Research report

Daphne Weemering

1/4/2022

Introduction

Randomized controlled trials (RCTs) are considered the gold standard in determining treatment efficacy. At first glance, these standard RCTs seem to earn their position. However, a drawback is that they require a relatively large sample size to establish the treatment effect with sufficient power. The larger the sample size, the more patients are potentially exposed to inferior treatment, making it desirable to limit this number. Additionally, standard RCTs are not feasible in populations of patients with rare diseases, since these populations only have a limited number of patients that could enter such a study.

A clinical trial methodology that reduces the number of patients necessary to find a treatment effect with sufficient power, is the N-of-1 trial. The N-of-1 trial is a randomized controlled multiple crossover trial where a single patient repeatedly receives both the treatment and control intervention in several cycles, where the order of the treatment is decided at random.[1] The clinical conditions under which a N-of-1 trial is useful, is when the disease under study is long-term and stable, avoiding the treatment differences to be obscured within- and between cycles due to disease progression. Also, the intervention should not modify the course of the disease and should have a rapid on- and offset of biological action of the medication.[3]

The result at the end of a N-of-1 trial is an estimate of the treatment effect for that one patient. However, when several N-of-1 trials are combined, the aggregation of the results allows for the estimation of the population treatment effect.[4] In this aggregated estimate, both the magnitude of the treatment effectiveness as well as the heterogeneity in response to the treatment, are taken into account.[4] These series of N-of-1 trials require a relatively smaller sample size because each subject serves as its own control, and because multiple observations per individual are obtained. Several recent articles[7] (with many more examples available) have shown the aggregation of separate N-of-1 trials to obtain a population treatment effect.

Sample size calculations constitute an important part of the design phase of any clinical trial. A priori sample size calculations are necessary to avoid under- or overpowering a study. For these calculations assumptions have to be made with regard to unknown parameters, such as the various nuisance parameters (i.e., variance components) and the average treatment effect in the population.[8] Overpowering a study potentially exposes too many patients to an inferior treatment and it is a waste of resources, whereas underpowering a study increases the risk of not finding an effect when there actually exists one in the population.

Sample size formulas have been derived for both random and fixed effects models.[9] Since the main objective of aggregating the results of N-of-1 trials is to make inferences to the population, random effects models are of main interest here. In series of N-of-1 trials, the between- and within-patient heterogeneity of the treatment effect are unknown at the start of the studies.[9] Taking estimates of nuisance parameters from other studies to calculate the required sample size can be unreliable due to differences in the study population or differences in study design.[10] Furthermore, the between-patient heterogeneity in treatment response is often not available because the kind of study to obtain these estimates is a trial incorporating such a component, such as a series of N-of-1 trials.[11] N-of-1 trials are not that common, and even if similar N-of-1 trials exist, these kinds of estimates are usually not reported.

To conquer the problem of incorrect assumptions for unknown parameters in sample size calculations, interim sample size reestimation can be considered. Interim sample size reestimation is a type of adaptive trial in which the data that is obtained after observing a prespecified number of patients, is used to estimate the unknown parameters. Those estimates measured at interim are used to recalculate the sample size. Then, the sample size can be adjusted accordingly in order to avoid having an under- or overpowered study.[12] A distinction can be made between interim sample size reestimation based on nuisance parameter estimates and based on treatment effect estimates.[8] This thesis will not cover the latter approach, but rather focuses on interim sample size reestimation based on nuisance parameter estimates.

The application of interim sample size reestimation has not yet been investigated in the context of N-of-1 trials and no specific methods and guidelines have been established. Also, it is yet unclear what the minimally required sample size is to carry out sample size reestimation in N-of-1 trials. With the use of simulation studies, this thesis aims at establishing guidelines for the minimally required number of patients for sample size reestimation, and to compare series of N-of-1 trials incorporating interim sample size reestimation with series of N-of-1 trials with a fixed sample size. Power and the expected sample size are important measures of

performance herein.

The remainder of this report will be structured as follows. In section 2, the specific methods will be discussed, including the population model and sample size calculations for N-of-1 trials. In section 3, the simulation studies will be explained. Section 4 will elaborate on the results.

