Simplification or simulation: power calculation in clinical trials

Chao Huang[1]; Pute Li[2]; Colin R. Martin[3]

[1] Hull York Medical School, University of Hull, UK;

[2] Pute Li, School of Professional Study, New York University

[3] Institute for Health and Wellbeing, University of Suffolk, UK

Corresponding author: Chao Huang, PhD, Hull York Medical School, University of Hull, 3rd Floor, Allam Medical Building, Cottingham Road, Hull, HU6 7RX, UK; chao.huang@hyms.ac.uk; Tel: + 44(0)1482463281

© 2021. This manuscript version is made available under the CC-BY-NC-ND 4.0 license: http://creativecommons.org/licenses/by-nc-nd/4.0/

Abstract:

Background and Objectives**:** A justifiable sample size is essential at trial design stage. Generally this task is completed by forming the main research question into a statistical procedure and then implementing the published formulae or software packages. When these standard statistical formulae/software packages become unavailable for studies with complex statistical procedures, some statisticians choose to fill this gap by assuming an alternative simplified sample size calculation. Monte Carlo simulations can also be deployed, particularly for complex trials. However, it is still unclear on how to determine the appropriate approach under certain practical scenarios.

Methods**:** We adopted real clinical trials as examples and investigated on simplification and simulation-based sample size calculation approaches.

Results**:** Compared to simplified sample size calculation, the simulation approach can better address the non-ignorable impact of baseline/follow-up outcome correlation on study power. For studies with multiple endpoints and multiple co-primary endpoints, the sample sizes calculated by simplification approach should be scrutinized.

Conclusions**:** Directly using the simplification approach for sample size calculation should be restricted. We recommend to utilize the simulation approach, particularly for complex trials, at least as a sensitivity checking and a useful triangulation to the simplification approach outlined.

Key words: power calculation; clinical trials; sample size by simplification; simulation approach;

Background

Sample size calculation in trials is crucial as an underpowered study may cause clinical important effects not to be detected, whereas an overpowered study can lead to waste in unnecessary recruitment resources and the unnecessary detection of statistically significant but not clinical important treatment effects. Existing sample size calculation formulae or software packages, such as N-query, Stata and packages in R language, are well placed for sample size calculation for conventional trials. However, with the rapid development of evidence-based medicine, more innovative trial designs were developed and utilized in health research (1,2).

Correspondingly, a variety of formulae-based sample size calculation approaches were developed for cluster-randomized trials (3,4), hierarchical longitudinal designs (5) and proportional hazards mixture cure model (6). However, formulae-based sample size calculation occasionally fails to cover innovative trial designs. Also, sometimes formulae-based sample size calculation fails to incorporate the related trial information, such as the correlation among repeated measures or multiple endpoints. When facing these difficulties, an alternative simpler statistical procedure for direct formulae/software calculation is sometimes chosen instead. Here we define this behaviour as simplified sample size calculation, i.e., when there is no direct sample size calculation formula/software for the statistical approach used in the analysis of primary outcome, one simplified statistical approach was used for sample size calculation.

On the other hand, simulation approaches had also been adapted for sample size calculations. With the assistance of computers, statisticians can repeatedly simulate the trial data and carry out the statistical analysis to accumulate the proportion of hypotheses rejections as an estimate of power. It avoids the complexity of direct formulae development and hence can comprehensively serve sample size calculations. Feiveson (7) provided a step-to-step process for simulation based sample size calculation with examples coded in Stata. Eng (8) discussed the usage of simulation approach in correlated data and ROC curves. Gastañaga et.al (9) considered the simulation approach for

longitudinal data with the non-zero with-in cluster correlation. Zhang et.al (10) utilized the simulation approach in power calculation for interrupted time series. Simulation approach was also used to investigate the impact of cluster size variation on study powers (11).

To this end, we investigate the sample size calculation by simplification and simulation approaches in clinical trials, in terms of the reliability aspect of each approach.

