Received: 22 July 2020 | Revised: 28 April 2023 | Accepted: 31 May 2023

DOI: 10.1002/sim.9820

RESEARCH ARTICLE

Statistics in Medicine WILEY

Sample size for partially nested designs and other nested or crossed designs with a continuous outcome when adjusted for baseline

Steven Teerenstra[1] | **Jessica Kasza**[2] | **Ruslan Leontjevas**[3,4] | **Andrew B. Forbes**[2]

[1]Department for Health Evidence, Section Biostatistics, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands

[2]School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia

[3]Department of Primary and Community Care, Radboud University, Medical Center, Nijmegen, The Netherlands

[4]Faculty of Psychology and Educational Sciences, Open University of The Netherlands, Heerlen, The Netherlands

Correspondence
Steven Teerenstra, Department for Health Evidence, Section Biostatistics, Radboud Institute for Health Sciences, Radboud University Medical Center, Internal Post 133, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands.
Email: steven.teerenstra@radboudumc.nl

In a randomized controlled trial, outcomes of different subjects may be independent at baseline, but correlated at a follow-up measurement due to treatment. This treatment-related clustering at follow-up can arise for instance because the treatment is given in a group or because subjects are treated individually but by the same therapist (therapist effect). There is substantial literature on the design and analysis of such trials when estimation of the intervention effect is based on a follow-up measurement (eg, directly after treatment or at a later time point). However, often the baseline measurement of the outcome is highly correlated with the follow-up measurement, and this information can be used in the analysis. For a randomized design with a baseline and a follow-up measurement, we compare sample size requirements for analyses with and without adjustment for this baseline measure. We show that adjusting for baseline reduces required sample size. This reduction depends on the variance of the difference between arms at baseline, the variance of this difference at follow-up, and the correlation between the two. From this, we derive sample size formulas for partially or fully nested designs, and cluster randomized trials with treatment as a partially or fully cross-classified factor. Also, we discuss situations where clusters are already present at baseline or where treatment by cluster interaction is present. For the partially nested design, we work out practical design considerations (eg, use of content-matter input, design factors and optimal allocation ratio) and investigate small sample properties of the sample size formula.

KEYWORDS

baseline adjustment, cross-classification, nesting, randomized controlled trial, sample size

1 | INTRODUCTION

In a randomized controlled trial, outcomes of different subjects may be independent at baseline, but correlated at a follow-up (ie, post-baseline) measurement due to treatment. This correlation at follow-up can for instance arise because treatment is given in a group setting. In such settings, interaction among group members, the presence of a domineering

group member and the skill of the group leader (eg, therapist) may influence cohesion, attendance patterns, attrition, engagement in treatment of the group members, and thus their outcomes.[1] Further, even if treatment is delivered individually, correlation can arise at follow-up if health care professionals (eg, acupuncturists, physical therapists, surgeons, etc.) treat multiple patients. Patients randomized to the same health care professional may be treated similarly because they all interact with that particular professional, and are exposed to their skills and attitudes. This is also known as the therapist effect.[2] In both cases, the correlation is induced by the treatment; such trials are also called individually randomized group therapy trials, and more recently have been referred to as individual randomized trials with post-randomization clustering.[3]

An example where the correlation arises from treatment in a group setting is a cluster randomized trial. In this case, outcomes of participants are nested within clusters. However, more general structures with participants nested within clusters are possible: nesting can be present in some of the arms and not in other arms (called a partially clustered or partially nested design[4,5]) or in all of the arms, but with different degrees of clustering (fully nested design). An example where the therapist effect induces correlation is when subjects receive a treatment from several therapists while each therapist treats different patients: subjects are then cross-classified with therapists. Also, the treatment in an arm may be delivered by a set of multiple therapists where each therapist provides a portion of the total treatment (multiple-membership models[6]).

Trials where nesting or therapist effects induce correlation are not infrequent in health care research. Firstly, there are many examples of partially nested trials: addition of telephone coaching to a physiotherapy program for knee osteoarthritis,[7] acupuncture compared with usual care by general practitioners for the treatment of persistent non-specific low back pain,[8] community leg ulcer clinics vs care as usual by district nurses,[9] community post-natal support workers on top of care as usual by midwives,[10] telephone befriending by volunteers in addition to usual health and social care.[11] In these trials, participants in the control arm are not clustered, while those in the treatment arm are nested within clusters. An example of a fully nested design is the comparison between nurse practitioners or general practitioners in primary care.[2,12] In this trial, the intervention arm clusters have a different cluster size and intracluster correlation than the control arm clusters. Furthermore, examples where the treatment to a subject arises from cross-classification with therapist are discussed by Moerbeek and Safarkhani[13] and Walwyn and Roberts.[14] In the case of full cross-classification (ie, each therapist delivers both the intervention and control), such designs are also known as crossed therapist designs. Finally, examples of multiple-membership models are provided by Roberts and Walwyn.[6]

When effect estimation is based only on a follow-up measurement (eg, taken immediately after the intervention or at a later time point), substantial literature is available on how to design and analyze partially nested/clustered designs,[1,2,4,5,15-20] and fully nested/clustered designs.[2] Such literature is also available for cross-classified designs.[13,14] However, guidance seems to be lacking for the design and analysis of such trials when a baseline measurement is included in the analysis.

In many cases, a baseline measurement of the outcome will be available, if only for assessing whether there is a systematic difference at baseline between the arms. Also, adjustment for baseline is often conducted in real practice, for example, Baldwin[1] (formula 35, p. 162) and Roberts and Roberts[2] (Table 4, p. 157) used baseline outcome values in their analyses. Typically, a baseline measurement is (easily) obtained as part of the recruitment procedure. Among the baseline covariates, the outcome score at baseline is often a strong predictor of the outcome scores at follow-up, and hence power could be increased or sample size could be reduced when the baseline measurement is incorporated in the analysis. In this article, we investigate how large the gain in precision / power or reduction in sample size can be when this baseline measurement is included. In all these situations, we assume that each subject has a baseline and follow-up measurement, so we have a cohort design. We do not consider issues of loss-to-follow-up, non-compliance with assigned treatment, or other similar complications.

As central point of this article in Section 2, we first collect the basic results underlying the sample size / power formula for a randomized, two-arm design with a baseline and a follow-up measurement. To illustrate the breadth of application, we then derive the standard error (SE) for some typical designs with nesting and/or cross-classification in Section 3. Section 4 shows how to calculate power and required sample size using the SE, taking the partially nested trial as the example. In particular, we reparameterize the SE formula so that content-matter knowledge (such as the test–retest reliability coefficient) can be used as input; we derive the design effect and the optimal allocation ratio; we investigate the small sample performance of the sample size formula, and we illustrate a sample size calculation. We close with a discussion in Section 5.

Technical details and software implementation code are provided in the Web materials (web-appendix and simulation program).

2 | GENERAL OBSERVATIONS FOR A RANDOMIZED, TWO-ARM DESIGN WITH BASELINE AND FOLLOW-UP MEASUREMENT

What is common to all the designs we consider is that each subject s is measured at baseline ($t = 0$), and randomized to the intervention ($g = 1$) or control ($g = 0$) arm, and finally measured at follow-up ($t = 1$). We can thus write the outcome y as a linear model

$$y_{gti} = \mu_{gt} + \epsilon_{gti},$$

where μ_{gt} are the means for each randomized group at each time point and the terms ϵ_{gti} capture correlations and variances due to repeated measures within subjects and due to other sources (eg, nesting or cross-classification of subjects at a time point). Estimation of the intervention effect and its SE in such a linear model with a general covariance matrix can of course be done using generalized least squares[21] (GLS), but the following approach is more intuitive and leads to shorter derivations later.

The two groups have a common baseline due to randomization, so that $\mu_{00} = \mu_{10} = \mu$. We can write $\mu_{10} = \mu + \tau_1$ where τ_1 is the time trend in the control arm and $\mu_{11} = \mu + \tau_1 + \delta$ where δ is the intervention effect. Every linear unbiased estimator is of the form $\widehat{\delta} = \sum_{gt} c_{gt} y_{gt\bullet}$, where the c_{gt} are some constants and $\bullet$ denotes averaging over the corresponding index. Using the means $y_{gt\bullet}$ suffices, because those means already capture all the information from the subjects' scores y_{gti} about the intervention effect in the applications we consider. Now, a linear estimator $\widehat{\delta}$ is unbiased if its expected value

$$E\left(\sum_{gt} c_{gt} y_{gt\bullet}\right) = c_{00}\mu + c_{10}\mu + c_{01}(\mu + \tau_1) + c_{11}(\mu + \tau_1 + \delta)$$

always equals δ. That means that

$$(c_{00} + c_{10} + c_{01} + c_{11}) \cdot \mu + (c_{01} + c_{11}) \cdot \tau_1 + (c_{11} - 1) \cdot \delta = 0$$

for all choices of δ, μ, τ_1. From this, we first see that $c_{11} = 1$, subsequentially that $c_{01} = -1$, and finally that $c_{10} = -c_{00}$. Writing $\mathbb{r} = c_{00}$, we get that each linear unbiased estimator is of the form

$$\widehat{\delta} = (1 - \mathbb{r}) \cdot \widehat{\delta}_{fu} + \mathbb{r} \cdot \widehat{\delta}_{change}, \tag{1}$$

where $\widehat{\delta}_{fu}$ is the difference between arms at follow-up ($c_{00} = c_{10} = 0; c_{11} = 1; c_{01} = -1$),

$$\widehat{\delta}_{fu} = y_{g=1,t=1,\bullet} - y_{g=0,t=1,\bullet},$$

and $\widehat{\delta}_{change}$ is the difference between arms in change from baseline ($c_{00} = 1; c_{10} = -1; c_{11} = 1; c_{01} = -1$),

$$\widehat{\delta}_{change} = \left[y_{g=1,t=1,\bullet} - y_{g=1,t=0,\bullet}\right] - \left[y_{g=0,t=1,\bullet} - y_{g=0,t=0,\bullet}\right].$$

The variance of a linear unbiased estimator (1) is thus

$$\text{var}(\widehat{\delta}) = (1 - \mathbb{r})^2 \cdot \text{var}\left(\widehat{\delta}_{fu}\right) + \mathbb{r}^2 \cdot \text{var}\left(\widehat{\delta}_{change}\right) + 2(1 - \mathbb{r}) \cdot \mathbb{r} \cdot \text{covar}\left(\widehat{\delta}_{fu}, \widehat{\delta}_{change}\right),$$

which is a quadratic function of $\mathbb{r}$. Minimizing this variance using the standard quadratic formula yields

$$\mathbb{r} = \frac{\text{covar}\left(\widehat{\delta}_{fu},\ \widehat{\delta}_{base}\right)}{\text{var}\left(\widehat{\delta}_{base}\right)} = corr\left(\widehat{\delta}_{fu},\ \widehat{\delta}_{base}\right) \cdot \sqrt{\frac{\text{var}\left(\widehat{\delta}_{fu}\right)}{\text{var}\left(\widehat{\delta}_{base}\right)}}, \tag{2}$$

where $\widehat{\delta}_{base} = y_{g=1,t=0,\bullet} - y_{g=0,t=0,\bullet} = \widehat{\delta}_{fu} - \widehat{\delta}_{change}$ is the difference between arms at baseline (which has an expectation equal to 0, but nonzero variance). The right hand side of (2) shows that the variance-minimizer $\mathbb{r}$ lies between $-\sqrt{c}$

and $+\sqrt{c}$ where $c = \text{var}\left(\widehat{\delta}_{fu}\right)/\text{var}\left(\widehat{\delta}_{base}\right)$ can be larger than 1. For example, $\mathbb{r} > 1$ in a randomized trial with no clustering or cross-classification when both the intervention and control condition increase the between-subject variance compared to the baseline condition and the within-subject variance is smaller than the between-subject variance at baseline (see the web-appendix). While having $\mathbb{r} > 1$ may seem strange at first, this phenomenon has been observed earlier by Samuel-Cahn.[22] In this article, the following linear combination was considered:

$$T = \mathbb{r}T_1 + (1 - \mathbb{r})T_2,$$

where T_1, T_2 are unbiased estimators for a parameter θ and $\mathbb{r} \in (-\infty, \infty)$ is chosen to result in minimal variance of T. One of the observations in this article was that

$$\mathbb{r} > 1 \text{ when } corr\,(T_1, T_2) > \sqrt{\text{var}(T_1)/\text{var}(T_2)},$$

which is equivalent to what is implied by Equation (2) (further details regarding this equivalence is provided in the web-appendix).

In certain situations, this variance minimizer $\mathbb{r}$ is in fact a correlation. In such situations, we will emphasize this by using the notation r instead.

The minimum variance that occurs for the choice of $\mathbb{r}$ as in (2) can be expressed as either an absolute or relative reduction compared to $var\left(\widehat{\delta}_{fu}\right)$:

$$\text{var}\left(\widehat{\delta}_{\text{minvar}}\right) = \text{var}\left(\widehat{\delta}_{fu}\right) - \mathbb{r}^2\text{var}\left(\widehat{\delta}_{base}\right) = \text{var}\left(\widehat{\delta}_{fu}\right) \cdot \left[1 - corr^2\left(\widehat{\delta}_{fu}, \widehat{\delta}_{base}\right)\right], \tag{3}$$

(see the web-appendix for additional details on the above calculations).

In conclusion, the estimator $\widehat{\delta}_{\text{minvar}}$, that is, the estimator in (1) with $\mathbb{r}$ chosen as in (2), is the best linear unbiased estimator (BLUE) for the linear model $y_{gts} = \mu_{gt} + \epsilon_{gts}$. As we focus on the planning stage of the trial, we take the variance–covariance structure of ϵ_{gts} as known. Because for known $\text{var}\left(\epsilon_{gts}\right)$ the generalized least-squares (GLS) estimator is the BLUE,[21] the estimator $\widehat{\delta}_{minvar}$ is the GLS estimator. In the applications that follow, we take ϵ_{gts} to be multivariately normally distributed and then the GLS estimator is also the maximum likelihood estimator (Chapter 2.3 in Kariya and Kurata[21]). Thus, we are actually deriving the asymptotic SE of the maximum likelihood estimator.

