The effect of heterogeneous variance on efficiency and power of cluster randomized trials with a balanced 2x2 factorial design

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2015, Vol. 24(5) 574–593

DOI: 10.1177/0962280215583683
smm.sagepub.com
SAGE

The effect of heterogeneous variance on efficiency and power of cluster randomized trials with a balanced 2×2 factorial design

Francesca Lemme,[1] Gerard JP van Breukelen,[2] Math JJM Candel[2] and Martijn PF Berger[1]

Abstract

Sample size calculation for cluster randomized trials (CRTs) with a 2×2 factorial design is complicated due to the combination of nesting (of individuals within clusters) with crossing (of two treatments). Typically, clusters and individuals are allocated across treatment conditions in a balanced fashion, which is optimal under homogeneity of variance. However, the variance is likely to be heterogeneous if there is a treatment effect. An unbalanced allocation is then more efficient, but impractical because the optimal allocation depends on the unknown variances. Focusing on CRTs with a 2×2 design, this paper addresses two questions: How much efficiency is lost by having a balanced design when the outcome variance is heterogeneous? How large must the sample size be for a balanced allocation to have sufficient power under heterogeneity of variance? We consider different scenarios of heterogeneous variance. Within each scenario, we determine the relative efficiency of a balanced design, as a function of the level (cluster, individual, both) and amount of heterogeneity of the variance. We then provide a simple correction of the sample size for the loss of power due to heterogeneity of variance when a balanced allocation is used. The theory is illustrated with an example of a published 2×2 CRT.

Keywords

Balanced design, optimal design, heterogeneous variance, 2×2 factorial, cluster randomized trial

1 Introduction

An important issue in the design phase of a randomized trial concerns the optimal allocation of units across the treatment conditions being compared in order to achieve a certain power to detect a treatment effect, or to obtain a certain precision for the treatment effect estimator. Randomized

[1]Department of Methodology and Statistics, Maastricht University, The Netherlands
[2]Department of Methodology and Statistics, CAPHRI School for Public Health and Primary Care, Maastricht University, The Netherlands

Corresponding author:
Francesca Lemme, Department of Methodology and Statistics, Maastricht University, The Netherlands.
Email: francesca.lemme@maastrichtuniversity.nl

trials are frequently run in populations with a multilevel data structure with individuals nested within clusters, for instance, pupils in schools or patients in general practices. In this case randomized treatment assignment can be done at the cluster level (cluster randomized trials or CRTs) or at the individual level (multicentre trials). Cluster randomization is proven less efficient than individual randomization, as the outcome variance between clusters in the same treatment arm increases the sampling variance of the treatment effect.[1–3] However, for some types of intervention individual randomization is logistically impossible, for example school health promotion programs delivered in the class rooms. Then, at best classes instead of individual pupils can be randomized. Further, randomization at the individual or even at the class level may induce serious treatment contamination through exchange of treatment information between treated and control subjects, leading to a reduced treatment contrast and, thereby, also a reduced power to detect a treatment effect. Randomization of entire schools may then be the best option.

As mentioned above, outcome variation between clusters unfortunately makes cluster randomization less efficient than individual randomization.[1,2] This makes an efficient study design important especially for CRTs. The sample size should be as large as possible in order to have maximum precision of treatment effect estimation, and maximum power to detect a treatment effect. However, the high costs of scientific studies, and the burden of study participation to clusters (i.e. schools or general practices), require the sample size to be as small as possible. Optimal design[4,5] helps researchers to find a balance between efficiency and power and sampling costs, by maximizing precision and power under budget and costs constraints. In the context of CRTs it has been shown how the optimal sample size at each design level (cluster, individual) depends on the ratio of cluster level outcome variance to individual level outcome variance as expressed by the intraclass correlation.[1–3] In the context of individual randomization it has been shown how the optimal treatment:control allocation ratio depends on the ratio of the treated to control outcome variance.[6] However to our knowledge, optimal design of a CRT with heterogeneous variance has not been studied yet. Moreover, the outcome variance at each design level and in each treatment arm is unknown at the design stage of a study, which is an obstacle to optimal design. Therefore, the aim of this paper is to study the optimal sample size and optimal treatment allocation ratio of a CRT under heterogeneity of variance at each design level (cluster, individual), to evaluate the efficiency of the commonly used balanced allocation relative to the optimal design, and to find a practical procedure for sample size calculation for a CRT with heterogeneous variances.

The issue of unknown but heterogeneous outcome variance at the design stage of a study has been addressed by a number of authors. For one-way ANOVA with unclustered data and heterogeneous variance, Jan and Shieh[7] examined methodology to have a minimum sample size such that the expected half-width of the confidence interval is within designated boundaries, and Wong and Zhu[8] developed optimal treatment allocation rules for various contrasts of interest. Furthermore, Guo and Luh[9] and Schouten[6] took into account the costs per included individual. For the case of CRTs with a two-arm design, equations to compute the optimal design have been published, for homogeneous variances, by Moerbeek et al.[1] and by Raudenbush[2] and, for heterogeneous variances, by Moerbeek and Wong[10] and Candel and van Breukelen.[11] However, Moerbeek and Wong,[10] consider clustering in one treatment arm only and Candel and van Breukelen[11] optimize only the number of clusters but not the cluster sizes. Moreover, none of the work published in the literature considers efficient treatment allocation within a nested design with more than two treatment conditions when the outcome variance is heterogeneous. Among nested designs, CRTs with a 2×2 factorial design have been adopted for various purposes in public health and medicine.[12–19] Therefore, we will focus our work on CRTs

with a 2×2 factorial design but our results will also apply to CRTs with two treatment groups as will be seen later.

In CRTs with a 2×2 factorial design two treatment factors are evaluated at the same time, by allocating clusters to one of four conditions. An example of a 2×2 factorial design is a trial comparing a new decision aid tool (Statin Choice) with a standard tool (Standard Pamphlet) to assist clinicians and patients with diabetes on sharing information about statin, and comparing two methods of delivery of the decision tool (before or after the visit).[19] A 2×2 factorial design was employed, with the following treatment conditions: statin choice during visit (SD), statin choice before visit (SB), standard pamphlet during visit (PD), and standard pamphlet before visit (PB). Usually, a balanced (or nearly balanced) allocation of units is adopted in 2×2 CRTs, as in the studies 12 to 19 in the reference list. This is optimal under homogeneity of variance,[20] but not under heterogeneity of variance. Further, as mentioned above, the optimal allocation depends on the unknown variances at each design level (cluster, individual) in each treatment arm, which makes optimal allocation unfeasible. Two obvious questions are then: (a) For a given study budget, and given costs per included cluster and individual, how much efficiency is lost when a balanced design is adopted instead of the design that is optimal under heterogeneity of variance? (b) Given the adoption of a balanced design, how much power is lost when the variances are heterogeneous instead of homogeneous and how can the sample size be adjusted for that loss? We address these two questions at each level (cluster, individual) for three different plausible scenarios of heterogeneous variance. Within each of these scenarios, we first use optimal design theory[4,5] to determine the optimal allocation of clusters to treatments as well as the relative efficiency of a balanced allocation, as a function of the amount of heterogeneity of the variance at each design level (cluster, individual). We then plot this relative efficiency of the balanced design against the amount of heterogeneity of variance, for each scenario and for different values of the intraclass correlation. Finally we extend a published formula for two-arm designs with a balanced allocation to 2×2 designs, and we show how the sample size can be adjusted for the loss of precision and power due to heterogeneous variance when a balanced allocation is used.

The outline of this article is as follows. In section 2.1 we describe the statistical model, the corresponding treatment effect estimators and their variances. In section 2.2 we then describe the design optimality criteria used, and we present the equation for the optimal allocations of units under heterogeneity of variance, as well as an equation for the relative efficiency of the balanced design, relative to the optimal allocation. In section 2.3 we introduce the three heterogeneous variance scenarios considered in this article, and an interpretable measure of heterogeneity of variance across the four treatment conditions. In section 3, we graphically show the relative efficiency of the balanced design as a function of the amount of heterogeneity of the variance across the four conditions. In section 4, we provide a simple correction of the sample size for the loss of power due to heterogeneity of variance when a balanced allocation is used, and we apply this correction to a published 2×2 CRT. Finally, in section 5 we discuss our results.

2 Methods

2.1 The statistical model

Consider that each cell of a 2×2 factorial design is represented by (rc) indicating a row-column combination of the design. So, in the Statin trial, cell $(rc) = (11)$ is the new decision aid tool delivered before the visit, cell $(rc) = (12)$ is the new decision aid tool delivered during the visit, $(rc) = (21)$ is the standard tool delivered before the visit and $(rc) = (22)$ is the standard tool delivered during the visit. Assume that a total of K clusters is included into the study and denote each cluster by subscript j

($j = 1, \ldots, , K$). Denote individuals within each cluster with subscript i ($i = 1, \ldots, n_{rc}$, where n_{rc} is the number of individuals per cluster in cell rc of the 2×2 design).