Methodology

Model and notations

Series of N-of-1 trials consist of multiple individuals with each individual having several measurements. Throughout this report, it is assumed that individuals receive each of treatment A and B once within each cycle. The order of receiving the treatments will be determined at random. It is also assumed that the outcome concerns a continuous measurement, that the disease under study is relatively stable over time, and that carryover effects are accounted for.

Previous research has shown that linear mixed models provide robust inferences for aggregated N-of-1 trials data.[2] More specifically, Araujo et al.[2] proposed the following model based on within-cycle treatment differences (see also Senn):[9]

$$d_{ij} = \tau_i + \epsilon_{ij}, i = 1, ..., n, j = 1, ...k$$
$$\tau_i \sim N(T, \psi^2)$$
$$\epsilon_{ij} \sim N(0, 2\sigma^2)$$

Let n denote the patient and k the number of cycles within each patient. Then i indexes the patient and j the cycle. d_{ij} is the observed treatment difference for patient i in cycle j. $\tau_1, ..., \tau_n$ are the random treatment effects with an assumed mean T and variance ψ^2. ϵ_{ij} are random within-patient within-cycle disturbance terms.

Sample size calculations in N-of-1 trials

Following the model given in equation equation 1, the average treatment effect in the population and the corresponding variance could be derived.[9] Senn[9] has made use of the fact that the summary measures approach offers the same results as linear mixed models for balanced data.[14] Therefore, the trial data is reduced to an average treatment effect for each patient:

$$\bar{d}_{i.} = \sum_{j=1}^{k} d_{ij}/k$$

If these estimates are averaged over all patients, the within-cycle treatment differences can be tested. Then, the variance of this average treatment effect estimate ($\hat{T}$) is:

$$\frac{\psi^2 + 2\sigma^2/k}{n}$$

The variance at the patient level, $\psi^2 + 2\sigma^2/k$, is an efficient estimate of the variance of the n observed summary measures and has $(n-1)$ degrees of freedom.[9] The sample size under hypothesized values of ψ^2, σ^2 and τ can be calculated using standard approaches for the one-sample t-test as can be specified with the `pwr` function from the `r` package[15] in `r`.[16]

Simulations

Simulation studies are used to answer the research questions. Special guidelines have been established for planning and reporting simulation studies based on the so called "ADEMP" structure (Aim, Data-generating

mechanisms, Estimands, Methods and Performance measures).[17] This section is structured following this structure and explains how the simulation studies are constructed.

Aim

The aim of the simulation studies is to (1) obtain the minimally required number of patients to reliably (i.e. in terms of statistical power) reestimate the sample size, and (2) compare series of N-of-1 trials with sample size reestimation to series of N-of-1 trials with a fixed sample size (fixed number of patients and cycles within patients).

Data-generating mechanism

The linear mixed model provided in equation equation 1 will be used to simulate data for a series of N-of-1 trials. The unknown parameters in this model are the variance of the treatment effect (ψ^2), the within-patient within-cycle variance (σ^2) and the average treatment effect (T). See table 1 for an overview of the parameter settings in the simulations. The initial sample size will be calculated based on a (potentially incorrect) scenario, and the data is simulated under the true scenario.

Some parameter values are varied and some are held constant, resulting in $3*3*3*3*3 = 243$ different scenarios: there are three different parameter settings for σ^2 and ψ^2 separately for calculating the initial sample size, three different fractions of the initial sample size (f) on which reestimation will be applied to, and three different parameter settings for σ^2 and ψ^2 separately for calculating the true sample size.

Table 1: Parameters in the simulation study

Description	Constant parameters	Varied parameters
Design parameters		
Fraction of initial sample size		$f = 0.25, 0.5, 0.75$
Cycles per patient	$k = 3$	
Power	$1 - \beta = 0.8$	
Two-sided significance level	$\alpha = 0.05$	
Clinically relevant difference	$\Delta = 1$	
Model parameters		
Within-patient within-cycle variance		$\sigma^2 = 0.25, 0.5, 1$
Variance of treatment effect		$\psi^2 = 0.5, 1, 2$
Average treatment effect	$T = \Delta = 1$	
Simulation parameter		
Number of simulations	$N = 1000$	
Outcome parameters		
Statistical power	Compare power of trial with sample size reestimation in simulation to target of 80%	
Sample size	Compare sample size under reestimation with sample size under true parameter values	

Estimands

The main estimand of the simulation studies will be the average treatment effect in the population. Statistical power and the expected sample size will be used to test the performance of the model under different parameter settings. The power will be measured as the fraction of iterations where the treatment effect is significant ($p < .05$). The expected sample size is calculated as the average sample size under reestimation, which will be compared with the sample size of the true model.