3 | APPLICATIONS TO NESTED AND CROSS-CLASSIFIED DESIGNS

We apply the results of Section 2 to the designs illustrated in Figure 1. Within each subfigure, the following are depicted from top to bottom: the situation at baseline (clustering or not), the situation just after randomization, and the situation after implementation of the intervention and control treatment. Each subject is depicted by a dot and the treatment condition of a subject is depicted by the filling (color) of the dot. Nesting of subjects is shown by clustering of the dots. If subjects are cross-classified, for example, between physiotherapists and coaches, one of the factors is depicted by clustering of dots (all the subjects treated by the same physiotherapist). The other factor is depicted as a filled circle with arrows pointing outward to dots (arrows connecting the coach to the subjects she or he coaches). For example, in Figure 1 (4), all clusters (left and right) receive physiotherapy from their physiotherapist; the clusters randomized to intervention (left) receive intervention style coaching, while the cluster randomized to control (right) receive control-style coaching. Each coach provides both intervention and control styles of coaching. In contrast, in Figure 1 (6), each cluster (indicated by the clustering in the middle) receives physiotherapy, but half of the subjects in each cluster are randomized to receive the intervention from the coaches (left).

To put the observations in Section 2 into practice, we need to calculate $\text{var}\left(\widehat{\delta}_{fu}\right)$, $\text{var}\left(\widehat{\delta}_{base}\right)$, and $\text{covar}\left(\widehat{\delta}_{fu}, \widehat{\delta}_{base}\right)$. These depend on the joint distribution of the baseline and follow-up measurements, that is, depend on the data-generating model chosen. We will use random intercepts at subject level to model repeated measurements of subjects and random intercepts at cluster level to model repeated measurements of clusters and/or cross-classification factors. To model nesting and/or cross-classification that is only present at baseline or only at follow-up, we will use random slopes. Thus, the crux is to define data-generating models with random intercepts and random slopes corresponding to the design of interest.

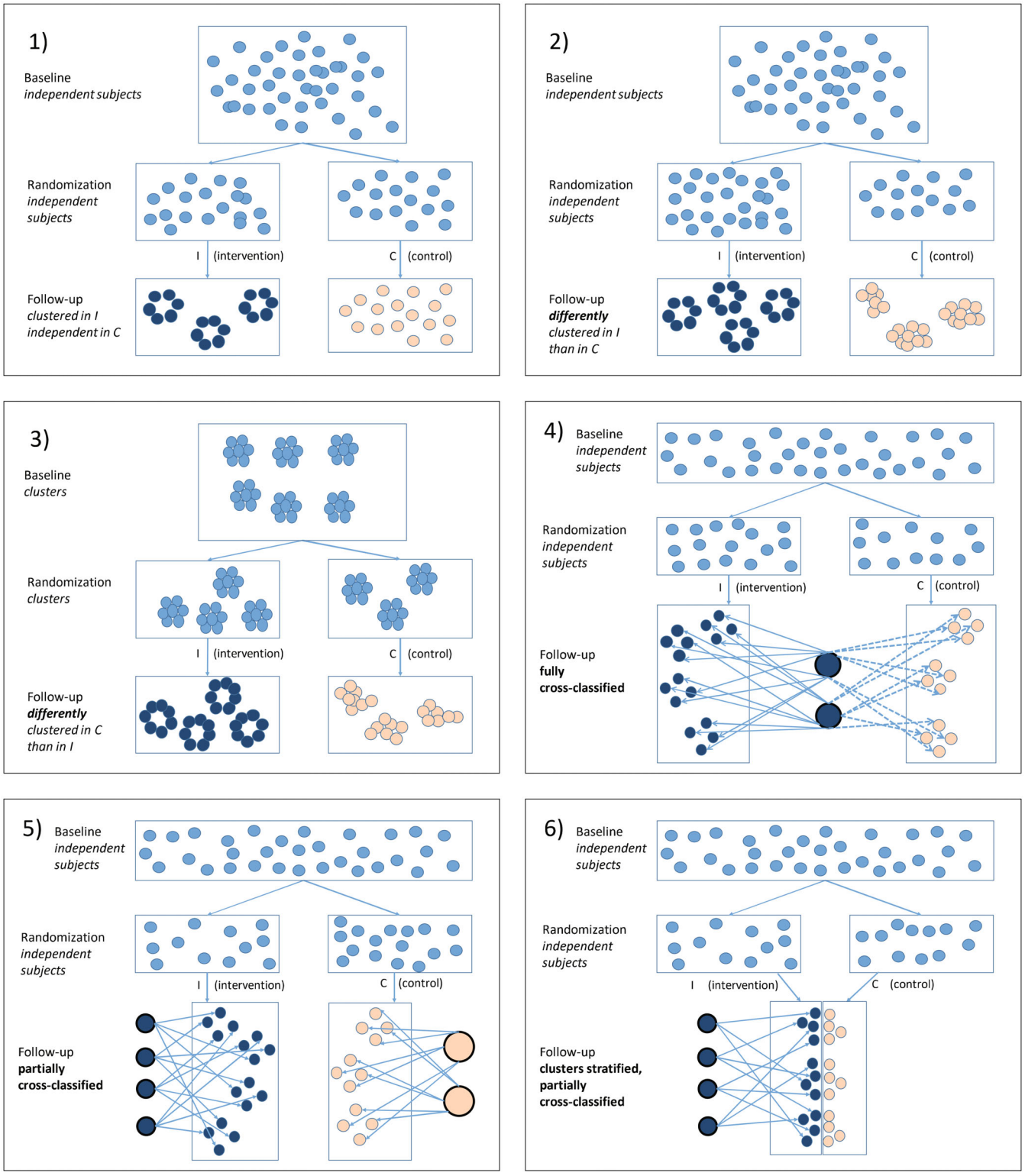

FIGURE 1 Two-arm designs with nesting and/or cross-classification at baseline or follow-up measurement. Nesting is shown by clustering of the dots (subjects). If subjects are cross-classified, one factor (eg, physiotherapist) is depicted by clustering of dots and the other (eg, coach) is depicted as a filled circle with arrows pointing outward to the dots (ie, subjects guided by the coach). In the sub-figures the following designs are shown: (1) Partially nested design with independent subjects at baseline (Section 3.1.2). (2) Fully nested design with independent subjects at baseline (Section 3.1.3). (3) Fully nested design with subjects in clusters at baseline (cluster size the same at baseline as at follow-up, Section 3.1.4). (4) Fully cross-classified design with independent subjects at baseline (Section 3.2.1). (5) Partially cross-classified design with independent subjects at baseline (Section 3.2.2). (6) A design combining nesting, stratified randomization, and partial cross-classification (Section 3.3).

3.1 | Designs with partial or full nesting at follow-up and no or full nesting at baseline

We start with a data-generating model that captures independence or nesting of subjects at baseline combined with partial or full nesting at follow-up in a general way. The outcome y_{gtij} of subject $j = 1, \ldots, n_g$ in cluster $i = 1, \ldots, k_g$ randomized to arm g is at baseline ($t = 0$) given by

$$y_{gtij} = \mu + \underbrace{c_{gi}}_{\sim N(0,\sigma_c^2)} + \underbrace{(c\tau)_{g,t=0,i}}_{\sim N(0,\sigma_{c\tau 0}^2)} + \underbrace{s_{gij}}_{\sim N(0,\sigma_s^2)} + \underbrace{(s\tau)_{g,t=0,ij}}_{\sim N(0,\sigma_{s\tau 0}^2)} \tag{4}$$

and at follow-up ($t = 1$) it is

$$y_{gtij} = \mu + \tau_1 + \delta \cdot \mathbb{I}_{[g=1]} + c_{gi} + \underbrace{(c\tau)_{g,t=1,i}}_{\sim N\left(0,\sigma_{gc\tau 1}^2\right)} + s_{gij} + \underbrace{(s\tau)_{g,t=1,ij}}_{\sim N\left(0,\sigma_{gs\tau 1}^2\right)}. \tag{5}$$

Here, we use Greek letters to denote fixed effects: μ is the baseline common to both arms (due to the randomization), τ_1 the change from baseline in the control arm and δ the intervention effect, that is only present at follow-up if the subject is in the intervention arm ($\mathbb{I}_{[g=1]} = 1$). Random effects denoted by roman letters are present at a cluster level (c) and at subject level (s). These random effects can be time invariant, that is, random intercepts c and s, or time varying, that is, $(c\tau)$ and $(s\tau)$ are present depending on the time point. We take these random effects as independently normally distributed and their specific variance is indicated by the $\sim N\left(0, \sigma^2\right)$-notation. The size σ^2 of the variance components is subscripted as gXt, where the first subscript denotes the arm g, the middle subscripts X the level ($c, s, c\tau$, and$s\tau$), and the last subscript the timepoint t. At baseline, there is no difference between arms and we drop the first subscript g. If a random effect is the same at follow-up as at baseline (c_{gi} and s_{gij}), then the $N\left(0, \sigma^2\right)$-notation is not repeated at follow-up for readability. Because the subjects (and if present, clusters) come from one population at baseline, $\text{var}\left((c\tau)_{g,t=0,i}\right) = \sigma_{c\tau 0}^2$ and $\text{var}\left((s\tau)_{g,t=0,i}\right) = \sigma_{s\tau 0}^2$ for all arms g. The number of subjects n_g per cluster and number of clusters k_g may be different per arm. As a final note on the notation: the $(s\tau)$ terms are usually denoted as residual error terms. However, we use this notation to emphasize that they are within-subject variances that are specific to the time point, instead of viewing them as random errors that are left if all other known sources of variation have been accounted for.

The above model induces nesting at baseline (denoted by the subscript "base"), and at follow-up in the control ("0") and intervention arm ("1") as described by the following intracluster correlations:

$$\rho_{base} = \left(\sigma_c^2 + \sigma_{c\tau 0}^2\right)/\sigma_{base}^2, \quad \rho_0 = \left(\sigma_c^2 + \sigma_{0c\tau 1}^2\right)/\sigma_0^2, \quad \text{and} \quad \rho_1 = \left(\sigma_c^2 + \sigma_{1c\tau 1}^2\right)/\sigma_1^2,$$

where

$$\sigma_{base}^2 = \sigma_c^2 + \sigma_{c\tau 0}^2 + \sigma_s^2 + \sigma_{s\tau 0}^2$$

is the total variance of a subject at baseline and

$$\sigma_0^2 = \sigma_c^2 + \sigma_{0c\tau 1}^2 + \sigma_s^2 + \sigma_{0s\tau 1}^2, \quad \sigma_1^2 = \sigma_c^2 + \sigma_{1c\tau 1}^2 + \sigma_s^2 + \sigma_{1s\tau 1}^2$$

are the total variance at follow-up in the control and intervention arms.

With some standard algebra $\text{var}\left(\widehat{\delta}_{base}\right)$ follows straightforwardly from (4), $\text{var}\left(\widehat{\delta}_{fu}\right)$ from (5), and $\text{covar}\left(\widehat{\delta}_{fu}\ ,\ \widehat{\delta}_{base}\right)$ follows from their common terms:

$$\text{var}\left(\widehat{\delta}_{base}\right) = \frac{\sigma_c^2}{k_1} + \frac{\sigma_c^2}{k_0} + \frac{\sigma_{c\tau 0}^2}{k_1} + \frac{\sigma_{c\tau 0}^2}{k_0} + \frac{\sigma_s^2}{k_1 n_1} + \frac{\sigma_s^2}{k_0 n_0} + \frac{\sigma_{s\tau 0}^2}{k_1 n_1} + \frac{\sigma_{s\tau 0}^2}{k_0 n_0},$$

$$\text{var}\left(\widehat{\delta}_{fu}\right) = \frac{\sigma_c^2}{k_1} + \frac{\sigma_c^2}{k_0} + \frac{\sigma_{1c\tau 1}^2}{k_1} + \frac{\sigma_{0c\tau 1}^2}{k_0} + \frac{\sigma_s^2}{k_1 n_1} + \frac{\sigma_s^2}{k_0 n_0} + \frac{\sigma_{1s\tau 1}^2}{k_1 n_1} + \frac{\sigma_{0s\tau 1}^2}{k_0 n_0},$$

$$\text{covar}\left(\widehat{\delta}_{fu}\ ,\ \widehat{\delta}_{base}\right) = \frac{\sigma_c^2}{k_1} + \frac{\sigma_c^2}{k_0} + \frac{\sigma_s^2}{k_1 n_1} + \frac{\sigma_s^2}{k_0 n_0},$$

so that

$$\mathbb{r} = \frac{\text{covar}\left(\hat{\delta}_{fu}\ ,\ \hat{\delta}_{base}\right)}{\text{var}\left(\hat{\delta}_{base}\right)} = \frac{\sigma_c^2 \cdot \left(\frac{1}{k_1} + \frac{1}{k_0}\right) + \sigma_s^2 \cdot \left(\frac{1}{k_1 n_1} + \frac{1}{k_0 n_0}\right)}{\left(\sigma_c^2 + \sigma_{c\tau 0}^2\right) \cdot \left(\frac{1}{k_1} + \frac{1}{k_0}\right) + \left(\sigma_s^2 + \sigma_{s\tau 0}^2\right) \cdot \left(\frac{1}{k_1 n_1} + \frac{1}{k_0 n_0}\right)}. \tag{6}$$

Rewritten in terms of intracluster correlation and total variances we get

$$\begin{aligned}\text{var}\left(\hat{\delta}_{\text{minvar}}\right) = \text{var}\left(\hat{\delta}_{fu}\right) - \mathbb{r}^2 \text{var}\left(\hat{\delta}_{base}\right) &= \left[1 + (n_1 - 1)\,\rho_1\right] \cdot \frac{\sigma_1^2}{k_1 n_1} + \left[1 + (n_0 - 1)\,\rho_0\right] \cdot \frac{\sigma_0^2}{k_0 n_0} \\ &- \mathbb{r}^2 \cdot \left\{ \left[1 + (n_1 - 1)\,\rho_{base}\right] \cdot \frac{1}{k_1 n_1} + \left[1 + (n_0 - 1)\,\rho_{base}\right] \cdot \frac{1}{k_0 n_0} \right\} \cdot \sigma_{base}^2. \end{aligned} \tag{7}$$

(See web-appendix for details of the above calculations). We now consider each of the six scenarios depicted in Figure 1.

