Sample Size considerations for Seba Aljomaa's RCT Project

Márton Kiss, MD

2024-11-27

Table of contents

Executive summary

Based on the variables discussed for the calculations, taking into consideration the possible effect of rounding, and after conducting a simulation study an appropriate sample size for the study would be between 114 and 171 subjects (total).

Total_Sample_Size	Power
114	0.801
117	0.808
120	0.818
123	0.826
126	0.834
129	0.841
132	0.848
135	0.854
138	0.86
141	0.865
144	0.87
147	0.874
150	0.878
153	0.882
156	0.886
159	0.889
162	0.892
165	0.895
168	0.897
171	0.9

Introduction

Assumptions for sample size considerations

The sample size estimation supposes two sided tests at a global α = 5% significance level (equivalent to one-sided testing with an α = 2.5% significance level). This is in line with the recommendations described in ICH E9 Section 3.5.

"The issue of one-sided or two-sided approaches to inference is controversial and a diversity of views can be found in the statistical literature. The approach of setting type I errors for one-sided tests at half the conventional type I error used in two-sided tests is preferable in regulatory settings. This promotes consistency with the two-sided confidence intervals that are generally appropriate for estimating the possible size of the difference between two treatments."

The general assumptions for the sample size considerations (where not otherwise indicated) are:

- Superiority testing (using two-sided tests)
- Global two-sided level of significance: **5%**
- Power: **80% - 90%**
- Dropout rate: **0%** (not applicable here since the intervention only applied once and takes ~5 minutes)
- Variability of the pain scores: **2.5** points (on a 0-10 point continuous scale)
- Expected value during intervention A (no distraction): **3** points
- Expected value during intervention B (iPAD distraction): 1 point
- Expected value during intervention C (VR distraction): 0 points
- Primary outcome: All comparisons between interventions A, B and C must be statistically significant