For two levels of nesting and cluster randomization, the model relating a continuous outcome to treatment condition in a 2×2 factorial design is:

$$y_{ij} = \beta_{0j} + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \varepsilon_{ij} \quad (1)$$

with

$$\beta_{0j} = \beta_0 + u_{0j}$$

where y_{ij} is the outcome for the ith individual within the jth cluster; x_{1j}, x_{2j}, x_{3j} are three centered treatment indicators, that is, coded as $-1/+1$ rather than $0/1$, for all individuals within the jth cluster, with x_1 and x_2 as main effect indicators and $x_3 = x_1 \times x_2$ as interaction indicator as shown in the upper part of Table 1; β_1, β_2, and β_3 are the regression coefficients for the three treatment indicators; β_0 is the grand mean of the outcome and β_{0j} is the mean outcome in the jth cluster; u_{0j} and ε_{ij} are random effects at the cluster and individual levels, both normally distributed with mean 0 and either homogeneous or heterogeneous variance across the four treatment conditions. Note that, due to the centering of the treatment indicators x_1 and x_2, the present model is an ANOVA model, with β_1 and β_2 the two "main effects" of treatment, as shown in Table 1.

In the remainder of this article, we express homogeneous variance across the four cells as $\mathrm{var}(u_{0j}) = \sigma_u^2$ and $\mathrm{var}(\varepsilon_{ij}) = \sigma_\varepsilon^2$, and heterogeneous variance as $\mathrm{var}(u_{0j}) = \sigma_{urc}^2$ and $\mathrm{var}(\varepsilon_{ij}) = \sigma_{\varepsilon rc}^2$, where subscript ($rc$) indicates one cell (row-column combination) of the 2×2 design.

Under homogeneity of sampling costs and outcome variances *within* conditions, equally sized clusters within the same condition are optimal for estimating treatment effects with

Table 1. Coding schemes for the predictors in equation (1) and relations between cell means μ_{rc} and regression parameters β_h (upper part of the table) and regression parameters β_h in equation (1) expressed as weighted sums of cell means μ_{rc}, and the weights (lower part of the table).

Cell	x_0	x_1	x_2	x_3	μ_{rc}
11	1	−1	−1	1	$\mu_{11} = \beta_0 - \beta_1 - \beta_2 + \beta_3$
12	1	−1	1	−1	$\mu_{12} = \beta_0 - \beta_1 + \beta_2 - \beta_3$
21	1	1	−1	−1	$\mu_{21} = \beta_0 + \beta_1 - \beta_2 - \beta_3$
22	1	1	1	1	$\mu_{22} = \beta_0 + \beta_1 + \beta_2 + \beta_3$

Contrast/parameter	$w_{(11)}$	$w_{(12)}$	$w_{(21)}$	$w_{(22)}$	β_h in terms of μ_{rc}
0	1/4	1/4	1/4	1/4	$\beta_0 = \frac{\mu_{11}+\mu_{12}+\mu_{21}+\mu_{22}}{4}$
1	−1/4	−1/4	1/4	1/4	$\beta_1 = \frac{(\mu_{22}+\mu_{21})-(\mu_{12}+\mu_{11})}{4}$
2	−1/4	1/4	−1/4	1/4	$\beta_2 = \frac{(\mu_{22}+\mu_{12})-(\mu_{21}+\mu_{11})}{4}$
3	1/4	−1/4	−1/4	1/4	$\beta_3 = \frac{(\mu_{22}+\mu_{11})-(\mu_{21}+\mu_{12})}{4}$

The cell means μ_{rc} in the last column of the upper part of the table are obtained as $\mu = \boldsymbol{X}\beta$, where $\boldsymbol{X}$ is a 4×4 matrix containing the coding schemes showed in the upper part of the table and β is a 4×1 vector containing the regression coefficients in equation (1); the regression coefficients β in the last column of the lower part of the table are obtained as $\beta = \boldsymbol{W}\mu$, where $\boldsymbol{W}$ is a 4×4 matrix containing the weights in the lower part of the table and $\boldsymbol{W} = \boldsymbol{X}^{-1}$.

maximum precision.[1,2] Model (1) can then be reduced to a one level model, with clusters as units and cluster means as data. The maximum likelihood (ML) estimator of the fixed treatment effects then is:

$$\hat{\beta}_h = \sum_{rc} w_{rc(h)} \hat{\mu}_{rc} \tag{2}$$

where $w_{rc(h)}$ is the weight for cell (rc) of the 2×2 design as shown in the lower part of Table 1 with respect to the hth regression coefficient ($h = 1, ..., 3$), and $\hat{\mu}_{rc}$ is the ML estimator of the expected outcome within cell (rc), which is computed as the average of the cluster means within that cell.[21] From equation (2) and from independence between the four cell mean estimators $\hat{\mu}_{rc}$, it then follows that the variance of the fixed-effects estimators $\hat{\beta}_h (h = 0, ... , 3)$ is:

$$var(\hat{\beta}_h) = \sum_{rc} w_{rc(h)}^2 var(\hat{\mu}_{rc}) = \sum_{rc} w_{rc(h)}^2 \times \frac{1}{k_{rc}} \times \left(\sigma_{0rc}^2 + \frac{\sigma_{\varepsilon rc}^2}{n_{rc}} \right) \tag{3}$$

where k_{rc} and n_{rc} are respectively the number of clusters and the number of subjects per cluster (henceforth called the cluster size) in cell (rc). Note that heterogeneous outcome variances σ_{0rc}^2 and $\sigma_{\varepsilon rc}^2$ are assumed in equation (3). The special case of homogeneous variance is obtained by substituting the homogeneous variances σ_u^2 and σ_ε^2 for the heterogeneous variances σ_{0rc}^2 and $\sigma_{\varepsilon rc}^2$. Further, note that $w_{rc(h)}^2 = 1/16$ for all $rc(h)$ (see Table 1).

2.2 The derivation of the optimal design

2.2.1 The optimality criterion

The first step of our work is to derive the design which minimizes the variance of the estimators of the fixed effects β_h in equation (1) under the constraint of a fixed total budget for sampling clusters and individuals, and under heterogeneity of variance. However, the present model has four fixed effects, and so, to find the optimal design, some scalar function of the 4×4 variance–covariance matrix $Cov(\hat{\beta})$ of the treatment effect estimators, called an optimality criterion, must be chosen and then minimized. In this article we consider that interest might either focus on a specific treatment contrast, for instance the interaction, or be equally divided between all effects. In the first case, we consider the variance of a single treatment effect estimator, which is an example of the c-criterion.[4,5] In the second case, we consider the sum of the variances $\text{var}(\hat{\beta}_h)$ of all four fixed effects in equation (1), which is known in optimal design as the A-criterion, and the sum of the variances of the three treatment effects, excluding $\hat{\beta}_0$, called A_s-criterion.[4,5] Now, in the present model all fixed-effect estimators have the same variance (see equation (3) and the weights w_{rc} in the lower part of Table 1). Consequently, the c-criterion is proportional to the A- and A_s-criteria for this model, and minimization of the c-criterion will give the same optimal design as minimization of the A- or the A_s-criterion.

2.2.2 The optimal sample size

As mentioned in section 2.2.1, we want to derive a design which minimizes the variance of the treatment effect estimators under the constraint of a fixed total budget. A cost function must therefore be defined. Assume that inclusion of a cluster into any treatment condition (rc) costs c currency units (e.g. Euros) and that inclusion of an individual within a cluster in any condition (rc)

costs p units. The total budget B needed to include within the (rc)-th condition k_{rc} clusters with n_{rc} individuals each, is then:

$$B = \sum_{rc} B_{rc} = \sum_{rc} k_{rc}(c + p \times n_{rc}) \tag{4}$$

where B_{rc} is the budget spent on treatment condition (rc) and the factor $(c + p \times n_{rc})$ is the total sampling cost per cluster of size n_{rc} within cell (rc).

Finding an optimal design means finding, for each treatment condition, the number of clusters, k_{rc}, and the number of subjects per cluster n_{rc}, which minimize the c- criterion introduced in section 2.2.1, given the constraint in equation (4). These optimal sample sizes per level can be shown to satisfy equation (5) (for details of the proof, see Appendix 1):

$$k_{rc} = \frac{B_{rc}}{c + p \times n_{rc}},\ n_{rc} = \sqrt{\frac{\sigma^2_{\varepsilon rc}}{\sigma^2_{0rc}}}\sqrt{\frac{c}{p}} \tag{5}$$

Equation (5) shows that the optimal cluster size n_{rc} within each treatment condition depends on the cluster-to-person cost ratio (c/p) and on the person-to-cluster variance ratio for that condition $(\sigma^2_{\varepsilon rc}/\sigma^2_{0rc})$, or equivalently, on the ratio $(1 - \rho_{rc})/\rho_{rc}$. Here, ρ_{rc} is the intraclass correlation (ICC) in cell (rc). This is defined as the proportion of the total variance due to variability at cluster level in cell (rc), that is $\rho_{rc} = \sigma^2_{0rc}/\sigma^2_{yrc}$, where $\sigma^2_{yrc} = \sigma^2_{0rc} + \sigma^2_{\varepsilon rc}$. In addition the optimal number of clusters k_{rc} depends on B_{rc}, the budget allocated to cell (rc), for which the optimal value is given by equation (13) in Appendix 1, and this again depends on the variances and costs.

