0 so that the expectation was the same for adjusted and unadjusted models.

Survival outcomes were generated from a Weibull distribution with baseline hazard functions based on the observed data. Survival times were generated as suggested by Bender et al[38] by

$$T = H_0^{-1}[-\ln(U)\exp(-(\beta X_{\text{treat}} + \gamma \boldsymbol{X}_{\text{strat}}))],$$

where H_0 is the cumulative baseline hazard function and $U \sim \text{Uniform}(0,1)$. This model for survival times implies a proportional hazards model

$$h(t|x) = h_0(t)\exp(\beta X_{\text{treat}} + \gamma \boldsymbol{X}_{\text{strat}}).$$

As with binary outcomes, coverage of the treatment effect was compared with β set to 0.

3.4.1. FASTER The FASTER trial (Function After Spinal Treatment: Exercise and Rehabilitation) was a 2×2 factorial trial which tested the effects of rehabilitation and an educational booklet on Oswestry disability index (ODI), a continuous outcome, in 316 participants following back surgery (either a discectomy for herniated disc or decompression for spinal stenosis)[35]. The type of surgery and the operating surgeon were used as balancing variables.

Our simulations simplify the study to a single factor parallel group trial. Two sets of simulations were performed: the first set used the type of surgery and surgeon as balancing factors, as in the original trial; the second excluded surgeon and instead used type of surgery and baseline ODI as balancing variables. Additionally, in order to investigate the effect of block size on the results, the second set of simulations used block sizes of 2, 8, and 32.

For all simulations the residual standard deviation was set to 18, 46% of participants were assumed to have a discectomy, and the regression coefficient for discectomy was -13. For the simulations involving surgeon, there were 23 surgeons in the study, 10 of which operated on less than 10 patients each. This was simplified in our simulations by combining all surgeons with less than 10 patients, resulting in 14 surgeons with the number of patients per surgeon ranging from 10 to 40. This was done to avoid overstratification. The effect of surgeon was included by generating a random effect for each surgeon based on a normal distribution. The variance was chosen so that the intraclass correlation coefficient for surgeon was 0.01, reflecting the observed data. All parameters used for simulation were estimated prior to combining surgeons with less than 10 patients. For the second set of simulations, baseline ODI was generated as $\sim N(46, 19^2)$ and its regression coefficient was set to 0.5.

3.4.2. MIST2 The MIST2 trial was a 2×2 factorial trial testing whether tPA (tissue plasminogen activator) or DNase (deoxyribunoclease) were effective in reducing the size of patients pleural effusion (a continuous outcome)[36]. Patients were randomised to one of four treatment groups using minimisation. The size of the baseline pleural effusion (greater or less than 30% of the hemithorax), whether the patient was purulent, and whether the infection was acquired via the community or in hospital were all used as minimisation factors. For simplicity, simulations were conducted assuming a parallel group design with two treatment groups.

The residual standard deviation was set to 16, 48% of patients were assumed to be purulent, and 12% of patients were assumed to have had a hospital acquired infection. The mean and standard deviation of the size of the baseline pleural effusion were set to 43 and 22 respectively. The regression coefficients were -0.6 for the size of the baseline pleural effusion, 0.9 for purulence, and 6.5 for a hospital acquired infection.

3.4.3. RE01 The RE01 trial compared interferon-α with medroxyprogesterone acetate in patients with metastatic renal carcinoma. Of the 347 patients in the study, 322 (93%) died. Data were simulated as time-to-event data to be analysed using a Cox model. Because of the high proportion of mortality, time to death was also simulated as a continuous variable to be analysed using linear regression on those patients with an observed outcome. Observed time to death was approximately log-normally distributed.

WHO score (0–2) and tumour grade (1–4) were chosen to be balancing variables. 26% of patients were assigned a WHO score of 0, 48% a score of 1, and 26% a score of 2. 11% of patients were assigned a tumour grade of 1, 43% a grade of 2, 31% a grade of 3, and 15% a grade of 4.

