Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Interim sample size reestimation for adequately powered series of N-of-1 trials

Daphne N. Weemering*

[1]Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands

Correspondence
*Corresponding author name, This is sample corresponding address. Email: authorone@gmail.com

Present Address
Padualaan 14, 3584 CH Utrecht, The Netherlands

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean ut elit odio. Donec fermentum tellus neque, vitae fringilla orci pretium vitae. Fusce maximus finibus facilisis. Donec ut ullamcorper turpis. Donec ut porta ipsum. Nullam cursus mauris a sapien ornare pulvinar. Aenean malesuada molestie erat quis mattis. Praesent scelerisque posuere faucibus. Praesent nunc nulla, ullamcorper ut ullamcorper sed, molestie ut est. Donec consequat libero nisi, non semper velit vulputate et. Quisque eleifend tincidunt ligula, bibendum finibus massa cursus eget. Curabitur aliquet vehicula quam non pulvinar. Aliquam facilisis tortor nec purus finibus, sit amet elementum eros sodales. Ut porta porttitor vestibulum. Integer molestie, leo ut maximus aliquam, velit dui iaculis nibh, eget hendrerit purus risus sit amet dolor. Sed sed tincidunt ex. Curabitur imperdiet egestas tellus in iaculis. Maecenas ante neque, pretium vel nisl at, lobortis lacinia neque. In gravida elit vel volutpat imperdiet. Sed ut nulla arcu. Proin blandit interdum ex sit amet laoreet. Phasellus efficitur, sem hendrerit mattis dapibus, nunc tellus ornare nisi, nec eleifend enim nibh ac ipsum. Aenean tincidunt nisl sit amet facilisis faucibus. Donec odio erat, bibendum eu imperdiet sed, gravida luctus turpis.

KEYWORDS:
N-of-1 trials; sample size reestimation; simulation study; statistical methods;

1 | INTRODUCTION

Randomized controlled trials (RCTs) are considered the gold standard in determining treatment efficacy in healthcare. At first glance, these standard RCTs seem to earn their position as the randomization of patients into a parallel experimental and control condition works quite well in balancing factors that are not under experimental control, allowing for unbiased estimation of the population treatment effect. A drawback, however, is that these standard RCTs require a relatively large sample size to establish the effectiveness of treatment with sufficient power. However, for the instances of finding the right intervention for patients with rare diseases, i.e., small patient populations, standard RCTs become therefore unfeasible.

A clinical trial design that offers the possibility to reduce the number of subjects 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 the experimental and control intervention in a random order[1]. As the experiment is conducted within a single patient, the advantage of the N-of-1 trial is that a patient-specific treatment effect estimate is obtained. Often, clinical evidence that is generated by standard RCTs turns out to have poor generalization and is therefore to a limited amount applicable

to patients in the general population[2,3]. This imposes challenges to the usefulness of standard RCT findings in evidence-based medicine. One-size-fits-all does not always apply in practice.

There are some essential clinical conditions that are suitable for the use of a N-of-1 trial. First, the medical condition for which the intervention in prescribed should be chronic and (relatively) stable over time in order to reduce the chance that the progression of the disease can obscure the treatment differences between and within the trial cycles[4,5]. Moreover, the intervention being tested in the N-of-1 trial should have a rapid on- and offset of biological action, and should have a short half-life to ensure that there is rapid washout as the cycles alternate[5]. Additionally, the effect of the intervention should be measured using a validated (clinical) outcome measure (e.g., choosing the right scale ensuring that real benefits and real burdens are being measured). Lastly, the intervention used in the study should not alter the underlying condition, as this will make it unable to interpret the results for an individual as the trial progresses[5]. This all necessitates careful selection of participants, short time cycles and relatively stable symptoms.

As results of a single N-of-1 trial are specific to an individual patient and can therefore not be generalized to the population, a N-of-1 trial does not compare itself with a standard RCT. However, combining several separate N-of-1 trials, under the condition that the trials are identical, creates the possibility to estimate the population-level treatment effect[6]. In the combined analysis of separate N-of-1 trials, now referred to a series of N-of-1 trials, both the magnitude of the average treatment effect as well as the heterogeneity in treatment response are taken into account[6]. Comparing multiple treatment cycles combined with the recognition that the variability in response within individuals is typically lower than the variability between individuals, a smaller sample size is required to detect an effect of treatment in series of N-of-1 trials compared to parallel RCTs[5]. This makes series of N-of-1 trials a valuable alternative to standard RCTs in the accumulation of a comprehensive evidence base in populations of people with rare diseases.

These series of N-of-1 trials have been performed in, among others, studying the effect of mexiletine on nondystrophic myotonia[7], studying the effectiveness of methylphenidate on fatigue in patients with end-stage cancer[8], and for investigating the usefulness of sildenafil on Raynaus-Phenomenon patients[9]. Reasons for choosing the N-of-1 trial methodology vary, in general but also specifically for these aforementioned studies. The latter study chose to conduct a series of N-of-1 trials due to the heterogeneity in treatment response that should be taken into account, whereas the first two studies chose for the N-of-1 trial methodology due to inability to achieve the required sample size for a standard RCT.

