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. 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.

As in any clinical study, a priori sample size determination is necessary to avoid under- or overpowering the study, for planning on allocating resources and for assessing the feasibility of the study. Sample size formulas have been derived for series of N-of-1 trials for both random and fixed effects models[10]. As the main objective of combining the results of separate N-of-1 trials is to make inferences regarding the population treatment effect, random effects models are most appropriate and of interest here. For the sample size calculations, assumptions have to be made with regard to the parameters in the model, such as the clinically relevant difference and the nuisance parameters. However, the nuisance parameters in the model, which concern the within- and between subject variance of the response to treatment for series of N-of-1 trials[11], are generally unknown at the start of the study. 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 (or trials) 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. Making unrealistic assumptions for these nuisance parameters can lead to substantial over- or underpowering, where the former exposes too many patients to potentially inferior treatment and the latter increases the risk of failing to identify a clinically relevant treatment effect due to a lack of power.

An appealing strategy for conquering the problem of incorrect assumptions for unknown parameters in sample size calculations is a two-stage design with interim sample size reestimation based on nuisance parameter estimates. With this design, the initially required sample size is calculated by making reasonable assumptions for the unknown nuisance parameters. Then, a portion of the data is collected up until a prespecified interim point along the trial and the unknown nuisance parameters are estimated. These estimates are then used to update the power analysis and to adjust the sample size. Subsequently, the study is continued until all required participants are observed, and finally, the hypothesis is tested with all the data[14]. Simulations studies have shown that this method has a high potential to protect from an incorrect sample size if the nuisance parameters where misspecified at the design stage of the study[15]. A distinction can be made between interim sample size reestimation based on nuisance parameter estimates and based on treatment effect estimates[16]. This thesis will not cover the latter approach.

A concern with interim sample size reestimation based on estimates of nuisance parameters, is the inflation of the type I error rate[17,15,18]. 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. However, it will be assessed whether the type I error rate becomes inflated for specific scenarios considered here, allowing future research to build upon these results.

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. With the use of simulation studies, guidelines for the minimally required sample size for sample size reesimation in series of N-of-1 trials will be established, and series of N-of-1 trials incorporating interim sample size reestimation will be compared 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 for series of N-of-1 trials incorporating interim sample size reestimation will be evaluated.

The remainder of this thesis will be structured as follows: In section 2, notation, sample size calculations for series of N-of-1 trials, and the model that is used for the simulation studies are discussed. In section 3 the design of the simulation studies in which power, expected sample size and the type I error rate are evaluated will be explained. Section 4 discusses the results of the simulations studies. And finally, this thesis is concluded with a discussion which is outlined in section 5.

2 | METHODOLOGY

First, the methodology of a one-stage design for series of N-of-1 trials will be discussed in section 2.1, in which notation, the model used and sample size calculations for series of N-of-1 trials are introduced. In section 2.2, the procedure for interim sample size reestimation in series of N-of-1 trials will be outlined.

2.1 | One-stage design

In a series of N-of-1 trials, n subjects in k cycles, with each cycle consisting of two periods, receive the experimental condition in one period and the control condition in the other period. The order of treatment administration within each cycle will be randomly determined. At the end of each period, the outcome Y_{ijt} is measured, indicating the outcome for patient i ($i = 1, ..., n$) in cycle j ($j = 1, ..., k$) who is given treatment t ($t = 1, 2$). It is assumed that the disease under study is stable over time, that carryover effects are absent because of a sufficient duration of the washout period, and that there are no missing data. Furthermore, it is assumed that the outcome is a continuous measure and that it is normally distributed according to the following model:

$$Y_{ijt} = \lambda_i + \beta_{ij} + \epsilon_{ijt} + Z_{ijt}\tau_i \tag{1}$$

In this model[11], $\lambda_i \sim N(\Lambda, \phi^2)$, $\beta_{ij} \sim N(0, \gamma^2)$, $\epsilon_{ijt} \sim N(0, \sigma^2)$ and $\tau_i \sim N(T, \psi^2)$. $Z_{ijt} = \frac{1}{2}$ or $-\frac{1}{2}$, dependent on whether $t = 1$ or 2 for patient i in cycle j. In this model, λ_i represents the random effect for subject i, β_{ij} indicates the cycle effects for subject i in cycle j, ϵ represents the i-th subject's random error for the j-th cycle and treatment t, and finally, τ_i indicates the treatment effect for subject i. In this model, all the nuisance parameters are assumed to be independent of each other.

