Behavior Research Methods (2021) 53:1731–1745
https://doi.org/10.3758/s13428-020-01529-7

The effect of missing data on design efficiency in repeated cross-sectional multi-period two-arm parallel cluster randomized trials

Mirjam Moerbeek[1]

Accepted: 17 December 2020 / Published online: 2 February 2021

Abstract

The reduced efficiency of the cluster randomized trial design may be compensated by implementing a multi-period design. The trial then becomes longitudinal, with a risk of intermittently missing observations and dropout. This paper studies the effect of missing data on design efficiency in trials where the periods are the days of the week and clusters are followed for at least one week. The multilevel model with a decaying correlation structure is used to relate outcome to period and treatment condition. The variance of the treatment effect estimator is used to measure efficiency. When there is no data loss, efficiency increases with increasing number of subjects per day and number of weeks. Different weekly measurement schemes are used to evaluate the impact of planned missing data designs: the loss of efficiency due to measuring on fewer days is largest for few subjects per day and few weeks. Dropout is modeled by the Weibull survival function. The loss of efficiency due to dropout increases when more clusters drop out during the course of the trial, especially if the risk of dropout is largest at the beginning of the trial. The largest loss is observed for few subjects per day and a large number of weeks. An example of the effect of waiting room environments in reducing stress in dental care shows how different design options can be compared. An R Shiny app allows researchers to interactively explore various design options and to choose the best design for their trial.

Keywords Cluster randomization · Dropout · Intermittently missing observations · Efficiency

Introduction

Over the past two decades, the cluster randomized trial (Campbell & Walters, 2014; Donner & Klar, 2000; Eldridge & Kerry, 2012; Hayes & Moulton, 2009; Murray, 1998) has become a standard design in the biomedical, health and behavioral sciences. As outcomes of subjects within the same cluster are correlated, cluster randomization has lower efficiency than individual randomization. Efficiency may be improved by increasing the sample size, but this is not always possible in practice, as the number of clusters and cluster size are often limited. Various alternative strategies have been proposed to increase efficiency, such as including covariates (Bloom, 2005; Bloom, Richburg-Hayes, & Black, 2007; De Hoop, Teerenstra, Van Gaal, Moerbeek, & Borm, 2012; Konstantopoulos, 2012; Moerbeek, 2006; Murray & Blitstein, 2003; Raudenbush, 1997; Raudenbush, Martinez, & Spybrook, 2007), taking a pretest measurement on the response variable (Murray, 2001; Murray & Blitstein, 2003; Murray, Hannan, Wolfinger, Baker, & Dwyer, 1998; Murray, Van Horn, Hawkins, & Arthur, 2006) and taking multiple measurements at baseline and endline (Copas & Hooper, 2020).

Another strategy involves implementing a multi-period design within a cluster randomized trial such that the study becomes longitudinal. The duration of the trial is split into periods, such as days, weeks or months. Within each period a treatment is implemented within each cluster, and subjects are measured on their outcome variables. Introducing multiple periods raises questions with respect to the design and analysis of cluster randomized trials. What is the optimal trade-off between study duration, the number of periods, the number of clusters and the number of subjects per cluster per period? What is the increased efficiency of a cohort versus a repeated cross-sectional design? What is the increased efficiency of a crossover or stepped-wedge design versus a parallel-group design? What is the appropriate model for data obtained from multi-period trials, and which is the best estimation method?

✉ Mirjam Moerbeek
m.moerbeek@uu.nl

[1] Department of Methodology and Statistics, Utrecht University, PO Box 80140, 3508 TC Utrecht, The Netherlands

How do we model the correlation between outcomes of subjects within the same or different period? How does the correlation structure influence efficiency? Over the past two decades, dozens of papers on the design and analysis of multi-period cluster randomized trials have appeared, especially in the biostatistical literature (Giraudeau, Ravaud, & Donner, 2008; Grantham, Kasza, Heritier, Hemming, & Forbes, 2019; Hooper & Bourke, 2015; Rietbergen & Moerbeek, 2011).

A question that has received too little attention thus far is: what are the effects of missing data on design efficiency in multi-period cluster randomized trials? Two types of missing data can be distinguished: those that are planned by the researcher and those that are not. The former are referred to as planned missing data designs (Rhemtulla, Jia, & Little, 2014; Wu, Jia, Rhemtulla, & Little, 2016). Such designs occur when, for generally practical reasons, the researcher plans not to record all the measures that would be desirable, but only some, for instance to lower the burden on the clusters. In this case the registration is necessarily carried out intermittently. For instance, a researcher can plan a trial with a duration of a certain number of weeks where measurements are only taken on the five workdays of each week. In such a trial design, intermittently missing data occur at the two weekend days of each week. It would then be of interest to compare such a design with one in which measurements are taken on all seven days each week.

The other type of missing data occurs unintentionally when a researcher plans to measure on certain days, but for reasons beyond his or her control, measurements cannot be taken on some of these days. Such unplanned missing data can describe an intermittent or monotone pattern (or a combination of the two), depending on the position of the missing values within the longitudinal study design. Intermittently missing data, also known as non-monotone or general missing data, are missing within a trajectory: there are missing observations between the observed. Monotone missing data are missing either at the beginning or at the end of a trajectory. The monotone missing data include the case of left- or right-censored follow-ups. If measurements cannot be taken on a certain day, then all the following (respectively, preceding) days are also missing (Genolini, Écochard, & Jacqmin-Gadda, 2013). Right-censored follow-ups occur when there is dropout. Dropout is the rule rather than the exception in longitudinal research. During the course of the trial, clusters may drop out for various reasons, for instance because they were randomized to the least interesting treatment condition, because they lost interest in the trial or because they are no longer willing to put effort in recruiting, treating and measuring subjects. The effect of dropout in longitudinal trials has been studied for trials in which there is no clustering (Galbraith & Marschner, 2002; Hedeker, Gibbons, & Waternaux, 1999; Moerbeek, 2008; Molenberghs & Verbeke, 2001; Vallejo, Ato, Fernández, & Livacic-Rojas, 2019) and for cluster randomized trials (Heo, 2014; Roy, Bhaumik, Aryal, & Gibbons, 2007).

This paper extends previous research on the effect of dropout on planned missing data design efficiency of repeated cross-sectional multi-period two-arm parallel cluster randomized trials. The design is repeated in cross-sectional fashion, meaning that different sets of subjects are measured during each period and hence each subject is measured only once. The design is parallel, meaning that each cluster receives one treatment condition, and does not change treatment during the course of the trial. All calculations can be done using a Shiny app that is available on the internet. This Shiny app will be explained in further detail later in this contribution. The focus is on trials where each period is one day of the week, and where clusters are followed for either one or multiple weeks. In such trials the implementation of treatment can be done during a single day, and outcomes are measured the same day. Examples are trials that evaluate methods to reduce anxiety in the waiting room of medical care practices, trials that evaluate nutrition and hydration on the day of exams, those that evaluate a new type of equipment or procedure during surgery, and trials that evaluate distraction methods during child vaccination. There exist trials in which the period is shorter (i.e. a fraction of the day) or longer (e.g. a week or a month). The Shiny app cannot be used for such trials.

The effect of dropout is studied using the Weibull survival function to model the probability of clusters dropping out during the course of planned missing data designs, by comparing various weekly measurement schemes. This function allows for various rates of dropout and also for constant, increasing or decreasing dropout probabilities over time, the same in both groups (experimental group and control group) or different. The dropout occurs at the level of the cluster, meaning that once a cluster drops out, no further data are recorded on any subjects within that cluster. The Shiny app does not allow for dropout at the individual level.

The multilevel model is used to relate a subject's outcome to study period and treatment condition. This model explicitly takes into account the nesting of subjects within clusters, and hence the correlation of outcomes of subjects within the same cluster. Furthermore, the correlation between outcomes of subjects within different time periods is modeled to decrease with increasing lag between these time periods. Design efficiency is measured by the variance of the treatment effect estimator. A high variance implies the design is inefficient, meaning that it has low statistical power for the test on treatment effect, whereas a low variance implies an efficient design with high statistical power. An example of a trial evaluating a method to reduce anxiety in the waiting room of dental practices is used to illustrate the findings in this paper. Throughout the paper it is assumed that the missing data mechanism is either missing completely at random (MCAR) or missing at random (MAR). In both cases the missingness is unrelated to the outcome variable. For MCAR, the missingness does not depend on other variables either, while for MAR it does depend on other variables, such as treatment condition. It is much more difficult to model informative missingness (missing not at random, MNAR), which is therefore outside the scope of this paper and not included in the Shiny app. Furthermore, the app is restricted to linear models for quantitative outcomes.

Statistical model

This section describes the statistical model for the repeated cross-sectional multi-period two-arm parallel cluster randomized trial. A graphical presentation of this trial is given in Fig. 1. The multilevel model, also known as the mixed or hierarchical model, is used to describe the relation between treatment condition and outcome (Goldstein, 2011; Hox, Moerbeek, & Van de Schoot, 2018; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). The outcome y_{hij} of subject $i = 1, \ldots, m$ in period $h = 1, \ldots, T$ in cluster $j = 1, \ldots, 2k$ is given by

$$y_{hij} = \beta_h + x_j\theta + u_{hj} + e_{hij}. \tag{1}$$

β_h is the period effect for period h, x_j denotes treatment condition (0 = control, 1 = intervention), and θ is the effect of treatment. Note that because all period effects are included in the model, a common intercept is not required.