Note that if the total budget B in equation (4) is increased, the number of subjects per cluster remains constant, whereas the total number of clusters increases. However, the relative distribution of clusters across treatment conditions does not change (see equation (5)).

2.2.3 *The relative efficiency measure*

As shown by equation (5), the optimal design depends on the cluster level and individual level outcome variances and is not balanced under heterogeneity of variance. In other words, the optimal number of clusters and, except in the case of a homogeneous ICC (which requires a very special case of heterogeneity of variance), also the optimal number of persons per cluster, varies between the four cells (treatment conditions). Now, the outcome variance in equation (3) is an unknown quantity in the design stage and a balanced design is commonly used. The question now is how efficient a balanced design is compared to the optimal design for heterogeneous variance. This can be addressed by considering the relative efficiency of the balanced design versus the optimal design for heterogeneous variance, which can be shown to be (for details, see Appendix 1):

$$RE = \frac{\left(\sum_{rc}\left(\sqrt{c \times \sigma^2_{0rc}} + \sqrt{p \times \sigma^2_{\varepsilon rc}}\right)\right)^2 / B}{\sum_{rc}\left(n \times \sigma^2_{0rc} + \sigma^2_{\varepsilon rc}\right)/(k \times n)} \tag{6}$$

The numerator of equation (6) is the c-criterion for the optimal design for heterogeneous variances. This is obtained by inserting equation (5) into equation (3) and then plugging in the optimal B_{rc} as derived in Appendix 1. The denominator of equation (6) is the c-criterion when the design is balanced. This is obtained from equations (3) and (5) by substituting the average variances across

the four conditions, that is $\sigma^2_{0(avg)} = \sum_{rc} \sigma^2_{0rc}/4$ and $\sigma^2_{\varepsilon(avg)} = \sum_{rc} \sigma^2_{\varepsilon rc}/4$, for the heterogeneous variances σ^2_{0rc} and $\sigma^2_{\varepsilon rc}$. Thus, the optimal and balanced design are based on the same average variance at the cluster level, the same average variance at the individual level, and on the same budget and costs. This allows a fair comparison between the two designs. Note that the numerator of equation (6) is simply the squared sum of eight terms, each involving one of the eight variances weighted by the corresponding cost term (c for cluster, p for person). The denominator of the equation is likewise the sum of eight variances, with weights $1/k$ for cluster variances and $1/(k \times n)$ for person variances. The practical implication of this is that the RE does not depend on which variance is observed in which cell (treatment condition). For instance, it can occur that the cell with the largest cluster level variance also has the largest individual level variance, or that it has the smallest individual level variance. In both cases the RE of the balanced design compared to the optimal design is the same.

Further, the relative efficiency measure in equation (6) depends on the ratio of individual to cluster level variance and on the ratio of cluster to individual level costs. We therefore consider three person-to-cluster variance ratios: 99, 19, and 9, implying, respectively, three realistic intraclass correlations (ICC, as defined in section 2.2.2), as suggested by reviews of ICC values in primary care trials:[22,23] $\rho = 0.01$, $\rho = 0.05$, and $\rho = 0.10$. Note that these ICC values are defined under homogeneity of variances, i.e. $\rho = \sigma^2_{0(avg)}/\sigma^2_{y(avg)}$, where $\sigma^2_{y(avg)} = \sigma^2_{0(avg)} + \sigma^2_{\varepsilon(avg)}$. Under heterogeneity of variances, the ICC itself is heterogeneous in general, and its value for each cell (rc) then follows from the constraint that the average variance (i.e. $\sigma^2_{0(avg)}$ or $\sigma^2_{\varepsilon(avg)}$) will be kept the same under all scenarios and all extents of heterogeneity.

Finally, we consider the same cluster-to-person cost ratios of 9, 19, and 99 as the person-to-cost variance ratios, because the costs per cluster and per person are rarely reported in publications of trials. These values cover a wide range of cost ratios.

2.3 The heterogeneous variance scenarios

Equation (5) in section 2.2.2 can be used to compute optimal designs for heterogeneous outcome variances. However, different types of heterogeneity of outcome variances may occur. For instance, the outcome variances in two cells of the 2×2 design may be equal to each other and lower than those in the other two cells of the design. Alternatively, the outcome variances may differ between all four cells. We consider three scenarios, related to the outcome variance at cluster level in cell (rc) of the 2×2 design:

(1) $\sigma^2_{0(11)} = \sigma^2_{0(12)} < \sigma^2_{0(21)} = \sigma^2_{0(22)}$

(2) $\sigma^2_{0(11)} < \sigma^2_{0(12)} = \sigma^2_{0(21)} < \sigma^2_{0(22)}$

(3) $\sigma^2_{0(11)} < \sigma^2_{0(12)} < \sigma^2_{0(21)} < \sigma^2_{0(22)}$

We also consider the same three scenarios for heterogeneity of person level variance. Note that in scenario 1, the outcome variance is affected by one treatment only, in scenario 2 it is affected to the same extent by both treatments, and in scenario 3 it is affected by both treatments but not to the same extent.

Each of the three scenarios may occur at either or at both levels of the data structure (i.e. one level heterogeneous and the other homogeneous or both levels heterogeneous) and with different extents

of heterogeneity. The heterogeneity is measured through the coefficient of variation. At cluster level, it is defined as:

$$CV_{\sigma_0^2} = \frac{\sqrt{\sum_{rc} (\sigma_{0rc}^2 - \sigma_{0(avg)}^2)^2 / 4}}{\sigma_{0(avg)}^2} \tag{7}$$

Analogously, the coefficient of variation for the individual level variance can be obtained from equation (7) by substituting σ_ε^2 for σ_0^2.

We assume equidistance of the heterogeneous outcome variances across the four conditions (see Appendix 2 for further details). This implies that the maximum value of the coefficient of variation is reached when the smallest variance is zero and the largest variance is twice the average variance. However, this maximum differs between scenarios. For the sake of comparability between scenarios, we therefore computed the ratio of the actual coefficient of variation, at either cluster ($CV_{\sigma_0^2}$) or individual ($CV_{\sigma_\varepsilon^2}$) level, to the maximum possible coefficient of variation ($CV_{\max}$), and used this ratio as measure of heterogeneity of variances. A more intuitive, though less complete measure of heterogeneity in case of more than two treatment arms, is the ratio of the largest to the smallest variance. To improve the interpretability of the $CV/CV_{\max}$ measure, Table 2 shows selected values of this measure with the corresponding ratio of largest to smallest variance, denoted by λ for each of the heterogeneity scenarios. As shown in the table, strong amounts of heterogeneity are covered already by $CV/CV_{\max}$ values up to 0.60, that is, by CV values up to 60% of their maximum value. Note that, in terms of the actual coefficient of variation for a given value of $CV/CV_{\max}$, scenarios 1 and 2 are the extreme cases, with scenario 3 in-between. After all, in scenario 1 all four variances are as far removed from the mean variance as possible, given the CV and $CV_{\max}$, whereas in scenario 2 two of the four variances are equal to the mean, and in scenario 3 these two are closer to the mean than in scenario 1 (see the definition of the three scenarios at the beginning of this section).

3 Results

In this section, we show the relative efficiency of a balanced design, relative to the optimal designs for heterogeneous variance, as a function of the level (cluster, individual, or both) and extents (measured by $CV/CV_{\max}$) of heterogeneity of the variance, for various cost ratios and ICCs, under variance scenario 1 and variance scenario 2. We do not show results related to variance

Table 2. Ratio of the largest and the smallest variance (λ) for selected values of the coefficient of variation CV as measure of heterogeneity of variances in scenarios 1 to 3.

	$CV/CV_{\max}$	0	0.30	0.35	0.40	0.50	0.60	0.80	0.85	0.90
Scenario (1)	λ	1	1.86	2.08	2.33	3	4	9	12.24	19
Scenario (2)	λ	1	1.86	2.08	2.34	3.02	4.03	9.17	12.65	19.76
Scenario (3)	λ	1	1.86	2.09	2.35	3.03	4.05	9.26	12.82	20.19

This table only shows selected values of $CV/CV_{\max}$. Higher values of $CV/CV_{\max}$ imply (almost) zero variance in the treatment condition with the smallest variance, which is improbable in reality.

scenario 3 as this scenario is in-between scenario 1 and scenario 2, see section 2.3. At the end of the section we then provide a practical application of our results.