Series of N-of-1 trials require a smaller sample size compared to standard two-arm RCTs, because every patient serves as its own control, creating the possibility to obtain multiple observations per individual[10]. A prior sample size calculations are necessary to avoid under- or overpowering the study, for planning on allocating resources and for assessing the feasibility of the study. For these a priori sample size calculations, assumptions have to made with regard to unknown parameters in the model. The nuisance parameters (i.e., the stochastic model components) are often unknown before the start of the studies. Incorrect estimates for these nuisance parameters can lead to substantial over- or underpowering. Overpowering a study potentially exposes too many patients to inferior treatment, whereas underpowering a study increases the risk of committing an error of the second kind, where one is not able to find an effect of treatment when there exists one in the population.

Sample size formulas have been derived for series of N-of-1 trials for both random and fixed effects models[11]. As the main objective of combining the results of separate N-of-1 trials is to make inferences to the population, random effects models are most appropriate and of interest here. In series of N-of-1 trials, the within- and between subject variance of the response to treatment are unknown at the start of the studies[11]. Sample size estimation in series of N-of-1 trials require estimates for these unknown parameters. Taking estimates of nuisance parameters from other studies can be unreliable because of differences in the study population, background conditions or study design[12]. Furthermore, an estimate of the between-subject variance 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[13]. Series of N-of-1 trials are not (yet) that common, and even if similar series of N-of-1 trials exist, these kinds of estimates are usually not reported in the literature.

An appealing strategy for conquering the problem of incorrect assumptions for unknown parameters in sample size calculations is the two-stage design. With this design, the required sample size is calculated by making reasonable assumptions for the unknown parameters. Then, data is obtained up until a predetermined interim point along the trial. The data that is obtained up until the interim point is used to estimate the parameters that were unknown prior to the studies. These estimates at interim are then used to reestimate the sample size, which can then be adjusted accordingly[14]. This two-stage design, now referred to as interim sample size reestimation, can avoid a study to become under- or overpowered in the scenario where wrong assumptions are made with regard to unknown parameters in sample size calculations. A distinction can be made between interim sample

size reestimation based on nuisance parameter estimates and based on treatment effect estimates[15]. This thesis will not cover the latter approach, but rather focuses on interim sample size reestimation based on nuisance parameter estimates.

A general concern with interim sample size reestimation, is the inflation of the type I error rate. Various studies have attempted to analytically calculate or control the type I error rate under various circumstances[16,17,18,19]. Under specific circumstances, this type I error rate becomes inflated, and this is not desirable. Investigating the influence of interim sample size reestimation for series of N-of-1 trials on the type I error rate is not the main objective of this thesis, but it will be assessed whether the type I error rate becomes inflated for specific scenarios considered here.

The application of interim sample size reestimation has not yet been investigated in the context of series of N-of-1 trials and no specific guidelines have been established. Moreover, the minimally required sample size for interim sample size reestimation in series of N-of-1 trials has also not yet been established. With the use of simulation studies, this thesis aims at establishing guidelines for the minimally required sample size for sample size reestimation in series of N-of-1 trials, and to compare series of N-of-trials incorporating interim sample size reestimation with series of N-of-1 trials having a fixed sample size. Power and expected sample size are important measures of performance herein. Furthermore, the type I error rate between series of N-of-1 trials with and without the incorporation of interim sample size reestimation will be assessed.

The remainder of this paper will be structured as follows: In section 2, the model used for the simulation studies and the notation are discussed. In section 3, sample size calculations for series of N-of-1 trials are discussed. In section 4 the setup of the simulation studies will be explained. Section 5 discusses the results of the simulations studies, and section 6 includes the conclusions that can be drawn based on the results, the limitations of this study and considerations for future research on this topic.

2 | MODEL, ASSUMPTIONS AND NOTATION

Simulation studies are performed to compare the reliability and efficacy of series of N-of-1 trials including sample size reestimation with series of N-of-1 trials without including the interim reestimation process of the sample size. For simulating the data, it is assumed that individuals receive treatment A and treatment B once within each cycle. The order of treatment administration within each cycle will be randomly determined. Furthermore, it is assumed that the outcome concerns a continuous measurement, that the disease under study is stable over time, and that carryover effects are absent because of a sufficient duration of the washout period. Lastly, it is assumed that there are no missing data.

Previous research[20,10,11] has shown that linear mixed models provide robust inferences for series of N-of-1 trial data. The so called summary measures approach[10] uses a linear mixed model, only this model calculates the mean difference in outcome under treatment A minus the mean difference in outcome under treatment B. The unit of analysis under the summary measures approach becomes the patient instead of the cycle, which would have been the unit of analysis under a mixed effects model for the original observations. If the data is balanced, the summary measures approach under this model will lead to the same result as a mixed model[11]. Using this approach results in a simpler model:

$$d_{ij} = \tau_i + \epsilon_{ij} \tag{1}$$

In this model, d_{ij} is the observed treatment difference for patient $i = 1, ..., n$ in cycle $j = 1, ..., k$), where treatment B is subtracted from treatment A (or vice versa, as long as it is consistently applied), $\tau_i \sim N(T, \psi^2)$, and $\epsilon_{ij} \sim N(0, 2\sigma^2)$. τ_i is the random treatment effect for patient i. This term has a common average T and a variance ψ^2, which indicates how much the individual treatment effects vary from each other. ϵ_{ij} are random within-patient within-cycle disturbance terms. These disturbance terms are assumed to be i.i.d. (independent and identically distributed) both across cycles and across patients. The variance term of ϵ_{ij}, $2\sigma^2$, is given to make this summary measures model compatible with a mixed model using the original observations[11].