The model explicitly takes the hierarchical data structure into account by including a random term at the subject level, $e_{hij} \sim N\left(0, \sigma_e^2\right)$, and another one at the level of the cluster-period, $u_{hj} \sim N\left(0, \sigma_u^2\right)$. The first implies that the outcomes of subjects within a cluster-period vary across the mean score within that cluster-period. The second implies that the mean score of a cluster in a certain period varies across the mean within that period across all clusters in the same treatment condition. These two random terms are assumed to be independent from each other; hence the variance of an outcome is simply the sum of the two variance components: $var\left(y_{hij}\right) = \sigma_e^2 + \sigma_u^2$. The intraclass correlation coefficient is the proportion variance at the cluster level: $\rho = \sigma_u^2 / \left(\sigma_e^2 + \sigma_u^2\right)$. It is the expected correlation between the outcomes of two randomly drawn subjects within the same cluster-period. The covariance between two outcomes from different clusters is always equal to zero, as outcomes of subjects from different clusters are assumed independent. The covariance between two outcomes within the same cluster-period is $cov\left(y_{hij}\ , y_{hi'j}\right) = \sigma_u^2$. The covariance between two outcomes within different cluster-periods h and h' is smaller and is defined as $cov\left(y_{hij}\ , y_{h'\ i'\ j}\right) = \sigma_u^2 r^{|h' - h|}$. Here we account for a non-uniform exponential decay structure: the covariance becomes smaller if the lag between the two cluster-periods increases (Grantham et al., 2019; Kasza, Hemming, Hooper, Matthews, & Forbes, 2019). This structure is also known as the first-order autoregressive structure. It should be noted that many other covariance structures are possible, such as Toeplitz, and the heterogeneous first-order autoregressive and Toeplitz structures. See for instance section 5.1 of Liu (2016) for an extensive description of such structures. Although these structures are also very

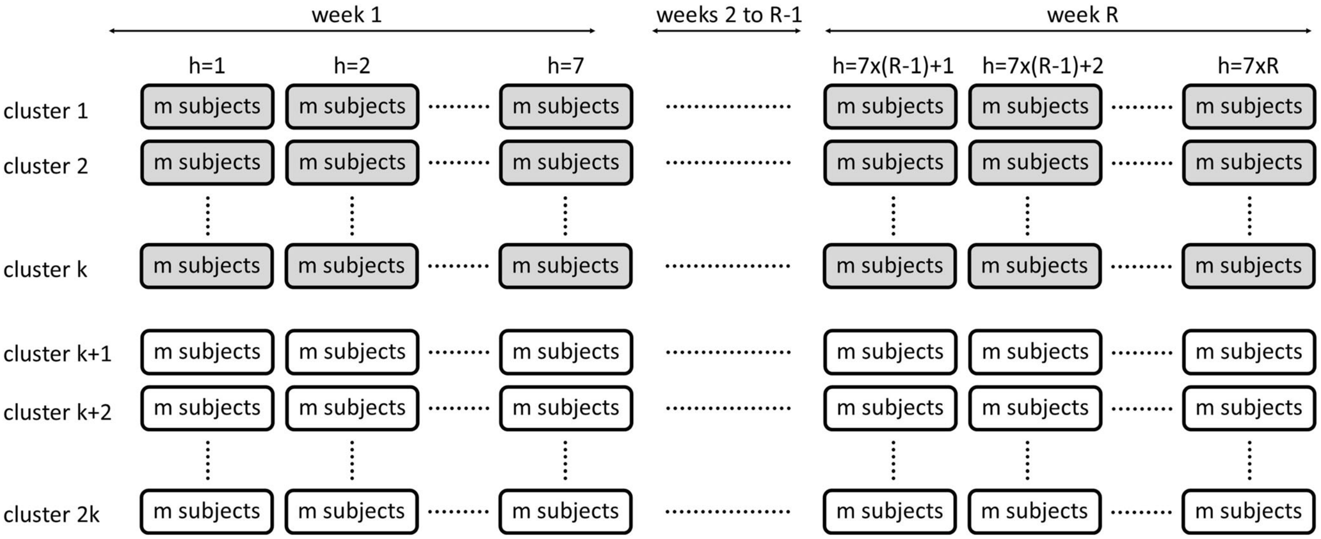

Fig. 1 Schematic representation of the repeated cross-sectional multi-period two-arm parallel cluster randomized trial. Each box represents a cluster-period. As the design is repeated cross-sectional, different sets of subjects are included in each of the cluster-periods

common in longitudinal data research, the first-order autoregressive structure is the only structure that is applied in this contribution and its Shiny app.

The first-order autoregressive structure is the most parsimonious of these structures, as it specifies only two covariance parameters to describe all variances and covariances within a cluster across all time periods: $\rho \in [0, 1]$ and $r \in [0, 1]$. As was explained above, the first is a correlation that only applies if the two subjects are within the same cluster-period. The second quantifies how much this correlation decreases if the two subjects are still in the same cluster but one period apart: $(1 - r) \times 100\%$ is the percentage reduction in correlation. For instance, if $r = 0.9$, then the covariance decreases by $(1 - r) \times 100\ \% = 10\%$ per period. Say the correlation is $\rho = 0.05$. Then the correlation between two subjects one period apart is 90% of 0.05, or 0.045; the correlation between two subjects two periods apart is 90% of 0.045, or 0.0405, and so forth. If $r = 1$, then the covariance structure is uniform, which implies all covariances are equal, irrespective of the lag between any two cluster-periods. This structure is also known as compound symmetry. It results in higher correlations between cluster-periods than the first-order autoregressive structure, and as a result, the variance of the treatment effect estimator is higher, especially when $(1 - r)$ is large (Grantham et al., 2019; Kasza et al., 2019). Therefore, the compound symmetry correlation structure results in lower power.

The model for cluster j in matrix notation is

$$\boldsymbol{y}_j = \boldsymbol{X}_j\gamma + \boldsymbol{Z}_j\boldsymbol{u}_j + \boldsymbol{e_{ij}}. \tag{2}$$

$\boldsymbol{y}_j$ is a vector of length $m \times T$ of responses in cluster j, and $\boldsymbol{X}_j$ is the $m \times T$ by $T + 1$ design matrix for the fixed parameters in cluster j. $\gamma = (\beta_1, \beta_2, \ldots, \beta_T, \theta)'$ is the vector of length $T + 1$ with the effects of period and treatment. $\boldsymbol{Z}_j$ is the $m \times T$ by T design matrix for the random cluster-level effects in cluster j. $\boldsymbol{e}_{ij}$ is the vector of length $m \times T$ with random subject-level effects. $\boldsymbol{e}_{ij} \sim N_{m\times T}\left(\boldsymbol{0}, \sigma_e^2\boldsymbol{I}_j\right)$ with $\boldsymbol{I}_j$ the $m \times T$ by $m \times T$ identity matrix in cluster j. $\boldsymbol{u}_j = (u_{j1}, u_{j2}, \ldots, u_{jT})'$ is the vector of length T with random cluster-level effects. $\boldsymbol{u}_j \sim N_T\left(0, \sigma_u^2\boldsymbol{R}_{j,hh'}\right)$ with $\boldsymbol{R}_{j,hh'} = cov\left(y_{hij}\ , y_{h'\ i'\ j}\right) = \sigma_u^2 r^{|h'-h|}$.

The covariance matrix of the outcomes in cluster j is calculated as

$$cov\left(\boldsymbol{y}_j\right) = \boldsymbol{V}_j = \sigma_u^2\boldsymbol{Z}_j\boldsymbol{R}_j\boldsymbol{Z}_j' + \sigma_e^2\boldsymbol{I}_j. \tag{3}$$

The regression coefficients are estimated as

$$\widehat{\gamma} = \left(\sum_{j=1}^{2k}\boldsymbol{X}_j'\widehat{\boldsymbol{V}}_j^{-1}\boldsymbol{X}_j\right)^{-1}\sum_{j=1}^{2k}\boldsymbol{X}_j'\widehat{\boldsymbol{V}}_j^{-1}\boldsymbol{y}_j \tag{4}$$

with associated covariance matrix

$$\widehat{cov}\left(\widehat{\gamma}\right) = \left(\sum_{j=1}^{2k}\boldsymbol{X}_j'\widehat{\boldsymbol{V}}_j^{-1}\boldsymbol{X}_j\right)^{-1}. \tag{5}$$

The element in row $T + 1$ and column $T + 1$ is the variance of the treatment effect estimator, $var\left(\widehat{\theta}\right)$, which has our primary attention in the remainder of this paper. This variance indicates how efficiently the treatment effect is estimated. A low variance is preferred, as it implies an efficient estimate and hence a high power for the test on treatment effect. High variance, on the other hand, implies an inefficient estimate and hence low power. A simple analytical expression for $var\left(\widehat{\theta}\right)$ is difficult to obtain, especially when the trial includes many cluster-periods or in the case of missing data. For that reason, $var\left(\widehat{\theta}\right)$ is calculated numerically in the software `R` (R Core Team, 2020), using the function `solve` to invert matrices. An R shiny app (Chang, Cheng, Allaire, Xie, & McPherson, 2016) is made available to calculate $var\left(\widehat{\theta}\right)$ of various designs and to compare these designs with each other. Shiny is a working environment for the development of web applications in the R language. The application is an interactive app for applied researchers and methodologists that allows easy access to a series of tools for evaluating the effect of planned missing data and dropout on design efficiency and statistical power in repeated cross-sectional multi-period two-arm parallel cluster randomized trials. The Shiny app will be described in further detail in Section 4.