Figures 1 and 2 show the RE of the balanced design under variance scenario 1 (Figure 1) and under variance scenario 2 (Figure 2). The two figures contain two plots each, one for a cluster-to-person cost ratio equal to 9 and one for a cost ratio equal to 99. Each plot contains RE functions for two ICC values (0.01, thin lines, and 0.10, thick lines) and all three possible types of heterogeneity of

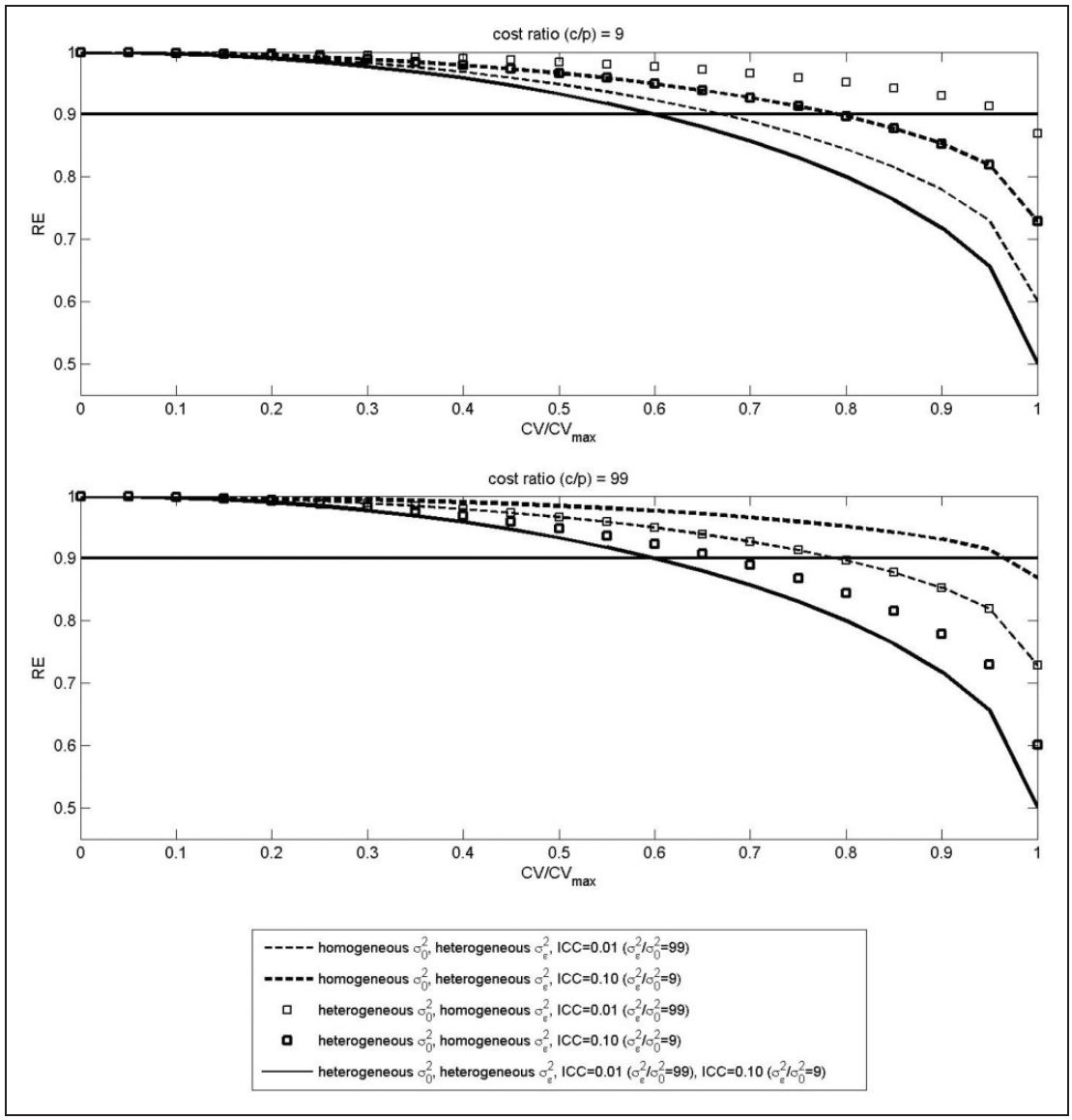

Figure 1. RE of a balanced design within variance scenario (1), for two cost ratios (9 and 99), two ICCs ($p = 0.01$ and $p = 0.10$,) and all three types of heterogeneity (cluster, individual, or both).

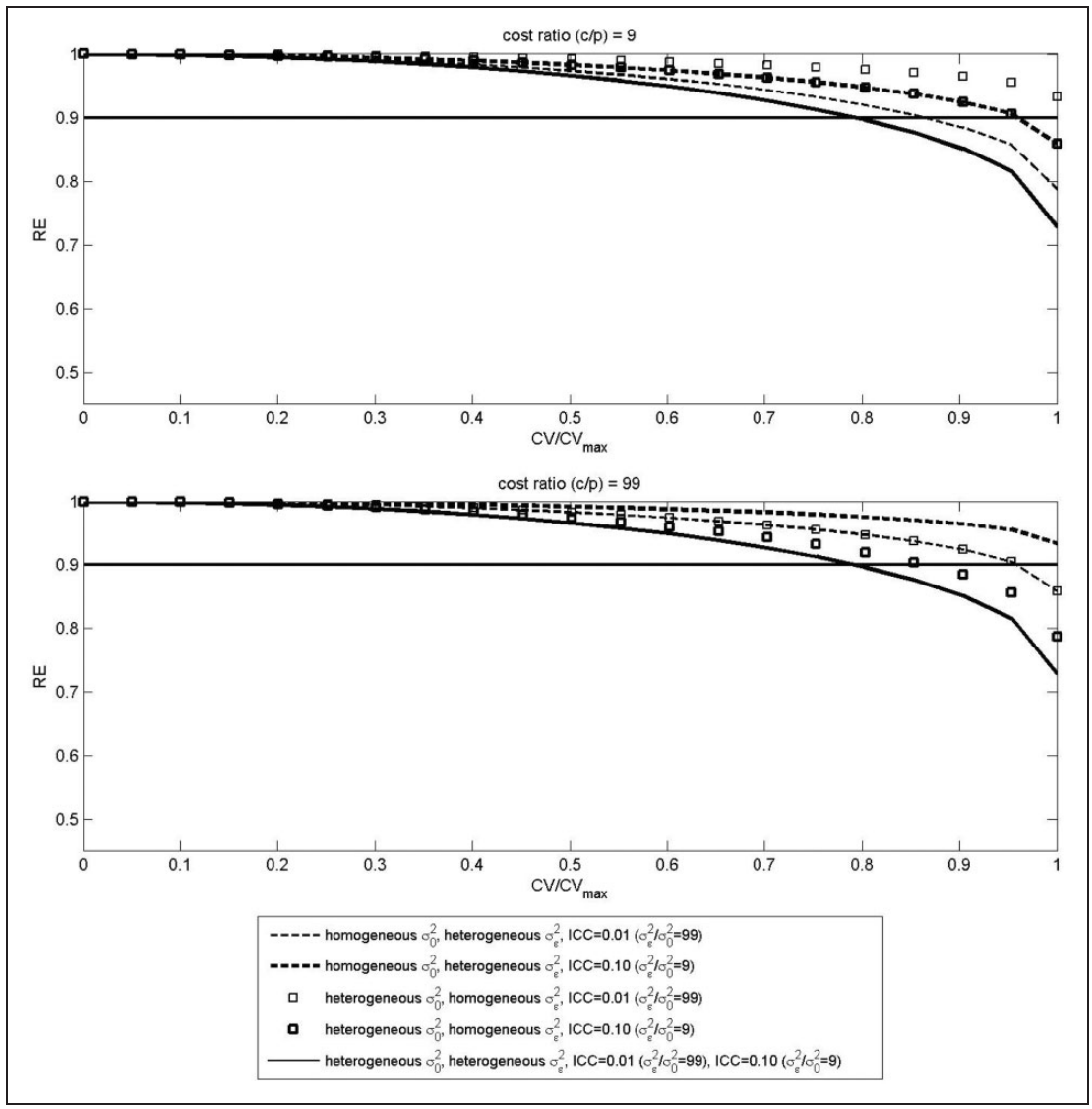

Figure 2. RE of a balanced design within variance scenario (2), for two cost ratios (9 and 99), two ICCs ($p = 0.01$ and $p = 0.10$) and all three types of heterogeneity (cluster, individual, or both).

variance within a scenario: heterogeneity at the cluster level only (squared markers), at the person level only (dashed lines), and at both levels (solid lines). Further, to help interpretation of the results, all plots in the two Figures contain a horizontal line indicating a relative efficiency of 0.90. Results for the intermediate cost ratio of 19 and ICC of 0.05 are not shown, as the RE results obtained for these values are in-between the RE results shown in Figures 1 and 2 for the two extreme values of the cost ratio and ICC.

Focus first on Figure 1. Note that there are four instead of six curves in each plot. This is due to two implications of the RE equation (6), of which formal proofs are available upon request.

First, one line with squared markers (heterogeneity at cluster level) and one dashed line (heterogeneity at individual level) overlay each other. This is because the RE does not depend on the level at which the heterogeneity occurs when the cluster-to-person cost ratio equals the person-to-cluster variance ratio and the outcome variance is heterogeneous at one level only. Secondly, each plot shows one instead of two continuous lines. This is because, when there is heterogeneity at both levels, the relative efficiency does not depend on the cost ratio and ICC anymore.

Let us now turn to the implications of Figure 1 for the relative efficiency of the balanced design under heterogeneity of variance. In both plots all curves are above a RE = 0.90 up to high extents of heterogeneity of the variance. For instance, when the variance is heterogeneous at both levels (solid line), the relative efficiency is above 90% when $CV/CV_{\max}$ is at most 0.60. This corresponds to having a largest-to-smallest variance ratio (λ) of at most 4. Further, all other curves are above the solid curve and so the case of heterogeneity at both levels is the worst case in terms of the RE of the balanced design.