Methods

Full powered individual randomised controlled trials, i.e., Phase III RCTs, were chosen as worked examples for comparison between simplification and simulation based sample sizes. For each case scenario, general trial information on the study design, models for simulation and simplification based sample size calculation were summarised in Table 1.Power calculation by simplification was firstly undertaken using nQuery version 8, with the given trial information and pre-specified simplification procedure. Then the simulated outcome data were formulated and generated by multivariable linear models following multivariate normal distribution. Details on the simulation formulation for each case scenario, together with their simulation codes, were placed in the supplementary materials. The study power of a given sample size was calculated as the proportion of significant p values in the primary outcome data analysis. We fulfil 10,000 times simulation for each scenario in order to get adequate accuracy in power calculation (12). With 90% power as the conventional requirement, the study sample size is the first one that reaches the required power. In order to assess the impact of correlations among multiple primary and co-primary endpoints, common setting of 0.2, 0.5 and 0.8 as weak, moderate and strong correlation were adopted. Simulation based power calculation was programmed and undertaken by R language version 3.6.1.

Results

Case 1. Baseline/follow-up correlation

For studies with repeated trial outcome measures, the correlation between baseline and follow-up measures of each participant, such as test-retest correlation (13), often exists. To reflect on this,

ANCOVA or regression approach with the follow-up measure regressed by the baseline measure are commonly employed. In this sense, sample size calculation based on unadjusted tests (independent t-test or Mann-Whitney test) is a simplified procedure, as it does not consider the baseline measure and potential baseline/follow-up correlation.

The CREAM study was a three-arm randomised controlled trial (14). Patients were randomised to three study groups: one placebo (control) group and two active treatment groups. The primary outcome, Patient Oriented Eczema Measure (POEM) score, was measured at baseline and 2-week follow-up. The initial sample size calculation was based on a clinically important difference of 3 in POEM score and a common standard deviation of 7. Based on independent t-test on POEM score at 2-week follow-up and with 0.025 significance level and 90% power, 137 patients per treatment group were required, giving a total of 411.

During an interim analysis, the data from the first 69 patients was used to revisit some of the parameters. Using the standard deviation from the baseline POEM scores (5.3) and the correlation between baseline and 2-week POEM scores (0.27) and the same clinically important difference for POEM, we use the simulation approach to re-calculate the sample size (Refer to supplementary materials). It was found that less number of patients were actually needed, i.e., 75 patients per group were required to reach 90% power, giving a total of 225 required for analysis.

Case 2. Multiple primary endpoints

The term multiple primary endpoints and multiple co-primary endpoints were specified in recent EMA guidance and FDA guidance (15, 16). For studies with multiple endpoints, a success in at least one endpoint is regarded as sufficient. To control the overall type I error, this multiple-comparison statistical inference is often carried out by splitting the significance level alpha for each primary endpoint under the union-intersection test. Simplification in sample size calculation occurs when direct corrections (Bonferroni correction) are used for alpha splitting, which ignores the potential correlation between primary endpoints (assuming independence).

We redesigned the aforementioned CREAM trial as a two–arm trial with 2-week, 4-week and 3-month POEM scores as multiple primary endpoints. Patients were randomized to two study groups: placebo (the control group) or active treatment group. Three statistical hypothesis testing (treatment vs control group comparison of POEM scores at 2-week, 4-week and 3-month) are required to conclude whether the treatment is effective or not. The treatment is regarded as successful if it is effective in at least one of these three tests. To control the overall type I error, the significance level alpha needs to be split by these tests. The simplification approach lies in evenly splitting alpha cross these endpoints ($\frac{\alpha}{3} = \frac{0.05}{3} = 0.0167$). To reach 90% power, 86 patients per group is required.

Depending on the level of correlations, we employed the simulation approach to calculate the sample size (Refer to supplementary materials). The corresponding power calculations were listed in the first half of Table 2. With 86 patients per group, the calculated study power were 99.6%, 98.4% and 95.7% for weak, moderate and strong correlation.

The over-estimation in sample size can be tuned by the simulation approach with an iterative process of adjusting for type I and type II error. For the case with weak correlation (rho=0.2), we plot the powers against the sample sizes, together with its corresponding split alpha (Figure 1). It was shown 46 patients per group can achieve 90% study power, with the simulated split alpha $\frac{\alpha}{3}$=0.0173 under its null hypothesis.

Case 3. Multiple co-primary endpoints

A trial with multiple co-primary endpoints is successful if there is a significant improvement for all the endpoints. This means that the collective power should be deduced under the intersection-union test. Simplification in its sample size calculation occurs when the independence of these co-primary endpoints was assumed, which ignores the potential correlation between co-primary endpoints.