2.1 | Further notation

In section 4 the simulations studies are explained in detail, and the notation used in that section will be clarified here. The nuisance parameter (ψ^2 and σ^2) values that are hypothesized (i.e. assumed) a priori are indicated with a subscript h: ψ_h^2 and σ_h^2. The nuisance parameters values that are used for data generation, and which may be considered the true values of these nuisance parameters, are indicated with a subscript t: ψ_t^2 and σ_t^2. The observed interim values for ψ^2 and σ^2 are indicated as ψ_{obs}^2 and

σ^2_{obs}.For the sample size n, the initial sample size is indicated as n_{init}, the fraction of the initial sample size on which sample size reestimation is based is indicated as n_{frac}, and the reestimated, final sample size is indicated as n_{final}.

3 | SAMPLE SIZE CALCULATIONS IN SERIES OF N-OF-1 TRIALS

From the model in equation 1, the average treatment effect in the population and the corresponding variance could be derived[11], which are necessary for calculating the required sample size. Following Senn[11], the summary measures approach is used. First, the data can be reduced to an average treatment effect for each individual patient:

$$\bar{d}_{i.} = \sum_{j=1}^{k} \frac{d_{ij}}{k} \tag{2}$$

The mean over all the n individual treatment differences ($\bar{d}_{i.}$) is then defined as:

$$\hat{T} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} d_{ij}}{nk} \tag{3}$$

$\bar{T}$ can be used to test the difference between two treatments (say treatment A and treatment B) at hand. This average over all the n separate patients will have a variance of

$$var(\hat{T}) = \frac{\psi^2 + 2\sigma^2/k}{n} \tag{4}$$

Then, the variance at the level of the individual patient is equal to $var(\bar{d}_{i.}) = \psi^2 + 2\sigma^2/k$. This variance, the variance of the n summary measures $\bar{d}_{i.}$, has $(n-1)$ degrees of freedom and can be used to calculate the sample size under hypothesized values for ψ^2 and σ^2 using a one sample t-test. The exact formula for the calculation of the sample size is based on a non-central t-distribution. Computation of the sample size therefore requires an iterative process and this can straightforwardly be done with the `pwr.t.test` function of the `pwr` package[21] in `R`[22].

4 | SIMULATIONS

Simulation studies are conducted to compare the method of interim sample size reestimation in series of N-of-1 trials with a series of N-of-1 trials that do not include interim sample size reestimation for varying values of the nuisance parameters. Special guidelines have been established for planning and reporting simulation studies based on the so called "ADEMP" structure (Aim, Data-generating mechanism, Estimands, Methods, Performance measures)[23] and this section is written in accordance with this structure.

4.1 | Aim

The aims of the simulation studies are twofold First, by means of simulation studies the minimally required sample size for reliable reestimation of the sample size at interim is sought under various scenarios for the nuisance parameters. By comparing the reestimated sample size with the sample size under the true nuisance parameter values (ψ^2_t and σ^2_t) statements about the reliability of the reestimation method can be made. Second, the aim is to compare series of N-of-1 trials incorporating interim sample size reestimation with series of N-of-1 trials having a fixed sample size on power and the type I error rate.

4.2 | Data-generating mechanism

The linear mixed model provided in equation (1) will be used to simulate data for series of N-of-1 trials. As ψ^2 and σ^2 are unknown prior to the study, assumptions have to be made with regard to these parameters. The scenarios that are considered in the simulation study are shown in table **??**. The data is generated under all possible combinations of $\psi^2_t = 0.5, 1, 2$ and $\sigma^2_t = 0.25, 0.5, 1$ ($3^2 = 9$ combinations of ψ^2_t and σ^2_t). For each generated data set, all combinations of the scenarios where $\psi^2_h = 0.5, 1, 2$ and $\sigma^2_h = 0.25, 0.5, 1$ are considered (resulting in $3^4 = 81$ combinations of ψ^2_t, σ^2_t, psi^2_h, and σ^2_h). Lastly, for each

of the 81 scenarios so far discussed, a different fraction ($f = 0.25, 0.5, 0.75$) of the initially calculated sample size is taken to base the interim sample size reestimation on. In total, $3^5 = 243$ scenarios are considered.

For the interpretation of the results in a subsequent section, labels are specified to the sizes of the variance parameters. For the between-subject variance, $\psi^2 = 0.5$ is considered small heterogeneity, $\psi^2 = 1$ moderate heterogeneity, and $\psi^2 = 2$ large heterogeneity. For the within-subject within cycle variance, $\sigma^2 = 0.25$ is considered small error, $\sigma^2 = 0.5$ moderate error, and $\sigma^2 = 1$ large error.