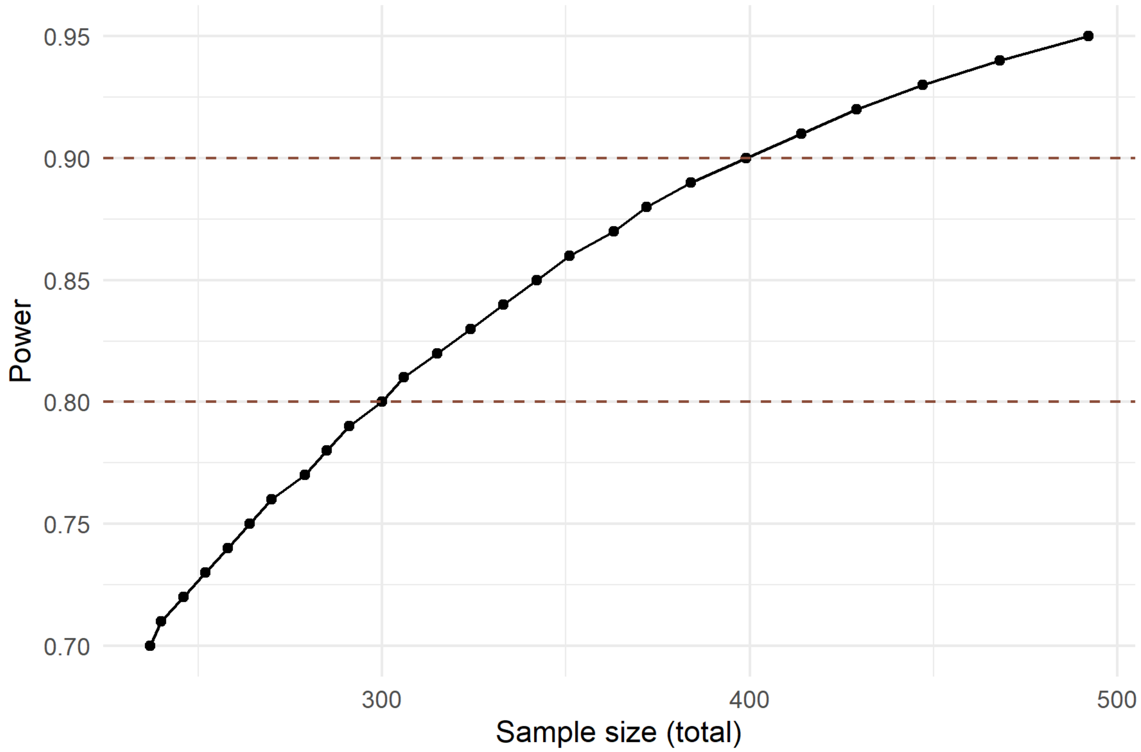
Detailed explanation

Simple approach

As a first approximation, it could be assumed, that the smallest expected difference (B vs. C, 1 point) would drive the sample size. in that case the sample size could be determined as a simple calculation for the difference between 2 means using an approach based on a t-distribution. The results indicate that **100** patients **per arm** are recommended or **300** subjects total.

```
pwrss::pwrss.t.2means(mu1 = 1, mu2 = 0, margin = 0, sd1 = 2.5,
    sd2 = 2.5, welch.df = TRUE, power = 0.8, alternative = "greater",
    alpha = 0.05/2, verbose = TRUE)

 Difference between Two means
 (Independent Samples t Test)
 H0: mu1 = mu2
 HA: mu1 > mu2
 ------------------------------
  Statistical power = 0.8
  n1 = 100
  n2 = 100
 ------------------------------
 Alternative = "greater"
 Degrees of freedom = 198
 Non-centrality parameter = 2.83
 Type I error rate = 0.025
 Type II error rate = 0.2
```



Effect of not evaluating the difference between the two distraction methods

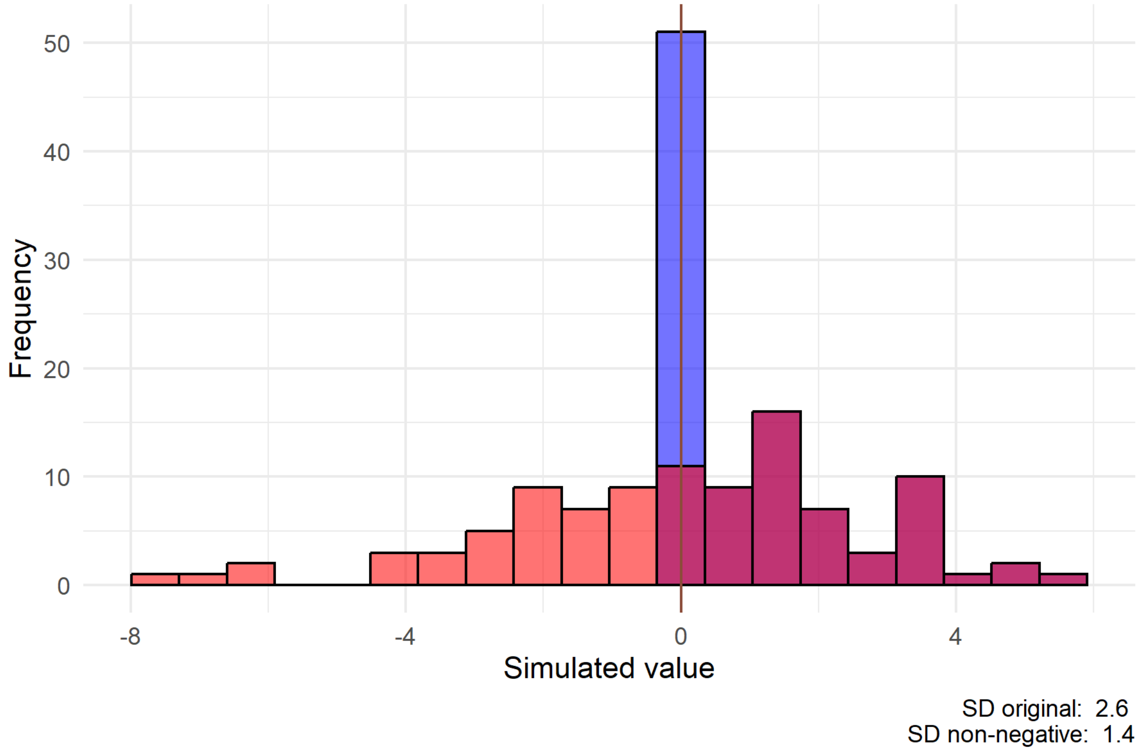
To note, if the primary endpoint would be stated as aiming to establish difference between arms A & B **and** A & C only (omitting the comparison between B & C for the primary endpoint), the required sample size would shrink considerably to **26** patients per arm or a total sample size of **78** for 80% Power.

```
pwrss::pwrss.t.2means(mu1 = 3, mu2 = 1, margin = 0, sd1 = 2.5,
    sd2 = 2.5, welch.df = TRUE, power = 0.8, alternative = "greater",
    alpha = 0.05/2, verbose = TRUE)

 Difference between Two means
 (Independent Samples t Test)
 H0: mu1 = mu2
 HA: mu1 > mu2
 ------------------------------
  Statistical power = 0.8
  n1 = 26
  n2 = 26
 ------------------------------
 Alternative = "greater"
 Degrees of freedom = 50
 Non-centrality parameter = 2.88
 Type I error rate = 0.025
 Type II error rate = 0.2
```