Now consider Figure 2 for variance scenario 2. This shows the same trends as Figure 1. The important difference between the two is that the lines in Figure 2 are higher than those in Figure 1, and so the balanced design is even more efficient for variance scenario 2 than for scenario 1. In particular, the RE for the worst case of heterogeneity at both levels is now above 90% for $CV/CV_{\max}$ up to almost 0.80, corresponding to λ equal to almost 9. As explained in section 2.3, this is because in scenario 1 all four cells have a nonaverage variance, against only two cells in scenario 2. Apart from this difference in RE, the plots in Figure 1 and in Figure 2 are very similar.

Furthermore, Figures 1 and 2 show two other results which are worth mentioning. Firstly, the two figures show that, when comparing lines *within* plots, as the ICC increases, the relative efficiency decreases under heterogeneity at cluster level (i.e. thick squared markers are below thin squared markers), but increases under heterogeneity at individual level (i.e. thick dashed lines are above thin dashed lines). This finding can be understood by considering that, keeping the total variance constant, an increasing ICC means an increasing cluster level variance and decreasing individual level variance, which in turn increases the effect of heterogeneity of variance at cluster level and decreases the effect of individual level variance.

Secondly, comparing *between* plots for different cost ratios, it can be seen that when heterogeneity is at cluster level only (squared markers), the relative efficiency decreases as the cost ratio increases, whereas the opposite occurs when heterogeneity is at individual level only (dashed lines). The explanation for this finding is that, as the cost ratio increases, n_{rc} increases and k_{rc} decreases (see equation (5)). This in turn increases the effect of cluster level variance and decreases the effect of individual level variance on the sampling variance of the fixed effects (see equation (3)). It likewise increases the effect of cluster level heterogeneity of variance, and decreases the effect of individual level heterogeneity of variance, on the relative efficiency of the balanced design.

Let us now consider the implications of our results to the statin trial by Jones et al.[19] Remember from the Introduction that this was a 2×2 CRT comparing two decision aid tools for patients with diabetes, new versus standard, and two methods of delivery of the tool, before versus during the patient's visit to the clinician introduced in section 1. One of the outcomes of interest was a summary acceptability score, obtained by averaging the scores from a 5-question, 7-point Likert scale of acceptability. The researchers reported an almost balanced allocation of patients in each condition, with $n = 26$ patients in each of the two statin choice conditions (SB and SD) and $n = 23$ patients in each of the two standard pamphlet conditions (PB and PD). Unfortunately the researchers only reported a total number of endocrinologists equal to 21 but did not report the allocation ratio of the endocrinologists into each condition. Given the reported balanced allocation at the patients' level and the use of a balanced allocation at both cluster and individual levels in all

other trials mentioned in section 1, we assume it was balanced at cluster level as well, with approximately $k = 5$ endocrinologists per condition. The variation of the summary acceptability score ranged between $\sigma_{11}^2 = 1.3^2$ in the SD group and $\sigma_{22}^2 = 2.7^2$ in the PB group. These outcome variances are sums of the cluster level and individual level variances[16] and the authors did not report either the variance per level or the ICC. We therefore assume the worst case of heterogeneity of variance at both levels, with the same extents of heterogeneity at the two levels. So the ratio between the largest and smallest variances is $\lambda = 2.7^2/1.3^2 = 4.3$, corresponding to $CV/CV_{\max} \approx 0.60$ (see Table 2). Consider now variance scenario 1, which is the worst case. Then, from Figure 1 this extent of heterogeneity gives a relative efficiency of the balanced design equal to $RE \approx 0.90$, meaning that the use of a balanced allocation gave only a small loss of efficiency compared with the optimal allocation.

4 Sample size computation for a balanced 2×2 cluster randomized trial

In section 3 we have seen that the popular balanced design is highly efficient as compared to the optimal design, up to substantial amounts of heterogeneity of the variance. The next question then becomes how to compute the sample size for a balanced 2×2 CRT with heterogeneous variance, to have sufficient power for a prespecified effect size and type I error risk α. This question is addressed in the present Section. Sample size formulae are available in literature for cluster randomized trials with two treatment arms and homogeneous variance,[3,24] but not for 2×2 cluster randomized trials with heterogeneous variance. In this section we show how the sample size computation for the two-arm design with homogeneous variance can be extended to the 2×2 design with heterogeneous variance. We assume heterogeneity of variance at both cluster and individual levels, as this was the worst case in terms of the RE of the balanced design. At the end of the section we illustrate this extension again using the statin choice example.

To compute the number of clusters k, with n subjects per cluster, needed in each arm of a two-arm CRT to achieve a desired power (1-γ) and a two-tailed type I error rate α for the treatment effect a commonly used equation is[24]:

$$k = \left(z_{1-\gamma} + z_{\alpha/2}\right)^2 \times \frac{(\sigma_{Y1}^2 + \sigma_{Y2}^2) \times [1 + (n-1)\rho]}{n \times (\mu_1 - \mu_2)^2} \tag{8}$$

where $z_{1-\gamma}$ and $z_{\alpha/2}$ are respectively the 100(1-γ)th and 100(1-α/2)th percentiles of the standard normal distribution corresponding to the desired power (1-γ) and a type I error risk α; $\mu_1 - \mu_2$ is the unknown treatment effect of interest; $\sigma_{Y1}^2 = \sigma_{0(1)}^2 + \sigma_{\varepsilon(1)}^2$ and $\sigma_{Y2}^2 = \sigma_{0(2)}^2 + \sigma_{\varepsilon(2)}^2$ are the total outcome variances in each of the two conditions, where σ_0^2 and σ_ε^2 are the cluster level and individual level variances and the numbers (1) and (2) in their subscript indicate the treatment arm to which they are related; $\rho = \sigma_{0(1)}^2/\sigma_{Y(1)}^2 = \sigma_{0(2)}^2/\sigma_{Y(2)}^2$ due to our assumption of the same amount of heterogeneity at both levels, which was seen to be the worst case for the balanced design; n is the number of patients included per cluster and can be optimized by using equation (5) under homogeneity of variance. Equation (8) gives the number of clusters needed per arm as a function of the cluster size n, of the ICC ρ, and of the outcome variance, which may differ between treatments. Further, it is based on a z-test applied to cluster means, and assumes known variances. However, with unknown variances, the treatment effect test is a t-test applied to cluster means.[3] Under heterogeneity of variance, the degrees of freedom of this test can be computed with the Welch or Satterthwaite formula,[25,26] which we express below as a function of two factors: the number of clusters k in each condition, and a measure of heterogeneity of the variance across the two

conditions, $\lambda = \sigma_{Y2}^2/\sigma_{Y1}^2$, where we assume that σ_{Y2}^2 is larger than σ_{Y1}^2, and both variances now apply to the cluster means (but λ is the same for individual data as for cluster means under the present worst case assumption of equal heterogeneity at cluster and individual level):

$$df = (k-1) \times \frac{(1+\lambda)^2}{(1+\lambda^2)} \tag{9}$$

If the number of clusters is small, say $k < 30$, the degrees of freedom of the t-statistic are so small that the percentage points of the t-distribution are noticeably larger than those of the z-distribution, which gives a loss of power. Hayes and Moulton[3] addressed this loss of power by adding one cluster per treatment arm directly in their sample size equation. We have extended this by computing the adjusted number of clusters (k^*) needed to achieve the desired power and type I error rate, as a function of the extents of heterogeneity λ. We have plotted the difference between the original k as calculated with equation (8) and the adjusted k^* in Figure 3, for two values of the planned number of clusters in each arm ($k = 10$ and $k = 20$), as a function of the heterogeneity of the variance λ ranging from 1 (homogeneity) to 20 ($CV/CV_{\max}$ almost one, see Table 2), assuming a two-tailed type I error risk $\alpha = 0.01$ or $\alpha = 0.05$, and a power $(1-\gamma) = 0.80$ or $(1-\gamma) = 0.90$.

Figure 3 (left plot) shows that, because of using a t-test instead of a z-test, with $k = 10$ clusters per condition and $\lambda = 1$ (homogeneous variances), regardless of whether the required power is 80% or 90%, one or two extra clusters are needed per condition, for, respectively, $\alpha = 0.05$ and $\alpha = 0.01$. When $\lambda > 1$ (heterogeneous variance), the number of additional clusters needed per condition increases with λ up to a maximum of approximately four clusters when $\alpha = 0.01$ and approximately two clusters when $\alpha = 0.05$. For a higher initially planned number of clusters per condition (right plot, $k = 20$), the number of additional clusters needed is somewhat smaller, as a larger number of clusters gives more degrees of freedom of the t-distribution (which implies that the percentage points of the t-distribution approach those of the z-distribution).

Let us now consider how the sample size computation for a two-arm design can be extended to a 2×2 design, by noting that each of the three treatment contrasts β_1 to β_3 in equation (1) compares the average outcome in a pair of conditions with the average outcome in the other pair of conditions. For instance, if the contrast of interest is β_1, then we compare $\mu_1 = (\mu_{11} + \mu_{12})/2$ with $\mu_2 = (\mu_{21} + \mu_{22})/2$, where μ_{11}, μ_{12}, μ_{21}, and μ_{22} are the outcomes in each cell of the 2×2 design, as defined in section 2.1. Likewise, each of the other two treatment contrasts (β_2 and β_3) comes down to the comparison between two pairs of cells (see lower part of Table 1 for details). Therefore each contrast in a 2×2 design can be regarded as a treatment effect in a two-arm design. In this case, however, the original k in equations (8) and (9) and the adjusted k^* following Figure 3, represent the number of clusters in a pair of conditions rather than the number of clusters in one treatment arm. Thus, we need one or two (instead of two or four as in the two-arm design) extra clusters per condition, depending on α.