TABLE 1 Parameters in the two simulation studies

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 nominal significance level	$\alpha = 0.05$	
Clinically relevant difference		$\Delta = 0, 1$
Model parameters		
Within-patient within-cycle variance (data generation)		$\sigma_t^2 = 0.25, 0.5, 1$
Variance of treatment effect (data generation)		$\psi_t^2 = 0.5, 1, 2$
Within-patient within-cycle variance (hypothesized)		$\sigma_h^2 = 0.25, 0.5, 1$
Variance of treatment effect (hypothesized)		$\psi_h^2 = 0.5, 1, 2$
Average treatment effect		$T = \Delta = 0, 1$
Simulation parameter		
Number of simulations	$N = 10000$	
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	
Type I error rate	Compare type I error rate of trial with sample size reestimation in simulation to target of 5%	

In the first simulation study, the clinically relevant difference, Δ, will be 1 both for data generation and for calculating the sample size, allowing for estimation of the power of series of N-of-1 trials incorporating interim sample size reestimation. In the second simulation study, Δ will be set to 0 for generating the data and assumed 1 for calculating the sample size in order to estimate the type I error rate. As α is the probability of falsely rejecting the null hypothesis when it is true[24], setting Δ for data generation equal to zero (which it would be under the null hypothesis) and subsequently calculating the sample size assuming $\Delta = 1$ allows for the estimation of the type I error rate.

4.3 | Estimands

The estimand (the parameter to be estimated) that is of interest in the simulation studies, is the average treatment effect in the population T.

4.4 | Methods

Linear mixed models will be fitted to the simulated data using the `lme4` package[25] in `R`[22]. In fitting the linear mixed models, restricted maximum likelihood (REML) will be applied to estimate the variance parameters, as this approach, to the contrary of

"regular" maximum likelihood (ML), produces unbiased estimates for the variance components in the model for smaller sample sizes[26].

The initial sample size, n_{init}, will be calculated as the sample size required to achieve 80% power under ψ_h^2, σ_h^2, the clinically relevant difference Δ, and the two-sided significance level $\alpha = 0.05$. Sample size calculations will be performed using the `pwr` package[21] in `R`. Equation 3 will be applied to calculate the standard deviation for the standardized effect size needed for the `pwr.t.test` function. A one-sample t-test is used to test the treatment difference. The sample size will be reevaluated after $f * n$ subjects (where f represents the fraction of the initial sample size) using ψ_{obs}^2 and σ_{obs}^2. Eventually, the results of the trials including interim sample size reestimation will be compared with trials having a fixed sample size.

4.5 | Performance measures

The performance of interim sample size reestimation will be evaluated by means of statistical power, the expected sample size and the type I error rate. Power is defined as $\frac{p \leq 0.05}{N}$, where p is the p-value of the test statistic and N is the number of iterations in the simulation. The same formula for the power can be used to estimate the type I error rate when for the generation of the data $\Delta = 0$. The power will be compared to a target of 80%, the type I error rate will be compared with the target of 5%. The sample size under reestimation (and under a fixed sample size) will be compared with the sample size under ψ_t^2 and σ_t^2. If interim sample size reestimation works properly, the reestimated sample size should be close to the sample size under the true model.

5 | RESULTS

The simulation studies allowed for the investigation of the question regarding the minimally required sample size for reliable reestimation of the sample size at interim, and for the comparison on the power and type I error rate between series of N-of-1 trials that incorporate interim sample size reestimation with series of N-of-1 trials that have a fixed sample size set in advance.

5.1 | Sample size

The average reestimated sample size under ψ_h^2 and σ_h^2 is approximately equal to the sample size under the true values for data generation, ψ_t^2 and σ_t^2, as displayed in figure 1, which displays the reestimated sample size under different combinations of ψ_h^2, σ_h^2, ψ_t^2, σ_t^2 and f. Table S1 in the supplementary materials includes the corresponding table to this figure.

Even though the reestimated sample size is approximately equal to the sample size under the true scenario for most of the hypothesized scenarios, the variation in the reestimated sample size is quite large in some cases. From figure 1, it quickly becomes clear that the larger the fraction of the initial sample size on which interim sample size reestimation is based, the smaller the variance in reestimated sample sizes. This makes sense, as using a larger fraction of the initial sample size to estimate the nuisance parameters ψ_{obs} and σ_{obs} increases the precision of these estimates and hence also the reestimated sample size.

Additionally, the size of the two nuisance parameters also seems to influence the variability in reestimated sample sizes. The size of ψ_t^2, the between-person variance, seems to have a greater influence on the variability in reestimated sample sizes than σ_t^2, the within-person within cycle variance. If there exists a lot of interperson heterogeneity in treatment effectiveness in the population, then this influences the effect size, and thus also the sample size, more than when there would be a lot of intraperson heterogeneity in the treatment effect.