Methods

Linear mixed models will be fitted to the simulated data using the `lme4` package.[18] Restricted maximum likelihood will be applied, since this method, to the contrary of standard maximum likelihood, produces

unbiased estimates for the random effects.[19]

The initial sample size (n) will be calculated as the sample size required to achieve 80% power under the hypothesized parameter values of ψ^2, σ^2, the clinically relevant difference (Δ) and a two-sided significance level of 5%. Sample size calculations will be done using the `pwr` package.[15] Equation 3 will be used to calculate the variance used for the standardized effect size needed for the `pwr` function. The sample size will be reevaluated after *fn* subjects (where *f* is the fraction of the initial sample size) using the observed estimates for ψ^2 and σ^2. The sample size of the model under reestimation will then be compared to the sample size of the model with the true parameter values. Then, complete trials including interim reestimation will be simulated and these results will be compared with N-of-1 trials with a fixed sample size.

Performance measures

The performance of sample size reestimation will be tested by means of statistical power and the expected sample size. The reliability of the approaches under different scenarios will be assessed by comparing the power of trials observed in the simulation to the target of 80 percent. The sample size under reestimation (and under fixed sample size) will be compared with the sample size based on the true parameter values. If reestimation works properly, the reestimated sample size should be close to the sample size under the true model.

Results

For each model scenario, 1000 linear mixed models were fitted and used to recalculate the sample size based on interim values of ψ^2 and σ^2. Hence, for each of the model scenarios, 1000 final sample sizes were calculated and the average and standard deviation of these 1000 values are summarized in table 2, together with the sample size under the true parameter settings for ψ^2 and σ^2. With this information, a preliminary answer can be given to the first research question regarding the minimally required sample size for reliable sample size reestimation. Notice that ψ_h^2 and σ_h^2 indicate hypothesized values, and ψ_t^2 and σ_t^2 indicate true values.

From table 2 it becomes clear that sample size reestimation works quite well on average. The largest discrepancy between true and mean reestimated sample sizes is maximally two patients (too many), which in practice would not immediately imply that the study becomes overly powered. Notice, however, that the variability is quite big for some scenarios. For some cases the standard deviation even exceeds the mean reestimated sample size. The large variability in reestimated sample sizes mainly occurs when ψ_h^2 and σ_h^2 are small and only 25 percent of the initial sample size is used to apply reestimation to. Specifically, when $\psi_h^2 = 0.5$ and $\sigma_h^2 = 0.25$ are taken for the calculation of the initial sample size, the variability in the reestimated sample size is highest compared to all other combinations of ψ_h^2 and σ_h^2.

In figure 1, the influence of different fractions of the initial sample to base sample size reestimation on and the influence of different hypothesized values for ψ^2 and σ^2 is shown in an example scenario. First of all, figure 1 shows that the variability is higher when *f* is small and when ψ_h^2 and σ_h^2 are relatively small. The larger the fraction of the initial sample size to base reestimation on, the smaller the variability in the mean reestimated sample size. The means of the reestimated sample sizes (indicated with the black dots) under all scenarios are very close to the true sample size (indicated with the red horizontal line). So even under great variability, the mean is close to the true value.