We now use the statin trial by Jones et al.[19] to illustrate the sample size calculation for a 2×2 CRT. Unfortunately the authors did not report their sample size computation. Therefore, we here report our sample size calculation based on information on the outcome measure used and its reported variability. Since the summary acceptability score ranged from 1 to 7, we assume that for each of the three treatment contrasts β_1 to β_3 in equation (1), an effect equal to 1 is clinically relevant (first column of Table 3). Assuming variance scenario 1, $\sigma_{11}^2 = \sigma_{12}^2 = 1.3^2$ and $\sigma_{21}^2 = \sigma_{22}^2 = 2.7^2$ then gives a variance for each pair of conditions being compared as shown in columns 2 and 3 of Table 3, which translates into an average variance of $\sigma_h^2 = 2.12^2$, for each contrast $h = 1, 2, 3$, and an effect size $d = (\mu_1 - \mu_2)/\sigma_h = 0.47$, again for each contrast.

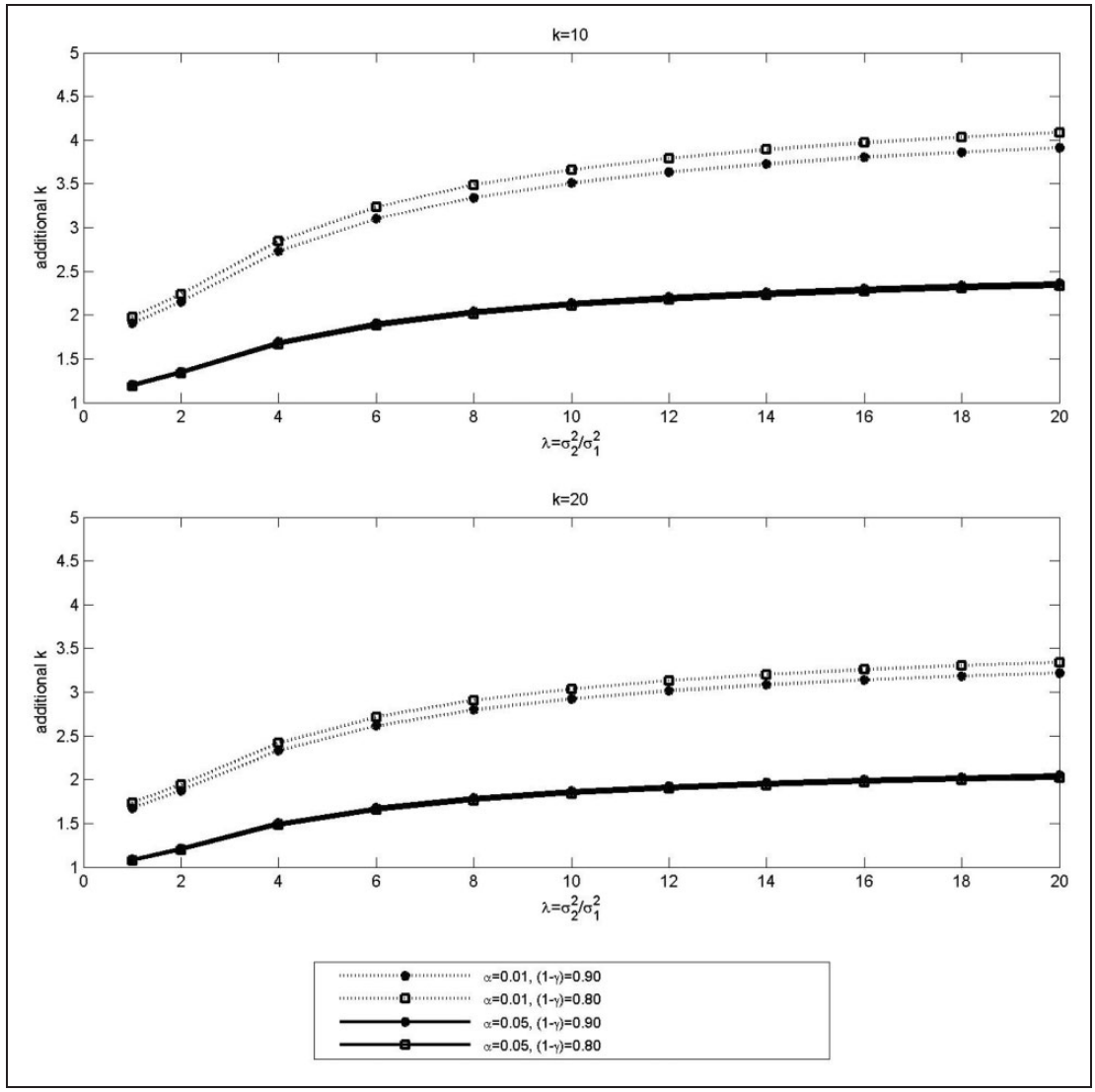

Figure 3. Additional number of clusters k needed in each arm of a two-arm CRT, for a t-test with dof from the Welch formula compared to a z-test, a two-tailed type I error $\alpha = 0.01$ and $\alpha = 0.05$ and a power $(1\text{-}\gamma) = 0.90$ and $(1\text{-}\gamma) = 0.80$, as a function of $\lambda = \sigma^2_{Y2}/\sigma^2_{Y1}$.

Column 4 of Table 3 shows the number of endocrinologists, each with 5 patients, needed in each pair of conditions to achieve a type I error rate $\alpha = 0.01$ and a power $(1 - \gamma) = 0.80$ and assuming an ICC $\rho = 0.05$, computed using (8) which is based on a z-test. Column 5 of Table 3 shows the extents of heterogeneity of the variance for each contrast, given by the ratio between the variances in columns 2 and 3. Column 6 shows the degrees of freedom of a t-test, computed using equation (9), as a function of the extents of heterogeneity in column 5. The last column of Table 3 shows the additional number of endocrinologists needed per pair of conditions for a t-test with the degrees of

Table 3. Unadjusted (k) and additional ((k^*-k)) number of endocrinologists, each with 5 patients, needed in each pair of conditions of the statin choice trial, assuming a clinically relevant effect size d = 0.47 for each treatment effect β_1 to β_3, a type I error rate $\alpha = 0.01$, a power $(1-\gamma) = 0.80$, an ICC = 0.05, and an extent of heterogeneity of variance λ as shown in column 5.

	σ_1^2	σ_2^2	k	$\lambda = \sigma_2^2/\sigma_1^2$	df	$(k^* - k)$
$\beta_1 = \frac{\mu_{11}+\mu_{12}}{2} - \frac{\mu_{21}+\mu_{22}}{2} = 1$	$\frac{\sigma_{11}^2+\sigma_{12}^2}{2} = 1.3^2$	$\frac{\sigma_{21}^2+\sigma_{22}^2}{2} = 2.7^2$	25.38	4.31	35.104	2.40
$\beta_2 = \frac{\mu_{11}+\mu_{21}}{2} - \frac{\mu_{12}+\mu_{22}}{2} = 1$	$\frac{\sigma_{11}^2+\sigma_{21}^2}{2} = 2.1^2$	$\frac{\sigma_{12}^2+\sigma_{22}^2}{2} = 2.1^2$	25.38	1	48.755	1.69
$\beta_3 = \frac{\mu_{11}+\mu_{22}}{2} - \frac{\mu_{12}+\mu_{21}}{2} = 1$	$\frac{\sigma_{11}^2+\sigma_{22}^2}{2} = 2.1^2$	$\frac{\sigma_{12}^2+\sigma_{21}^2}{2} = 2.1^2$	25.38	1	48.755	1.69

freedom shown in column 6. So, according to the z-test approach in equation (8), $k = 25.38$ endocrinologists are needed per pair of conditions in the 2×2 design, that is, 13 endocrinologists per cell. Next, taking into account the heterogeneity of variance as shown in column 5, an additional number of $(k^* - k) = 2.40$ and $(k^* - k) = 1.69$ endocrinologists is needed per pair of conditions to test β_1 respectively to test β_2 and β_3, see column 7 of Table 3 and see Figure 3. Thus we need 27.78 endocrinologists per pair of conditions to estimate β_1 and 27.07 endocrinologists per pair of conditions to estimate β_2 and β_3, implying 14 endocrinologists per cell and 56 in total, to have sufficient power for all three effects of interest. Of course a different effect size or a different extent of heterogeneity of variance would change these numbers.

5 Discussion

At the design stage of a study a balanced allocation of units, which is optimal for homogeneous variance, is usually adopted. The outcome variance may however be heterogeneous across the treatment conditions implying a loss of efficiency (power and precision) of the balanced design as compared to the optimal design. For 2×2 CRTs, we therefore derived and plotted the efficiency loss of a balanced design as a function of the level (cluster, individual, both) and extent of heterogeneity of the variance, within three heterogeneous variance scenarios. We also provided guidelines for the additional number of clusters needed in each condition of a balanced 2×2 CRT to compensate the efficiency loss due to heterogeneity of variance as compared with homogeneity of variance on which sample size calculations are usually based.