Furthermore, the results show greater variability in reestimated sample size when ψ_h^2 is relatively small (in this case when $\psi_h^2 = 0.5$ compared to when σ_h^2 is small. This greater variability in ψ_h^2 occurs regardless of the true underlying values for the nuisance parameters. However, the larger the discrepancy between the true and hypothesized values for ψ^2, the larger the variation in reestimated sample sizes. Noteworthy are the results that show that even when the hypothesized and the true values for the between-person variance are equal to each other, it still occurs that the variance in reestimated sample sizes is largest when ψ_h^2 (and thus also ψ_t^2) is small.

5.2 | Power

Despite the fact that variability can be quite high for some values of the nuisance parameters overall power for these cases might still be adequate. Figure 2 depicts the power of all the different scenarios in the simulation study. These results for the power

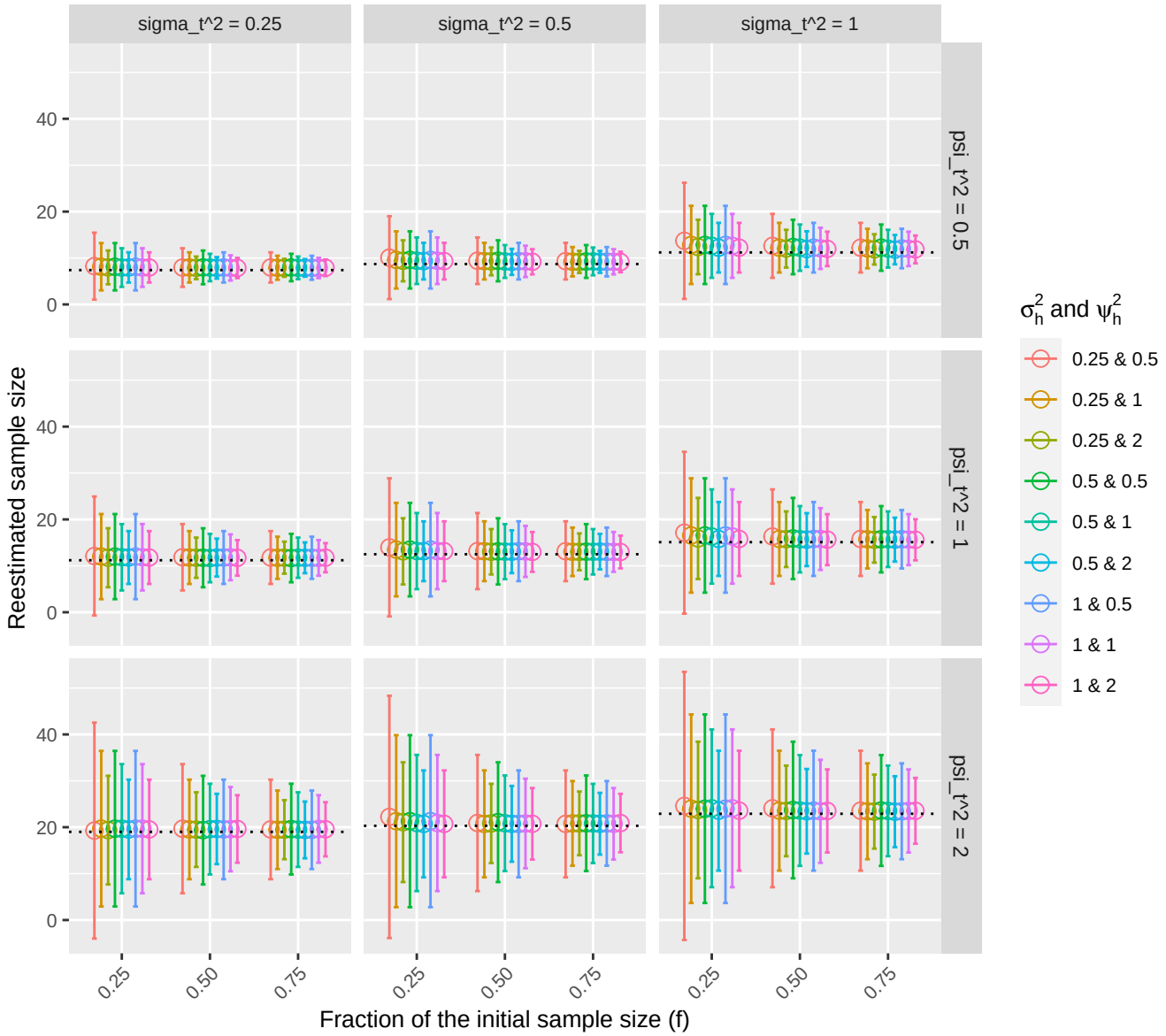

FIGURE 1 Reestimated sample sizes in simulated series of N-of-1 trials under different data, hypothesized values for the nuisance parameters, and fractions of the initial sample size. The dotted line indicates the sample size under the true values for the nuisance parameters.

can also be found in table S2 in the supplementary materials. As the values for ψ_h^2 and σ_h^2 showed to have different effects on the reestimated sample size, these are separated by means of shape and color in the figure. First, the scenarios for the combinations of the true and hypothesized nuisance parameters where f is equal to 0.25 are considered. Evidently, these scenarios showed the greatest variability in reestimated sample sizes. Figure 2 shows that a part of these scenarios also result in an underpowered study. Especially the scenarios where the hypothesized between-person variance is small (0.5) and the between-person in the population is large (2) lead to a lack of statistical power. For most of these cases the reduction in power remains limited, but the cases where $\psi_h^2 = 0.5$, $\sigma_h^2 = 0.25$ and where $\psi_t^2 = 2$) lead to a power of 0.7. Here again the limited influence of the true value for the within-person variance becomes clear.