3.1.1 | No clustering at baseline

In both the partially nested design shown in Figure 1 (1) (further discussed in 3.1.2) and the fully nested design shown in Figure 1 (2) (further discussed in 3.1.3), there is no clustering at baseline, so $\rho_{base} = 0$, and

$$0 = \rho_{base} \cdot \sigma_{base}^2 = \sigma_c^2 + \sigma_{c\tau 0}^2 \geq \sigma_c^2 \geq 0,$$

that is, $\sigma_c^2 = 0$ and with the same argument $\sigma_{c\tau 0}^2 = 0$. Thus, $\mathbb{r}$ in (6) reduces to $\sigma_s^2 / \left(\sigma_s^2 + \sigma_{s\tau 0}^2\right)$. We observe that $\sigma_s^2 / \left(\sigma_s^2 + \sigma_{s\tau 0}^2\right)$ is the correlation r between a baseline measurement and follow-up measurement for a subject who *remains in the same situation as at baseline*, that is, when Equation (5) equals (4), that is, when the model for baseline and follow-up is

$$y_{gtij} = \mu + s_{gij} + (s\tau)_{gtij} \text{ for } t = 0, 1 \text{ with } s_{gij} \sim N\left(0, \sigma_s^2\right) \text{ and } (s\tau)_{gtij} \sim N\left(0, \sigma_{s\tau 0}^2\right).$$

In this sense, the variance minimizer $\mathbb{r}$ can be interpreted as the correlation r between repeated measurements of a subject in the situation as at baseline. We will discuss this interpretation more fully in Approach 1 in Section 4.1.

3.1.2 | Partially nested design at follow-up with independent subjects at baseline (Figure 1 (1))

When there is no clustering at baseline ($\rho_{base} = 0$) and the control condition does not induce clustering ($\rho_0 = 0, n_0 = 1$), then (7) reduces to

$$\text{var}\left(\hat{\delta}_{\text{minvar}}\right) = \left[1 + (n_1 - 1)\,\rho_1\right] \cdot \frac{\sigma_1^2}{k_1 n_1} + \frac{\sigma_0^2}{k_0} - r^2 \cdot \left\{ \frac{1}{k_1 n_1} + \frac{1}{k_0} \right\} \cdot \sigma_{base}^2 \ , \text{with } \ r = \frac{\sigma_s^2}{\left(\sigma_s^2 + \sigma_{s\tau 0}^2\right)}. \tag{8}$$

By Section 3.1.1., r is the correlation between two repeated measurements on subjects in the situation as at baseline. Formula (8) reduces to the sample size for analysis of covariance (ANCOVA)[23] if there is also no clustering in the intervention arm ($\rho_1 = 0, n_1 = 1$).

3.1.3 | Fully nested design at follow-up with independent subjects at baseline (Figure 1 (2))

On the other hand, when both conditions induce clustering, then (7) gives

$$\text{var}\left(\hat{\delta}_{\text{minvar}}\right) = \left[1 + (n_1 - 1)\,\rho_1\right] \cdot \frac{\sigma_1^2}{k_1 n_1} + \left[1 + (n_0 - 1)\,\rho_0\right] \cdot \frac{\sigma_0^2}{k_0 n_0} - r^2 \cdot \left\{ \frac{1}{k_1 n_1} + \frac{1}{k_0 n_0} \right\} \cdot \sigma_{base}^2, \text{ with } r = \sigma_s^2 / \left(\sigma_s^2 + \sigma_{s\tau 0}^2\right). \tag{9}$$

Again by Section 3.1.1, r is the correlation between two repeated measurements on subjects in the situation as at baseline.

3.1.4 | Fully nested design at follow-up with subjects nested at baseline in the same clusters as at follow-up (Figure 1 (3))

If the clusters of size n exist at baseline ($\rho_{base} \neq 0$) and are randomized, then they have by definition at follow-up the same size as at baseline ($n_1 = n_0 = n$), but the degree of clustering can be changed by the intervention and control conditions ($\rho_1 \neq \rho_{base}, \rho_0 \neq \rho_{base}, \rho_1 \neq \rho_0$). We then have

$$\begin{aligned} \text{var}\left(\widehat{\delta}_{\text{minvar}}\right) = &[1+(n-1)\rho_1]\cdot\frac{\sigma_1^2}{k_1 n} + [1+(n-1)\rho_0]\cdot\frac{\sigma_0^2}{k_0 n} \\ &- r^2\cdot\left\{[1+(n-1)\rho_{base}]\cdot\frac{1}{k_1 n} + [1+(n-1)\rho_{base}]\cdot\frac{1}{k_0 n}\right\}\cdot\sigma_{base}^2, \end{aligned} \tag{10}$$

where

$$r = \frac{\sigma_c^2 + \sigma_s^2/n}{\left(\sigma_c^2 + \sigma_{c\tau 0}^2\right) + \left(\sigma_s^2 + \sigma_{s\tau 0}^2\right)/n},$$

is the correlation between two repeated measurements on a cluster average[24] in the measurement model as at baseline, that is, when Equation (5) equals (4). Using the same algebraic calculation as in the appendix of Teerenstra et al.,[24] we can write

$$r = \frac{n\rho_{base}}{n\rho_{base} + (1-\rho_{base})}\rho_c + \frac{(1-\rho_{base})}{n\rho_{base} + (1-\rho_{base})}\rho_s. \tag{11}$$

where $\rho_c = \sigma_c / \left(\sigma_c^2 + \sigma_{c\tau 0}^2\right)$ is the autocorrelation at cluster level[24] and $\rho_s = \sigma_s^2 / \left(\sigma_s^2 + \sigma_{s\tau 0}^2\right)$ is autocorrelation at subject level,[24] both in the measurement model as at baseline, see (4). When ICCs and variances are the same at baseline and follow-up, $\rho_1 = \rho_0 = \rho_{base}$ and $\sigma_{s\tau 0} = \sigma_{1s\tau 1} = \sigma_{0s\tau 1}$, we retrieve the ANCOVA formula for cluster randomization,[24] see formula (5) therein.

3.2 | Designs with random cross-classification factors

For designs that include random cross-classification factors,[14] the procedure is the same. We have to state a data-generating model for the baseline and follow-up measurement and derive the variances and covariance of $\widehat{\delta}_{base}$ and $\widehat{\delta}_{fu}$ from that. The only difference is that the model specification is more involved and consequently the calculations require more algebra. Therefore, we include these calculations in the web-appendix. Further, there are numerous designs possible with random cross-classification factors. We only consider two designs for illustration.

3.2.1 | Cluster randomized trial with treatment as a random fully cross-classification factor[13] (Figure 1 (4))

Independent subjects at baseline with total variance σ_{base}^2 are randomized to one of k_0 control clusters at follow-up (total subject variance σ_0^2) or to one of the k_1 intervention clusters (total subject variance σ_1^2). All clusters are fully cross-classified with q professionals: each professional treats n of the subjects in each cluster, so the cluster size is qn. We model the outcome at baseline of subject $j = 1, \ldots, n$ that is cross-classified at follow-up between cluster $i = 1, \ldots, k_g$ of arm g and professional $p = 1, \ldots, q$ as:

$$y_{g,t=0,ipj} = \mu + \underbrace{s_{gipj}}_{\sim N(0,\sigma_s^2)} + \underbrace{(s\tau)_{t=0,ipj}}_{\sim N(0,\sigma_{s\tau 0}^2)}, \tag{12}$$

and at follow-up as

$$y_{g,t=1,ipj} = \mu + \tau_1 + \delta \mathbb{I}_{[g=1]} + \underbrace{(d\delta)_{g=1,t=1,p} \cdot \mathbb{I}_{[g=1]}}_{\sim N(0,\sigma^2_\delta)} + \underbrace{(c\tau)_{g,t=1,i}}_{\sim N(0,\sigma^2_{c\tau 1})} + \underbrace{(d\tau)_{t=1,p}}_{\sim N(0,\sigma^2_{d\tau 1})} + \underbrace{(cd\tau)_{g,t=1,ip}}_{\sim N(0,\sigma^2_{cd\tau 1})} + s_{gipj} + \underbrace{(s\tau)_{g,t=1,ipj}}_{\sim N\left(0,\sigma^2_{gs\tau 1}\right)},$$

where compared to (5) new random effects are present to capture variation of the treatment effect over professionals ($d\delta$), variation between professionals ($d\tau$), and professional by cluster interaction ($cd\tau$). The clustering is taken to be the same in both arms, as typically the subjects are randomized to similar clusters and the difference arises due to the treatment the professionals implement in the cluster. However, this assumption could be relaxed.

At baseline, the total variance of a subject is

$$\sigma^2_{base} = \sigma^2_s + \sigma^2_{s\tau 0}.$$

In the control arm, the cross-classification induces correlations in the follow-up measurements at the cluster level, at the professional level and at interaction of cluster by professional level given by

$$\rho_{0c\tau 1} = \frac{\sigma^2_{c\tau 1}}{\sigma^2_0}, \quad \rho_{0d\tau 1} = \frac{\sigma^2_{d\tau 1}}{\sigma^2_0}, \quad \text{and} \quad \rho_{0cd\tau 1} = \frac{\sigma^2_{cd\tau 1}}{\sigma^2_0},$$

where

$$\sigma^2_0 = \sigma^2_{c\tau 1} + \sigma^2_{d\tau 1} + \sigma^2_{cd\tau 1} + \sigma^2_s + \sigma^2_{0s\tau 1}$$

is the total variance at the subject level. Here, we are subscripting the correlations with gXt, where the first subscript denotes the arm g, the middle subscripts the level (cluster by time $c\tau$, professional by time $d\tau$, or cluster by professional by time $cd\tau$), and the last subscript the timepoint t. We can interpret these correlations at follow-up as follows.

- $\rho_{0c\tau 1}$ is the correlation between two subjects in the same cluster but treated by different professionals;
- $\rho_{0d\tau 1}$ is the correlation between two subjects treated by the same professional but in different clusters;
- $\rho_{0cd\tau 1}$ is the difference $\rho_{0cd\tau 1} = corr - (\rho_{0c\tau 1} + \rho_{0d\tau 1})$ where $corr$ is the correlation between two subjects in the same cluster and same professional.

In the intervention arm at follow-up, correlations at treatment (by professional and time) interaction level, cluster level, at the professional level and at the interaction of cluster by professional level are present given by

$$\rho_{1d\delta\tau 1} = \frac{\sigma^2_\delta}{\sigma^2_1}, \quad \rho_{1c\tau 1} = \frac{\sigma^2_{c\tau 1}}{\sigma^2_1}, \quad \rho_{1d\tau 1} = \frac{\sigma^2_{d\tau 1}}{\sigma^2_1}, \quad \text{and} \quad \rho_{1cd\tau 1} = \frac{\sigma^2_{cd\tau 1}}{\sigma^2_1},$$

where now

$$\sigma^2_1 = \sigma^2_\delta + \sigma^2_{c\tau 1} + \sigma^2_{d\tau 1} + \sigma^2_{cd\tau 1} + \sigma^2_s + \sigma^2_{1s\tau 1}$$

is the total variance at the subject level.

From (12), we easily see that

$$\text{var}\left(\hat{\delta}_{base}\right) = \frac{\sigma^2_s + \sigma^2_{s\tau 0}}{qn} \cdot \left(\frac{1}{k_1} + \frac{1}{k_0}\right) \quad \text{and} \quad \text{covar}\left(\hat{\delta}_{base}, \hat{\delta}_{fu}\right) = \frac{\sigma^2_s}{qn} \cdot \left(\frac{1}{k_1} + \frac{1}{k_0}\right),$$

as there is only one common term between baseline and follow-up: s_{gipj}. Then, the variance-minimizer $\mathbb{r}$ equals

$$r = \frac{\text{covar}\left(\hat{\delta}_{base}, \hat{\delta}_{fu}\right)}{\text{var}\left(\hat{\delta}_{base}\right)} = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_{s\tau 0}}$$

which is again the correlation r between two repeated measurements in the same situation as at baseline.

When calculating $\text{var}\left(\hat{\delta}_{fu}\right)$ from (13), we have by design that each professional treats a part of each control cluster and a part of each intervention cluster. Thus, first the intervention effect per professional is calculated in which the main random effect of professional, $(d\tau)$, cancels out. However, a cluster by professional random interaction $(cd\tau)$ may be present[13] with variance $\sigma^2_{cd\tau 1}$. Next, we average over professionals and get

$$\text{var}\left(\hat{\delta}_{fu}\right) = \frac{\sigma^2_\delta}{q} + \sigma^2_{c\tau 1} \cdot \left(\frac{1}{k_1} + \frac{1}{k_0}\right) + \frac{\sigma^2_{cd\tau 1}}{q} \cdot \left(\frac{1}{k_1} + \frac{1}{k_0}\right) + \frac{\sigma^2_s}{qn} \cdot \left(\frac{1}{k_1} + \frac{1}{k_0}\right) + \frac{1}{qn}\left(\frac{\sigma^2_{1s\tau 1}}{k_1} + \frac{\sigma^2_{0s\tau 1}}{k_0}\right),$$

which corresponds to formula (5) of Moerbeek and Safarkani,[13] when substituting $c\tau 1 \to u$, $(cd\tau 1) \to w$, $k_1 = k_0 \to \frac{n_{2A}}{2}$, $q \to n_{2B}$, $\sigma^2_s + \sigma^2_{1s\tau 1} = \sigma^2_s + \sigma^2_{0s\tau 1} \to \sigma^2_\varepsilon$.

We get a digestible form of the SE for the special case that the subject by time (residual) variances at follow-up and baseline are the same $(\sigma^2_{1s\tau 1} = \sigma^2_{0s\tau 1} = \sigma^2_{s\tau 0})$:

$$\text{var}\left(\hat{\delta}_{\text{minvar}}\right) = \frac{\sigma^2_\delta}{q} + \sigma^2_0 \cdot \left[\left(\rho_{0c\tau 1} + \frac{\rho_{0cd\tau 1}}{q}\right) \cdot \left(\frac{1}{k_1} + \frac{1}{k_0}\right)\right] + \left(1 - r^2\right) \cdot \frac{\sigma^2_{base}}{qn} \cdot \left(\frac{1}{k_1} + \frac{1}{k_0}\right), \text{with } r = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_{s\tau 0}}. \quad (14)$$

As only subjects are common to baseline and follow-up, there is only a reduction of the subject variance and the $\rho_{0c\tau 1}$ term captures the cluster variance that is not reduced.

We can similarly derive formulas for the case the clusters already exist at baseline and then also the cluster variance is reduced (web-appendix).