Our results can be summarized as follows: Firstly, the relative efficiency of the balanced design is lower when the variance is heterogeneous at both cluster and individual levels, than when it is heterogeneous at one level only, within each of the three variance scenarios and for any extent of heterogeneity. Secondly, even when the outcome variance is heterogeneous at both levels, the loss in efficiency of the balanced design as compared with the optimal design, is less than 10%, up to high extents of heterogeneity of the variance, that is, up to a largest-to-smallest variance ratio of 4 under variance scenario 1, and a variance ratio of almost 9 under variance scenarios 2 and 3. Finally, when a balanced design is adopted and the sample size is calculated based on homogeneous variances, the number of clusters per condition in the 2×2 design must be increased with 1 or 2, depending on the anticipated extent of heterogeneity of variance and on whether the desired type I error rate is $\alpha = 0.05$ or $\alpha = 0.01$. In conclusion, the balanced 2×2 design is quite robust against heterogeneity of variance at each design level (cluster, individual), and the loss of efficiency due to heterogeneity of variance is easily compensated by one or two additional clusters per treatment condition.

The finding that the balanced design is quite efficient even under heterogeneity of variance is good news for sample size planning in 2×2 CRTs. Of course, such planning still requires some knowledge of the outcome variance at each design level (cluster, individual), since the variance of the treatment effect strongly depends on that variance, whether homogeneous or heterogeneous (see equation (8)). Unfortunately, of all published 2×2 CRT we reviewed, almost none reported this information. Reporting guidelines for CRTs, such as the CONSORT[27] guidelines should perhaps recommend that such information is reported.

Our work focuses on CRTs with a 2×2 factorial design, yet our results can also be applied to CRTs with a two-arm design. For instance, variance scenario 1 has the smallest two variances and the largest two variances equal (see section 2.3). This can be seen as a 2-arm design with one arm having variance equal to the smallest two, and the other arm having variance equal to the largest two variances in a 2×2 design. Further, as explained in section 4, each contrast in a 2×2 design compares the average outcome in a pair of conditions with the average outcome in the other pair of conditions and can be regarded as a treatment effect in a two-arm design.

The extents of heterogeneity of the outcome variance are measured through the coefficient of variation, which is defined as the ratio between the standard deviation of the four heterogeneous outcome variances across the treatment conditions and the average of the four variances, at either cluster or individual level. This should not be confused with the coefficient of variation of the true outcome mean between clusters in the same treatment condition, defined as the ratio between the cluster level variance (σ_0^2) and the true outcome mean (μ) across all clusters in a treatment condition, and can be used to compute the sample size for a CRT.[3] This gives the same result as our equation (8) (for details, see equations (7.9) and (7.12) in Hayes and Moulton[3]).

Our computations are based on the assumption of equal clusters size within the same treatment condition, which is optimal for estimating treatment effect with maximum precision, given homogeneity of costs and of the intraclass correlation (see section 2.1). However, due to variation of the actual size of the organizational units under study (e.g. schools or general practices) and to the nonresponse and dropout, cluster sizes do vary in practice and this leads to some loss of efficiency. There is evidence in literature[28] that this loss can be approximated by a very simple function of the coefficient of variation (CV = mean/standard deviation) of the cluster size and is usually about 10% or less, which can be compensated by sampling 11% more clusters.

We have restricted our work to three heterogeneous variance scenarios, in order to keep the number of combinations within reasonable boundaries. Unfortunately, as the outcome variance per treatment condition is rarely reported, there is lack of empirical evidence from the literature on the most likely scenarios of heterogeneity of variance. Therefore we have chosen three scenarios that are plausible in the sense that these follow from the presence of one treatment effect on the outcome variance (scenario 1), or of two treatment effects on the variance, which can either be equally large (scenario 2) or not (scenario 3). In scenario 1, the smallest two variances are equal as are the largest two. In scenario 2 the two middle variances are equal to each other and to the average variance across all four cells. Therefore, scenario 1 is more extreme than scenario 2. Scenario 3 assumes equidistant variances and is thus in-between scenario 1 and scenario 2, not only conceptually, but also in terms of results (not shown here). By releasing the assumption of equidistance other possible realistic scenarios may be obtained from scenario 3. However, these scenarios are then again in-between scenario 1 and scenario 2. Similarly, we assumed either heterogeneity of variance at one level (cluster or individual) only, or the same extent of heterogeneity at both levels within each scenario. Having a different extent of heterogeneity at each level is in-between the extremes of heterogeneity at one level and equal extent of heterogeneity at both levels, and can be expected to

give results similar to those in Figures 1 and 2. In short, the present results cover a wide range of cases of heterogeneous variance and many other cases are in-between the present ones.

Further, throughout, we used the variance of one treatment effect estimator (c-criterion) as optimality criterion, and we saw that two other criteria, respectively the sum of the variances of the three treatment effects (A_s-criterion), and the sum of the variances of all four fixed effects (A-criterion), were proportional to our c-criterion for the model in equation (1), thus giving the same results as the c-criterion. A drawback of the A- and As-criteria, well-known in optimal design literature, is that they are not generally invariant to transformations of the scale of the linear predictors. In other words, rescaling one or more predictors may result in different optimal A- and A_s-designs. However, this is not a serious problem for the present paper. Each predictor contrasts one pair of treatment conditions with the other pair. So changing the scaling of one predictor would need a corresponding rescaling of the other predictors, at least if each contrast is of equal interest. This would give the same optimal design as before the rescaling.

Two limitations of our work have to be mentioned. First of all, all derivations were based on the assumption of homogeneity of costs across the four conditions. Ongoing work considers heterogeneous costs and outcome variances simultaneously. Secondly, our work concerns continuous outcomes. A logical extension would be to also consider binary outcomes. Previous studies[29,30] suggest that the optimal design results for the mixed linear model in equation (1) also apply to the mixed logistic models but with two modifications: (1) the outcome variance within each treatment condition at the individual level $\sigma^2_{\varepsilon rc}$ must be replaced with the expression $\delta^2 = \left(\pi_{(rc)} \times \left(1 - \pi_{(rc)}\right)\right)^{-1}$ where $\pi_{(rc)}$ is the probability of the outcome in the (rc)-th condition in a 2×2 design, which is always heterogeneous unless there is no treatment effect at all; (2) for mixed logistic regression the number of clusters as computed with equations (5) and (8), must be increased by 10% and up to 25% according to simulations (for details, see Candel and van Breukelen[29] and Moerbeek et al.[30]). However, the mathematics of the mixed logistic regression model are more complicated than the mathematics of the mixed linear model. Moreover, closed forms like equations (2) and (3) give at best a crude approximation, which must be complemented with simulation studies. Likewise, other categorical data give heterogeneity. For instance, count data are often modeled with the Poisson distribution. Then the outcome variance equals the mean, implying that any treatment effect on the mean implies heterogeneity of variance. Therefore, our work needs to be extended not only to binary data but also to other categorical data including count data.

Conflict of interest

None declared.

Funding

This research was supported by Research Grant number 400-09-396 from the Netherlands Organization for Scientific Research (NWO).

References

1. Moerbeek M, Van Breukelen GJP and Berger MPF. Design issues for experiments in multilevel populations. *J Educ Behav Stat* 2000; **25**: 271–284.
2. Raudenbush SW. Statistical analysis and optimal design for cluster randomized trials. *Psychol Meth* 1997; **2**: 173–185.
3. Hayes RJ and Moulton LH. *Cluster randomized trials*. Boca Raton, FL: CRC Press LLC, 2009.
4. Berger MPF and Wong WK. *An introduction to optimal designs for social and biomedical research*. New York, NY: Wiley, 2009.