Effects of the scale

The 0-10 point scale introduces a problem since if we suppose that the expected value of intervention C is 0 points, its variability would be 0 which is highly unlikely. A possible correction would be to recode all negative simulated values to 0. This would result in a smaller apparent variability in arm C, but it would still capture the essence of the problem. An example is shown below, where for 100 simulated values, the negative values of the original (red) distribution are recoded to 0 (blue; overlapping sections are purple). This approach was used in the simulations described later.

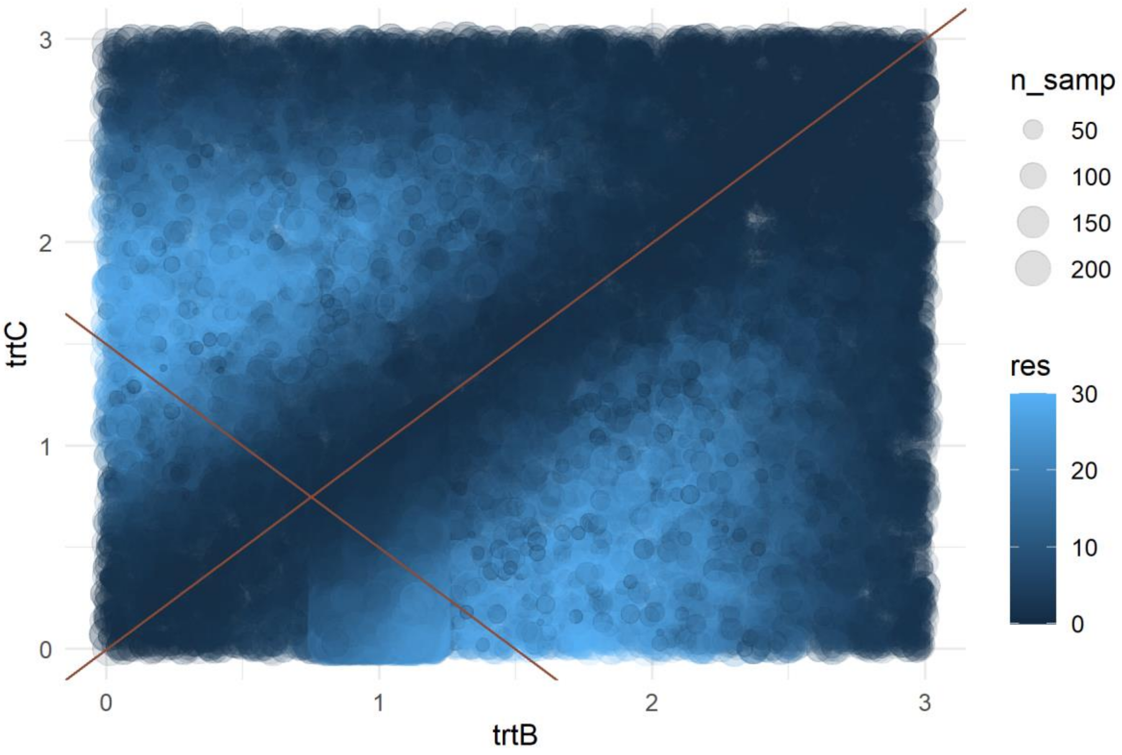


Simulation results (taking into consideration of the rounding effects)