For the time-to-event simulations the hazard ratio relative to a WHO score of 0 was 1.5 for a WHO score of 1 and 3.5 for a WHO score of 2. The hazard ratio relative to tumour grade 1 was 1.4 for a tumour grade of 2, 1.2 for a tumour grade of 3, and 1.8 for a tumour grade of 4. For generating Weibull survival times the shape and scale parameters[39] were set to 0.95 and 0.04 respectively.

For the linear regression model the residual standard deviation was set to 1. The regression parameters were -0.5 for a WHO score of 1, -1.3 for a score of 2, -0.2 for a tumour grade of 2, -0.3 for a grade of 3, and -0.5 for a grade of 4.

3.4.4. GBSG The GBSG trial was a 2×2 factorial comparing three vs. six cycles of chemotherapy and additional hormonal treatment with tamoxifen in patients with primary node positive breast cancer. The primary endpoint was tumour recurrence or death. Approximately one third of patients in this study were non-randomised. Of the 686 patients involved in the study, 299 died. Outcomes were simulated as binary, and were analysed using logistic regression.

Balancing factors were chosen to be progesterone receptor status (a continuous variable), the number of positive nodes, and tumour grade (0-3). For simulations, 12% of patients were assigned a tumour grade of 1, 65% of patients a grade of 2, and 23% of patients a grade of 3. The number of positive nodes was analysed using the transformation log(positive nodes/10), and was generated as $\sim N(-1.1, 0.9^2)$. Progesterone receptor status was generated from a normal distribution with mean 0 and standard deviation 11, then was squared to give it a skewed distribution similar to the original dataset.

Relative to tumour grade 1, the odds ratio was 2.3 for a tumour grade of 2, 2.1 for a tumour grade of 3, and 1.8 for a one unit increase in the log-transformed number of positive nodes. The odds ratio was set to 0.98 for a one unit increase in progesterone receptor status. The baseline odds of tumor recurrence or death were set to 0.90.

3.4.5. PBC The PBC parallel group trial investigated the effect of D-penicillamine vs. placebo in patients with primary biliary cirrhosis (PBC). The outcome was overall survival time.

Both binary and time-to-event outcomes were simulated. Of the 312 patients in the study, 125 died. Balancing factors were chosen to be age, bilirubin level, albumin level, sex, and histological stage (1 to 4).

For all simulations, 89% of patients were assigned to be female. 5% were assigned to histological stage 1, 22% to stage 2, 38% to stage 3, and 35% to stage 4. Age, albumin, and the log-transformed values of bilirubin were generated from $\sim N(50, 10.6^2)$ for age, $\sim N(3.5, 0.4^2)$ for albumin and $\sim N(0.58, 1)$ for log-bilirubin.

For simulations involving binary outcomes, the odds ratio was 1.05 for age, 0.6 for albumin, 3.1 for the log-transformed value of bilirubin, 0.5 for sex, and, relative to histological stage 1, 3.4 for histological stage 2, 5.0 for stage 3, and 6.6 for stage 4. The baseline odds were 0.07.

For time-to-event outcomes the hazard ratio was 1.03 for age, 0.4 for albumin, 2.5 for the log-transformed value of bilirubin, 0.75 for sex, and, relative to histological stage 1, 3.0 for histological stage 2, 4.2 for stage 3, and 5.5 for stage 4. For generating Weibull survival times the shape and scale parameters[39] were set to 1.58 and 0.03 respectively.

3.4.6. Simulation results Unadjusted analyses after stratification or minimisation led to standard errors that were biased upwards, confidence intervals with non-nominal coverage, type I error rates that were too low, and a reduction in power. For simulations involving the original study sample size (Table 1), the coverage rate for unadjusted analyses ranged from 95.7 to 97.8%. These correspond to type I error rates of 2.2 to 4.3%. The per cent bias in the standard errors ranged from 2.3 to 17.4%. In contrast, adjusted analyses gave unbiased standard errors and confidence intervals with nominal coverage. The coverage rate for adjusted analyses varied from 94.7 to 95.3%, and the per cent bias in standard errors ranged from -2.4 to 1.1%. In some cases, unadjusted analyses led to a serious reduction in power. The PBC trial with a time-to-event outcome had 78% power using an adjusting analysis, and only 53% power using an unadjusted analysis, a 25% reduction. Similarly, MIST2 saw a 21% reduction in power.