An example

Anxiety in patient-centered care is linked to negative health outcomes, such as longer recovery periods, lower pain thresholds and resistance to treatment. Various methods aimed at reducing pre-procedure waiting anxiety have been proposed: music, aromatherapy, interior design features, play opportunities and media distractions (Biddiss, Knibbe, & McPherson, 2014).

Leather, Beale, Santos, Watts, and Lee (2003) compared two different types of waiting room environments in a United Kingdom neurology outpatient waiting area, a so-called nouveau environment and a traditional environment. These environments differed with respect to various features including general layout, color scheme, floor covering and lighting. The two types of environments were compared on outcome measures such as self-reported stress and anxiety, satisfaction ratings and pulse readings. The authors concluded that the physical design of the hospital environment is an important and integral part of the therapeutic milieu.

Suppose this trial is to be replicated in another setting in another country, say in dentistry care in the Netherlands. It is obvious that multiple dental practices are to be recruited so that a sufficient number of patients can be enrolled in a limited amount of time, and so that the results will be generalizable to all dental practices in the country. The redesign of the waiting areas may be expensive, so a trial like this will obviously run for multiple weeks to justify these trial costs. In theory, a crossover or stepped-wedge trial design would be possible. The advantage of clusters changing their treatment during the course of a trial is increased efficiency. However, in this specific trial this advantage may be outweighed by increased costs due to the redesign of the waiting areas during each crossover. Also, redesigning multiple waiting areas during the course of a weekend may be difficult to achieve, and it may also result in increased dropout rates. The parallel-group design is therefore the most obvious choice.

Shiny app

The Shiny app is available online at https://utrecht-university.shinyapps.io/missing_data_CRT/. It can be used to plan a repeated cross-sectional multi-period two-arm parallel cluster randomized trial and in four different situations:

1. When there is no data loss. In this case, data are recorded for seven days a week for at least one week, and dropout is absent. Section 5 shows how design efficiency is influenced by the number of weeks and number of subjects per cluster-period, for various values of ρ and r.
2. Trials where the registration is carried out during a certain number of days of the week. In other words, the trial is planned under the planned missing data design umbrella, and there is no data loss due to dropout. Section 6 considers four conditions of a planned missing data design with measurements on five or fewer days of the five workdays (Monday through Friday) and compares these to the complete data design (i.e. the design from situation 1).
3. Trials for which measurements are taken seven days of the week and where dropout occurs. In Section 7, the Weibull survival function is used to describe the amount of dropout and whether the risk of dropout increases, decreases or remains constant during the course of the trial. Designs with various dropout patterns will be compared with the complete data design.
4. Trials with intermittently planned missing observations with dropout. In Section 8 the example from the previous section is revisited. Measurements are taken Mondays through Fridays, and designs with either 4 or 8 weeks and with either 10 or 15 clusters per condition are compared in a situation where dropout occurs.

The Shiny app allows for the specification of five designs. Whenever acceptable values have been entered, it returns the results for those five designs. The user has to pay attention to the designs that are most convenient in his/her trial by changing the default input. The top part of the Shiny app allows the user to specify the designs, the correlation parameters, details about the statistical test on treatment effect, and the dropout pattern in both experimental conditions. The right part of the bottom shows the variance of the treatment effect estimator, power and design efficiency in graphical and table format.

The left part at the top allows the user to specify the design parameters. The range of the number of subjects per day is specified at the top. Below that, five different designs can be specified by selecting the days of the week in which measurements are taken, along with the number of weeks ($R \geq 1$) and number of clusters per condition ($k \geq 1$). It should be noted that designs should be specified such that measurements are taken on at least two days. So a design with one week and one day will generate a warning. Note that the five default sets of days of the week are those that are used in Section 6. However, the user can select any number and combination of days of the week. In the middle part at the top, the correlation parameters $\rho \in [0, 1]$ and $r \in [0, 1]$ should be specified, along with the details for the test on treatment effect (effect size Cohen's $d \neq 0$, type I error rate $\alpha \in [0, 1]$ and whether the test is one- or two-sided). In the right part at the top, the parameters of the Weibull survival function should be specified ($\omega \in [0, 1]$ and $\gamma > 0$), along with the maximum duration of the trial ($t_{max} \geq 1$). These parameters will be explained in further detail in Section 7; for now it is important to understand that the parameter ω should be equal to zero in the case where there is no dropout (situations 1 and 2) and between 0 and 1 if there is dropout (situations 3 and 4). The app allows different dropout patterns for the two experimental conditions. The hazard and survival probability functions are displayed in the second and third tabs. The user may consult Table 11.1 in Moerbeek and Teerenstra (2016) for a priori estimates of ρ. Unfortunately, such an extensive overview of estimates does not exist for r, ω or γ, and the user is encouraged to search the literature for similar studies, or use experts' opinions or expectations.

Once all input has been specified, the submit button at the bottom left should be pressed. Calculating the output may take a while, especially when the number of subjects per cluster-period is large; the progress is shown at the bottom right. The output is given in three graphs (variance of the treatment effect estimator, power for the test on treatment effect, and efficiency of designs 2—5 as compared to design 1). The same output is given in table format; the number of subjects per cluster-period is shown in the first column, the variance of the treatment effect estimator of designs 1—5 shown in the next five columns, the power of these designs in the next five columns, and their efficiencies as compared to design 1 in the last five columns.

Trials without missing data

This section summarizes how the variance of the treatment effect estimator behaves in trials without missing data. For more extensive results the reader is referred to related papers in the medical statistical literature (Grantham et al., 2019; Kasza et al., 2019). The variance depends on the intraclass coefficient ρ and decay parameter $1-r$ and on the design $\xi=(k, m, T)$, where k is the number of clusters per condition, m is the number of subjects per cluster-period and T is the number of periods. The design can also be written as $\xi=(k, m, R)$, where R is the number of consecutive weeks over which measurements are taken. In the case where measurements are taken on each of the seven days of the week, the number of periods is $T=7\times R$.

Figure 2 shows the variance of the treatment effect estimator as a function of the number of subjects per cluster-period m, number of weeks R, the intraclass correlation coefficient ρ and decay parameter $1-r$. The variances are calculated for trials with $k=5$ clusters per treatment condition. As explained previously, the variance is a measure of efficiency: the higher the variance, the less efficient the design.

Figure 2 shows that the variance increases with increasing intraclass correlation ρ. This is not surprising, as the same relation holds for cluster randomized trials with just one period: the higher the intraclass correlation, the higher the correlations between the outcomes of subjects in the same cluster, the less information there is in the data, and hence the higher the variance of the treatment effect estimator. Figure 2 also shows that the variance decreases with increasing decay $1-r$. A higher decay implies a lower correlation between outcomes in the same cluster, and hence a lower variance, but this parameter is most often not under the control of the researcher. By using the Shiny app, the reader can easily verify that highest variance is obtained with the compound symmetry correlation structure, for which $1-r=0$.

Figure 2 also shows that the variance decreases with increasing number of subjects per cluster-period m, especially when the number of weeks R is low. However, this effect becomes negligible when the number of subjects per cluster-period becomes larger than 5, and the variance approaches a limit when the number of subjects per cluster-period further increases. A similar relation holds for cluster randomized trials with just one period: increasing the cluster size has some effect when cluster size is small, but the effect becomes minimal at larger cluster sizes (Hemming, Girling, Sitch, Marsh, & Lilford, 2011; Moerbeek & Teerenstra, 2016; Raudenbush, 1997). Furthermore, Fig. 2 shows that the variance decreases with increasing number of weeks, especially when the number of weeks is small. This is also obvious, since a larger study duration implies that more measurements are taken and hence a lower variance is achieved.

As follows from Eq. (5), the variance is inversely related to the number of clusters, a relation that also holds for cluster randomized trials with one period (Moerbeek & Teerenstra, 2016; Raudenbush, 1997). For instance, doubling the number of clusters results in a variance that is half as large. Changing the number of clusters implies rescaling the vertical axis in Fig. 2, while the effects of the other two design factors (m and R) remain unchanged for any given ρ and $1-r$. For that reason, other values of the number of clusters are not considered in this figure or the figures in the next two sections.

Trials with intermittently planned missing observations without dropout

The results in the previous session are based on trials in which measurements can be taken on all seven days of the week. This may indeed be the case in, for example, a trial with patients nested within hospitals where two different methods to relieve stress and anxiety in the emergency room are compared. In other trials, however, the clusters may be health professionals such as psychotherapists or dentists, who most often do not work seven days a week. The aim of this section is to compare planned missing data designs with fewer than seven days of the week to the design with seven days of the week. As we do not consider dropout, the parameter ω should be fixed to zero in the Shiny app for both conditions.