From these aforementioned points, the following preliminary answer to the research question could be given. Since there is still a lot of variability in reestimated sample sizes when the fraction of the initial sample size to base reestimation on is 0.25 or 0.5, it is desirable to take 0.75 of the initial sample size to base reestimation on. However, this comes with a risk, because if the initial sample size is estimated to large compared to the true sample size, taking 75% of the initial sample size might cause you to exceed the reestimated required sample size at interim already. Especially when ψ^2 and σ^2 are hypothesized to be relatively large, the risk of this occurring becomes higher. Using 50% of the initial sample size to base reestimation on when hypothesized values are large can avoid the study to become excessively overpowered, and would be recommended in that scenario. When hypothesized values are medium to small ($\psi_h^2 = 0.5$ to 1 and $\sigma_h^2 = 0.25$ to 0.5) it would be recommended to use 75% of the initial sample size to base reestimation on.

In the further course of this project entire trials including sample size reestimation will be run, and the power of these complete trials will be compared with the target of 80%. The first research question regarding the minimally required sample size reestimation cannot be completely answered with the information acquired up until this point, because it might be that the overall power is close to or at the target, even when the variability in reestimated sample sizes is quite large. Therefore, no firm conclusions can be drawn yet.

Table 2: True sample sizes and mean reestimated sample sizes (with standard deviation in brackets) under varying fractions of the initial sample size (f) and varying hypothesized values for ψ^2 and σ^2.

f	$\psi_t^2 = 0.5; \sigma_t^2 = 0.25$			$\psi_t^2 = 0.5; \sigma_t^2 = 0.5$			$\psi_t^2 = 0.5; \sigma_t^2 = 1$		
	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
True sample size		7.38			8.65			11.23	
$\psi_h^2 = 0.5; \sigma_h^2 = 0.25$	8.24 (7.27)	8.03 (4.26)	7.98 (3.33)	10.06 (8.91)	9.52 (5.17)	9.37 (4.07)	13.74 (12.22)	12.68 (7.04)	12.33 (5.51)
$\psi_h^2 = 0.5; \sigma_h^2 = 0.5$	8.23 (5.06)	8.00 (3.65)	7.89 (2.90)	9.79 (5.94)	9.46 (4.50)	9.23 (3.52)	13.01 (8.06)	12.55 (6.15)	12.11 (4.77)
$\psi_h^2 = 0.5; \sigma_h^2 = 1$	8.23 (5.06)	7.80 (3.33)	7.81 (2.51)	9.79 (5.94)	9.37 (4.07)	9.13 (3.12)	13.01 (8.06)	12.33 (5.51)	11.92 (4.35)
$\psi_h^2 = 1; \sigma_h^2 = 0.25$	8.25 (4.93)	7.80 (3.33)	7.81 (2.51)	9.79 (5.94)	9.37 (4.07)	9.13 (3.12)	13.01 (8.06)	12.33 (5.51)	11.92 (4.35)
$\psi_h^2 = 1; \sigma_h^2 = 0.5$	8.03 (4.26)	7.89 (2.90)	7.86 (2.43)	9.52 (5.17)	9.23 (3.52)	9.19 (2.96)	12.68 (7.04)	12.11 (4.77)	11.95 (4.00)
$\psi_h^2 = 1; \sigma_h^2 = 1$	8.03 (4.26)	7.91 (2.71)	7.89 (2.43)	9.52 (5.17)	9.21 (3.32)	9.27 (2.78)	12.68 (7.04)	12.03 (4.49)	12.01 (3.83)
$\psi_h^2 = 2; \sigma_h^2 = 0.25$	8.00 (3.65)	7.91 (2.71)	7.94 (1.97)	9.46 (4.50)	9.19 (2.96)	9.28 (2.46)	12.55 (6.15)	11.95 (4.00)	11.97 (3.38)
$\psi_h^2 = 2; \sigma_h^2 = 0.5$	7.98 (3.33)	7.86 (2.25)	7.91 (1.87)	9.37 (4.07)	9.17 (2.81)	9.20 (2.34)	12.33 (5.51)	11.91 (3.89)	11.82 (3.23)
$\psi_h^2 = 2; \sigma_h^2 = 1$	7.80 (3.33)	7.96 (2.26)	8.02 (1.82)	9.37 (4.07)	9.27 (2.78)	9.33 (2.24)	12.33 (5.51)	12.01 (3.83)	12.00 (3.05)