Interestingly the results for the GBSG dataset were more modest than for other datasets. An unadjusted analysis after stratification gave coverage rates ranging from 95.3 to 95.8% depending on the sample size. One possible explanation is the effect sizes of the balancing variables were not large enough to cause serious bias in the standard errors. It is worth noting that the impact of the balancing variables will also depend on the within-patient variance. In these simulations the data were generated using a logistic distribution, which has variance $\frac{\pi^2}{3}$. If a Probit model had been used instead, the variance would

have been set at 1, which would have led to more extreme results. Repeating the simulations using a Probit model gave coverage rates ranging from 95.9 to 96.4% depending on sample size. This underlines the difficulties of determining how large the effect of a balancing variable needs to be to result in biased standard errors for binary and time-to-event outcomes.

The FASTER trial with the type of surgery and baseline ODI as balancing variables was simulated using block sizes of 2, 8, and 32 in order to investigate what impact block size would have on coverage rates. The results were similar for all three block sizes. For block sizes of 2, 8 and 32, unadjusted analyses gave coverage rates ranging from 97.3 to 97.7%, 97.0 to 97.6%, and 97.1 to 97.4% respectively. As the block size gets extremely large, and the number of patients in each block is small compared to the overall block size, stratified block randomisation will approach simple randomisation. In this case unadjusted analyses will give valid results. However, it appears that for commonly used block sizes this will not be the case, and adjusted analyses are needed.

With small sample sizes for binary or time-to-event outcomes, adjusted analyses resulted in coverage rates that were too low (Table 2). This was the case under both simple and balanced randomisation. For example, the PBC data with a binary outcome and sample size of 100 had coverage rates of 94.1% and 94.3% after simple and stratified randomisation respectively. The PBC data with a time-to-event outcome and sample size 100 had coverage rates of 93.6% and 94.2%, and the RE01 data with a time-to-event outcome and sample size of 100 had coverage rates of 94.3% and 94.0%. This may be because the adjusted analyses included too many variables in relation to the number of observed events. Unadjusted analyses after simple randomisation did give nominal coverage rates, even with small sample sizes. This would suggest that balancing variables need to be chosen carefully in small trials with binary or time-to-event outcomes, and the number of balancing variables should be kept to a minimum.

4. Literature review

A literature review was carried out in order to assess current practice in analysing clinical trials that had allocated treatments using stratification or minimisation. Four major medical journals were searched: the Lancet, the British Medical Journal, the Journal of the American Medical Association, and the New England Journal of Medicine. Articles appearing in the three month period from March to May 2010 were included.

The titles for each article were searched, and titles that identified the study as *cohort*, *cross-sectional*, *case-control*, *systematic review* or *meta-analysis*, or in some other way indicated it was not a randomised trial were discarded. The abstracts for all other studies were read and all randomised trials were identified.

Only parallel-group, individually randomised trials were considered. All randomised trial articles were searched to determine what method of randomisation had been used, what the primary outcome was (continuous, binary, time to event, or rate/count), and whether the primary analysis had been adjusted for balancing factors.

Seventy-two trials were identified. Of these, seven were excluded: three were cluster-randomised trials, one was a crossover trial, two were single arm studies, and one was a secondary analysis of a study that had been previously reported within the review period. In total 65 trials met the inclusion criteria.

Fourty-one trials (63%) balanced on centre or on a patient-level prognostic variable (or both). Centre was used as a balancing variable in 35 trials (54%), and 24 trials (37%) used at least one patient-level prognostic factor as a balancing factor. Nine trials did not use any balancing variables in their randomisation, and in 15 articles (23%) the method of randomisation was unclear. In trials that used a balancing variable in their randomisation, 29 used stratified permuted blocks, 10 used minimisation, one used urn randomisation and one used another dynamic allocation method.

In total, 14 of the 41 trials (34%) that balanced on either centre or a prognostic variable adjusted for all balancing variables in the analysis. Only 13 of the 35 trials (37%) which balanced on centre adjusted for centre in the analysis, and only 7 of the 24 trials (29%) which used at least one prognostic variable in the randomisation appropriately adjusted for these variables in the analysis.