Apart from the cases where underpowering occurs, there are also some cases where an excess of statistical power in evident. For the scenarios where $\psi_t^2 = 0.5$, all the combinations for the hypothesized nuisance parameters lead to overpowered studies. Especially the scenarios where the true value for the between-subject variance is small and the hypothesized values for this parameter are large will lead to an excess of power. The profusion of power becomes even more evident when a larger fraction

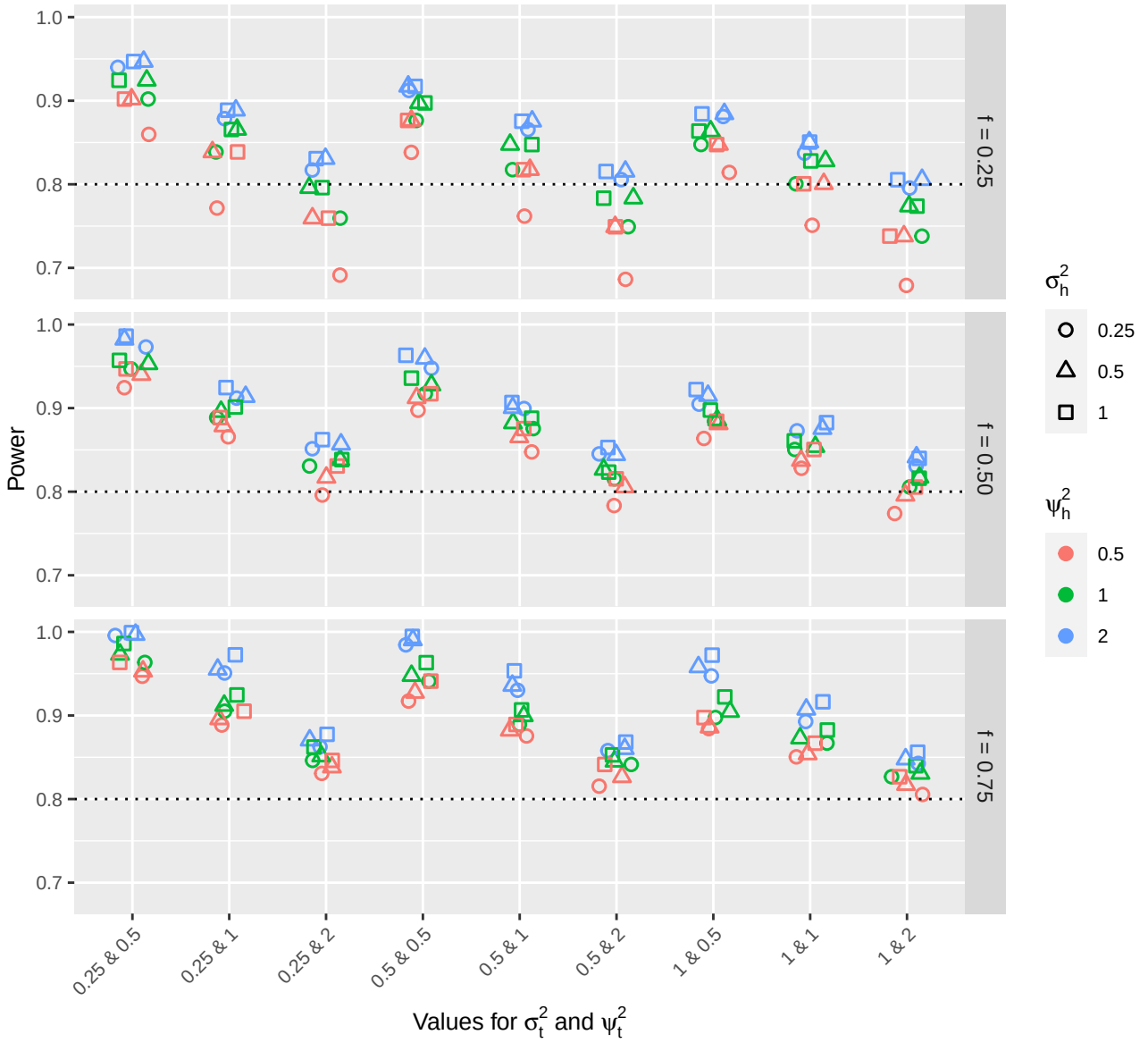

FIGURE 2 Power in simulated series of N-of-1 trials under different data and hypothesized values for the nuisance parameters. The x-axis indicates the different combinations of the nuisance parameters for data generation (ψ^2 and σ^2, respectively). The dotted line indicates a power of 0.8.

of the initial sample size is taken to base interim sample size reestimation on. When $f = 0.75$, the combination of a small ψ_t^2 and a large ψ_h^2 will even lead to studies with a power close to 1. By taking the 75% of the initial sample size, almost all the scenarios considered here will lead to overpowered studies, with the exception of the scenarios where $\psi_h^2 = 0.5$. Even though overpowering a study might be not as bad as underpowering a study, as the result in case of an underpowered study might be less reliable, overpowering results in quite a waste of resources and is therefore also not very desirable.