f	$\psi_t^2 = 1; \sigma_t^2 = 0.25$			$\psi_t^2 = 1; \sigma_t^2 = 0.5$			$\psi_t^2 = 1; \sigma_t^2 = 1$		
	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
True sample size		11.23			12.52			15.11	
$\psi_h^2 = 0.5; \sigma_h^2 = 0.25$	12.06 (12.98)	11.87 (7.61)	11.85 (5.95)	13.81 (14.63)	13.31 (8.57)	13.19 (6.73)	17.39 (17.87)	16.32 (10.40)	16.06 (8.21)
$\psi_h^2 = 0.5; \sigma_h^2 = 0.5$	12.31 (9.00)	11.80 (6.53)	11.73 (5.22)	13.76 (9.96)	13.25 (7.38)	13.03 (5.85)	16.85 (11.97)	16.22 (9.04)	15.79 (7.09)
$\psi_h^2 = 0.5; \sigma_h^2 = 1$	12.31 (9.00)	11.85 (5.95)	11.55 (4.42)	13.76 (9.96)	13.19 (6.73)	12.85 (5.05)	16.85 (11.97)	16.06 (8.21)	15.58 (6.30)
$\psi_h^2 = 1; \sigma_h^2 = 0.25$	12.31 (9.00)	11.85 (5.95)	11.55 (4.42)	13.76 (9.96)	13.19 (6.73)	12.85 (5.05)	16.85 (11.97)	16.06 (8.21)	15.58 (6.30)
$\psi_h^2 = 1; \sigma_h^2 = 0.5$	11.87 (7.61)	11.73 (5.22)	11.73 (4.33)	13.31 (8.57)	13.03 (5.85)	13.04 (4.87)	16.32 (10.40)	15.79 (7.09)	15.69 (5.96)
$\psi_h^2 = 1; \sigma_h^2 = 1$	11.87 (7.61)	11.79 (4.84)	11.88 (3.99)	13.07 (8.57)	13.07 (5.44)	13.19 (4.55)	16.32 (10.40)	15.74 (6.66)	15.83 (5.63)
$\psi_h^2 = 2; \sigma_h^2 = 0.25$	11.80 (6.53)	11.73 (4.33)	11.83 (3.44)	13.25 (7.38)	13.04 (4.87)	13.17 (3.98)	16.22 (9.04)	15.69 (5.96)	15.83 (4.97)
$\psi_h^2 = 2; \sigma_h^2 = 0.5$	11.85 (5.95)	11.69 (3.96)	11.82 (3.28)	13.19 (6.73)	12.98 (4.53)	13.10 (3.76)	16.06 (8.21)	15.61 (5.64)	15.69 (4.71)
$\psi_h^2 = 2; \sigma_h^2 = 1$	11.85 (5.95)	11.88 (3.98)	11.93 (3.19)	13.19 (6.73)	13.19 (4.55)	13.28 (3.63)	16.06 (8.21)	15.83 (5.63)	15.96 (4.49)