We redesigned the CREAM study as a two-arm trial with 2-week, 4-week and 3-month POEM scores as three co-primary endpoints. Patients were randomized to two study groups: placebo (the control

group) or active treatment group. Three statistical hypothesis testing (treatment vs control group comparison of POEM scores at 2-week, 4-week and 3-month) are required to conclude whether the treatment is effective or not. The treatment is regarded as effective only when it is tested to be effective at all three tests. The simplification approach indicates that each primary endpoint has to be at least 96.6% to ensure the 90% overall power of the study ($96.6\%^3$=90.1%). To reach the overall power, 91 patients per group are required by simplification approach.

Depending on the level of correlations, we employed the simulation approach to calculate the sample size (Refer to supplementary materials). The corresponding powers were presented in the second half of Table 2. With 91 patients per group, the calculated study power were 90.5%, 92.0% and 93.4% for weak, moderate and strong correlation.

Conclusion

Using simplified statistical approach for sample size calculation is not rarely seen in clinical trial practice, despite it is not well recognized. To the best of our knowledge, this is the first research that formally address and investigate the impact of this behaviour.

Our results showed that there is a growing trend in power when the correlation between baseline and follow-up outcome measures increases. Simplified sample size calculation does not consider this correlation and hence provides a conservative sample size, which can cause unnecessary recruitment burden. Unless the prior study information endorse no/weak correlation (such as studies with long follow-up period), it is recommendable to avoid simplification based sample size calculation approach. Under certain circumstances, this issue can be resolved by multiplying a deflating factor to the sample size calculated by two sample t-test (17). However, it is only valid for cases with bivariate normal distribution, constant baseline/follow correlation and variances in two treatment groups. Simulation based approach, on the contrary, can serve this purpose more comprehensively. An alternative approach is to replace the primary outcome by the change from baseline to follow-up and then still undertake two sample t-test. This approach is preferable to

simplified power calculation, as the correlation is incorporated into the baseline/follow-up change to some extent. However, comparing to the simulation approach which fully takes the baseline and follow-up measures into the analysis model, the baseline/follow-up change is actually a data compression hence subject to information loss. Therefore, the statistical property of this alternative approach is inferior to the simulation approach.

For studies with multiple primary endpoints, the sample sizes calculated by directly splitting alpha largely over-power the study, as the Bonferroni correction is quite conservative for multiple comparison inference and it doesn't consider the potential correlations among primary endpoints.

For studies with multiple co-primary endpoints, the sample sizes calculated by simplification approaches are over-powered when correlation exists. Therefore, it is accurate only when the co-primary endpoints can be regarded as non-correlated, such as studies with one clinical endpoint and one safety endpoint (18). For some studies with two or three co-primary continuous endpoints, formula in adjusting Type II error were developed to accommodate correlation (19). A statistical package mpe was develop in R language for this adjustment (20). Again, simulation approach can serve this purpose more comprehensively. The flip side of intersection union test for multiple co-primary endpoints is its impact on type I error. Several methods are available to improve the power for co-primary endpoints with an adjusted Type I error (21), whilst these approaches are only applicable to certain cases (two or three continuous outcomes and superiority trials). As we demonstrated in Figure 1, simulation approaches can achieve more accurate sample size calculation with iterative tuning process on type I error and type II error.

Our three case scenarios showed that the calculated sample size by simplification approach gave conservative sample size calculation (over-powered), whilst simplification approach sometimes provides under-powered sample size calculation. One typical example is the factorial trial. For a two by two factorial RCT, if the interaction term between two main treatments was ignored (simplification) in sample size calculation, the study sample size will be underestimated (22). Another example is some mediation and moderation models with covariates. In a mediation model with two

covariates and one mediator, the sample size calculated on the basis of the X-Y direct effect only will be under-powered (23).

For the most recent innovative trials, the simulation approach has already been widely accepted and used. Hence the behaviour of using simplified sample size calculation is less seen. Despite that, practical burdens in using simulation based approach hinders its usage in common practice, because a certain level of prior trial information is required to set up the simulation model. If there are available data from existing early phase studies (such as feasibility or pilot studies), we would be able to use these data as prior information. For the trials without these data, we may refer to the existing literature to infer the simulation model specification, and sometimes we need to assume a range of values on the model parameter, such as correlation, which will lead to a sample size range to check the robustness of calculated sample size against different parameter settings. For trials without any prior information, we will need to simulate the sample size with various model settings and go with the most conservative option. This sensible checking approach can avoid taking the simplified sample size calculation for granted, without exploring the possibilities of getting more accurate sample size calculation via simulation approach.