5. Discussion

We have shown that treatment allocation by stratification leads to correlation between treatment groups. If this correlation is ignored by not adjusting for the stratification variables in the analysis then the standard error for the treatment effect is biased upwards, leading to confidence intervals that are too wide, type I error rates that are too low, and a reduction in power. We have shown using simulation that this affects continuous, binary, and time-to-event outcomes, and is an issue for both stratification and minimisation. These issues lead to a loss of interpretability: we can no longer interpret a 95% confidence interval as nominal, or p-values as the probability of observing a result as or more extreme under the null hypothesis. The number of balancing variables in small trials with a time-to-event or binary outcome should be kept to

a minimum, as adjusted models may lead to type I error rates that are too high (though an unadjusted analysis in these situations still leads to type I error rates that are too low).

Stratification and minimisation are common in randomised trials. 63% of trials we reviewed used stratification or minimisation to balance prognostic factors or study centres between treatment groups. However, as common as these randomisation techniques are, their implications remain poorly understood. Only 34% of trials using stratification or minimisation properly adjusted for those variables in the primary analysis. Additionally, in 23% of papers the method of randomisation was not clear. Given that the appropriate method of analysis depends on the method of randomisation, it is essential for papers to follow the CONSORT guidelines[40] and to explain clearly what method of randomisation was used, and whether the analysis adjusted for any variables used in randomisation.

Stratification and minimisation lead to the same issues as crossover trials, matched-pair designs, longitudinal studies, and cluster randomised trials, in that all of these designs introduce correlation between observations. It is well known that adjusted analyses are required for all these designs, however not all parallel group trials recognise these issues with balancing covariates. It is interesting to note that crossover trials, matched-pair designs, and stratification/minimisation introduce correlation between treatment groups which, if ignored in the analysis, will lead to standard errors that are biased upwards. However, cluster randomised trials and longitudinal studies will introduce correlation within treatment groups, which will lead to standard errors which are biased downwards if clustering is ignored in the analysis. Depending on the type of clustering (either between- or within-treatment groups) results will be conservative or anti-conservative respectively.

It has been argued that adjustment for stratification variables is not always sensible on the basis that they may be unrelated to outcome[19]. This is conceptually correct, as omitting a stratification variable that is unrelated to outcome from the analysis will not lead to biased standard errors. However, balancing on patient-level factors is recommended only if the balancing variable is known to be related to outcome[41] (or, in cases where little prior information is available on which variables are prognostic, if the balancing variable is suspected to be related to outcome). Determining which balancing variables are not related to outcome within a trial will rely on post-hoc analysis or preliminary testing (where the method of analysis depends on the results of a preliminary significance test). Statistical significance tests and p-values will be of little use, as a stratification variable could have a large p-value, but still have an effect size large enough to lead to biased standard errors. Looking at the effect size of stratification variables will likewise be of little assistance, as it will be unclear how small an effect size can safely be ignored. Additionally, with multiple balancing variables the effect of each may appear negligible, but the cumulative effect may be larger. Preliminary testing has proven to give poor results in other areas. Freeman showed that preliminary testing for a carry over effect in crossover trials leads to biased estimates of the treatment difference, as well as increasing the probability of making a type I error[42]. Raab et al[13] give a good overview of the issues for post-hoc covariate selection in clinical trials. Given that the only effect of adjusting for a stratification variable which is unrelated to outcome is the loss of a degree of freedom, it is recommended that when stratification or minimisation has been used all analyses which estimate the treatment effect are pre-specified to be adjusted for balancing factors. If the stratification factors are associated with the outcome, this analysis will give correct inference. If they are not related to the outcome, this analysis will give similar results to an unadjusted analysis.

The reasons why so many practitioners do not adjust after stratification are unclear. There are a number of potential explanations. People may like the simplicity of unadjusted analyses and feel they are easier to interpret. Perhaps the idea that trials are special because they do not suffer from the systematic imbalances of observational studies makes a simple analysis appealing. Some people are inherently mistrustful of adjusted analyses as they feel the investigators may have used a variable selection technique that leads to biased results, or performed several different analyses and only presented those which were most favourable.