| f	$\psi_t^2 = 2; \sigma_t^2 = 0.25$			$\psi_t^2 = 2; \sigma_t^2 = 0.5$			$\psi_t^2 = 2; \sigma_t^2 = 1$		
	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
True sample size		19.02			20.32			22.93	
$\psi_h^2 = 0.5; \sigma_h^2 = 0.25$	19.73 (24.35)	19.73 (14.28)	19.68 (11.21)	21.41 (26.05)	21.09 (15.28)	21.00 (12.01)	24.91 (29.35)	23.93 (17.21)	23.75 (13.54)
$\psi_h^2 = 0.5; \sigma_h^2 = 0.5$	20.55 (17.25)	19.55 (12.22)	19.50 (9.93)	21.94 (18.12)	20.95 (13.11)	20.79 (10.52)	24.87 (20.02)	23.82 (14.84)	23.42 (11.77)
$\psi_h^2 = 0.5; \sigma_h^2 = 1$	20.55 (17.25)	19.68 (11.21)	19.16 (8.26)	21.94 (18.12)	21.00 (12.01)	20.44 (8.88)	24.87 (20.02)	23.75 (13.54)	23.06 (10.16)
$\psi_h^2 = 1; \sigma_h^2 = 0.25$	20.55 (17.25)	19.68 (11.21)	19.16 (8.26)	21.94 (18.12)	21.00 (12.01)	20.44 (8.88)	24.87 (20.02)	23.75 (13.54)	23.06 (10.16)
$\psi_h^2 = 1; \sigma_h^2 = 0.5$	19.73 (14.28)	19.50 (9.93)	19.51 (9.13)	21.10 (15.28)	20.79 (10.52)	20.81 (8.70)	23.93 (17.21)	23.42 (11.77)	23.41 (9.82)
$\psi_h^2 = 1; \sigma_h^2 = 1$	19.73 (14.28)	19.69 (9.13)	19.82 (7.48)	21.09 (15.28)	20.95 (9.75)	21.14 (8.02)	23.93 (17.21)	23.52 (10.99)	23.77 (9.14)
$\psi_h^2 = 2; \sigma_h^2 = 0.25$	19.55 (12.22)	19.51 (8.13)	19.66 (6.39)	20.95 (13.11)	20.81 (8.70)	20.99 (6.92)	23.82 (14.84)	23.41 (9.82)	23.70 (7.99)
$\psi_h^2 = 2; \sigma_h^2 = 0.5$	19.68 (11.21)	19.44 (7.42)	19.70 (6.12)	21.00 (12.01)	20.70 (7.97)	21.00 (6.60)	23.75 (13.54)	23.30 (9.11)	23.57 (7.56)
$\psi_h^2 = 2; \sigma_h^2 = 1$	19.68 (11.21)	19.82 (7.48)	19.84 (5.96)	21.00 (12.01)	21.14 (8.02)	21.23 (6.40)	23.75 (13.54)	23.77 (9.14)	23.92 (7.28)
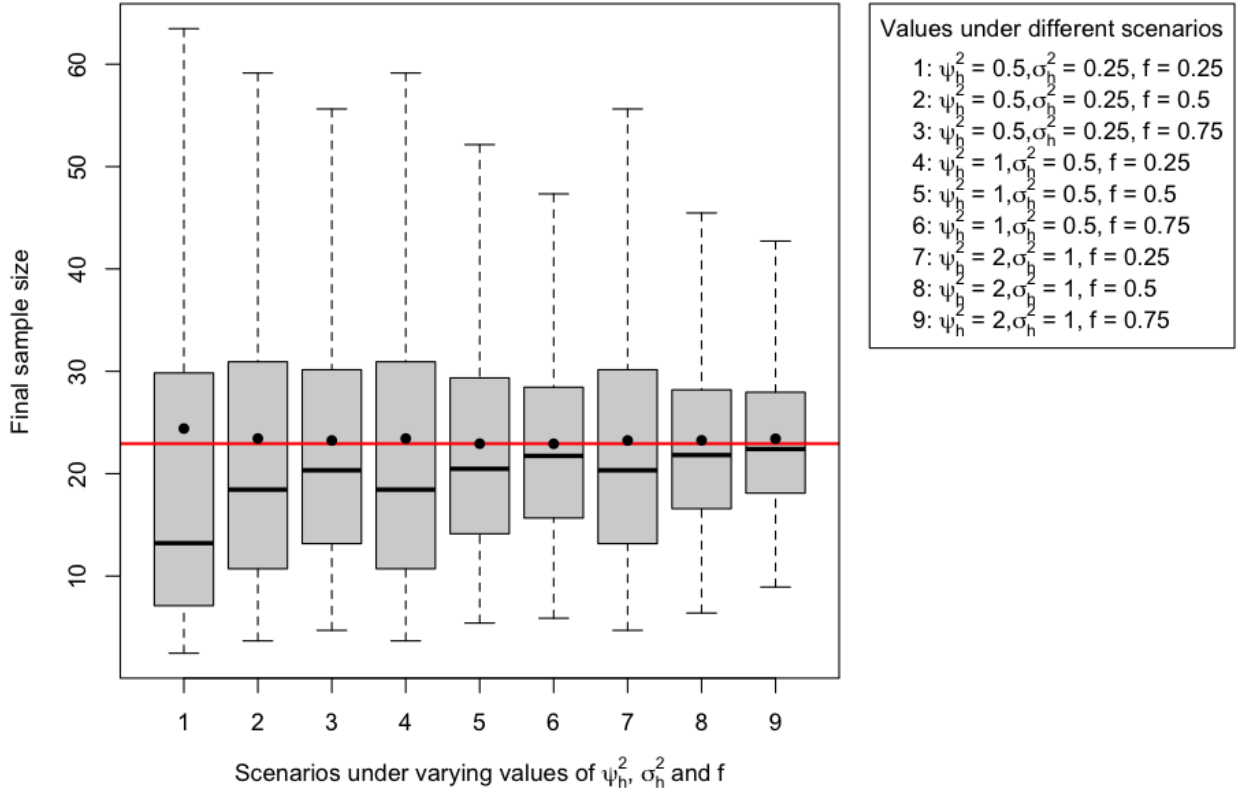
Figure 1: Mean reestimated sample sizes (mean indicated with black dot) and the variability under different ψ_h^2, σ_h^2 and f compared to true model with $\psi_t^2 = 2$ and $\sigma_t^2 = 1$.