The results under the scenarios where $f = 0.5$ are a bit more moderate then the results of the scenarios discussed before. Again, the cases where $\psi_h^2 = 0.5$ and where $\psi_t^2 = 2$ result in slight underpowered studies, and the scenarios where ψ_h^2 is large result in quite overpowered studies, with the exception of the scenario where the true value for the between-person variance is also large. However, excess power stays present for most scenarios.

A last noteworthy point on the power in these simulations is the fact that overpowering also occurs when the hypothesized and true values of the nuisance parameters are equal to each other. Even when the "guess" for the nuisance parameters with right at the start of the studies, an excess of power is inevitable, at least for the scenarios discussed here.

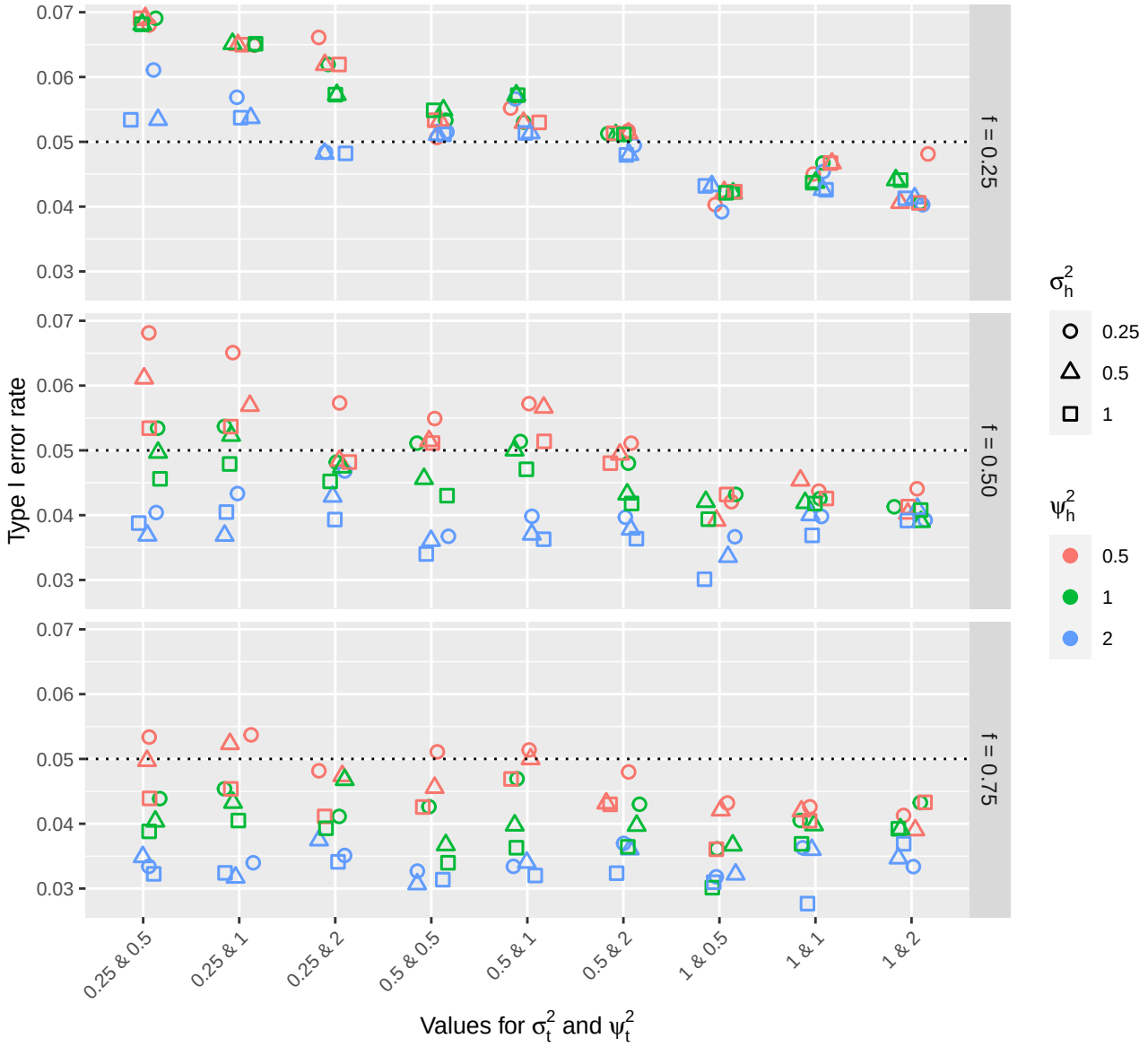

FIGURE 3 Type I error rate in simulated series of N-of-1 trials under different data and hypothesized values for the nuisance parameters. The x-axis indicates the different combinations of the nuisance parameters for data generation (ψ^2 and σ^2, respectively). The dotted line indicates the nominal α level of 0.05.