These are all valid concerns. In principle we agree that a simple appropriate analysis is preferable. However, after balancing an unadjusted analysis is no longer valid. The simplest appropriate analysis then becomes one which is adjusted for the balancing variables. Additionally, an unadjusted analysis does not guard against presenting the most favourable of several analyses, as the unadjusted model itself may be the most favourable. Therefore, the model should be pre-specified, and with a published protocol or analysis plan for support it should not be hard to demonstrate this. If people strongly prefer an unadjusted analysis they should avoid using any balancing covariates in the randomisation scheme, and use simple or permuted block randomisation instead.

In some situations we balance on factors that are not expected to be associated with outcome, and are therefore not normally included in the analysis. One example is time. Permuted block randomisation and minimisation can both be seen as balancing treatment over time[43]. If time is associated with the outcome then an analysis which ignores this may lead to biased standard errors. For example, if the inclusions/exclusion criteria are changed part way through a trial to allow sicker patients to enroll, then the time period (before or after the inclusion/exclusion changes were made change) is associated with outcome and should be adjusted for in the analysis.

Another situation where this might occur is when multiple prognostic variables are balanced on. Because permuted blocks balance within each strata, this has the effect of balancing on covariate interactions. This can also occur in minimisation when covariate interactions are added as balancing variables. If there is a non-negligible interaction between two covariates

that have been balanced on, an analysis that is not adjusted for these interactions will lead to biased standard errors.

In both of the above situations we do not necessarily expect that time or interactions between balancing variables will be associated with outcome, and therefore would not normally adjust for them in the analysis. However, if these variables are expected to be associated with outcome they should be adjusted for in the primary analysis. Even if they are not expected to be associated with outcome, a sensitivity analysis could be undertaken by adjusting for these variables to see if the results are substantially affected.

In many studies, recruiting centre or site is used as a balancing variable[11]. In some ways, this is no different to balancing on a patient-level characteristic, such as disease stage or gender; if centre is associated with outcome, than it must be adjusted for in the analysis in order to obtain valid results. However, typical methods of adjustment require fitting a parameter for each centre, which can lead to biased or inefficient results when the number of centres is large. More research is needed to determine the best way to adjust for centre in these situations.

A limitation of this work is that the simulations for binary and time-to-event outcomes were not based on trials that had used stratified randomisation or minimisation. However, balancing variables were chosen on the basis that they seemed candidates for use in practice (for example WHO stage, tumour grade, number of positive nodes). Given that balancing variables should only be chosen if they are highly prognostic, it is likely that the effect sizes seen in our simulations would be similar to those seen in actual trials using stratification or minimisation.

The consequences of an unadjusted analysis after stratification or minimisation are not fully understood. This is not surprising as most textbooks and articles describing these methods of randomisation do not mention that using balancing methods in the randomisation process has implications for the analysis. The solution to this issue is two-fold. Articles, textbooks and statistics courses explaining minimisation and stratification should highlight the issues these methods pose for the analysis. Secondly, reviewers of medical journals should be made aware of these issues, and should request that analyses be adjusted for balancing variables when treatment allocation has been carried out using stratification or minimisation.

Acknowledgements

We thank the FASTER and MIST2 trial teams for providing their data. We are grateful to Caroline Doré, Rebecca Walwyn, Richard Morris, Tony Johnson, Daniel Bratton and Louise Choo for helpful comments. We would also like to thank an anonymous referee and associate editor for their comments which greatly helped to improve the article.