3.2.2 | Cluster randomized trial with intervention as a random partially cross-classified factor[13] (Figure 1 (5))

As in the previous subsection, subjects at baseline with total variance σ^2_{base} are randomized to one of k_0 control clusters at follow-up (total subject variance σ^2_0) or to one of the k_1 intervention clusters (total subject variance σ^2_1). However, now the control clusters are cross-classified with $\pi_0 q$ professionals and the intervention clusters are cross-classified with the remaining $\pi_1 q$ professionals (partial cross-classification). In particular, each professional delivers always the control condition or always the intervention condition. This may help avoid contamination within professionals, but the main effect of professional no longer cancels out.[13]

Suppose each of the professionals $p = 1, \ldots, \pi_0 q$ treats n_0 subjects of each control cluster $i = 1, \ldots, k_0$ (so the control cluster size is $\pi_0 q n_0$). The analogue holds for the professionals $p = \pi_0 q + 1, \ldots, q$ and intervention clusters $i = 1, \ldots, k_1$ which are each of size $\pi_1 q n_1$.

The data-generating model is the same as in Section 3.2.1, but the term $(d\delta)_{g=1,t=1,p}$ is absent because the intervention effect is no longer estimated within each professional. As a consequence, r and $var\left(\hat{\delta}_{base}\right)$ are the same as in Section 3.2.1 and the same expressions for correlations and total variances apply (after removal of σ^2_δ).

When we calculate the difference at follow-up, we average the outcome over all control clusters and over all intervention clusters separately. The result is

$$\begin{aligned}\text{var}\left(\hat{\delta}_{fu}\right) = \sigma^2_{c\tau 1} \cdot \left(\frac{1}{k_1} + \frac{1}{k_0}\right) + \frac{\sigma^2_{d\tau 1}}{q} \cdot \left(\frac{1}{\pi_1} + \frac{1}{\pi_0}\right) + \frac{\sigma^2_{cd\tau 1}}{q} \cdot \left(\frac{1}{k_1\pi_1} + \frac{1}{k_0\pi_0}\right) \\ + \frac{\sigma^2_s}{q} \cdot \left(\frac{1}{k_1\pi_1 n_1} + \frac{1}{k_0\pi_0 n_0}\right) + \frac{1}{q} \cdot \left(\frac{\sigma^2_{1s\tau 1}}{k_1\pi_1 n_1} + \frac{\sigma^2_{0s\tau 1}}{k_0\pi_0 n_0}\right)\end{aligned}$$

In the special case that the subject by time (residual) variances are the same at baseline and follow-up in both arms $(\sigma^2_{s\tau 0} = \sigma^2_{0s\tau 1} = \sigma^2_{1s\tau 1})$, we get in terms of the intracluster correlations and total variances

$$\begin{aligned}\text{var}\left(\hat{\delta}_{\text{minvar}}\right) = \sigma^2_0 \cdot \left[\rho_{0c\tau 1} \cdot \left(\frac{1}{k_1} + \frac{1}{k_0}\right) + \frac{\rho_{0d\tau 1}}{q} \cdot \left(\frac{1}{\pi_1} + \frac{1}{\pi_0}\right) + \frac{\rho_{0cd\tau 1}}{q}\left(\frac{1}{k_1\pi_1} + \frac{1}{k_0\pi_0}\right)\right] \\ + \left(1 - r^2\right) \cdot \sigma^2_{base} \cdot \frac{1}{q} \cdot \left(\frac{1}{k_1\pi_1 n_1} + \frac{1}{k_0\pi_0 n_0}\right), \text{with } r = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_{s\tau 0}}\end{aligned} \quad (15)$$

that is, only the subject variance components get reduced and r is the correlation between two repeated measurements in the situation as at baseline. The web-appendix discusses also the situation with unequal cluster / residual variances and clustering at baseline.

3.3 | Application with nesting, stratified randomization and cross-classification (Figure 1 (6))

As a more complex application, we look at the trial of Bennell et al.[7] Independent subjects at baseline with total variance σ^2_{base} are treated by one of q physiotherapists at follow-up. Half of the subjects of a physiotherapist, say n subjects, are randomized to the control treatment and have a total variance σ^2_0. The remaining half of the subjects (n subjects) receive the intervention (telephone coaching) as an add-on and have total variance σ^2_1. The main effect for physiotherapist is removed when calculating the intervention effect, but there could be physiotherapist by intervention (coach) interaction with variance σ^2_δ. There are k_1 coaches each treating n_1 subjects (coming from different physiotherapists, $k_1 n_1 = qn$) with within-coach ICC ρ_1. There are $k_0 = qn$ control subjects (also coming from different physiotherapists). These control subjects are, so to say, their own 'coach' and each 'control coach' treats $n_0 = 1$ subjects. We refer to the web-appendix for the details of the calculations.

Let y_{pgtij} be the outcome at time t of subject j that at follow-up is nested in physiotherapist $p = 1, \ldots, q$ and treated by coach i. If the subject gets the intervention ($g = 1$) then (s)he corresponds to one of the coaches $i = 1, \ldots k_1$. If the subject gets the control, then the subject is her/his own 'coach', and counts as one of the 'control coaches' $i = 1, \ldots, k_0 = nq$.

At baseline we have

$$y_{pgtij} = \mu + s_{pgij} + \underbrace{(s\tau)_{p,g=0,t=0,ij}}_{\sim N(0,\sigma^2_{s\tau 0})} \tag{16}$$

and at follow-up

$$y_{pgtij} = \mu + \tau_1 + \delta \mathbb{I}_{[g=1]} + (d\tau)_p + \left[\underbrace{(c\tau)_{i,g=1,t=1}}_{\sim N(0,\sigma^2_{1c\tau 1})}\right] \mathbb{I}_{[g=1]} + \left[\underbrace{(cd\delta\tau)_{p,g=1,t=1,i}}_{\sim N(0,\sigma^2_\delta)}\right] \mathbb{I}_{[g=1]} + s_{pgij} + \underbrace{(s\tau)_{p,g=1,t=1,ij}}_{\sim N(0,\sigma^2_{1s\tau 1})}, \tag{17}$$

where random effects for physiotherapist $(d\tau)$ are present in both arms. In the intervention arm at follow-up, we have additionally a random effect for coach, $(c\tau)$ (describing the clustering of subjects treated by the same coach) and a random interaction of intervention effect due to coach and physiotherapist $(cd\delta\tau)$ that describes the variation of intervention effect over the different 'cells' (ie, combinations of coach and physiotherapist). A term $(cd\tau)$ is absent as this would be indistinguishable from $(cd\delta\tau)$ by design.

As half of the subjects will get the control and half the intervention, we have at baseline

$$\text{var}\left(\hat{\delta}_{base}\right) = \sigma^2_{base} \cdot \left(\frac{1}{k_1 n} + \frac{1}{k_0}\right) \quad , \quad \sigma^2_{base} = \sigma^2_s + \sigma^2_{s\tau 0}$$

with $k_1 n_1 = k_0 = qn$ as each of q physiotherapists have $2n$ subjects of which half receive the intervention and half the control. The only common random effect between baseline and follow-up is again the subject random intercept, s_{pgij}, so $\text{covar}\left(\hat{\delta}_{fu}, \hat{\delta}_{base}\right) = (1/(k_1 n) + 1/k_0) \cdot \sigma^2_s$ and the variance-minimizer $\mathbb{r}$ is also the repeated measurements correlation r in the situation as at baseline:

$$r = \sigma^2_s / \left(\sigma^2_s + \sigma^2_{s\tau 0}\right).$$

The variance of the intervention effect estimator $\hat{\Delta}_{pi}$ in physiotherapist p due to coach i at follow-up is

$$\text{var}\left(\hat{\Delta}_{pi}\right) = \sigma^2_\delta + \sigma^2_{c\tau} + 2\frac{\sigma^2_s}{\#J_{pi}} + \frac{\sigma^2_{1s\tau 1}}{\#J_{pi}} + \frac{\sigma^2_{0s\tau 1}}{\#J_{pi}},$$

where $\#J_{pi} = \#J_{pi,g=0} = \#J_{pi,g=1}$ is the number of control or intervention subjects in the physiotherapist p and coach i combination. For the intervention effect, we need to take the average of the physiotherapists' intervention effects. For that, we discuss two cases.

Nested case. If the physiotherapists are nested in coaches, then each of the k_1 coaches has q/k_1 physiotherapists and provides the intervention to all n intervention subjects of each of those physiotherapists. Thus, $\#J_{pi} = n$ and each coach treats $n_1 = nq/k_1$ subjects. We average the intervention effect $\widehat{\Delta}_{pi}$ across the physiotherapists within coach and then average that across coaches. This results in

$$\text{var}\left(\widehat{\delta}_{\text{minvar}}\right) = \frac{\sigma_\delta^2}{q} + \left[1 + (n_1 - 1)\,\rho_1\right] \cdot \frac{\sigma_1^2}{k_1 n_1} + \frac{\sigma_0^2}{k_0} - r^2 \cdot \left\{ \left(\frac{1}{k_1 n_1} + \frac{1}{k_0} \right) \cdot \sigma_{base}^2 \right\}, \text{with } n_1 = \frac{nq}{k_1}. \tag{18}$$

Fully cross-classified case. If the physiotherapists are fully cross-classified with the k_1 coaches, then each of the k_1 coaches gives, for each of the q physiotherapists, the intervention to only n/k_1 from the n intervention subjects of that physiotherapist. Thus, $\#J_{pi} = n/k_1$ and $n_1 = nq/k_1$.

We calculate the intervention effect $\widehat{\Delta}_p$ in a given physiotherapist by averaging over all the k_1 coaches as these together implemented the intervention. Note that the intervention effects in physiotherapist p and p' are now correlated due to the common coach effects:

$$\text{var}\left(\widehat{\Delta}_p\right) = \frac{\sigma_\delta^2}{k_1} + \frac{\sigma_{c\tau}^2}{k_1} + 2 \cdot \frac{\sigma_s^2}{n} + \frac{\sigma_{1s\tau 1}^2}{n} + \frac{\sigma_{0s\tau 1}^2}{n} \quad , \quad \text{covar}\left(\widehat{\Delta}_p \ , \ \widehat{\Delta}_{p'}\right) = \frac{\sigma_{c\tau}^2}{k_1} \ .$$

Taking the average of these interventions effects across physiotherapists while accounting for their covariances, we get as result

$$\text{var}\left(\widehat{\delta}_{\text{minvar}}\right) = \frac{\sigma_\delta^2}{k_1 q} + \left[1 + (n_1 - 1)\,\rho_1\right] \cdot \frac{\sigma_1^2}{k_1 n_1} + \frac{\sigma_0^2}{k_0} - r^2 \cdot \left[\left(\frac{1}{k_1 n_1} + \frac{1}{k_0} \right) \cdot \sigma_{base}^2 \right], \text{ with } n_1 = \frac{nq}{k_1}. \tag{19}$$

In the cross-classified case, the impact of the variation σ_δ^2 of the intervention effect across physiotherapists by coach combinations is thus less than in the nested case.

4 | EXAMPLE: THE PARTIALLY NESTED DESIGN

For various designs, we have shown in Section 3 how to derive the SE $se = \sqrt{var\left(\widehat{\delta}_{minvar}\right)}$.

From this, the power for an intervention effect δ at a two-sided significance level α follows from $power = \Phi\left(\delta/se - z_{1-\alpha/2}\right)$ where Φ is the cumulative standard normal distribution function and z_γ the γ-quantile of the standard normal distribution. Thus, with the SE we can investigate the impact of several design choices. For example:

1. How can content-knowledge be incorporated in the power / sample size calculation?
2. Which factors determine the amount of sample size reduction by including the baseline?
3. What is the best allocation of subjects when the total sample size is fixed (ie, how many clusters, etc.)?
4. Can sample size calculation be simplified using design effects?
5. What is the small sample performance of the asymptotic SE?

As an example, we will address these questions for the partially nested design (shown in Figure 1 (1)) in the coming subsections. Also we provide a web program to perform sample size calculations (Section 4.6) and a case study (Section 4.9). The web-appendix provides code for several statistical programs to analyze a trial and a program to perform simulation studies.

4.1 | Content-knowledge as input for sample size and power calculations

Formula (8) can be used for power and sample size calculation but requires knowledge of σ_{base}^2, r, and ρ_1, σ_1^2, and σ_0^2. Typically, the SD in the baseline σ_{base}^2 and control condition σ_0^2 will be known. Preferably, r (the correlation between two

TABLE 1 Data-generating model for outcome y_{gtij} in subject j in cluster i at time t in arm g in case of the partially nested randomized trial with independent baseline measurements.

	Measurement	
Arm	**Baseline ($t = 0$)**	**Follow-up ($t = 1$)**
Control ($g = 0$)	$\mu + s_{gij} + \underbrace{(s\tau)_{g,t=0,ij}}_{\sim N(0,\sigma^2_{s\tau 0})}$	$\mu + \tau_1 + s_{gij} + \underbrace{(s\tau)_{g=0,t=1,ij}}_{\sim N(0,\sigma^2_{0s\tau 1})}$
	$(i = 1, \ldots, k_0; j = 1)$	$(i = 1, \ldots, k_0; j = 1)$
Intervention ($g = 1$)	$\mu + s_{gij} + \underbrace{(s\tau)_{g,t=0,ij}}_{\sim N(0,\sigma^2_{s\tau 0})}$	$\mu + \tau_1 + \delta + \underbrace{(c\tau)_{g=1,t=1,i}}_{\sim N(0,\sigma^2_{1c\tau 1})} + s_{gij} + \underbrace{(s\tau)_{g=1,t=1,ij}}_{\sim N(0,\sigma^2_{1s\tau 1})}$
	$(i = 1, \ldots, k_1; j = 1, \ldots, n_1)$	$(i = 1, \ldots, k_1; j = 1, \ldots, n_1)$
	$s_{gij} \sim N\left(0, \sigma^2_s\right)$	

Note: Fixed effects in Greek letters are μ (=average at baseline in both arms), τ_1(=the change from baseline in the control arm), and δ (=the intervention effect). Random effects in roman letters include a time invariant random effect for subjects (s_{gij}) and time varying random effects indicated by the suffix τ and the level: s for subject level and c for cluster level. These are $(s\tau)_{gtij}$ (ie, residuals) and $(c\tau)_{gti}$. Variances σ^2_{gXt} are indexed by the arm g and time t and whether they are subject invariant ($X = s$) or time varying at the subject level ($X = s\tau$) or at the cluster level ($X = c\tau$).

repeated measurements in the situation as at baseline) is informed by past trials or observational data with repeated measurements in a compatible setting (ie, sufficient similarity in outcome, interval between measurements, and population). Otherwise, if the outcome has been validated, then the test–retest reliability will be known and this provides information about r. The issue of setting a value for ρ_1, that is, the ICC of the clustering in the intervention arm, occurs also in planning cluster randomized trials and can be dealt with similarly (eg, chapter 11 in Reference 25). The new issue is thus getting values for σ^2_1. Ideally, data from subjects' outcomes from several clusters in the intervention condition (or similar clusters in a similar condition) would be available (eg, from pilot data) and σ^2_1 could be estimated directly. If not, then the following approaches could use content-matter input.