5.3 | Type I error rate

Finally, the effect of interim sample size reestimation on the type I error rate was examined. In section 5.2, the effect of interim sample size reestimation on the statistical power was discusses. As the type I error rate and power are interrelated with each other, it is interesting to see if the excess of power that was discovered in the previous section will also translate in an increase in the rejection region or if it stays constant.

Figure 3 depicts the type I error rate for the different scenarios of the true and hypothesized nuisance parameters per fraction of the initial sample size on which reestimation was based. These results can also be found in tabular format in table S3 in the supplementary materials. First, the scenarios where $f = 0.25$. Here, the size of σ_t^2 in combination with the size of ψ_h^2 seem of most influence on the type I error rate. For smaller values of σ_t^2, the type I error rate becomes inflated for the smaller values of ψ_h^2, resulting in type I error rates around 0.07. Larger values of ψ_h^2 under $\sigma_t^2 = 0.25$ remain closer to the nominal level of α. When $\sigma_t^2 = 0.5$ or 1, the results also remain close to the nominal α level.

For the scenarios where $f = 0.5$, inflation of the type I error rate only occurs when the values for σ_t^2 are small to moderate in combination with values for ψ_h^2 being small. Other combinations of true and hypothesized values for the nuisance parameters result in adequate to lower type I error rates compared to the nominal α level. When $f = 0.75$, all the scenarios lead to adequate and lower type I error rates. A decrease in the type I error rate can result in a reduction of power, but as the results in section 5.2 indicate, this is not an issue for the scenarios that are considered here.

6 | DISCUSSION

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean ut elit odio. Donec fermentum tellus neque, vitae fringilla orci pretium vitae.

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. Greenfield S, Kravitz R, Duan N, Kaplan SH. Heterogeneity of treatment effects: implications for guidelines, payment, and quality assessment. *The American journal of medicine* 2007; 120(4): S3–S9. doi: 10.1016/j.amjmed.2007.02.002

3. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank Quarterly* 2004; 82(4): 661–687. doi: 10.1111/j.0887-378X.2004.00327.x

4. Johnston BC, Mills E. N-of-1 randomized controlled trials: an opportunity for complementary and alternative medicine evaluation. *Journal of Alternative & Complementary Medicine* 2004; 10(6): 979–984. doi: 10.1089/acm.2004.10.979

5. 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. doi: 10.1016/j.jclinepi.2010.05.009

6. 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. doi: 10.1016/S0895-4356(96)00429-5

7. 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. doi: 10.1001/jama.2018.18020

8. 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. doi: 10.1016/j.jpainsymman.2015.03.009

9. 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. doi: 10.7326/M18-0517

10. Araujo A, Julious S, Senn S. Understanding variation in sets of N-of-1 trials. *PloS one* 2016; 11(12): e0167167. doi: 10.1371/journal.pone.0167167

11. Senn S. Sample size considerations for n-of-1 trials. *Statistical methods in medical research* 2019; 28(2): 372–383. doi: 10.1177/0962280217726801

12. Zucker DM, Denne J. Sample–size redetermination for repeated measures studies. *Biometrics* 2002; 58(3): 548–559. doi: 10.1111/j.0006-341X.2002.00548.x

13. Senn S. Mastering variation: variance components and personalised medicine. *Statistics in medicine* 2016; 35(7): 966–977. doi: 10.1002/sim.6739

14. Proschan MA. Two-stage sample size re-estimation based on a nuisance parameter: a review. *Journal of biopharmaceutical statistics* 2005; 15(4): 559–574. doi: 10.1081/BIP-200062852

15. Proschan MA. Sample size re-estimation in clinical trials. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 2009; 51(2): 348–357. doi: 10.1002/bimj.200800266

16. Stein C. A two-sample test for a linear hypothesis whose power is independent of the variance. *The Annals of Mathematical Statistics* 1945; 16(3): 243–258.

17. Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in medicine* 1990; 9(1-2): 65–72. doi: 10.1002/sim.4780090113

18. Kieser M, Friede T. Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in medicine* 2000; 19(7): 901–911.

19. Gao P, Ware JH, Mehta C. Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of Biopharmaceutical Statistics* 2008; 18(6): 1184–1196. doi: 10.1080/10543400802369053

20. Chen X, Chen P. A comparison of four methods for the analysis of N-of-1 trials. *PloS one* 2014; 9(2): e87752. doi: 10.1371/journal.pone.0087752

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

22. R Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2021.

23. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in medicine* 2019; 38(11): 2074–2102. doi: 10.1002/sim.8086

24. Neyman J, Pearson ES. IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 1933; 231(694-706): 289–337. doi: 10.1098/rsta.1933.0009

25. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 2015; 67(1): 1–48. doi: 10.18637/jss.v067.i01

26. Corbeil RR, Searle SR. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics* 1976; 18(1): 31–38. doi: 10.2307/1267913