References

1. Lachin JM. Properties of simple randomization in clinical trials. *Controlled Clinical Trials* 1988; **9**:312–326, doi:doi:10.1016/0197-2456(88)90046-3. URL http://dx.doi.org/doi:10.1016/0197-2456(88)90046-3.
2. Kalish LA, Begg CB. Treatment allocation methods in clinical trials: a review. *Statistics in Medicine* 1985; **4**:129–144, doi:10.1002/sim.4780040204. URL http://dx.doi.org/10.1002/sim.4780040204.
3. Fisher RA. The arrangement of field experiments. *Journal of the ministry of agriculture of Great Britain* 1926; **33**:83–94.
4. Efron B. Forcing a sequential experiment to be balanced. *Biometrika* Dec 1971; **58**(3):403–417, doi:10.1093/biomet/58.3.403. URL http://dx.doi.org/10.1093/biomet/58.3.403.
5. Wei LJ, Lachin JM. Properties of urn randomisation in clinical trials. *Controlled Clinical Trials* 1994; **9**:345–364, doi:10.1016/0197-2456(88)90048-7. URL http://dx.doi.org/10.1016/0197-2456(88)90048-7.
6. Atkinson AC. Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika* Apr 1982; **69**(1):61–67, doi:10.1093/biomet/69.1.61. URL http://dx.doi.org/10.1093/biomet/69.1.61.
7. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 1975; **31**:103–115.
8. Pond GR, Tang PA, Welch SA, Chen EX. Trends in the application of dynamic allocation methods in multi-arm cancer clinical trials. *Clinical Trials* 2010; **7**:227–234, doi:10.1177/1740774510368301. URL http://dx.doi.org/10.1177/1740774510368301.
9. Scott NW, McPherson GC, Ramsay CR, Campbell MK. The method of minimization for allocation to clinical trials a review. *Controlled Clinical Trials* Dec 2002; **23**(6):662–674, doi:10.1016/S0197-2456(02)00242-8. URL http://dx.doi.org/10.1016/S0197-2456(02)00242-8.
10. Schulz KF, Grimes DA. Unequal group sizes in randomised trials: guarding against guessing. *The Lancet* Mar 2002; **359**(9310):966–970, doi:10.1016/S0140-6736(02)08029-7. URL http://dx.doi.org/10.1016/S0140-6736(02)08029-7.
11. Committee for Proprietary Medicinal Products. Points to consider on adjustment for baseline covariates. *Statistics in Medicine* 2004; **23**:701–709, doi:10.1002/sim.1647. URL http://dx.doi.org/10.1002/sim.1647.
12. ICH E9 Expert Working Group. Statistical principles for clinical trials: ICH harmonised tripartite guideline. *Statistics in Medicine* 1999; **18**:1905–1942, doi:10.1002/(SICI)1097-0258(19990815)18:15%3C1903::AID-SIM188%3E3.0.CO;2-F. URL http://dx.doi.org/10.1002/(SICI)1097-0258(19990815)18:15%3C1903::AID-SIM188%3E3.0.CO;2-F.
13. Raab GM, Day S, Sales J. How to select covariates to include in the analysis of a clinical trial. *Controlled Clinical Trials* 2000; **21**:330–342, doi:10.1016/S0197-2456(00)00061-1. URL http://dx.doi.org/10.1016/S0197-2456(00)00061-1.
14. Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. *Journal of Clinical Epidemiology* 1999; **52**:19–26, doi:doi:10.1016/S0895-4356(98)00138-3. URL http://dx.doi.org/doi:10.1016/S0895-4356(98)00138-3.
15. Senn S. *Statistical issues in drug development*. Wiley, Chichester, 2007.
16. Redmond C, Colton T. *Biostatistics in clinical trials*. Wiley: Chichester, 2001.
17. Piantadosi S. *Clinical Trials: A Methodologic Perspective*. Wiley, New Jersey, 2005.
18. Lachin JM, Matts JP, Wei LJ. Randomization in Clinical Trials: Conclusions and Recommendations. *Controlled clinical trials* 1988; **9**:365–374, doi:10.1016/0197-2456(88)90049-9. URL http://dx.doi.org/10.1016/0197-2456(88)90049-9.
19. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in medicine* 2002; **21**:2917–2930, doi:10.1002/sim.1296. URL http://dx.doi.org/10.1002/sim.1296.
20. Wang D, Bakhai A. *Clinical Trials: A Practical Guide to Design, Analysis, and Reporting*. Remedica: London, 2006.
21. Girling D, Parmar M, Stenning S, Stephens R, Stewart L. *Clinical Trials in Cancer: Principles and Practice*. Oxford University Press: Oxford, 2003.
22. Pocock SJ. *Clinical Trials: A Practical Approach*. Wiley: Chichester, 1983.
23. Altman DG. *Practical statistics for medical research*. Chapman and Hall: London, 1991.
24. Meinert CL. *Clinical Trials: Design, Conduct, and Analysis*. Oxford University Press: Oxford, 1986.
25. Altman DG, Bland JM. Treatment allocation by minimisation. *BMJ* Apr 2005; **330**(7495):843+, doi:doi:10.1136/bmj.330.7495.843. URL http://dx.doi.org/doi:10.1136/bmj.330.7495.843.
26. Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. *Lancet* 2002; **359**:515–519, doi:doi:10.1016/S0140-6736(02)07683-3. URL http://dx.doi.org/doi:10.1016/S0140-6736(02)07683-3.
27. Rovers MM, Straatman H, Zielhuis GA. Comparison of balanced and random allocation in clinical trials: A simulation study. *European Journal of Epidemiology* 2000; **16**:1123–1129, doi:10.1023/A:1010907912024. URL http://dx.doi.org/10.1023/A:1010907912024.
28. Weir CJ, Lees KR. Comparison of stratification and adaptive methods for treatment allocation in an acute stroke clinical trial. *Statistics in Medicine* 2003; **22**:705–726, doi:10.1002/sim.1366. URL http://dx.doi.org/10.1002/sim.1366.
29. Forsyth AB. Validity and power of tests when groups have been balanced for prognostic factors. *Computational Statistics and Data Analysis* 1987; **5**:193–200, doi:doi:10.1016/0167-9473(87)90015-6. URL http://dx.doi.org/doi:10.1016/0167-9473(87)90015-6.
30. Hagino A, Hamada C, Yoshimura I, Ohashi Y, Sakamoto J, Nakazato H. Statistical comparison of random allocation methods in cancer clinical trials. *Controlled Clinical Trials* 2004; **25**:572–584.
31. Birkett NJ. Adaptive allocation in randomized controlled trials. *Controlled Clinical Trials* 1985; **6**:146–155, doi:10.1016/0197-2456(85)90120-5. URL http://dx.doi.org/10.1016/0197-2456(85)90120-5.
32. Parzen M, Lipsitz SR, Dear KBG. Does clustering affect the usual test statistics of no treatment effect in a randomized clinical trial? *Biometrical Journal* 1998; **40**(4):385–402, doi:10.1002/(SICI)1521-4036(199808)40:4%3C385::AID-BIMJ385%3E3.0.CO;2-%23. URL http://dx.doi.org/10.1002/(SICI)1521-4036(199808)40:4%3C385::AID-BIMJ385%3E3.0.CO;2-%23.
33. Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Controlled Clinical Trials* 1998; **19**:249–256, doi:10.1016/S0197-2456(97)00147-5. URL http://dx.doi.org/10.1016/S0197-2456(97)00147-5.
34. Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* 1991; **58**:227–240.
35. McGregor AH, Doré CJ, Morris TP, Morris S, Jamrozik K. Function after spinal treatment, exercise and rehabilitation (FASTER): improving the functional outcome of spinal surgery. *BMC musculoskeletal disorders* Jan 2010; **11**(1):17+, doi:10.1186/1471-2474-11-17. URL http://dx.doi.org/10.1186/1471-2474-11-17.
36. Rahman NM, Maskell NA, West A, Teoh R, Arnold A, Mackinlay C, Peckham D, Davies CWH, Ali N, Kinnear W, *et al.*. Intrapleural Use of Tissue Plasminogen Activator and DNase in Pleural Infection. *New England Journal of Medicine* Aug 2011; **365**(6):518–526, doi:10.1056/NEJMoa1012740. URL http://dx.doi.org/10.1056/NEJMoa1012740.