Approach 1: Reparameterization of σ^2_1 in terms of $r_{bs,1fu}$ and $r = r_{bs,bs}$

The correlation between the baseline measurement and follow-up measurement of a subject randomized into the intervention arm is (see Table 1)

$$r_{bs,1fu} = \frac{\sigma^2_s}{\left(\sqrt{\sigma^2_s + \sigma^2_{s\tau 0}}\right) \cdot \left(\sqrt{\sigma^2_{1c\tau 1} + \sigma^2_s + \sigma^2_{1s\tau 1}}\right)} = \frac{\sigma^2_s}{\sqrt{\sigma^2_{base}} \cdot \sqrt{\sigma^2_1}}.$$

On the other hand, if a subject were measured at follow-up in the same condition as at baseline, that is, the residual (within-subject) variance is not changed and no cluster random effect is added, then the data-generating model is from Table 1:

$$\mu + s_{gij} + (s\tau)_{g=0,t,ij},$$

where $t = 0, 1$ and

$$s_{gij} \sim N\left(0, \sigma^2_s\right), \quad (s\tau)_{g=0,t,ij} \sim N\left(0, \sigma^2_{s\tau 0}\right).$$

In this data-generating model, the correlation between those repeated measurements is $r = r_{bs,bs} := \sigma^2_s / \sigma^2_{base}$ with $\sigma^2_{base} = \sigma^2_s + \sigma^2_{s\tau 0}$.

When we combine these two expressions for $r = r_{bs,bs}$ and $r_{bs,1fu}$, we get

$$\frac{\sigma^2_1}{\sigma^2_{base}} = \left(r / r_{bs,1fu}\right)^2. \tag{20}$$

We observe that $r = r_{bs,bs} = \sigma_s^2 / \left(\sigma_s^2 + \sigma_{s\tau 0}^2\right)$ has the definition of a test-retest reliability coefficient: the correlation between two repeated measurements on a subject, when (we expect that) the condition of that subject has not changed.

- The condition here is the situation as at baseline, that is, the (no-)treatment condition as present at baseline.
- The interval between the two measurements is the interval between the baseline and follow-up measurement in the trial.

These two observations mean that the value of r can be informed by known values of the test–retest reliability coefficient, for example those established when the outcome scale was validated. How much the values of known test-retest reliability coefficients are similar to r depends on how much the repeated measurement interval and the condition (and population) are similar between the trial and the study where the test–retest reliability coefficients were established.

Approach 2: Reparameterization of σ_1^2 in terms of $r_{1fu,1fu}$ and $r = r_{bs,bs}$,

If a subject is measured repeatedly in the situation as at follow-up in the intervention arm, that is, according to the repeated measurements model (see Table 1):

$$\mu + \tau_1 + \delta + c_{g=1,i} + s_{ij} + (s\tau)_{g=1,t,ij},$$

where $t = 0, 1$

$$c_{g=1,i} \sim N\left(0, \sigma_{1c}^2\right), \quad s_{gij} \sim N\left(0, \sigma_s^2\right), \quad (s\tau)_{g=1,t,ij} \sim N\left(0, \sigma_{1s\tau 1}^2\right),$$

then the correlation between those repeated measurements is $r_{1fu,1fu} = \left(\sigma_{1c\tau 1}^2 + \sigma_s^2\right) / \sigma_1^2$ with $\sigma_1^2 = \sigma_{1c\tau 1}^2 + \sigma_s^2 + \sigma_{1s\tau 1}^2$ (see Table 1). Together with the correlation between repeated measurements of a subject in the situation as at baseline $r = r_{bs,bs} := \sigma_s^2 / \sigma_{base}^2$ (as above), we can express $R = \sigma_1^2 / \sigma_{base}^2$ as follows:

$$R = \frac{r}{r_{1fu,1fu}} \cdot \frac{\sigma_{1c\tau 1}^2 + \sigma_s^2}{\sigma_s^2} = \frac{r}{r_{1fu,1fu}} \cdot \left(\frac{\sigma_{1c\tau 1}^2}{\sigma_s^2} + 1\right) = \frac{r}{r_{1fu,1fu}} \cdot \left(\frac{\rho_1 \sigma_1^2}{r \sigma_{base}^2} + 1\right) = a \cdot (b \cdot R + 1),$$

with $a = \frac{r}{r_{1fu,1fu}}$ and $b = \frac{\rho_1}{r}$. This linear equation has as solution $R = \frac{a}{1-ab} = \left[a^{-1} - b\right]^{-1}$, so

$$\frac{\sigma_1^2}{\sigma_{base}^2} = \left[\frac{r_{1fu,1fu}}{r} - \frac{\rho_1}{r}\right]^{-1} = \frac{r}{r_{1fu,1fu} - \rho_1}. \tag{21}$$

Approach 2′: assuming $r_{1fu,1fu} = r = r_{bs,bs}$ (a special case of Approach 2)

It may be that content matter specialists know the test–retest reliability coefficient of the outcome and base $r = r_{bs,bs}$ on that, but find it difficult to specify the correlation $r_{bs,1fu}$ between baseline and follow-up measurement in the intervention arm (Approach 1) or the repeated measurement correlation $r_{1fu,1fu}$ in the situation as at follow-up in the intervention arm (Approach 2). As a starting point the assumption

$$r_{bs,bs} = r_{1fu,1fu} \tag{22a}$$

could be taken. Then (21) becomes

$$\sigma_1^2 / \sigma_{base}^2 = r / (r - \rho_1), \tag{23}$$

so the variance in the intervention arm increases compared to baseline. We now discuss two alternative ways to look at this assumption (detail of the calculations in the web-appendix). Firstly, we can reformulate the assumption (22a) as follows:

$$\sigma_{1s\tau 1}^2 = \sigma_{s\tau 0}^2 \cdot \left(1 + \frac{\sigma_{1c\tau 1}^2}{\sigma_s^2}\right), \tag{22a'}$$

which implies that $\sigma_{1s\tau 1}^2 > \sigma_{s\tau 0}^2$, that is, the subject by time variance (residual variance) increases when a subject goes from an unclustered to a clustered situation.

Yet another way to look at this assumption (22a) is the following. As $\left(r/r_{bs,1fu}\right)^2 = \sigma_1^2/\sigma_{base}^2$ always (Approach 1) and the assumption of equal repeated measurement correlation (Approach 2′) implies $\sigma_1^2/\sigma_{base}^2 = r/(r-\rho_1)$, we are actually assuming that

$$r_{bs,1fu}^2 = r^2 - r\rho_1, \tag{22a''}$$

so the correlation between baseline (independent subjects) and follow-up (clustered subjects) is smaller than the correlation between repeated measurements at baseline.

When $r \approx \rho_1$ and Approach 2′ is taken, then $\sigma_1^2/\sigma_{base}^2 = r/(r-\rho_1)$ implies that $\sigma_1^2/\sigma_{base}^2$ is large, so that there is little power gained by including the baseline value in the analysis, see (8) in Section 3.1.2. If $r \leq \rho_1$, then the assumption of Approach 2′ makes no sense. In that case, however, we may not be inclined to include the baseline outcome in the analysis for another reason: the gain in power is limited, when r is so small.

4.2 | Factors determining the amount of sample size reduction due to including the baseline

As the sample size is proportional to the variance of the treatment effect, we can calculate the reduction in sample size as follows, using (8):

$$\frac{\text{var}\left(\widehat{\delta}_{fu}\right) - \text{var}\left(\widehat{\delta}_{\text{minvar}}\right)}{var\left(\widehat{\delta}_{fu}\right)} = \frac{r^2}{\left(\frac{k_0}{k_1 n_1 + k_0}\right) \cdot \left[1 + (n_1 - 1)\,\rho_1\right] \cdot \frac{\sigma_1^2}{\sigma_{base}^2} + \left(\frac{k_1 n_1}{k_1 n_1 + k_0}\right) \cdot \frac{\sigma_0^2}{\sigma_{base}^2}}, \tag{24}$$

see web-appendix. The reduction increases if the test–retest reliability, so r^2, is larger or if the clustering effect $[1 + (n_1 - 1)\,\rho_1] \cdot \sigma_1^2$ is smaller. This is illustrated in Figure 2 for various intracluster correlations ρ_1, cluster sizes n_1 and correlations r when $\sigma_0^2 = \sigma_{base}^2$ (eg, equal variance in the control condition at follow-up as at baseline), allocation is equal, that is, $k_0/(k_1 n_1 + k_0) = k_1 n_1/(k_1 n_1 + k_0) = 1/2$, and under the assumption of approach 2′ in Section 4.1: $\sigma_1^2/\sigma_{base}^2 = r/(r-\rho_1)$.

The denominator of (24) is a weighted average of the control arm contribution, $\sigma_0^2/\sigma_{base}^2$, and the intervention arm contribution $[1 + (n_1 - 1)\,\rho_1] \cdot \sigma_1^2/\sigma_{base}^2$, so in-between those two values. Clustering due to the intervention treatment would typically increase the variance compared to the control treatment, $[1 + (n_1 - 1)\,\rho_1] \cdot \sigma_1^2 \geq \sigma_0^2$. As a result, the relative reduction (24) would be typically $\leq r^2\left(\sigma_{base}^2/\sigma_0^2\right)$. For useful gains in precision, the test–retest reliability r needs to be large, although theoretically, a control treatment that reduces variance compared to baseline would also help.

If the number of control subjects is made large ($k_0 \to \infty$), then the relative reduction in (24) approaches $r^2 \cdot \left(\sigma_{base}^2/\sigma_1^2\right) / [1 + (n_1 - 1)\,\rho_1]$, so is limited by the clustering effect. For large clusters in the intervention arm ($n_1 \to \infty$), the relative reduction in (24) approaches

$$r^2 \cdot \frac{1}{\rho_1 \cdot \frac{k_0}{k_1} \cdot \left(\frac{\sigma_1^2}{\sigma_{base}^2}\right) + \left(\frac{\sigma_0^2}{\sigma_{base}^2}\right)}$$

so then it pays off to have more clusters, $k_1 > k_0$. If the number of intervention clusters is made large ($k_1 \to \infty$), the relative reduction approaches $r^2 \cdot \left(\sigma_{base}^2/\sigma_0^2\right)$ which is the maximum as argued above.

4.3 | Optimal allocation of sample size when the total sample size and cluster sizes are fixed

Take a fixed cluster size n_1. When we keep the total sample size $N = k_1 n_1 + k_0$ fixed and write the fraction in the intervention arm as $x = k_1 n_1/N$, then we can write the variance of $\widehat{\delta}_{\text{minvar}}$ in (8) as follows:

$$\text{var}\left(\widehat{\delta}_{minvar}\right) = \frac{A_1/N}{x} + \frac{A_0/N}{(1-x)},$$

where $A_1 = [1 + (n_1 - 1)\,\rho_1] \cdot \left(\sigma_1^2/\sigma_{base}^2\right) - r^2$ and $A_0 = \left(\sigma_0^2/\sigma_{base}^2\right) - r^2$ are constants.

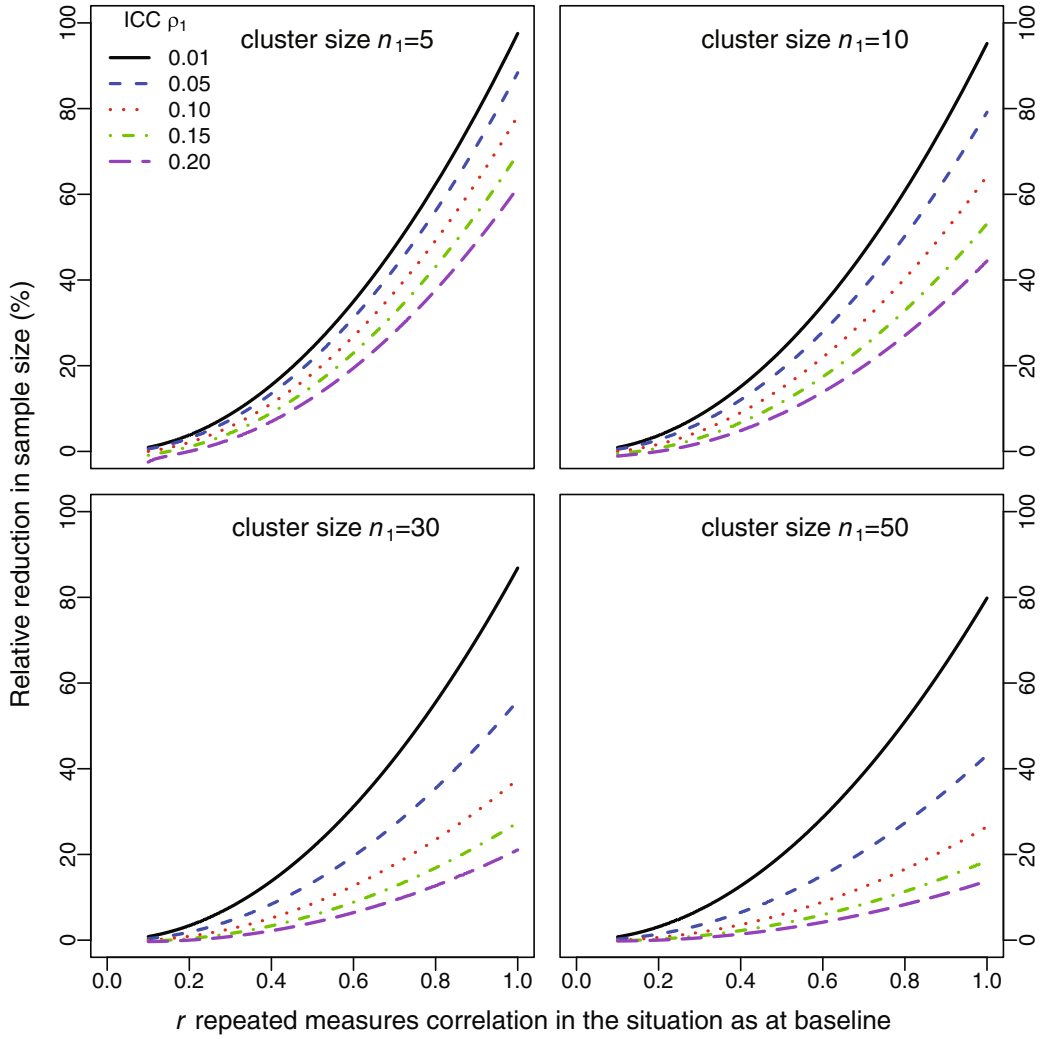

FIGURE 2 Relative reduction in sample size by adjusting for the baseline measurement in partially clustered designs. Relative reduction as a function of the repeated measures correlation as at baseline, $r = r_{bs,bs}$, for various intracluster correlations ρ_1 (different lines types) and various cluster sizes n_1 (different subfigures). The reduction is calculated as $r^2 / \left[\frac{1}{2}\left[1 + (n_1 - 1)\,\rho_1\right] \cdot \frac{r}{(r - \rho_1)} + \frac{1}{2}\right]$. This is derived from Formula (24) with (a) equal variance in the control condition at follow-up as at baseline; (b) equal allocation; (c) equal repeated measures correlation in the intervention arm at follow-up as at baseline (Approach 2′ in Section 4.1).