The following five measurement schemes are taken into account:

$$S_1=\{Mo, Tu, We, Th, Fr, Sa, Su)$$

$$S_2=\{Mo, Tu, We, Th, Fr)$$

$$S_3=\{Mo, Tu, Th, Fr)$$

$$S_4=\{Mo, Tu, We, Th)$$

$$S_5=\{Mo, Tu, Th)$$

With scheme S_1, measurements are taken on all seven days, and with scheme S_2, measurements are taken on the five workdays only. With the other three schemes, measurements are taken on either three or four days of the workweek. These schemes are rather typical in the Netherlands, where many elementary schools are closed on Wednesday and/or Friday afternoons and during the two days of the weekend. Parents who have elementary school children and do not work full-time most often have a day off each Wednesday or Friday or even both.

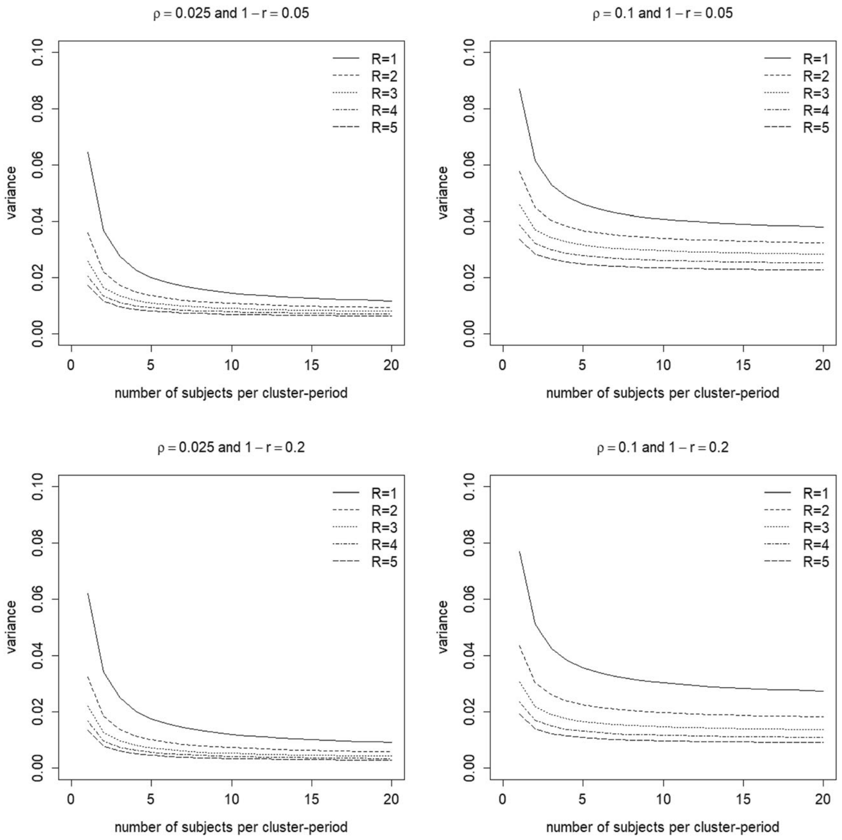

Fig. 2 Variance of the treatment effect estimator as a function of the number of measurements per cluster-period (m, horizontal axis within each graph), number of weeks (R, lines within each graph) and for four combinations of the intraclass correlation coefficient ρ and decay parameter $1-r$ (separate graphs). The number of clusters per condition is $k = 5$

The efficiency of measurement scheme S_s relative to scheme S_t is calculated as

$$RE = \frac{var\left(\widehat{\theta}\right)_{S_t}}{var\left(\widehat{\theta}\right)_{S_s}}, \tag{6}$$

where the numerator and denominator are the variance of the treatment effect estimator obtained with schemes S_t and S_s, and $s, t = 1, 2, 3, 4, 5$, respectively. The inverse of the relative efficiency indicates how often measurement scheme S_s should be replicated to do as well as S_t. In most practical situations, relative efficiencies above 0.8 or 0.9 are favored. $RE = 0.9$ implies that a trial with scheme S_s should include $\left(\frac{1}{0.9}-1\right) \times 100\% = 11\%$ extra clusters to do as well as a trial with scheme S_t; for $RE = 0.8$ an increase of 25% is needed.

The two top panels of Fig. 3 show the efficiency of measurements schemes S_2–S_5 relative to measurement scheme S_1 as a function of the number of subjects per cluster-period m and for two different values of the number of weeks R. This figure holds for $(\rho, 1-r) = (0.025, 0.05)$. As is obvious, lower efficiency is achieved when measurements are taken on fewer days. The relative efficiencies of measurement schemes S_3 and S_4 are almost the same, since both include four days. The lag between the first and the last day is larger for scheme S_3 than for S_4, and hence it has a slightly larger relative efficiency. A similar finding holds for longitudinal intervention studies with repeated measurements within subjects (Moerbeek, 2008): higher efficiency is achieved when study duration increases (while keeping the number of days constant). The loss in efficiency may be considerable: in most scenarios considered in Fig. 3 it is below 0.9, and in extreme cases it may be as low as 0.47.

For any measurement scheme S_2 through S_5, the loss of efficiency is larger for small numbers of subjects per cluster-period m and for small numbers of weeks R. This is unfortunate since, as was shown in Fig. 2, the highest variance is achieved with the lowest number of subjects per cluster-period and lowest number of weeks. In other words, the combination (m, R) with highest variance suffers the most from not measuring on all seven days of the week.

In the two top panels of Fig. 3, the scheme with measurements on all seven days is used as reference for calculating the relative efficiency. It is of course possible to use any other measurement scheme as reference, for instance the scheme with measurements on the five workdays (scheme S_2). The

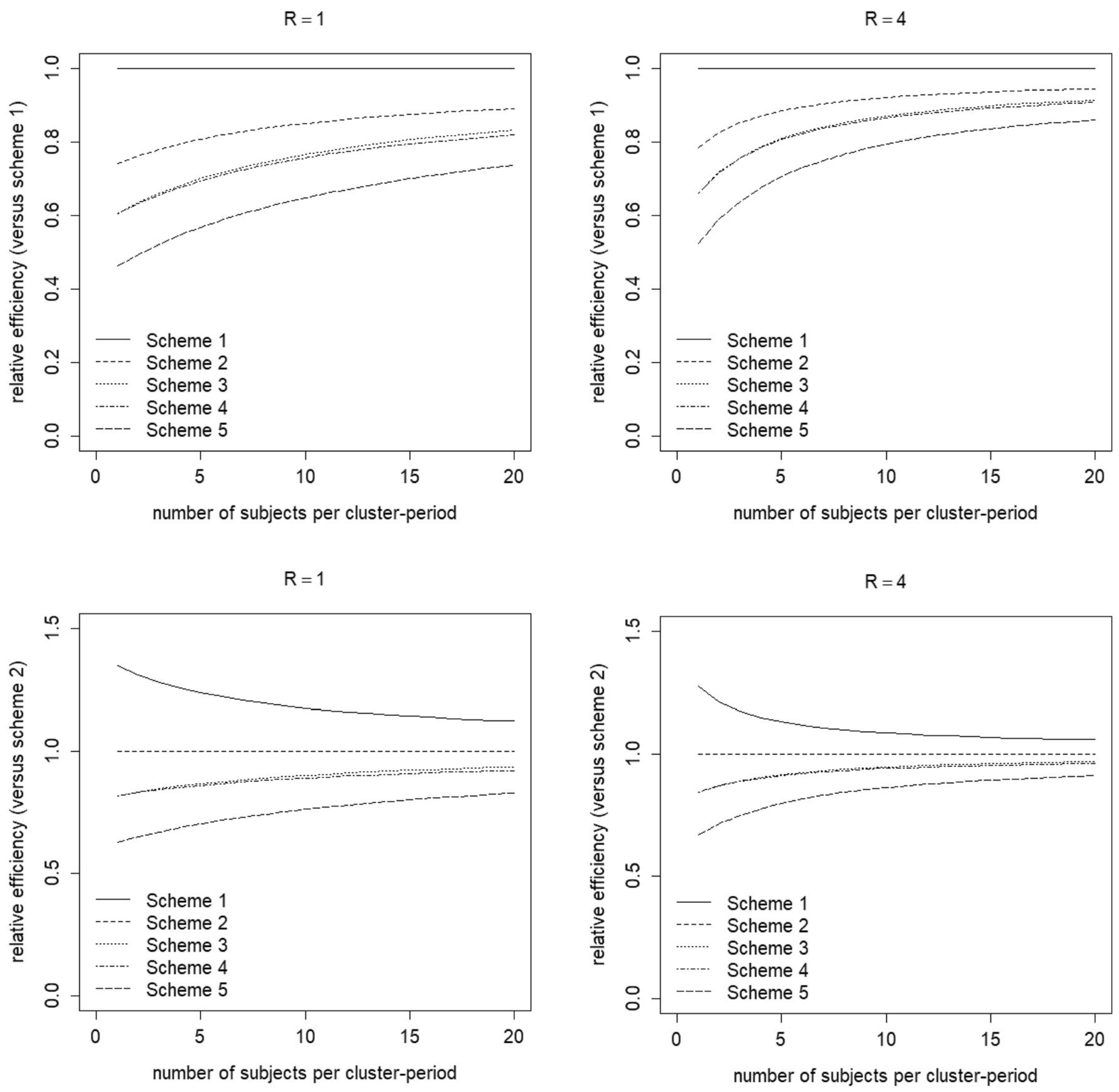

Fig. 3 Efficiency of the treatment effect estimator for various measurement schemes (separate lines) as compared to scheme 1 (top panels) and scheme 2 (bottom panels), as a function of the number of measurements per cluster-period m (horizontal axis) and the number of weeks R (separate panels). Intraclass correlation coefficient $\rho = 0.025$ and decay parameter $1 - r = 0.05$

results of this comparison are presented in the two bottom panels of Fig. 3. The schemes with four days (S_3 and S_4) have relative efficiencies above 0.8, so taking measurements on four rather than five workdays of the week requires a slight increase in the number of clusters. As is obvious, the measurement scheme with just three days performs worst; the relative efficiency can be as low as 0.65.