37. Royston P, Sauerbrei W. *Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Wiley: Chichester, 2008. URL http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470028424.html.
38. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* 2005; **24**, doi:10.1002/sim.2059. URL http://dx.doi.org/10.1002/sim.2059.
39. Collett D. *Modelling Survival Data in Medical Research*. Chapman and Hall: London, 1994.
40. Moher D, Hopewell D, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *British Medical Journal* 2010; **240**, doi:doi:10.1136/bmj.c869. URL http://dx.doi.org/doi:10.1136/bmj.c869.
41. Pocock SJ. *Clinical Trials: a Practical Approach*. Wiley: Chichester, 1983.
42. Freeman PR. The performance of the two-stage analysis of two-treatment, two-period crossover trials. *Statistics in Medicine* 1989; **8**(12):1421–1432, doi:10.1002/sim.4780081202. URL http://dx.doi.org/10.1002/sim.4780081202.
43. Matts JP, Lachin JM. Properties of permuted-block randomization in clinical trials. *Controlled Clinical Trials* Dec 1988; **9**(4):327–344, doi:10.1016/0197-2456(88)90047-5. URL http://dx.doi.org/10.1016/0197-2456(88)90047-5.

Table 1. Coverage, power and bias in standard errors after stratification or minimisation

				Unadjusted analysis			Adjusted analysis		
Study	Outcome	n	Method	Coverage (%)	Power	% bias in SE*	Coverage (%)	Power	% bias in SE*
FASTER†	Continuous	316	Strat	96.2	76.1	5.8	94.8	79.5	0.1
FASTER‡	Continuous	316	Strat	97.5	65.6	14.4	95.1	80.2	0.0
			Min	97.2	65.5	13.0	94.8	79.6	−0.3
MIST2	Continuous	210	Strat	97.4	58.6	13.7	95.1	80.0	0.2
			Min	97.7	58.5	14.8	95.2	80.7	1.4
RE01	Continuous	322	Strat	97.0	74.5	11.8	95.0	81.4	0.7
			Min	97.0	74.1	11.3	95.1	81.1	0.4
	Survival	347	Strat	96.6	67.8	8.1	94.8	78.5	−1.3
			Min	96.5	68.4	8.0	94.9	78.5	−0.7
PBC	Binary	312	Strat	96.8	65.2	9.6	95.0	78.8	−1.9
			Min	97.5	65.5	13.7	94.8	79.0	−1.8
	Survival	312	Strat	97.5	53.0	13.8	94.7	78.0	−1.8
			Min	97.6	53.6	17.4	95.1	79.2	−2.4
GBSG	Binary	686	Strat	95.8	77.7	2.3	95.3	80.8	− 0.4
			Min	95.7	77.7	3.0	94.9	80.5	−0.5

*% bias in SE is defined as $100\left(\frac{SE_{\text{model}}-SE_{\text{empirical}}}{SE_{\text{empirical}}}\right)$

†Balanced using surgeon and type of surgery

‡Balancing using baseline ODI and type of surgery

Table 2. Coverage after simple and stratified randomisation for different sample sizes

			Simple		Stratified	
Study	Outcome	n	Unadjusted	Adjusted	Unadjusted	Adjusted
FASTER*	Continuous	100	95.1	95.1	97.4	94.8
		200	95.2	95.1	97.0	94.8
		500	94.9	94.8	97.6	95.1
		1000	95.4	95.3	97.4	94.9
MIST2	Continuous	100	95.3	94.8	97.2	94.9
		200	94.9	95.3	97.2	94.9
		500	95.2	94.9	97.3	94.4
		1000	95.2	95.2	97.7	95.2
RE01	Continuous	100	94.5	94.7	96.5	94.8
		200	95.3	95.6	97.0	95.1
		500	94.9	94.9	96.8	95.1
		1000	94.7	94.7	96.4	94.2
PBC	Binary	100	94.4	94.1	96.3	94.3
		200	95.2	94.7	96.8	94.6
		500	95.0	94.8	97.1	94.7
		1000	94.8	94.8	97.6	94.9
GBSG	Binary	100	95.0	94.6	95.4	94.9
		200	94.7	94.6	95.5	94.7
		500	95.3	95.2	95.7	95.2
		1000	94.9	94.8	95.3	94.9
RE01	Survival	100	95.1	94.3	96.0	94.0
		200	94.7	94.3	96.3	94.3
		500	95.1	95.0	96.8	94.8
		1000	94.9	94.7	96.7	95.0
PBC	Survival	100	95.2	93.6	97.2	94.2
		200	95.3	94.6	97.4	94.7
		500	94.9	94.6	97.5	94.9
		1000	94.9	94.4	97.9	94.7

*Balanced using baseline ODI and type of surgery