To determine the minimum of this function of x, we set its derivative (with respect to x) equal to 0 and get $x^2/(1-x)^2 = A_1/A_0$, so $x = \sqrt{A_1}/\left(\sqrt{A_1} + \sqrt{A_0}\right)$and

$$\text{var}\left(\hat{\delta}_{\text{minvar},optimal}\right) = \frac{\left(\sqrt{A_1} + \sqrt{A_0}\right)^2}{N}, \quad \text{with } A_1 = \left[1 + (n_1 - 1)\,\rho_1\right] \cdot \left(\sigma_1^2/\sigma_{base}^2\right) - r^2 \text{ and } A_0 = \left(\sigma_0^2/\sigma_{base}^2\right) - r^2 \tag{25}$$

and

$$A_{opt} = \frac{k_1 n_1}{k_0} = \frac{\sqrt{A_1}}{\sqrt{A_0}} = \sqrt{\frac{\left[1 + (n_1 - 1)\,\rho_1\right] \cdot \left(\sigma_1^2/\sigma_{base}^2\right) - r^2}{\left(\sigma_0^2/\sigma_{base}^2\right) - r^2}}. \tag{26}$$

Typically, the clustering effect of the intervention increases the variance at follow-up more than the control treatment, so $\left[1 + (n_1 - 1)\,\rho_1\right] \cdot \left(\sigma_1^2/\sigma_{base}^2\right) \geq \left(\sigma_0^2/\sigma_{base}^2\right)$ and $A_{opt} \geq 1$.

This reconfirms that the larger the clustering effect $\left[1 + (n_1 - 1)\,\rho_1\right]\sigma_1^2$, the more it will pay off to allocate more subjects to the arm with clustering.

4.3.1 | Optimal vs equal allocation

Suppose we optimally allocate N subjects in total (again cluster size n_1 is fixed, so A_1 and A_0 are constants). The total sample size N_{equal} that would achieve the same power satisfies

$$\frac{A_1}{N_{equal}/2} + \frac{A_0}{N_{equal}/2} = \text{var}\left(\widehat{\delta}_{\text{minvar},equal}\right) = \text{var}\left(\widehat{\delta}_{\text{minvar},optimal}\right) = \frac{\left(\sqrt{A_1} + \sqrt{A_0}\right)^2}{N}.$$

When we rewrite this (see web-appendix), we get that

$$\text{reduction sample size by optimal allocation} = \frac{N_{equal} - N}{N_{equal}} = \frac{1}{2} - \frac{A_{opt}}{1 + A_{opt}^2}.$$

To achieve a certain reduction, we solve the corresponding quadratic equation and get for the situation that $[1 + (n_1 - 1)\,\rho_1]\,\sigma_1^2 \geq \sigma_0^2$ (so $A_{opt} \geq 1$) that

$$\left.\begin{array}{c}\text{Reductions greater than } 10\%, 20\%, 30\%, 40\% \text{are achieved if } A_{opt} \geq c, \text{ie,} \\ [1 + (n_1 - 1)\,\rho_1] \cdot \left(\sigma_1^2/\sigma_{base}^2\right) \geq c^2\left(\sigma_0^2/\sigma_{base}^2\right) - \left(c^2 - 1\right) \cdot r^2, \\ \text{where } c = 2, 3, \left(2\tfrac{1}{2} + \tfrac{1}{2}\sqrt{21}\right) \approx 4.8, (5 + 2\sqrt{6}) \approx 9.9, \text{respectively.}\end{array}\right\} \quad (27)$$

4.4 | Design factors to calculate sample size

To derive design factors, we compare the variances of the intervention effect estimator (8) under equal allocation of the total sample size N (ie, $n_1 k_1 = k_0 = N/2$) or optimal allocation (25) to the variance of the t test at baseline. The t test has $N/2$ independent subjects in each arm with variance σ_{base}^2, so variance $4\sigma_{base}^2/N$. This leads to the following design factors:

$$DE_{equal} = \frac{1}{2}\left\{[1 + (n_1 - 1)\,\rho_1] \cdot \left(\sigma_1^2/\sigma_{base}^2\right) + \left(\sigma_0^2/\sigma_{base}^2\right) - 2r^2\right\}, \quad (28)$$

and

$$DE_{optimal} = \frac{1}{4}\left(\sqrt{[1 + (n_1 - 1)\,\rho_1] \cdot \left(\sigma_1^2/\sigma_{base}^2\right) - r^2} + \sqrt{\left(\sigma_0^2/\sigma_{base}^2\right) - r^2}\right)^2, \quad (29)$$

where the input parameters can be obtained as described in Section 4.1.

When there is no clustering in the intervention arm ($\rho_1 = 0$), and variance at follow-up is equal to that at baseline ($\sigma_1^2 = \sigma_0^2 = \sigma_{base}^2$), both these formulas collapse to those obtained when adjusting for baseline outcome measures in completely unclustered trials using analysis of covariance (ANCOVA).[23]

Calculating k_1 and k_0 from n_1. If the cluster size n_1 is fixed, we first calculate the total sample size $N_{tot,t-test}$ for a two independent samples t test (with sufficient power for the effect of interest at the desired significance level) using the SD in the baseline condition. Many programs and formulas available give the sample size per arm and if so, this has to be doubled to obtain the total sample size. This must then be multiplied by the design effect reflecting the chosen allocation to the intervention vs control arm (ie, equal or optimal allocation). From this total sample size, that is, $N_{tot} = DE \cdot N_{tot,t-test}$, the number of clusters k_1 in the intervention arm and control subjects k_0 can be determined:

$$k_1 = N_{tot}/(2n_1) \text{ and } k_0 = N_{tot}/2 \text{ for equal allocation.}$$

If $k_1 n_1/k_0 = A_{opt}$ is the optimal allocation, then

$$k_1 = N_{tot} \cdot A_{opt}/\left[n_1\left(1 + A_{opt}\right)\right] \text{ and } k_0 = N_{tot}/\left(1 + A_{opt}\right) \text{ for optimal allocation.}$$

Calculating n_1 from k_1. If the number of clusters k_1 is fixed and we use equal allocation, then the cluster size n_1 can be determined from $2k_1n_1 = k_1n_1 + k_0 = DE_{equal} \cdot N_{tot,t-test}$, and using (28) this results in

$$n_{1,equal} = N_{tot,t-test} \cdot \left[1 - 2r^2 + (1-\rho_1) \cdot \left(\frac{\sigma_1^2}{\sigma_{base}^2}\right)\right] / \left[4k_1 - \rho_1 \left(\frac{\sigma_1^2}{\sigma_{base}^2}\right) N_{tot,t-test}\right]. \quad (30)$$

In particular, designs with k_1 clusters are only feasible if $k_1 \geq \rho_1 N_{tot,t-test}\sigma_1^2 / \left(4\sigma_{base}^2\right)$.

If one aims for optimal allocation, the formula does not take a simple form and calculating the power over a range of total sample sizes will likely be faster.

4.5 | Small sample performance

We investigated the small sample properties of the asymptotic SE across a number of scenarios with small number of clusters and small to large intracluster correlation (simulation program in the web materials). We varied the number of clusters in the intervention condition as: k_1=2, 3, 4, 5, 6, 8, 10, 15, 20, and 25 and each cluster had $n_1 = 4$ subjects with an intracluster correlation $\rho_1 = 0.05, 0.10,$ and 0.20. Data were generated using the model described in Table 1. We took the residual variance at baseline and in the control arm at follow-up to be equal. The residual variance in the intervention condition at follow-up was a multiple of that at baseline: $\sigma_{1s\tau 1}^2/\sigma_{s\tau 0}^2 = 0.5, 1, 2$. Equal allocation was used, so that the number of subjects in the control condition was $k_0 = 4k_1$. The repeated measures correlation as at baseline was set at $r = 0.7$. Under the alternative hypothesis, we set the true intervention effect in the simulations as follows:

$$\delta = \sqrt{\text{var}\left(\widehat{\delta}_{minvar}\right)} \cdot \left[\Phi_{t,df}^{-1}(1-\beta) + \Phi_{t,df}^{-1}(1-\alpha/2)\right] \quad \text{with } \beta = 0.2 \text{ and } \alpha = 0.05,$$

that is, the size of intervention effect that would have 80% power at a two-sided significance level of 0.05 according to formula (8) and the quantile function $\Phi_{t,df}^{-1}$ of the t-distribution with df degrees of freedom. Then observing a rejection rate close to 80% in simulations with this intervention effect means that the formula based on the asymptotic SE predicts power well.

Degrees of freedom based on the effective sample size $df = k_1n_1/\left[1 + (n_1 - 1)\,\rho_1\right] + k_0 - 2$ did not perform well. Degrees of freedom based on the Satterthwaite's approximation (with the total variances and effective sample sizes at follow-up as input) performed much better, but not meaningfully better than degrees of freedom based on the total number of clusters. Therefore, we restrict the reported simulation results below to the choice $df = k_1 + k_0 - 2$.

We simulated 2000 partially nested randomized trials. The resulting precision of 0.01 $(= 1.96 \cdot \sqrt{0.05 \cdot 0.95/2000})$ in the estimated type I error and the precision of 0.018 $(= 1.96 \cdot \sqrt{0.80 \cdot 0.20/2000})$ in the estimated power give a clear picture of how our proposed formulas behave when ICC, number of clusters in the intervention arm, and the ratio of the residual error variance in the intervention arm to that in the control arm are varied. Results are displayed in Figures 3 and 4.

For comparison, we provided a mixed effect analysis (labeled "posttest" in the figures) based on only the follow-up measurements with Kenward-Rogers degrees of freedom for the fixed effects, a random effect for cluster, and a residual covariance structure allowing for different residual variances in the intervention compared to the control arm. This analysis has good type I error control, but is too conservative for a small number of clusters as was already found in other studies.[5,19]

Next, we fitted a mixed effects model (labeled "ANCOVA indiv") with Kenward-Rogers degrees of freedom for the fixed effects, a random slope for only the intervention arm at follow-up, a random intercept for subjects, and a residual covariance structure allowing for different residual variance at baseline, control arm at follow-up, and intervention arm at follow-up. This individual level analysis is aligned with the data-generating model and gains power compared to the "posttest" analysis, but type I error seems not so well controlled. Its non-convergence rate decreases from around 13% for $k_1 = 2$ to below 1% for $k_1 \geq 8$.

Finally, we performed a repeated measures analysis on cluster means (labeled "ANCOVA clusavg"). In the control arm, each subject was its own cluster. In the intervention arm, a 'cluster' at baseline consisted of all independent subjects that were at follow-up in the same cluster. Then we fitted a generalized least-squares model with the means of a cluster at baseline and follow-up as repeated measurements and Kenward–Rogers degrees of freedom for the fixed effects. We specified a heterogeneous compound symmetry structure that allowed variances and correlation among the repeated

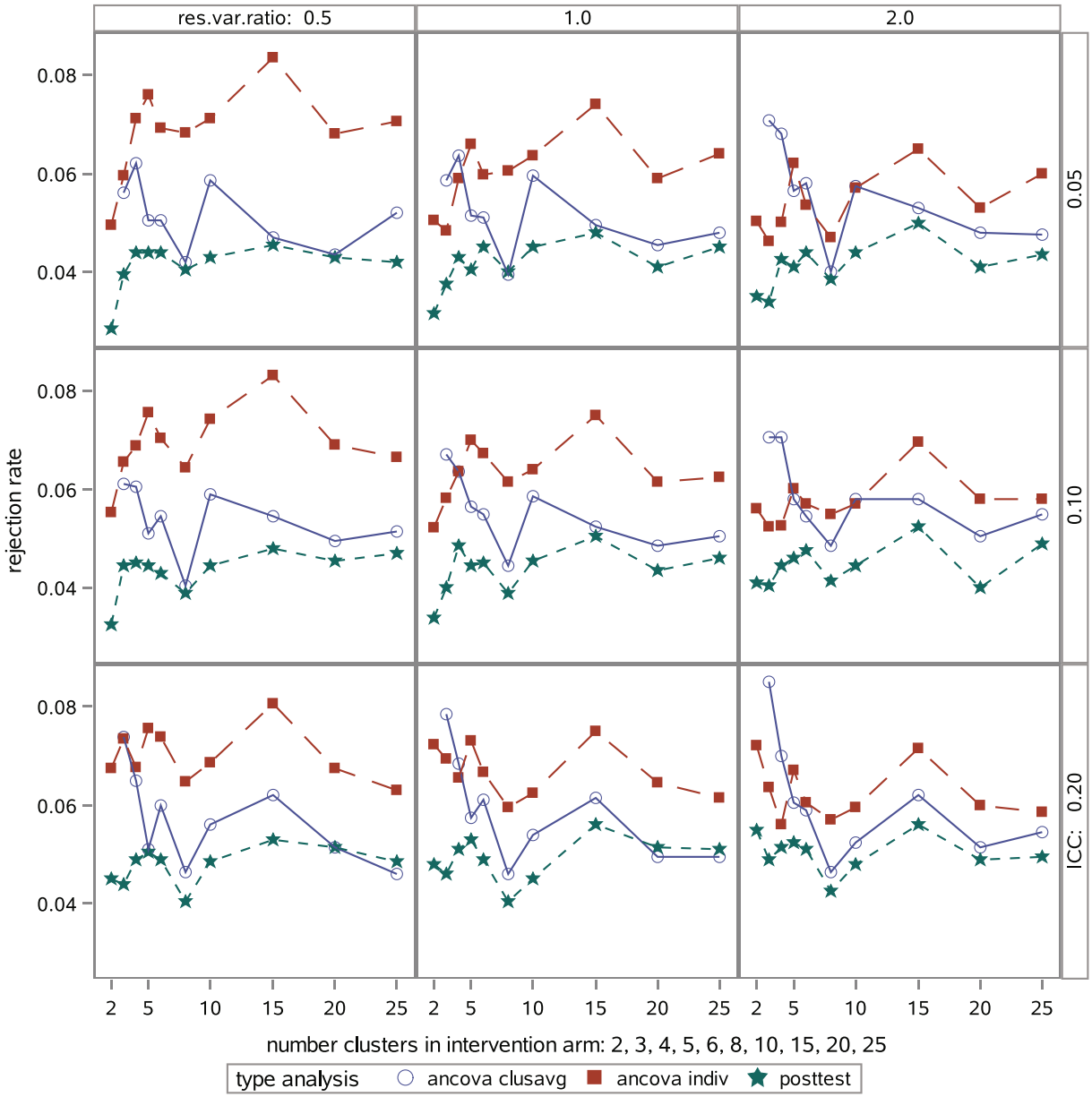

FIGURE 3 Estimated type I error. Rejection rate under the null hypothesis of no effect is based on 2000 trials in each scenario. The subject by time (residual) variance in the control arm at follow-up is equal to that at baseline ($\sigma^2_{sr0} = \sigma^2_{0sr1}$), the repeated measures correlation in the situation as at baseline is $r = 0.7$, each cluster in the intervention arm has $n_1 = 4$ subjects, and equal allocation is used. Res.var.ratio is the ratio of the residual variance in the intervention arm to that in the residual variance in the control arm ($\sigma^2_{1sr1}/\sigma^2_{0sr1}$ in Table 1). Three analyses were compared. One individual level analysis using only the follow-up measurement ("posttest"), and two analyses using both the follow-up and baseline measurement: an individual level analysis ("ANCOVA indiv") and a cluster level analysis ("ANCOVA clusavg") described in Section 4.5.