Interdependence of the expected values of the treatment arms

The simulation study was conducted varying all appropriate parameters to some degree. Evaluating three comparisons simultaneously is a complex task, as the expected values of the treatment arms are interdependent. The results of the simulation study from this perspective are shown in the figure below where 'bubbles' are colored according to how often they resulted in a successful study (out of 30 tries). The size of the bubble is proportional to the number of subjects in the study. The results indicate that if (at least) two treatment arms had a similar expected value, the chances for success went down considerably.

This highlights the importance of the supposed differences between the treatment arms.



Impact of variability

The impact of the change in the variability of the results on the Power of the study is highlighted below. In case the sample size is chosen to be **>=137**, the study's power would be robust against the variability being >=3 points.

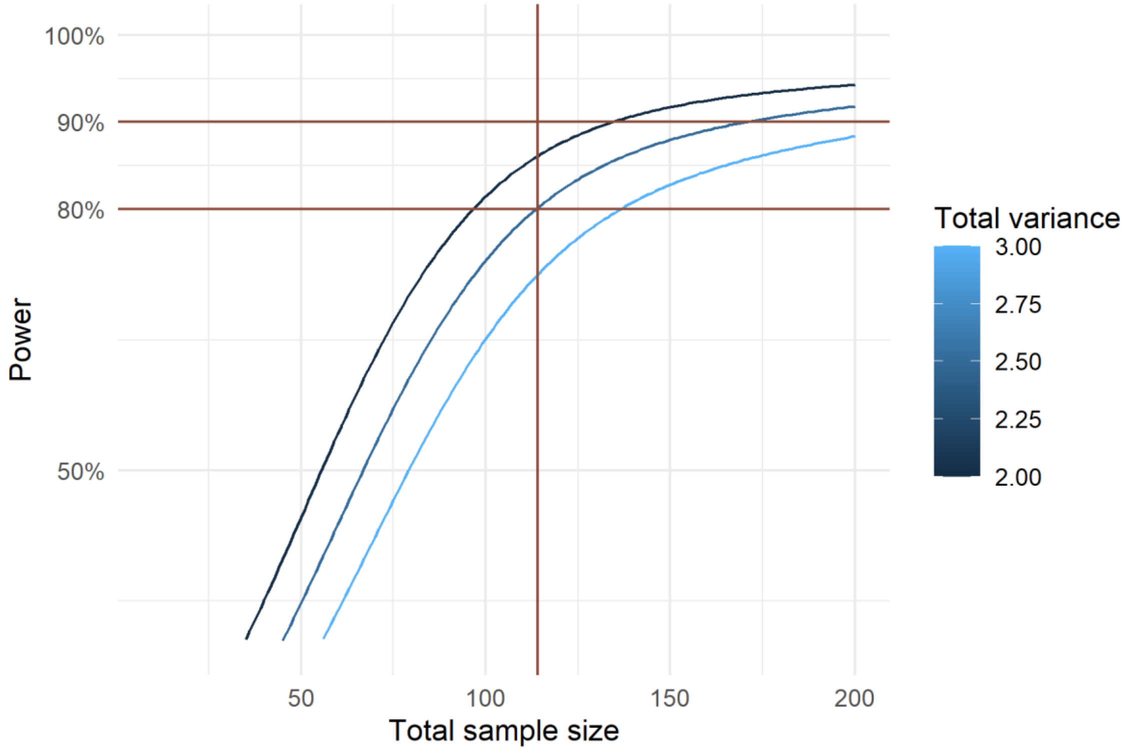
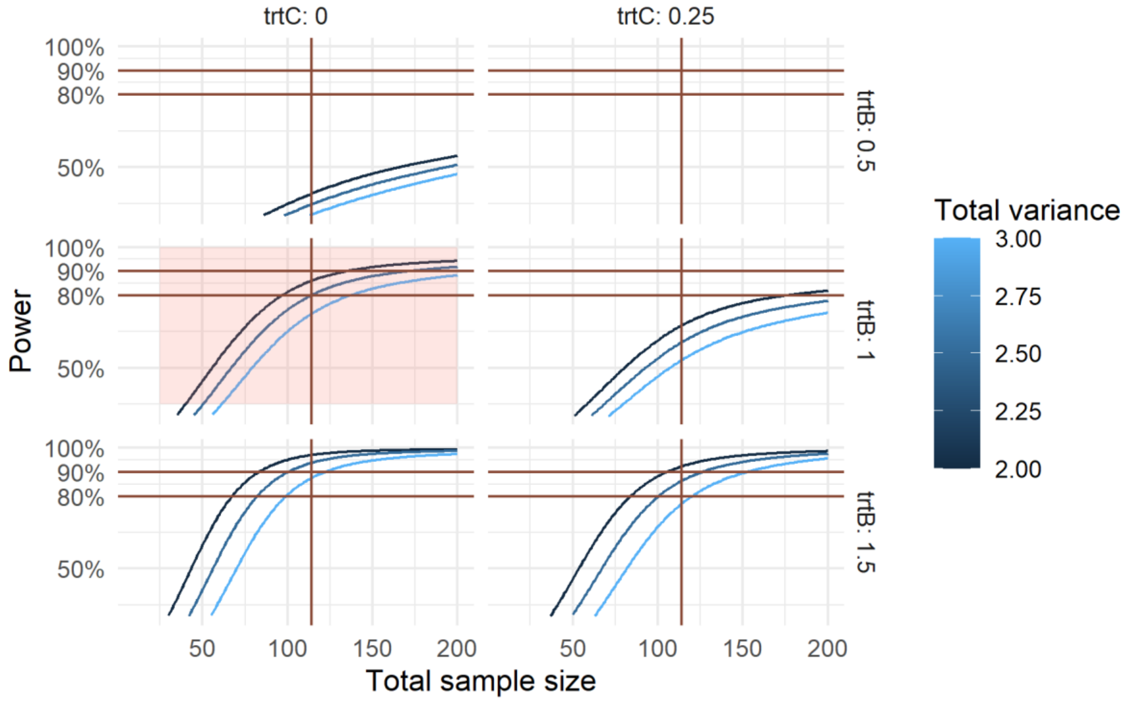