Figures S3a–S3d in the online supplement present results for four different combinations (ρ, $1 - r$). For a given decay, the loss in efficiency is slightly higher if the intraclass correlation is lower. Furthermore, for a given intraclass correlation, the loss in efficiency is slightly higher if the decay is higher. As we saw in the previous section (Fig. 2), the largest variance of the treatment effect estimator was observed for high intraclass correlation and low decay. Thus the combination (ρ, $1 - r$) with the highest variance suffers the least from measuring on fewer days of the week. The figures in the supplement also show that for the other combinations (ρ, $1 - r$), the relative efficiencies of schemes S_3 and S_4 differ somewhat more than for (ρ, $1 - r$) = (0.025,0.05).

In many cases the health professionals that are included will not work on exactly the same days, meaning that a trial uses a mixture of various measurement schemes: $S_s = \sum_l w_l S_l$, with w_l the weight for scheme S_l such that $0 < w_l < 1$ and $\sum_l w_l = 1$. The variance of the treatment effect estimator is calculated as explained in section 2, and plugged into Eq. (6) to calculate the relative efficiency.

Trials with dropout

Dropout occurs when clusters drop out of the study and do not return in later periods. Dropout may vary across the two experimental conditions. It may be higher in the control condition in the case where clusters in this condition are less motivated because they were not randomized to a new and promising intervention and are therefore less willing to recruit patients for a long amount of time. On the other hand, dropout may be higher in the intervention condition if the intervention puts a large burden on the clusters and subjects or when it has harmful side effects.

It is assumed that dropout can occur on any day, not necessarily at the end of a week. For the sake of simplicity, we assume dropout only occurs at the end of a day, not during a day. This implies that dropout is discrete rather than continuous. The effect of dropout depends on how many clusters drop out and when they do so. We use basic concepts of discrete-time survival analysis to model dropout patterns (Singer & Willett, 1993, 2003). The survival probability function gives a cluster *j*'s probability of staying in the trial up to at least day *t*:

$$S\left(t_{jt}\right) = P\left(T_j \geq t\right). \tag{7}$$

The discrete random variable T_j measures the elapsed study time (i.e. number of days). The associated hazard probability function gives the probability of cluster *j* experiencing the event on day *t*, conditional on not having experienced the event up till then

$$h\left(t_{jt}\right) = P\left(T_j = t \middle| T_j \geq t\right). \tag{8}$$

It is calculated as

$$h\left(t_{jt}\right) = \frac{S\left(t_{j(t-1)}\right) - S\left(t_{jt}\right)}{S\left(t_{j(t-1)}\right)}. \tag{9}$$

To calculate the effect of dropout, the vector $K = (k_1, k_2, \ldots, k_T)'$, with k_h being the number of clusters with exactly *h* measurements, needs to be known in both treatment conditions. This vector is random; the associated probability vector is $p = (p_1, p_2, \ldots, p_T)'$. A cluster's probability p_h of having exactly *h* measurements is calculated from the survival function: $p_h = S(t_h) - S(t_{h+1})$ for $h = 1, \ldots, T-1$ and $p_T = S(t_T)$. This is the probability of dropping out between t_h and t_{h-1}. For each possible vector *K*, a probability can be calculated, and the variance of the treatment effect estimator can be calculated from (5). The expected variance is then the sum of probability × variance over all possible vectors *K*. This procedure is hardly useful in settings where the number of periods, and hence the number of vectors *K*, is large. The variance of the treatment effect can be approximated by using a sampling procedure (Verbeke & Lesaffre, 1999). The vector *K* is sampled a large number of times from the multinomial distribution with probability vector *p*. For each draw, the variance of the treatment effect estimator is calculated and the mean of the variance across all draws is used to calculate the effect of dropout. The drawback of this procedure is that it may be time-consuming, especially if the number of draws is large. A yet further approximation is made by not using a sampling procedure, but replacing the vector *K* by its expectation $\mathrm{E}(K) = k \times p$. This procedure has been compared to the sampling procedure; both produced very similar results (Galbraith & Marschner, 2002).

Many different survival functions exist; we use the Weibull survival function (Galbraith & Marschner, 2002; Moerbeek, 2008), which is given by

$$S\left(t_{jt}\right) = \exp\left(-\lambda t^{\gamma}\right). \tag{10}$$

Time is rescaled by dividing by t_{max}, which is the maximum duration the trial can take. t_{max} may be based on financial or practical considerations, such as the maximal duration most clusters are willing to participate in the trial, but it may also be set by the trial's funding organization. The parameter λ is replaced by $-\log(1-\omega)$, where $\omega \in [0, 1]$ is the proportion of clusters that drop out at some time during the course of a trial with t_{max}. The survival function then becomes

$$S\left(t_{jt}\right) = (1-\omega)^{t^{\gamma}}. \tag{11}$$

Figure 4 shows the survival and hazard probability functions for a trial with a duration of 28 days (i.e. measurement scheme S_1 and $R = 4$ weeks) for various values ω and γ and with equal dropout across the two experimental conditions. When $\gamma < 1$ the hazard probability decreases during the course of the study, when $\gamma = 1$ it is constant and when $\gamma > 1$ it increases. As is obvious, the hazard probability becomes larger and the survival probability becomes smaller when ω becomes larger.

Figure 5 shows the effect of dropout for a trial with measurement scheme S_1 as a function of the number of subjects per cluster-period *m* and for two different numbers of weeks *R*. The efficiency is displayed relative to a trial without dropout ($\omega = 0$). Three values ω and three values γ are considered. Note that ω is the proportion dropout in a trial with a maximal duration $t_{max} = 28$ days (i.e. $R = 4$), as displayed in Fig. 4. The results in Fig. 5 hold for $(\rho, 1-r) = (0.025, 0.05)$.

For each γ, the loss in efficiency is larger when ω is larger. This is obvious, since the greater the number of clusters that drop out, the larger the variance of the treatment effect estimator and hence the lower the efficiency of the trial. Furthermore, for large γ, the effect of ω is smaller than for small γ, especially for small *R*. This finding can be explained as follows. The survival probability function hardly varies across the three values of ω in a trial with just seven days ($R = 1$) and $\gamma = 2$, see the panel at the bottom left in Fig. 4.