measurements to be different between arms. For $k_1 = 2$ intervention clusters, this model does not converge, probably because there are four data points (two intervention cluster means with baseline and follow-up measurement) out of which four parameters (mean outcome at follow-up in the intervention arm and relatedly the change from baseline in the intervention arm, variance at baseline and follow-up of the cluster mean, and their correlation) have to be estimated. For $k_1 = 3$, the non-convergence rate is $< 0.25\%$ and for $k_1 \geq 4$, the model always converged. Note that this model specification allows the correlation between baseline and follow-up, and variances at baseline and follow-up to be different for the intervention compared to the control arm. This covariance structure specification thus does not fully exploit the structure of the generated data, as the equality of the total subject variance at baseline (both arms) and at follow-up in the control condition implies a relation between the covariance parameters. Despite this, "ANCOVA clusavg" has the best power and

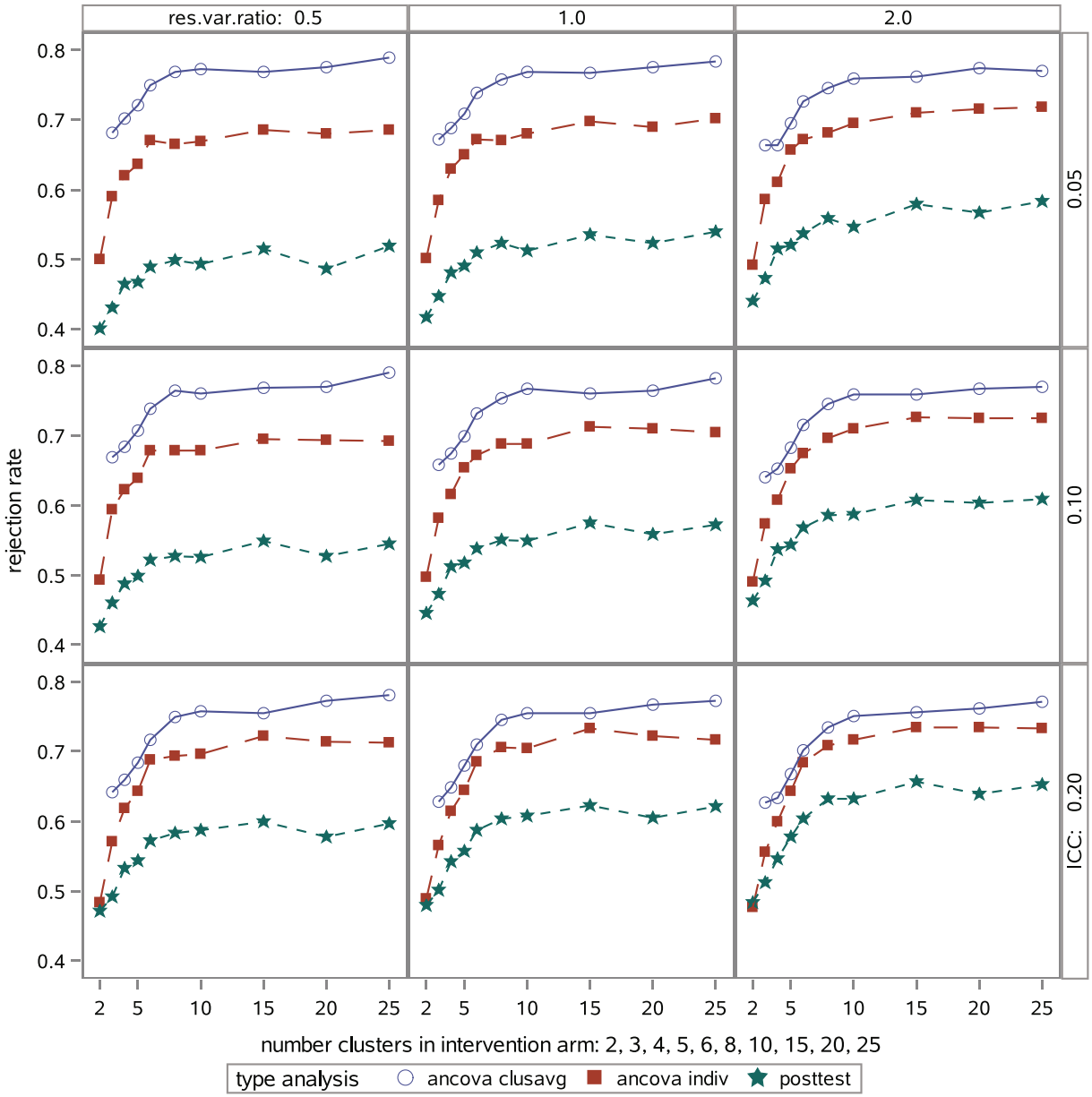

FIGURE 4 Estimated power. Rejection rate under the alternative hypothesis is based on 2000 trials in each scenario and an effect that is predicted to have 80% power from the derived formula for the SE. The subject by time (residual) variance in the control arm at follow-up is equal to that at baseline ($\sigma^2_{s\tau 0} = \sigma^2_{0s\tau 1}$), the test–retest in the situation as at baseline is $r = 0.7$, each cluster in the intervention arm has $n_1 = 4$ subjects, and equal allocation is used. Res.var.ratio is the ratio of the residual variance in the intervention arm to that in the residual variance in the control arm ($\sigma^2_{1s\tau 1}/\sigma^2_{0s\tau 1}$ in Table 1). Three analyses were compared. One individual level analysis using only the follow-up measurement ("posttest"), and two analyses using both the follow-up and baseline measurement: an individual level analysis ("ANCOVA indiv") and a cluster level analysis ("ANCOVA clusavg") described in Section 4.5.

generally good type I error control (except possibly for $k_1 \leq 4$ clusters). From $k_1 = 8$ clusters onward, its power levels seem reasonably close to 80%.

4.6 | Power calculations

A web program to calculate power based on the asymptotic SE is available at https://monash-biostat.shinyapps.io/DifferentialClustering/. For small samples, it would be wise to perform a simulation study. For this we provide a SAS program in the web materials.

4.7 | Analysis codes

We provide code to analyze data of a partially and fully nested trial in SAS, Stata, R, and SPSS, in the web-appendix.

4.8 | Summary of design advice

Generally, outcomes with a larger test–retest reliability (so larger r) lead to smaller sample sizes as will intervention treatments that induce a lower intracluster correlation ρ_1 or result in a lower variance σ_1^2. The same holds for control treatments that have smaller variance σ_0^2. Typically, we cannot influence these factors. What we can do is strive for many small clusters in the intervention arm. This is particularly important because many clusters can get us closer to the asymptotic maximum relative reduction of $r^2\left(\sigma_{base}^2/\sigma_0^2\right)$.

When the clustering effect $[1+(n_1-1)\,\rho_1]\cdot\sigma_1^2$ is not too far away from σ_0^2, equal allocation of sample size to the two arms will be quite efficient. More precisely stated, when the cluster size n_1 satisfies the condition

$$n_1 \leq \frac{1}{\rho_1}\left[\frac{4\left(\sigma_0^2/\sigma_{base}^2\right)-3r^2}{\sigma_1^2/\sigma_{base}^2}-1\right]+1, \tag{31}$$

reductions in sample size due to optimal instead of equal allocation will not be more than 10%, see (27). We expect clusters size n_1 to be smaller than in cluster randomized trials, because the more intense interaction of therapist with their subjects means the number of subject will rather be limited. Therefore, we expect condition (31) to be often satisfied and thus equal allocation to be satisfactory. We can then calculate the sample size from the design effect (28). Alternatively, we can then first calculate the total sample size for the partially nested design (with only a follow-up measurement) and then subtract r^2 times the total sample size for a t test.

When equal allocation is not efficient, it makes sense to consider optimal allocation and use the corresponding design factor (29).

4.9 | Case study example of a sample size calculation

Bennell et al.[7] conducted a trial that randomized patients with knee osteoarthritis to an activity program (control arm) or to telephone coaching (delivered by telephone coaches) in addition to that activity program (intervention arm). The activity program was delivered by physiotherapists. Because each physiotherapist treated patients in both arms of the trial and randomization was stratified by physiotherapist, the physiotherapist main effect cancels out, but there may be a physiotherapist by treatment interaction. For this trial, we expect that the effect of the telephone coaches is to decrease the variability of the patient's adherence to the activity program, for all physiotherapists. That is, the coaches will mitigate differences in therapist effects. Therefore, we illustrate the sample size calculation for an analysis without treatment by physiotherapist interaction. Since only patients in the intervention arm were treated by telephone coaches, there may be clustering effects in the intervention arm, so this is a partially clustered trial. We illustrate sample size and power calculations with one of the primary endpoints of the trial: change from baseline to 6 months in knee pain (a numeric rating scale 0-10). Of note, the factual sample size calculation used a different effect size and incorporated also the power for the other primary endpoint, hence is different from the illustrations presented here.

Suppose the effect of interest is $\delta = 1.3$ and the SD is $\sigma_{base} = 2.2$. The *total* sample size to detect this effect at a two-sided significance level $\alpha = 0.05$ with power $1-\beta = 0.8$ for an individually 1:1 randomized, parallel group design on only the measurement at the end of the trial is

$$N_{tot,indiv} = 4\cdot\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2\cdot\frac{\sigma_{base}^2}{\delta^2} \approx 4\cdot(1.96+0.84)^2\cdot\frac{(2.2)^2}{(1.3)^2} \approx 90.$$

The correlation between repeated measurements in the situation as at baseline is $r = 0.29$ based on other trials. The clustering within a coach is expected to result in an intracluster correlation $\rho_1 = 0.05$ and the control condition is not expected to change the variance compared to baseline: $\sigma_0^2 = \sigma_{base}^2$. We further assume that the repeated measurements correlation in the situation as at follow-up in the intervention condition is the same as the repeated measurement correlation

in the situation as at baseline: $\sigma_1^2/\sigma_{base}^2 = r/(r-\rho_1)$, see Approach 2′ in Section 4.1. According to (31), at least 44 patients per coach would be needed to have some (ie, 10%) reduction in sample size by using an optimal allocation. In this case the number of patients per coach is much lower, and 1:1 allocation will not suffer from a meaningful loss of power. Thus, 1:1 allocation is used. If each coach can support $n_1 = 5$ patients then the design effect for a 1:1 randomized partial clustered design assuming equal repeated measures correlation in the clustered and unclustered condition is

$$DE_{equal} = \frac{1}{2}\left\{[1+(n_1-1)\rho_1]\cdot\left(\frac{r}{r-\rho_1}\right)+1-2r^2\right\} = \frac{1}{2}\left\{[1+(5-1)\cdot 0.05]\cdot\left(\frac{0.29}{0.29-0.05}\right)+1-2(0.29)^2\right\} \approx 1.14,$$

so $N_{tot} = 1.14 \times 90 \approx 103$. With 110 patients in total, $k_0 = 55$ patients in the control group and $k_1 = 11$ coaches each treating $n_1 = 5$ patients would be expected to provide at least 80% power. To check this, 1000 simulated trials showed 81.2% power (type I error 0.054) using the "ANCOVA indiv" analysis and 80.9% power (type I error 0.055) for the "ANCOVA clusavg" analysis (see Section 4.5 for the description of analyses). With 100 patients in total, $k_0 = 50$ patients in the control group and $k_1 = 10$ coaches with $n_1 = 5$ patients, the "ANCOVA indiv" analysis has 76.3% power (type I error 0.044) and the "ANCOVA clusavg" has 75.0% power (type I error 0.048).

5 | DISCUSSION

For randomized two-arm designs with a baseline and a follow-up (ie, any post-baseline) measurement that have a multivariate normal distribution, we have shown that the sample size required for an analysis based on only the follow-up measurement can be reduced by adjusting for the baseline outcome measurement in the analysis. The variance of the treatment effect estimator takes the form

$$\text{var}(\widehat{\delta}) = \text{var}\left(\widehat{\delta}_{fu}\right) - \mathbb{r}^2\text{var}\left(\widehat{\delta}_{base}\right),$$

where $\widehat{\delta}_{base}$ and $\widehat{\delta}_{fu}$ are the estimators of the difference between the arms at baseline and at follow-up and $\mathbb{r} = \text{covar}\left(\widehat{\delta}_{fu}, \widehat{\delta}_{base}\right)/\text{var}\left(\widehat{\delta}_{base}\right)$. Thus, reductions in sample sizes (for any design) will only be small when $\text{var}\left(\widehat{\delta}_{fu}\right)$ is large (eg, due to clustering and/or cross-classification effects at follow-up) compared to $var\left(\widehat{\delta}_{base}\right)$ (which may be relatively small when for example no clustering/cross-classification is present at baseline).