5. Atkinson AC, Donev AN and Tobias RD. *Optimum experimental designs, with SAS*. Oxford: Clearedon Press, 2007.
6. Schouten HJA. Sample size formula with a continuous outcome for unequal group sizes and unequal variances. *Stat Med* 1999; **18**: 87–91.
7. Jan S and Shieh G. Determining sample sizes for precise contrast analysis with heterogeneous variances. *J Educ Behav Stat* 2014; **39**: 91–116.
8. Wong WK and Zhu K. Optimum treatment allocation rules under a variance heterogeneity model. *Stat Med* 2008; **27**: 4581–4595.
9. Guo JH and Luh WM. Efficient sample size allocation with cost constraints for heterogenous-variance group comparison. *J Appl Stat* 2013; **40**(12): 2549–2563.
10. Moerbeek M and Wong WK. Sample size formulaes for trials comparing group and individual treatments in a multilevel model. *Stat Med* 2008; **27**: 2850–2864.
11. Candel MJJM and van Breukelen GJP. Sample size calculation for treatment effects in randomized trials with fixed cluster sizes and heterogeneous intraclass correlations and variances. *Stat Meth Med Res*. DOI: 10.1177/0962280214563100.
12. Ausems M, Mesters I, Van Breukelen GJP, et al. Short-term effects of a randomised computer-based out-of-school smoking prevention trial aimed at Dutch elementary school children. *Prev Med* 2002; **34**: 581–589.
13. Eccles M, Steen N, Grimshaw J, et al. Effect of audit feedback, and reminder messages on primary-care radiology referrals: A randomized trial. *Lancet* 2001; **357**: 1406–1409.
14. Ayles HM, Sismanidis C, Beyers N, et al. ZAMSTAR, The Zambia South Africa TB and HIV Reduction study: Design of a 2x2 factorial community randomized trial. *Trials* 2008; **9**: 63.
15. Cheater FM, Baker R, Reddish S, et al. Cluster randomized controlled trial of the effectiveness of audit and feedback and educational outreach on improving nursing practice and patient outcomes. *Med Care* 2006; **44**: 542–551.
16. Tazeen HJ, Hatcher J, Poulter N, et al. Community-based interventions to promote blood pressure control in a developing country. *Ann Int Med* 2009; **151**: 593–601.
17. Ravaud P, Giraudeau B, Logeart I, et al. Management of osteoarthritis (OA) with an unsupervised home based exercise programme and/or patient administered assessment tools. A cluster randomised controlled trial with a 2x2 factorial design. *Ann Rheum Dis* 2004; **63**: 703–708.
18. Cals JWL, De Bock L, Beckers PJH, et al. Enhanced communication skills and C-reactive protein point-of-care testing for respiratory tract infection: 3.5-year follow-up of a cluster randomized trial. *Ann Fam Med* 2013; **11**: 157–164.
19. Jones LA, Weymiller AJ, Shah N, et al. Should clinicians deliver decision aids? Further exploration of the statin choice randomized trial results. *Med Decis Making* 2009; **29**: 468–474.
20. Lemme F, van Breukelen GJP and Berger MPF. Efficient treatment allocation in two-way nested design. *Stat Meth Med Res* 2013; Published online: 12 Sept 2013. DOI: 10.1177/0962280213502145.
21. Searle SR and Pukelsheim F. Effects of intraclass correlation on weighted averages. *Amer Statist* 1986; **40**: 103–105.
22. Adams G, Gulliford MC, Ukoumunne OC, et al. Patterns of intracluster correlation from primary care research to inform study design and analysis. *J Clin Epidemiol* 2004; **57**: 785–794.
23. Eldridge SM, Ashby D, Feder GS, et al. Lessons for cluster randomized trials in the twenty-first century. *Clin Trials* 2004; **1**: 80–90.
24. Van Breukelen GJP and Candel MJJM. Calculating sample sizes for cluster randomized trials: We can keep it simple and efficient! *J Clin Epidemiol* 2012; **65**: 1212–1218.
25. Welch BL. The significance of the difference between two means when the population variances are unequal. *Biometrika* 1938; **29**: 350–362.
26. Satterthwaite FE. Synthesis of variance. *Psychometrika* 1941; **6**: 309–316.
27. Consolidated Standard of Reporting Trials (CONSORT), http://www.consort-statement.org/ (2010, accessed 26 March 2015)
28. Van Breukelen GJP, Candel MJJM and Berger MPF. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Stat Med* 2007; **26**: 2589–2603.
29. Candel MJJM and van Breukelen GJP. Sample size adjustments for varying cluster sizes in cluster randomized trials with binary outcomes analyzed with second-order PQL mixed logistic regression. *Stat Med* 2010; **29**: 1488–1501.
30. Moerbeek M, van Breukelen GJP and Berger MPF. Optimal experimental designs for multilevel logistic models. *J Roy Stat Soc D: Statist* 2001; **50**: 17–30.

Appendix 1

Computation of the optimal sample size allocation and budget split for the c-criterion

Using equation (4), express the number of clusters k_{rc} in each cell (rc) of the 2×2 design as a function of B_{rc}, c, p, and n_{rc} and substitute it into equation (3). Equation (10) is then obtained:

$$var(\hat{\beta}_h) = \sum_{rc} w^2_{rc(h)} \times \frac{\left(n_{rc} \times \sigma^2_{0rc} + \sigma^2_{\varepsilon rc}\right) \times (c + p \times n_{rc})}{B_{rc} \times n_{rc}} \tag{10}$$

where $w^2_{rc(h)}$ is the squared weight in the lower part of Table 1 within any cell (rc) of the 2×2 design.

Equation (10) is the sum of four terms, one per cell (rc) of the 2×2 design. Each term is minimized by n_{rc} in equation (5), given B_{rc}, and so this also gives the minimum of the sum

in (10). Substitute the optimal n_{rc} given by equation (5) into equation (10). Equation (11) is then obtained:

$$var(\hat{\beta}_h) = \sum_{rc} w^2_{rc(h)} \times \frac{\left(\sqrt{c \times \sigma^2_{0rc}} + \sqrt{p \times \sigma^2_{\varepsilon rc}}\right)^2}{B_{rc}} \tag{11}$$

As the equation shows, the c-criterion now depends on the costs, on the cluster level and individual level variance and the available budget in each cell of the 2×2 design.

Equation (11) is then minimized, to obtain an equation for the optimal budget split across conditions B_{rc}, as a function of the allocation of the total budget across the four cells of the 2×2 design, subject to the constraint of a fixed total budget $B = \sum_{rc} B_{rc}$, using the Lagrange multiplier. So, we minimize the function:

$$f(B_{rc},\ \lambda) = \sum_{rc} w^2_{rc(h)} \times \frac{\left(\sqrt{c \times \sigma^2_{0rc}} + \sqrt{p \times \sigma^2_{\varepsilon rc}}\right)^2}{B_{rc}} + \lambda\left(B - \sum_{rc} B_{rc}\right) \tag{12}$$

where λ is the Lagrange multiplier and the term within brackets is zero due to the constraint. This gives equation (13) for the optimal budget split across conditions:

$$B_{rc} = B \times \frac{\left(\sqrt{c \times \sigma^2_{0rc}} + \sqrt{p \times \sigma^2_{\varepsilon rc}}\right)}{S}, \text{ where } S = \sum_{rc=(11)}^{(22)} \left(\sqrt{c \times \sigma^2_{0rc}} + \sqrt{p \times \sigma^2_{\varepsilon rc}}\right) \tag{13}$$

After substituting equation (13) into (11), the c-optimality criterion can be rewritten into a function of total budget, costs and outcome variances only:

$$var(\hat{\beta}_h) = \frac{1}{16} \times \frac{\left(\sum_{rc}\left(\sqrt{c \times \sigma^2_{0rc}} + \sqrt{p \times \sigma^2_{\varepsilon rc}}\right)\right)^2}{B} \tag{14}$$

where the constant term 1/16 is due to the $w_{rc(h)}$ values given in Table 1. Apart from this constant, equation (14) is the numerator of equation (6). The same constant 1/16 also appears in $\text{var}\left(\hat{\beta}\right)$ for the balanced design, and so these constants cancel out in equation (6).

Appendix 2

Measuring the amount of heterogeneity of the variance

Define the heterogeneous variance in cell (rc), at either cluster (σ^2_{0rc}) or individual ($\sigma^2_{\varepsilon rc}$) levels as σ^2_{rc} (i.e. $\sigma^2_{0rc} = \sigma^2_{rc}$ and $\sigma^2_{\varepsilon rc} = \sigma^2_{rc}$) and the average of the four heterogeneous variances as $\sigma^2_{(avg)}$ (i.e. $\sigma^2_{(avg)} = \sigma^2_{0(avg)}$ and $\sigma^2_{(avg)} = \sigma^2_{\varepsilon(avg)}$). The amount of heterogeneity, at either or both levels of the data structure, can then be measured through the coefficient of variation, expressed as in equation (6) and which we denote in this section as CV, with $CV = CV_{\sigma^2_0}$ at the cluster level, and $CV = CV_{\sigma^2_\varepsilon}$ at the person level.

Within each heterogeneous variance scenario 1 to 3, assume the below:

(a) the average of the four heterogeneous variances, $\sigma^2_{(avg)}$, remains constant as the coefficient of variation increases from zero (homogeneity) to its maximum value;
(b) the heterogeneous variances σ^2_{rc} are equidistant.

The variance within each cell of the 2×2 design can be derived from (7), under assumptions (a) and (b), and expressed as function of the coefficient of variation CV and of the average outcome variance σ^2_{avg}. The smallest and largest variances are then expressed as:

$$\sigma^2_{rc} = \sigma^2_{(avg)} \times \left(1 \pm \frac{CV}{CV_{\max}}\right) \tag{15}$$

where $CV_{\max}$ is the maximum value reached by the coefficient of variation: 1 in scenario (1), 0.71 in scenario (2) and 0.75 in scenario (3). Assuming that σ^2_{11} is the smallest variance and σ^2_{22} is the largest variance, the minus-sign in the right part of the equation applies to σ^2_{11} and the plus-sign applies to σ^2_{22}. Note that assumptions (a) and (b) imply that the intermediate variances are obtained in scenario (2) as $\sigma^2_{12} = \sigma^2_{21} = \sigma^2_{(avg)}$, and in scenario (3) σ^2_{12} and σ^2_{21} obey (15) apart from multiplying the $CV/CV_{\max}$ ratio by 1/3.