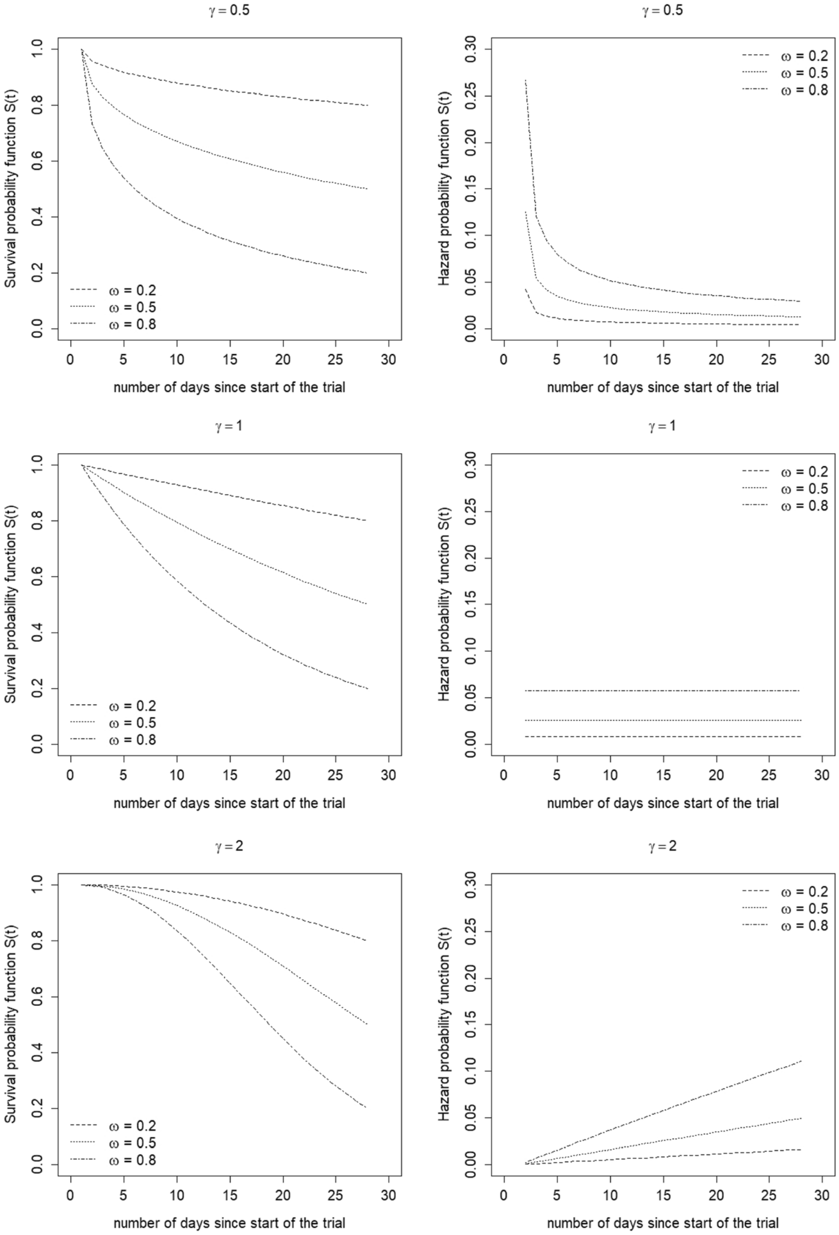

Fig. 4 Survival and hazard probability functions for ω= 0.2, 0.5 and 0.8 (separate lines within each graph) and for γ= 1/2 (top panel), 1 (middle panel) and 2 (bottom panel)

Hence, the relative efficiency hardly depends on the value of ω. The smaller the γ, the more the survival probability functions of the three values ω differ in a trial with seven days, and hence the larger the effect of ω on the relative efficiency. Increasing study duration implies that the survival functions for various values ω differ even more. Hence, the effect of ω on the relative efficiency becomes stronger.

Figure 5 also shows that, for each ω, the efficiency is larger when γ is smaller. This is also obvious, since a smaller γ implies that the hazard probability is largest at the beginning of the trial. In other words, smaller γ implies that there are more clusters with fewer days, so the variance becomes larger and the trial becomes less efficient.

The loss in efficiency is largest with few subjects per cluster-period m. This implies largest loss in efficiency for those trials that already have a large variance. Furthermore, the effect of m becomes smaller for large values of m.

Finally, a larger loss in efficiency is observed for a larger number of weeks R. A larger R implies a longer trial duration and hence a larger amount of dropout during the course of the trial. For instance, consider a trial with 28 days ($R = 4$) where half of the clusters have dropped out by the end of the trial ($\omega = 0.5$).

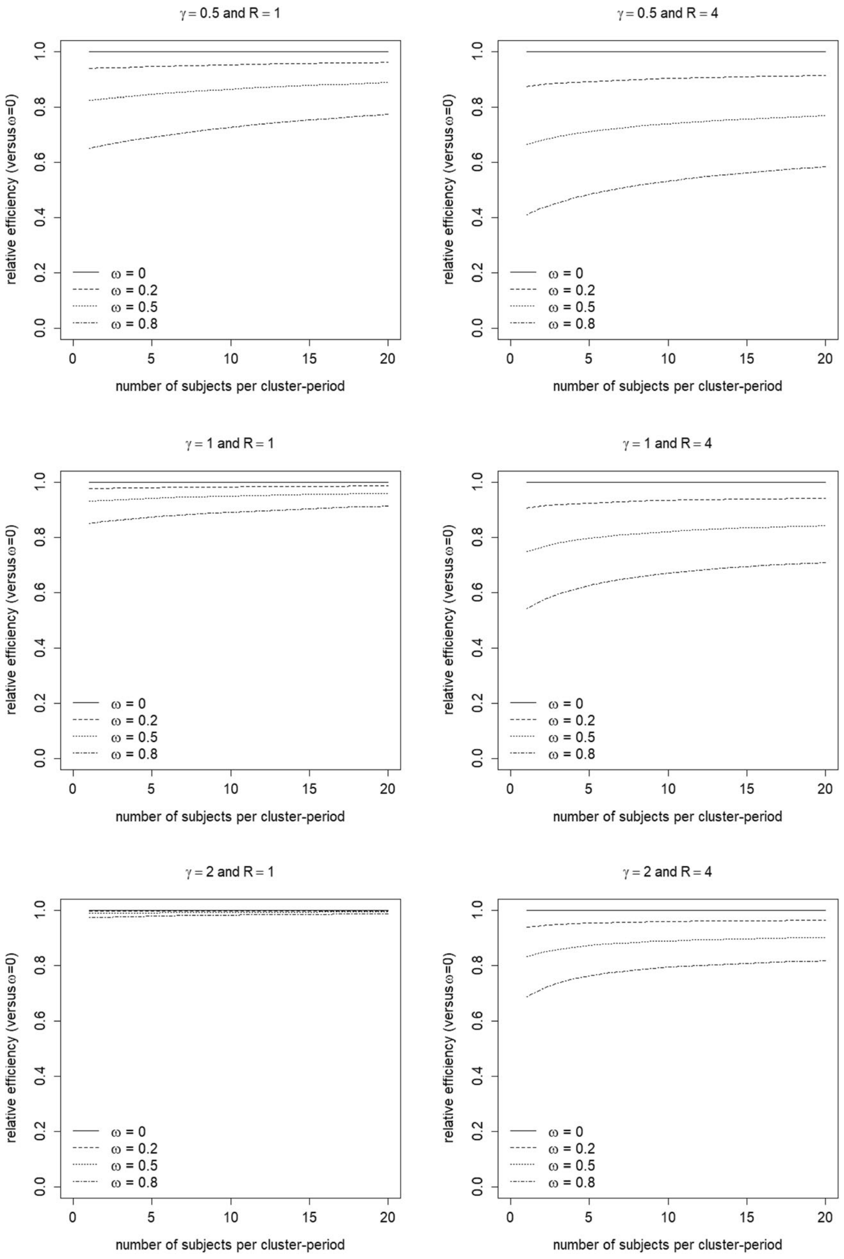

Fig. 5 Efficiency of the treatment effect estimator for various dropout patterns (ω, separate lines and γ, separate panels) as compared to no dropout, as a function of the number of measurements per cluster-period m (horizontal axis) and the number of weeks R (separate panels). Intraclass correlation coefficient $\rho = 0.025$ and decay parameter $1 - r = 0.05$

Reducing study duration to a quarter ($R = 1$) would result in a dropout proportion of 0.28 ($\gamma = 1/2$), 0.14 ($\gamma = 1$) or 0.03 ($\gamma = 2$). So a shorter study duration implies that fewer clusters drop out, hence a smaller loss of efficiency. Therefore, studies that have higher variance of the treatment effect estimator due to a short study duration suffer the least from dropout.

The Online Supplement shows relative efficiencies for four combinations of the intraclass correlation coefficient ρ and decay parameters $1 - r$, see Figs. S5a–S5d. The effects of ρ and $1 - r$ are as in the previous section: for a given decay $1 - r$, the loss in efficiency is slightly larger for a smaller ρ, and for a given ρ the loss in efficiency is slightly larger for a larger $1 - r$. Thus the combination (ρ, $1 - r$) with the highest variance suffers the least from dropout.

Trials with intermittently planned missing observations with dropout

Here we revisit the example in Section 3 to demonstrate the evaluation of designs of trials with intermittently planned missing observations with dropout. The data should become available within at most 8 weeks. It is likely that some dental practices may drop out. Let us suppose that the dropout rate is $\omega = 0.2$ in traditional waiting areas and $\omega = 0.1$ in nouveau waiting areas. Dropout may be expected to increase slightly during the course of the trial as dental practices lose motivation to ask their patients to participate. Such a dropout pattern is achieved by choosing $\gamma > 1$; let us assume that $\gamma = 2$. To calculate the efficiency of the design, a priori estimates of the intraclass correlation coefficient and decay parameter must be specified. Let us assume that $\rho = 0.05$ and $1 - r = 0.05$.

In the Netherlands, dental practices are typically closed during the weekend. We compare four designs with a different number of dental practices per condition and different numbers of weeks. All designs include all five workdays:

$$\begin{aligned} \mathit{Design}\ 1 &: k = 10, R = 4, \mathit{Scheme} \\ &= (\mathit{Mo}, \mathit{Tu}, \mathit{We}, \mathit{Th}, \mathit{Fr}) \\ \mathit{Design}\ 2 &: k = 15, R = 4, \mathit{Scheme} \\ &= (\mathit{Mo}, \mathit{Tu}, \mathit{We}, \mathit{Th}, \mathit{Fr}) \\ \mathit{Design}\ 3 &: k = 10, R = 8, \mathit{Scheme} \\ &= (\mathit{Mo}, \mathit{Tu}, \mathit{We}, \mathit{Th}, \mathit{Fr}) \\ \mathit{Design}\ 4 &: k = 15, R = 8, \mathit{Scheme} \\ &= (\mathit{Mo}, \mathit{Tu}, \mathit{We}, \mathit{Th}, \mathit{Fr}) \end{aligned}$$

The study should be designed so that it has a power of at least 80% in a two-sided test with a type I error rate $\alpha = 0.05$. Leather et al. (2003) found a zero effect for patients' rating on anxiety, and small effects for disability (Cohen's $d = 0.22$) and pain (Cohens $d = 0.27$). In the calculations that follow we assume a small effect size Cohen's $d = 0.2$.