The factor $\mathbb{r}$ turned out to be interpretable as a repeated measures correlation in the designs we considered. When scrutinizing the calculations, we see that this interpretation holds in these designs because the factors related to taking the difference between arms at baseline and follow-up cancel each other out. This will not generally be the case: $\mathbb{r}$ can be larger than 1 (hence not interpretable as a correlation) as noted in Section 2. Also in cases where $\mathbb{r}$ can be interpreted as a repeated measures correlation, although this is often the correlation between the repeated measures on a subject in the same situation as at baseline, this interpretation does not always hold. An example of this is the following: suppose subjects are nested in clusters at baseline (so a subject has total variance $\sigma_{c,base}^2 + \sigma_s^2 + \sigma_{s\tau 0}^2$) and are randomized, stratified by cluster, to new clusters. Half of the subjects in a baseline cluster are thus randomized to new control clusters and the other half to new intervention clusters. This means that the cluster effect at baseline with variance $\sigma_{c,base}^2$ is not present at follow-up. In this situation, the factor $\mathbb{r}$ takes the form $\sigma_s^2/\left(\sigma_s^2 + \sigma_{s\tau 0}^2\right)$ which can be interpreted as a repeated measures correlation. However, this interpretation needs to be "conditional on the (baseline) cluster" as the cluster variance term $\sigma_{c,base}^2$ is not present. In other words, first a cluster has to be chosen and then $\mathbb{r}$ is the repeated measures correlation of a randomly chosen subject in that cluster (in contrast to the 'unconditional' repeated measures correlation, in which a subject is randomly chosen from the whole sample at baseline).

Despite the above discussion that the factor $\mathbb{r}$ in the expression $\text{var}(\widehat{\delta}) = \text{var}\left(\widehat{\delta}_{fu}\right) - \mathbb{r}^2\text{var}\left(\widehat{\delta}_{base}\right)$ need not be a correlation in general, the variance reduction obtained by adjusting for the baseline measurement can always be expressed in term of a correlation using $\text{var}\left(\widehat{\delta}_{fu}\right)\cdot\left[1 - corr^2\left(\widehat{\delta}_{fu}, \widehat{\delta}_{base}\right)\right]$, see equation (3). However, this correlation is generally not a repeated measures correlation.

Regarding the choice of correlation parameters for sample size calculations, we recommend using the input of content matter specialists. However, it may be that researchers are not sure of the size of repeated measurements correlation in the situation as at follow-up in the intervention condition ($r_{1fu,1fu}$). As a starting point, we propose assuming that it is the

same as that at baseline $r_{bs,bs}$ (Approach 2′) for partially nested trials. Empirical evidence would be welcome to assess if and when this is a reasonable assumption. In any case, this implies that the variance in the intervention arm is larger than that in the control arm at follow-up. This is at least a more conservative approach than planning a trial based on equal variance in the intervention and control arm at follow-up. Nevertheless, it would be wise to vary $r_{1fu,1fu}$ also to higher (and lower) values than $r_{bs,bs}$. For the analysis, it is recommended to let the (co)variances in both arms at follow-up vary independently. This is because Figure 1 in Reference 2 showed that fixing the variance in the intervention arm to be larger than in the control arm may lead to bias. The analysis code in the appendix therefore allows the total variance in the intervention arm to vary independent from that in the control arm.

Extensions that could be investigated in the future research include the following:

More than three levels. We restricted our investigations to designs with at most three levels (measurements in subjects in clusters) in this article, but it is likely that the approach of Sections 2 and 3 could be applied to designs with more levels.

More than two measurements. For two measurements, the joint distribution can quite generally be captured by a random intercept random slope model. Using this model for more than two measurements would generalize results for cluster randomized trials with repeated measurements,[26] where the time trend is captured by dummy variables. However, a random intercept to capture the within-subject (and/or within-cluster) correlation implies a constant correlation regardless of the distance between measurements (similar as has been observed in[26]). The only way to vary the within-correlation between different time points is to use different residual variances at each time point. Thus, many types of joint distributions would not be covered by such models. Therefore, other models such as the exponential decay model could be considered.

Nonconcurrent time points of measurement between subjects. For cluster randomized trials with repeated measurements and a continuous linear time trend,[27] results are available to evaluate the gain in power by including a baseline measurement. For partially and fully nested designs, the SE for the intervention effect on slope was derived in a model with linear time trend.[28] Including a baseline measurement could be seen in such models as extending the time axis and the ensuing gain in power could be investigated. In the mentioned results, the same covariance structure at baseline and in both treatment conditions is assumed. Future research could extend this to a non-linear time trend and allow for (co)variances to be different between intervention and control arm and different for different time points.

More than two arms. The Body Project trial which investigated eating disorder prevention interventions is an example of partially nested trial with 4 arms.[29] For pairwise comparisons, the results of this article could be applied. However, for an overall test, extension of the current work would be welcome.

Other types of correlation. We focused on situations with independent subjects at baseline and showed extensions with nesting at baseline. Other situations at baseline such as cross-classification or multiple membership[6,14] at baseline and/or follow-up would be possible. We think these could be handled with the methods in Section 2 as well.

Varying cluster sizes. For analyses based on only the follow-up measurement, the loss of power is limited in case of cluster randomized trials,[30,31] as well as in partially nested trials.[16] Whether this also holds when adjusting for baseline in the designs considered in this article would need further investigation.

Multiple covariates. More and different covariates than the baseline outcome could be accounted for. Then, the GLS SE could be calculated from a matrix multiplication $var(\widehat{\beta}) = \left(X^t V^{-1} X\right)^{-1}$ where X is the design matrix and V the variance covariance matrix of the responses. Moerbeek et al. provided an example of this for one covariate.[32] On the other hand, Winkens et al.[33] showed that for individual randomization, a repeated measure model analysis that constrains the mean outcome to be the same between arms at baseline will generally outperform an ANCOVA when the covariance structures are different between arms. This supports our use of a repeated measurements model, as in our situation variances are different between arms (at follow-up) by design.

Costs. If costs are attached to obtaining a measurement (at baseline and/or follow-up), recruiting a subject, or a cluster (professionals), then optimal allocation in terms of costs could be derived similar as in situations with no clustering[34] or no baseline measurement.[15]

Regarding design advice, we make the following observations. With the trial of Bennell et al.,[7] we discussed the fact that stratifying the randomization may be more efficient and we showed how to tailor the sample size formula to a more complex hierarchical structure. There are more examples of this, especially if the interventions are add-on. In Morrell's trial in postnatal care,[10] all women received care by midwives and those in the intervention arm were also visited by a support worker. The randomization was conducted with sequentially numbered, sealed opaque envelopes that were prepared in advance with random digit tables. Likely, this randomization was not stratified by midwife but this would be more efficient. Thomas's acupuncture trial[8] was similar. Acupuncture was added-on to usual care by the general practitioner, but the method of randomization was not specified, and, therefore, it was not likely to have been stratified by

general practitioner. In another study, Morrell's leg ulcer care trial was conducted in eight clinics. Randomization was stratified by clinic to either care in the clinic or care by district nurses. Thus, we have two types of clusters at follow-up (clinic or district nurse) where one is a sub-cluster of the cluster at baseline. The situation was similar in Venning's trial[12] comparing care by the general practitioner vs care by the nurse practitioner. Randomization was by practice and the clusters at follow-up in one of the conditions were sub-clusters of those at baseline, that is, care by the general practitioner. Looking at these examples, we advise to be alert on such additional aspects of the hierarchical structure, tailor the sample size formula to it (similar to what was done in Section 3.3) and, if possible, to stratify randomization to cancel out part of the additional clustering.

In summary, our article has described how to assess the impact of including the outcome at baseline on sample size or power with a focus on situations with correlated measurements at follow-up. For the most common situation including the partially or fully nested trial with independent baseline measurements, syntax for the statistical analysis and an accompanying Rshiny app for power analysis are provided online.

DATA AVAILABILITY STATEMENT

The SAS® program code for producing simulated data to estimate the type I error and power are included as supplementary data file.

ORCID

Steven Teerenstra https://orcid.org/0000-0003-4103-7451
Jessica Kasza https://orcid.org/0000-0002-8940-0136

REFERENCES

1. Baldwin SA, Bauer DJ, Stice E, Rohde P. Evaluating models for partially clustered designs. *Psychol Methods*. 2011;16:149-165.
2. Roberts C, Roberts SA. Design and analysis of clinical trials with clustering effects due to treatment. *Clin Trials*. 2005;2:152-162.
3. Lange K, Kasza J, Sullivan TR, Yelland LN. Partially clustered designs for clinical trials: unifying existing designs using consistent terminology. *Clin Trials*. 2023;20:99-110.
4. Bauer DJ, Sterba SK, Hallfors DD. Evaluating group-based interventions when control participants are ungrouped. *Multivariate Behav Res*. 2008;43:210-236.
5. Flight L, Allison A, Dimairo M, Lee E, Mandefield L, Walters SJ. Recommendations for the analysis of individually randomised controlled trials with clustering in one arm – a case of continuous outcomes. *BMC Med Res Methodol*. 2016;16:165.
6. Roberts C, Walwyn R. Design and analysis of non-pharmacological treatment trials with multiple therapists per patient. *Stat Med*. 2013;32:81-98.
7. Bennell KL, Egerton T, Bills C, et al. Addition of telephone coaching to a physiotherapist-delivered physical activity program in people with knee osteoarthritis: a randomised controlled trial protocol. *BMC Musculoskelet Disord*. 2012;13:246.
8. Thomas KJ, MacPherson H, Thorpe L, et al. Randomised controlled trial of a short course of traditional acupuncture compared with usual care for persistent non-specific low back pain. *BMJ*. 2006;333:623.
9. Morrell CJ, Walters SJ, Dixon S, et al. Cost effectiveness of community leg ulcer clinics. *BMJ*. 1998;316:1487-1491.
10. Morrell CJ, Spiby H, Stewart P, Walters SJ, Morgan A. Costs and effectiveness of community postnatal support workers: randomised controlled trial. *BMJ*. 2000;321:593-598.
11. Mountain GA, Hind D, Gossage-Worral R, et al. 'Putting life in years' (PLINY) telephone friendship groups research study: pilot randomised controlled trial. *Trials*. 2014;15:141.
12. Venning P, Durie A, Roland M, Roberts C, Leese B. Randomized controlled trial comparing cost effectiveness of general practitioners and nurse practitioners in primary care. *BMJ*. 2000;320:1048-1053.
13. Moerbeek M, Safarkhani M. The design of cluster randomized trials with random cross-classifications. *J Educ Behav Stat*. 2018;43:159-181.
14. Walwyn R, Roberts C. Therapist variation within randomised trials of psychotherapy: implications for precision, internal, and external validity. *Stat Methods Med Res*. 2010;19:291-315.
15. Moerbeek M, Wong WK. Sample size formulae for trials comparing group and individual treatments in a multilevel model. *Stat Med*. 2008;27:2850-2864.
16. Candel M, van Breukelen G. Varying cluster sizes in trials with clusters in one treatment arm: sample size adjustments when testing treatment effects with linear mixed models. *Stat Med*. 2009;28:2307-2324.
17. Candel M, van Breukelen G. D-optimality of unequal versus equal cluster sizes for mixed effects linear regression analysis of randomized trials with clusters in one treatment arm. *Comput Stat Data Anal*. 2010;54:1906-1920.
18. Sterba SK. Partially nested desgins in psychotherapy trials: a review of modeling developments. *Psychother Res*. 2017;27:425-436.
19. Candlish J, Teare M, Dimairo M, Flight L, Mandefield L, Walters S. Appropriate statistical methods for analysing partially nested randomised controlled trials with continuous outcomes: a simulation study. *BMC Med Res Methodol*. 2018;18:105.
20. Roberts C, Batistatou E, Roberts SA. Design and analysis of trials with a partially nested design and a binary outcome measure. *Statist Med*. 2016;35:1616-1636.

21. Kariya T, Kurata H. *Generalized Least Squares*. Chichester: John Wiley & Sons; 2004.
22. Samuel-Cahn E. Combining unbiased estimators. *Am Statist*. 1994;48:34-36.
23. Borm GF, Franssen J, Lemmens WL. A simple sample size formula for analysis of covariance in randomized clinical trials. *J Clin Epidemiol*. 2007;60:1234-1238.
24. Teerenstra S, Eldridge S, Graff M, De Hoop E, Borm GF. A simple sample size formula for analysis of covariance in cluster randomized trials. *Stat Med*. 2012;31:2169-2178.
25. Moerbeek M, Teerenstra S. *Power Analysis of Trials with Multilevel Data*. Boca Raton: Chapman and Hall/CRC Press; 2016.
26. Hooper R, Teerenstra S, de Hoop E, Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Stat Med*. 2016;35:4718-4728.
27. Heo M, Leon AC. Sample size requirements to detect an intervention in longitudinal cluster randomized clinical trials. *Stat Med*. 2009;28:1017-1027.
28. Magnusson K, Andersson G, Carlbring P. The consequences of ignoring therapist effects in trials with longitudinal data: a simulation study. *J Consult Clin Psychol*. 2018;9:711-725.
29. Stice E, Shaw H, Burton E, Wade E. Dissonance and healthy weight eating disorder prevention programs: a randomized efficacy trial. *J Consult Clin Psychol*. 2006;74:263-275.
30. Girling A. Relative efficiency of unequal cluster sizes in stepped wedge and other trial designs under longitudinal or cross-sectional sampling. *Stat Med*. 2018;37:4652-4664.
31. van Breukelen G, Candel M, Berger M. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Stat Med*. 2007;26:2589-2603.
32. Moerbeek M, Van Breukelen G, Berger M. Optimal experimental designs for multilevel models with covariates. *Commun Stat Theory Methods*. 2001;30:2683-2697.
33. Winkens B, Van Breukelen G, Schouten H, Berger M. Randomized clinical trials with a pre- and a post-treatment measurement: repeated measures versus ANCOVA models. *Contemp Clin Trials*. 2007;28:713-719.
34. Green PG, Lin W, Gerber C. Optimal allocation of interviews to baseline and endline surveys in place-based randomized trials and quasi-experiments. *Eval Rev*. 2018;42:391-422.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Teerenstra S, Kasza J, Leontjevas R, Forbes AB. Sample size for partially nested designs and other nested or crossed designs with a continuous outcome when adjusted for baseline. *Statistics in Medicine*. 2023;42(19):3568-3592. doi: 10.1002/sim.9820