References

1. Guyatt G, Sackett D, Taylor DW, Ghong J, Roberts R, Pugsley S. Determining optimal therapy—randomized trials in individual patients. *New England Journal of Medicine.* 1986;314(14):889-892.

2. Araujo A, Julious S, Senn S. Understanding variation in sets of n-of-1 trials. *PloS one.* 2016;11(12):e0167167.

3. Nikles J, Mitchell GK, Schluter P, et al. Aggregating single patient (n-of-1) trials in populations where recruitment and retention was difficult: The case of palliative care. *Journal of clinical epidemiology.* 2011;64(5):471-480.

4. Zucker D, Schmid C, McIntosh M, D'agostino R, Selker H, Lau J. Combining single patient (n-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. *Journal of clinical epidemiology.* 1997;50(4):401-410.

5. Stunnenberg BC, Raaphorst J, Groenewoud HM, et al. Effect of mexiletine on muscle stiffness in patients with nondystrophic myotonia evaluated using aggregated n-of-1 trials. *Jama.* 2018;320(22):2344-2353.

6. Mitchell GK, Hardy JR, Nikles CJ, et al. The effect of methylphenidate on fatigue in advanced cancer: An aggregated n-of-1 trial. *Journal of pain and symptom management.* 2015;50(3):289-296.

7. Roustit M, Giai J, Gaget O, et al. On-demand sildenafil as a treatment for raynaud phenomenon: A series of n-of-1 trials. *Annals of internal medicine.* 2018;169(10):694-703.

8. Proschan MA. Sample size re-estimation in clinical trials. *Biometrical Journal: Journal of Mathematical Methods in Biosciences.* 2009;51(2):348-357.

9. Senn S. Sample size considerations for n-of-1 trials. *Statistical methods in medical research.* 2019;28(2):372-383.

10. Zucker DM, Denne J. Sample–size redetermination for repeated measures studies. *Biometrics.* 2002;58(3):548-559.

11. Senn S. Mastering variation: Variance components and personalised medicine. *Statistics in medicine.* 2016;35(7):966-977.

12. Proschan MA. Two-stage sample size re-estimation based on a nuisance parameter: A review. *Journal of biopharmaceutical statistics.* 2005;15(4):559-574.

13. Chen X, Chen P. A comparison of four methods for the analysis of n-of-1 trials. *PloS one.* 2014;9(2):e87752.

14. Senn S, Stevens L, Chaturvedi N. Repeated measures in clinical trials: Simple strategies for analysis using summary measures. *Statistics in medicine.* 2000;19(6):861-877.

15. Champely S, Ekstrom C, Dalgaard P, et al. Package 'pwr.' *R package version.* 2018;1(2).

16. R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing; 2021. https://www.R-project.org/

17. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in medicine.* 2019;38(11):2074-2102.

18. Bates D, Maechler M, Bolker B, Walker S. *Package 'Lme4'.*; 2021. https://cran.r-project.org/web/packages/lme4/lme4.pdf

19. Corbeil RR, Searle SR. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics.* 1976;18(1):31-38.