Figure 6 shows the power as a function of the number of patients per day. Obviously, higher power is achieved when more patients are observed per day, but for all designs, the power levels off to a certain limit when the number of patients per day increases. The lowest power is observed for design 1. With this design, sufficient power cannot be achieved even when as many as 20 patients per day are included. Designs 2 and 3 have comparable power levels. Design 3 has slightly higher power when at most four patients per day are included, while design 2 has slightly higher power for at least five patients per day. The number of patients to be included per day to achieve at least 80% power is 9 and 11 for designs 2 and 3, respectively. The largest power is achieved with design 4: only two patients per day are needed to achieve at least 80% power.

Based on some criterion, the best design can be chosen from designs 2-4. If, for example, trial duration is to be minimized, then design 2 is the best choice. If, as another example, recruiting dental practices is difficult, then design 3 should be chosen. If, as yet another example, the total number of patients should be minimized, then design 4 should be chosen.

The impact of using a measurement scheme with fewer days may also be studied. For $\mathit{Scheme} = (\mathit{Mo}, \mathit{Tu}, \mathit{Th}, \mathit{Fr})$, the required number of patients per day is 11 (design 2), 13 (design 3) and 3 (design 4). For $\mathit{Scheme} = (\mathit{Mo}, \mathit{Tu}, \mathit{Th})$, the number of patients per day is 15 (design 2), 18 (design 3) and 3 (design 4). Therefore, including fewer days in the design results in a larger number of subjects to be measured per day to achieve the desired power level.

Conclusions and discussion

Missing data result in a loss of efficiency. In the case of planned missing data designs, measurements are not taken on certain days of the week. The fewer the number of days in which measurements are taken, the lower the efficiency. In addition, lower efficiency is obtained if the lag between the first and last day of the week in which measurements are taken becomes smaller. Furthermore, the loss in efficiency in planned missing data designs is largest for small numbers of subjects per period and small numbers of weeks. In the case of unplanned missing data resulting from dropout, clusters drop out of the trial and do not return in later periods. The higher the amount of dropout, the larger the loss of efficiency, especially if dropout is largest at the beginning of the trial. The loss in efficiency due to dropout is largest for small numbers of subjects per period and large numbers of weeks. In both planned and unplanned missing data designs, the loss of efficiency increases when the size of the intraclass correlation decreases and the decay parameter increases.

The relation between sample size, number of weeks and missing data patterns, on the one side, and variance of the treatment effect estimator, efficiency and power on the other side cannot be captured by a simple mathematical relationship. For this reason, it may be explored by using the Shiny app. Various designs may be specified and the power for each of them can be calculated. Among all designs with sufficient power, the best one may be chosen based on some criterion, such as shortest study duration, smallest number of clusters or smallest number of measurements.

To use the Shiny app, the values of the intraclass correlation coefficient and decay parameter should be specified. The values of such parameters are often unknown in the design phase of a trial. An a priori guess may be based on researchers' expectations, expert opinion or findings in the literature. Over the past two decades, estimates of intraclass correlation coefficients in

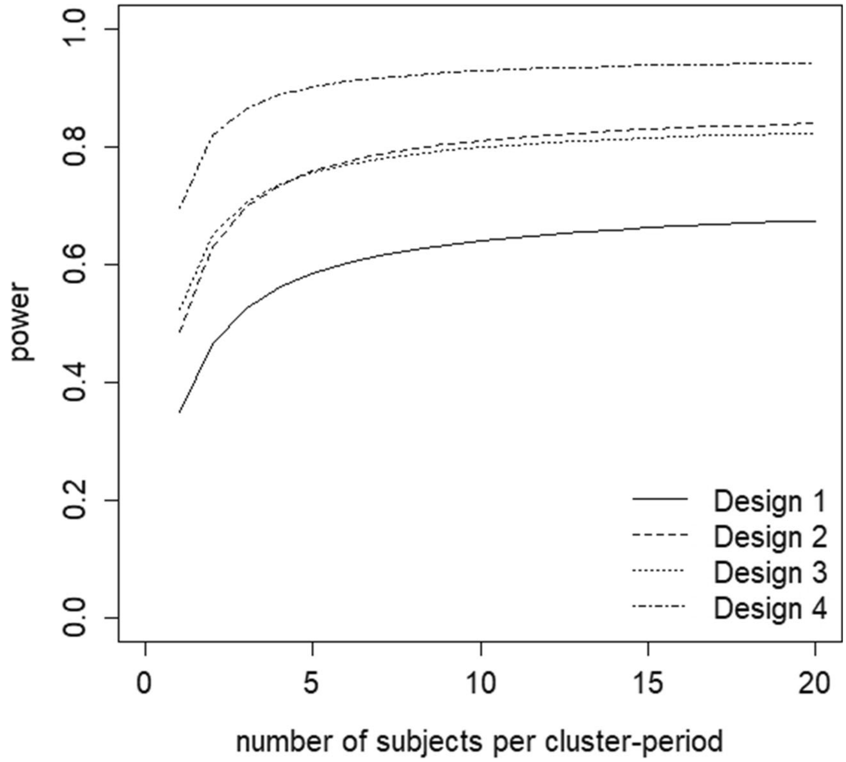

Fig. 6 Power levels as a function of the number of subjects for the four designs in the waiting room example

cluster randomized trials with one period have been published; see Table 11.1 of Moerbeek and Teerenstra (2016) for an overview of such papers. It is important that such estimates also become available for multi-period cluster randomized trials, along with estimates of the decay parameter. Furthermore, it would be very helpful if the dropout that occurred within a trial would be described in detail to facilitate planning future trials. Not only should the total amount of dropout during the course of a trial be given, but the hazard probability function should be given as well. Finally, the size of the effect must be specified. When a priori estimates of Cohen's d, ρ, $1-r$, ω and γ cannot be found in the literature, then the researcher may be able to come up with a range of plausible values of these parameters. Then, the worst-case scenario (i.e. smallest d, largest ρ, smallest $1-r$, largest ω and smallest γ) may be used in the Shiny app.

In this paper the Weibull function was used to model the probability of dropout in each of the time periods. This is a very flexible function, as it allows for constant, increasing or decreasing hazard over time. Of course, there exist many other survival functions, and it would even be possible to let the researcher specify the hazard probability in each of the time periods. However, this is outside the scope of this paper. Furthermore, the probability of dropout was assumed to depend on treatment condition and time elapsed since the start of the study. There may be other factors that influence dropout, for instance the number of days per week in which measurements are taken or the number of subjects who have to be measured per day. It would be interesting to explore more complicated survival and hazard functions in future research.

This paper is restricted to a repeated cross-sectional design, meaning that each subject is measured only once. Higher efficiency is achieved by using a cohort design. For such a design, the statistical model needs to be extended to accommodate repeated measures within subjects. Sample size calculations for such a model are given in Hooper, Teerenstra, De Hoop, and Eldridge (2016), but that study is restricted to a compound symmetry structure and does not take missing data into account. Future research should focus on exponential decay at both the cluster and subject level. It is also possible to include multiple cohorts, and this may be done in a serial or parallel manner. In the first case, cohorts are observed one after the other, while measurements across multiple days are taken within each cohort. In the second case, subjects are measured on a fixed day of the week during the course of multiple weeks. Furthermore, it is also possible to implement a combination of a cohort and repeated cross-sectional design.

The focus of this paper is on a parallel-arm design, meaning that all subjects within the same cluster receive the same treatment. Such a design is often chosen if there is a risk of contamination of the control condition (Moerbeek, 2005). Such contamination is likely to occur in trials where the cluster is a therapist. Although in theory it would be possible to let each therapist offer the intervention and control, in practice it would be difficult for therapists not to let clients in the control condition benefit from the intervention. The parallel-arm design may also be preferred for financial or practical reasons. In the example in the previous section, it would not be cost-effective to redesign the waiting area during the course of the trial in all dental practices. If such objections do not exist, a design in which clusters are exposed to both treatment conditions may be chosen to increase efficiency. With a crossover trial, clusters cross back and forth between the control and intervention conditions. In the example in the previous section, this might have been possible if the trial evaluated the effect of music or aromatherapy. The effect of dropout in a two-period crossover cluster randomized trial has been studied previously (Moerbeek, 2020), and an extension should be made to designs with more than one crossover. Another type of trial in which all clusters receive both treatment conditions is the stepped-wedge design. Here all clusters start in the control, and there is a sequential rollout of the intervention across all clusters. Kasza and Forbes (2019) studied the information content in each of the cluster-periods and found that the most information-rich are those cluster-periods that occur immediately before and after the switches. It would also be interesting to study the effect of dropout in stepped-wedge cluster randomized trials.

In summary, this paper presents new results concerning the effect of dropout on planned missing data design efficiency in repeated cross-sectional multi-period two-arm parallel cluster randomized trials. The Shiny app enables researchers to evaluate their design with respect to power and to compare its efficiency with competing designs.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-020-01529-7.