As the first research addressing and comparing the simulated and simplified sample size calculations, we concentrated on full powered individual randomised controlled trials. For some innovative trials, such as adaptive design, the more likely situation is the appropriate calculation is not available in software, hence the simulation approach is naturally used. Considering the adaptive design is a wide concept with various study designs and allows pre-planned changes to an ongoing trial, it is worth regarding it as a separate topic to further investigate. In summary, the simplification approach can cause over-powered and under-powered study sample sizes, while the direction and degree of miss-shot is case-dependent. Further work is thus important to evaluate robustly not only the contribution, ready applicability and ease of use of the simplification approach, but also potential limitations in terms of the impact of more complex models on statistical power. In general, we should restrict ourselves in using the simplification approach alone for sample size calculation. We

also reckon that more and more updated power calculation approaches for innovative trial designs have been incorporated into existing statistical software. Therefore, we encourage trial statisticians to regularly check these updates and adopt them in practice. We also recommend to utilize the simulation approach, particular for complex trials, at least as a sensitivity checking and a useful triangulation to the simplification approach outlined. This practice can reassure the power calculation from different aspects, hence enhance the replicability of scientific findings (24).

References:

1. Dorsey ER, Venuto C, Venkataraman V, Harris DA, Kieburtz K. Novel Methods and Technologies for 21st-Century Clinical Trials A Review. JAMA Neurol. 2015; 72(5), 582–588. doi:10.1001/jamaneurol.2014.4524
2. Gagne Joshua J, Thompson Lauren, O'Keefe Kelly, Kesselheim Aaron S. Innovative research methods for studying treatments for rare diseases methodological review. BMJ. 2014; 349, 6802.
3. Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. Int J Epidemiol. 1999; 28(2), 319-26.
4. Clare Rutterford, Andrew Copas, Sandra Eldridge. Methods for sample size determination in cluster randomized trials, International Journal of Epidemiology. 2015; 44, Issue 3 (1), 1051–1067.
5. Roy A, Bhaumik DK, Aryal S, Gibbons RD. Sample size determination for hierarchical longitudinal designs with differential attrition rates. Biometrics. 2007; 63(3), 699-707.
6. Cai C, Wang S, Lu W, Zhang J. NPHMC: an R-package for estimating sample size of proportional hazards mixture cure model. Comput Methods Programs Biomed. 2014; 113(1), 290-300.
7. A. H. Feiveson. Power by simulation. The Stata Journal. 2002, 2, 107–124.
8. John Eng. Sample Size Estimation: A Glimpse beyond Simple Formulas. RADIOLOGY. 2014. doi:10.1148/radiol.2303030297
9. Gastañaga VM, McLaren CE, Delfino RJ. Power calculations for generalized linear models in observational longitudinal studies: a simulation approach in SAS. Comput Methods Programs Biomed. 2006; 84(1), 27-33.
10. Zhang F, Wagner AK, Ross-Degnan D. Simulation-based power calculation for designing interrupted time series analyses of health policy interventions. J Clin Epidemiol 2011; 64(11), 1252-61.
11. Stephen A. Lauer, Ken P. Kleinman, Nicholas G. Reich. The Effect of Cluster Size Variability on Statistical Power in Cluster-Randomized Trials. PloS one. 2015. 10.1371/journal.pone.0119074