However, under the assumption that the data is balanced, a simpler model for the treatment differences for subject i in cycle j can be derived from equation 1 by subtracting the values from the first period in every cycle from the second period in the cycle, and subsequently divide by $(Z_{ij1} - Z_{ij2})$[11]:

$$d_{ij} = \tau_i + \epsilon_{ij} \tag{2}$$

Here, d_{ij} represents the observed treatment difference for patient i in cycle j, where treatment 1 is consistently subtracted from treatment 2. In this model, $\tau_i \sim N(T, \psi^2)$, as before and $\epsilon_{ij} \sim N(0, 2\sigma^2)$. τ_i is the random treatment effect for patient i, and ϵ_{ij} are random within-subject within-cycle disturbance terms. These disturbance terms are assumed to be i.i.d. (independent and identically distributed) both across cycles and across patients. Furthermore, τ_i and ϵ_{ij} are assumed to be independent of each other. The variance term of ϵ_{ij}, $2\sigma^2$, is given to make this model compatible with a mixed model using the original observations[10].

2.1.1 | Sample size calculations in N-of-1 trials

From the model of the treatment differences in equation 2, the average treatment effect in the population and the corresponding variance can be derived, both necessary for calculating the required sample size. Following Senn [10], an average treatment effect for each patient can be obtained:

$$\bar{d}_{i.} = \frac{\sum_{j=1}^{k} d_{ij}}{k} \tag{3}$$

The average over all the n averages is equal to $\hat{T} = \sum_{i=1}^{n} \sum_{j=1}^{k} d_{ij}/nk$, which can be used to test the differences between the two treatments under investigation. This estimate has a variance of $var(\hat{T}) = \psi^2 + 2\sigma^2/kn$. The variance at the patient level, the variance of the n estimates of $\bar{d}_{i.}$, is then defined as:

$$var(\bar{d}_{i.}) = \psi^2 + 2\sigma^2/k \tag{4}$$

This estimate has $(n-1)$ degrees of freedom [10] and can be used to calculate the sample size under hypothesized values for the nuisance parameters, ψ^2 and σ^2, now denoted as ψ_h^2 and σ_h^2, using a one sample t-test. For the calculation of the sample size, the standard deviation of the estimate in equation 4 is used. 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 [19] in `R` [20].

2.2 | Two-stage design

To cope with the problem around the a priori uncertainty regarding the nuisance parameters in the model, interim sample size reestimation is applied. For this process, the following steps are taken [15,17,21]:

1. Specify the clinically relevant difference (Δ), the type I error rate (α), the desired power ($1-\beta$), the proportion of the initial sample size on which interim sample size reestimation will be based (f) and the a priori hypothesized estimates for the nuisance parameters (ψ_h^2 and σ_h^2).
2. Use ψ_h^2 and σ_h^2 to estimate the initial sample size n_{init} that yields the desired level of power. All sample sizes are rounded up.
3. Use $f n_{init} = n_{frac}$, the initial sample size on which interim sample size reestimation is based, to estimate $\hat{\sigma}^2$ and $\hat{\psi}^2$.
4. Then, use $\hat{\sigma}^2$ and $\hat{\psi}^2$ to find the new total sample size, n_{final}, that is needed to achieve the target power and subsequently observe the additional $n_{final} - n_{frac}$ patients. If $n_{final} - n_{frac} \leq 0$, n_{frac} is used as the final sample size. Also, if the effect size (Cohen's d, mean treatment difference divided by the standard deviation) for recalculating the sample size at interim becomes larger than 10 due to small $\hat{\sigma}^2$ and $\hat{\psi}^2$, the sample size becomes too small. In that case, the effect size will be set equal to 10.
5. Test the hypothesis on all the $n_{frac} + (n_{final} - n_{frac})$ observations.

3 | SIMULATIONS

4 | RESULTS

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

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

11. 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

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. 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

16. 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

17. 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.

18. Birkett MA, Day SJ. Internal pilot studies for estimating sample size. *Statistics in medicine* 1994; 13(23-24): 2455–2463. doi: 10.1002/sim.4780132309

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

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

21. Coffey CS, Muller KE. Exact test size and power of a Gaussian error linear model for an internal pilot study. *Statistics in Medicine* 1999; 18(10): 1199–1214.