Open practices statements The Shiny app is available at https://utrecht-university.shinyapps.io/missing_data_CRT/.

The source code is available at https://github.com/MirjamMoerbeek/CRT_missing_data

References

Biddiss, E., Knibbe, T. J., & McPherson, A. (2014). The effectiveness of interventions aimed at reducing anxiety in health care waiting spaces: A systematic review of randomized and nonrandomized trials. *Anesthesia and Analgesia*, *119*(2), 433–448. https://doi.org/10.1213/ANE.0000000000000294

Bloom, H. S. (2005). *Learning more from social experiments. Evolving analytic approaches* (H. S. Bloom, ed.). New York: Russell Sage.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, *29*(1), 30–59. https://doi.org/10.3102/0162373707299550

Campbell, M. J., & Walters, S. J. (2014). *How to design, analyse and report cluster randomised trials in medicine and health related research*. Chichester: Wiley.

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2016). *shiny: Web Application Framework for R*. Retrieved from https://cran.r-project.org/package=shiny. Accessed 17 June 2020

Copas, A. J., & Hooper, R. (2020). Cluster randomised trials with different numbers of measurements at baseline and endline: Sample size and optimal allocation. *Clinical Trials*. https://doi.org/10.1177/1740774519873888

De Hoop, E., Teerenstra, S., Van Gaal, B. G., Moerbeek, M., & Borm, G. (2012). The "best balance" allocation led to optimal balance in cluster-controlled trials. *Journal of Clinical Epidemiology*, *65*(2), 132–137. https://doi.org/10.1016/j.jclinepi.2011.05.006

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London: Edward Arnold.

Eldridge, S., & Kerry, S. (2012). *A practical guide to cluster randomised trials in health services research*. Chichester: Wiley.

Galbraith, S., & Marschner, I. C. (2002). Guidelines for the design of clinical trials with longitudinal outcomes. *Controlled Clinical Trials*, *23*(3), 257–273. https://doi.org/10.1016/s0197-2456(02)00205-2

Genolini, C., Écochard, R., & Jacqmin-Gadda, H. (2013). Copy mean: A new method to impute intermittent missing values in longitudinal studies. *Open Journal of Statistics*, *03*(04), 26–40. https://doi.org/10.4236/ojs.2013.34a004

Giraudeau, B., Ravaud, P., & Donner, A. (2008). Sample size calculation for cluster randomized cross-over trials. *Statistics in Medicine*, *27*(27), 5578–5585.

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). Chichester: Wiley.

Grantham, K. L., Kasza, J., Heritier, S., Hemming, K., & Forbes, A. B. (2019). Accounting for a decaying correlation structure in cluster randomized trials with continuous recruitment. *Statistics in Medicine*, *38*(11), 1918–1934. https://doi.org/10.1002/sim.8089

Hayes, R. J., & Moulton, L. H. (2009). *Cluster randomised trials*. Boca Raton: CRC Press.

Hedeker, D., Gibbons, R. D., & Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition: Comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics*, *24*(1), 70–93. https://doi.org/10.3102/10769986024001070

Hemming, K., Girling, A. J., Sitch, A. J., Marsh, J., & Lilford, R. J. (2011). Sample size calculations for cluster randomisation controlled trials with a fixed number of clusters. *BMC Medical Research Methodology*, *11*, 102. https://doi.org/10.1186/1471-2288-11-102

Heo, M. (2014). Impact of subject attrition on sample size determinations for longitudinal cluster randomized trials. *Journal of Biopharmaceutical Statistics*, *24*, 507–522. https://doi.org/10.1080/10543406.2014.888442

Hooper, R., & Bourke, L. (2015). Cluster randomised trials with repeated cross sections: Alternatives to parallel group designs. *British Medical Journal*, *350*. https://doi.org/10.1136/bmj.h2925

Hooper, R., Teerenstra, S., De Hoop, E., & Eldridge, S. (2016). Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Statistics in Medicine*, *35*(26), 4718–4728. https://doi.org/10.1002/sim.7028

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2018). *Multilevel analysis. Techniques and applications*. New York: Routledge.

Kasza, J., & Forbes, A. B. (2019). Information content of cluster–period cells in stepped wedge trials. *Biometrics*, *75*(1), 144–152. https://doi.org/10.1111/biom.12959

Kasza, J., Hemming, K., Hooper, R., Matthews, J. N. S., & Forbes, A. B. (2019). Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. *Statistical Methods in Medical Research*, *28*(3), 703–716. https://doi.org/10.1177/0962280217734981

Konstantopoulos, S. (2012). The impact of covariates on statistical power in cluster randomized designs: Which level matters more? *Multivariate Behavioral Research*, *47*(3), 392–420. https://doi.org/10.1080/00273171.2012.673898

Leather, P., Beale, D., Santos, A., Watts, J., & Lee, L. (2003). Outcomes of environmental appraisal of different hospital waiting areas. *Environment and Behavior*, *35*(6), 842–869. https://doi.org/10.1177/0013916503254777

Liu, X. (2016). *Methods and applications of longitudinal data analysis*. London: Academic Press.

Moerbeek, M. (2005). Randomization of clusters versus randomization of persons within clusters: Which is preferable? *The American Statistician*, *59*(1), 72–78. https://doi.org/10.1198/000313005X20727

Moerbeek, M. (2006). Power and money in cluster randomized trials: When is it worth measuring a covariate? *Statistics in Medicine*, *25*(15), 2607–2617. https://doi.org/10.1002/sim.2297

Moerbeek, M. (2008). Powerful and cost-efficient designs for longitudinal intervention studies with two treatment groups. *Journal of Educational and Behavioral Statistics*, *33*(1), 41–61. https://doi.org/10.3102/1076998607302630

Moerbeek, M. (2020). The cluster randomized crossover trial: The effects of attrition in the AB/BA design and how to account for it in sample size calculations. *Clinical Trials*. https://doi.org/10.1177/1740774520913042

Moerbeek, M., & Teerenstra, T. (2016). *Power analysis of trials with multilevel data*. Boca Raton: CRC Press.

Molenberghs, G., & Verbeke, G. (2001). A review on linear mixed models for longitudinal data, possibly subject to dropout. *Statistical Modelling*, *1*(4), 235–269. https://doi.org/10.1177/1471082X0100100402

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.

Murray, D. M. (2001). Statistical models appropriate for designs often used in group-randomized trials. *Statistics in Medicine*, *20*(9–10), 1373–1385. https://doi.org/10.1002/sim.675

Murray, D. M., & Blitstein, J. L. (2003). Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review*, *27*(1), 79–103. https://doi.org/10.1177/0193841X02239019

Murray, D. M., Hannan, P. J., Wolfinger, R. D., Baker, W., & Dwyer, J. H. (1998). Analysis of data from group-randomized trials with repeat observations on the same groups. *Statistics in Medicine*, *17*(14), 1581–1600. https://doi.org/10.1002/(sici)1097-0258(19980730)17:14<1581::aid-sim864>3.0.co;2-n

Murray, D. M., Van Horn, M. L., Hawkins, J. D., & Arthur, M. W. (2006). Analysis strategies for a community trial to reduce adolescent ATOD use: A comparison of random coefficient and ANOVA/ANCOVA models. *Contemporary Clinical Trials*, *27*(2), 188–206. https://doi.org/10.1016/j.cct.2005.09.008

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized studies. *Psychological Methods*, *2*(2), 173–185. https://doi.org/10.1037/1082-989X.2.2.173

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods*. Thousand Oaks: Sage Publications.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, *29*(1), 5–29. https://doi.org/10.3102/0162373707299460

Rhemtulla, M., Jia, F., & Little, T. D. (2014). Planned missing designs to optimize the efficiency of latent growth parameter estimates. *International Journal of Behavioral Development*, *38*(5), 423–434. https://doi.org/10.1177/0165025413514324

Rietbergen, C., & Moerbeek, M. (2011). The design of cluster randomized crossover trials. *Journal of Educational and Behavioral Statistics*, *36*(4), 472–490.

Roy, A., Bhaumik, D. K., Aryal, S., & Gibbons, R. D. (2007). Sample size determination for hierarchical longitudinal designs with differential attrition rates. *Biometrics*, *63*(3), 699–707. https://doi.org/10.1111/j.1541-0420.2007.00769.x

Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational and Behavioral Statistics*, *18*(2), 155–195. https://doi.org/10.2307/1165085

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: {Modeling} change and event occurrence*. Oxford: Oxford University Press.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modelling*. London: Sage.

Vallejo, G., Ato, M., Fernández, M. P., & Livacic-Rojas, P. E. (2019). Sample size estimation for heterogeneous growth curve models with attrition. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1059-y

Verbeke, G., & Lesaffre, E. (1999). The effect of dropout on the efficiency of longitudinal experiments. *Journal of the Royal Statistical Society, Series C*, *48*(3), 363–375. https://doi.org/10.1111/1467-9876.00158

Wu, W., Jia, F., Rhemtulla, M., & Little, T. D. (2016). Search for efficient complete and planned missing data designs for analysis of change. *Behavior Research Methods*, *48*(3), 1047–1061. https://doi.org/10.3758/s13428-015-0629-5

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.