Figure 4

Impact of the expected values of the treatment arms

The impact of different expected treatment effects is shown below. The results highlight the impact of a higher-than-expected mean value in arm C being unfavorable, and the impact of a lower-than-expected mean value in arm B being unfavorable, and the impact if the mean value in arm B would be greater than expected (this being favorable).



Red emphasis put on the subplot presented above

Remarks

MD5 checksum of the database used

Other information regarding the document's compilation

Analyses were conducted using the R Statistical language (version 4.4.1; R Core Team, 2024) on Windows 11 x64 (build 22631), using the packages lubridate (version 1.9.3; Grolemund G, Wickham H, 2011), DHARMa (version 0.4.7; Hartig F, 2024), ggplot2 (version 3.5.1; Wickham H, 2016), dplyr (version 1.1.4; Wickham H et al., 2023) and kableExtra (version 1.4.0; Zhu H, 2024).

References

- *Grolemund G, Wickham H (2011). "Dates and Times Made Easy with lubridate." Journal of Statistical Software, 40(3), 1-25. https://www.jstatsoft.org/v40/i03/.*
- *Hartig F (2024). DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. R package version 0.4.7, https://CRAN.R-project.org/package=DHARMa.*
- *R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.*
- *Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.*
- *Wickham H, François R, Henry L, Müller K, Vaughan D (2023). dplyr: A Grammar of Data Manipulation. R package version 1.1.4, https://CRAN.R-project.org/package=dplyr.*
- *Zhu H (2024). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.4.0, https://CRAN.R-project.org/package=kableExtra.*

Time of compilation

2024-11-27 11:37:41.448943