12. Benjamin F Arnold, Daniel R Hogan, John M Colford Jr and Alan E Hubbard. Simulation methods to estimate design power: an overview for applied research. BMC Medical Research Methodology. 2017; 11, 94.
13. Hamer M, Gibson EL, Vuononvirta R, Williams E, Steptoe A. Inflammatory and hemostatic responses to repeated mental stress: individual stability and habituation over time. Brain Behav Immun. 2006; 20(5), 456-9.
14. Francis NA, Ridd MJ, Thomas-Jones E, Butler CC, Hood K, Shepherd V, Marwick CA, Huang C, Longo M, Wootton M, Sullivan F; CREAM Trial Management Group. Oral and Topical Antibiotics for Clinically Infected Eczema in Children: A Pragmatic Randomized Controlled Trial in Ambulatory Care. Ann Fam Med. 2017 Mar;15(2):124-130. doi: 10.1370/afm.2038. PMID: 28289111; PMCID: PMC5348229.
15. European medicines agency. Guideline on multiplicity issues in clinical trials. 2017. https://www.ema.europa.eu/en/documents/scientific-guideline/draft-guideline-multiplicity-issues-clinical-trials_en.pdf
16. Food and Drug Administration (FDA). Multiple Endpoints in Clinical Trials Guidance for Industry. 2017. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials-guidance-industry
17. Lei Clifton, Jacqueline Birks, David A. Clifton, Comparing different ways of calculating sample size for two independent means: A worked example. Contemporary Clinical Trials Communications. 2019, Volume 13.
18. Butler CC, Gillespie D, White P, Bates J, Lowe R, Thomas-Jones E, et al. C-Reactive Protein testing to guide antibiotic prescribing for COPD exacerbations. N Engl J Med. 2019;381(2):111–20.
19. Sugimoto, T. and Sozu, T. and Hamasaki, T. A convenient formula for sample size calculations in clinical trials with multiple co-primary continuous endpoints. Pharmaceut. Statist. 2012; 11: 118-128. doi:10.1002/pst.505
20. Matthias Kohl and Srinath Kolampally. mpe: Multiple Primary Endpoints. 2017. R package version 1.0.
21. Hamasaki T, Evans SR, Asakura K. Design, data monitoring, and analysis of clinical trials with co-primary endpoints: A review. J Biopharm Stat. 2018;28(1):28-51. doi: 10.1080/10543406.2017.1378668.
22. Alan A Montgomery, Tim J Peters, Paul Little. Design, analysis and presentation of factorial randomised controlled trials BMC Medical Research Methodology. 2003; 3, Number 1.
23. Shao-Hsien Liu, Christine M Ulbricht, Stavroula A Chrysanthopoulou, Kate L Lapane. Implementation and reporting of causal mediation analysis in 2015: a systematic review in epidemiological studies. BMC Research Notes. 2016; 9, 354. doi:10.1186/s13104-016-2163-7
24. Colin F. Camerer, Anna Dreber, Felix Holzmeister, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour. 2018; 2, pages 637–644.

Table 1. Study design, primary outcome and models for simplified and simulation based sample size calculation

	Study design	Phase	Primary outcome	Hypothesis testing	Simplified sample size calculation	Simulation based sample size calculation*
Case 1. Baseline/follow-up correlation	Individual randomised controlled trial	Phase III	The primary outcome is the POEM score at 2-week follow-up, which was also measured at baseline.	The treatment is regarded as successful if it is effective at 2-week follow-up.	Two sample t-test, ignoring the baseline/follow-up correlation.	Regression approach with POEM score at 2-week follow-up as the outcome measure, adjusting for its baseline measurement.
Case 2. Multiple primary endpoints	Individual randomised controlled trial	Phase III	Multiple primary endpoints, i.e., the POEM score measured at 2-week, 4-week and 3-month follow-up.	The treatment is regarded as successful if it is effective in at least one of the three primary endpoints.	Directly splitting alpha for t-test at each primary endpoint, ignoring the correlation among three primary endpoints.	Simulation on union-intersection test with correlation among primary endpoints accommodated.
Case 3. Multiple co-primary endpoints	Individual randomised controlled trial	Phase III	Multiple co-primary endpoints, i.e., POEM score measured at 2-week, 4-week and 3-month follow-up.	The treatment is regarded as successful if it is effective in all three co-primary endpoints.	Directly multiplying power for t-test at each co-primary endpoint, ignoring the correlation among three co-primary endpoints.	Simulation on intersection-union test with correlation among co-primary endpoints accommodated.

*The simulation formulation and simulation codes for each case scenario were placed in the supplementary materials.

Table 2. Power calculation by simplification approach for multiple primary endpoints and co-primary endpoints, along with corresponding powers by simulation approach.

	Sample size	Power by simplification	Power by simulation		
			Weak correlation (rho=0.2)	Moderate correlation (rho=0.5)	Strong correlation (rho=0.8)
Multiple primary endpoints	86	90%	99.6%	98.4%	95.7%
Multiple co-primary endpoints	91	90%	90.5%	92.0%	93.4%

Figure 1. Sample size, power and its split alpha for the case with three primary endpoints and weak correlation.