Robin Boudry, Maarten Rahier, Tom Schipper, Laurens van Paemel. Groep 9  
  
**Project Prinstat 2018-2019  
Part 2 : Simulation study**

**Data cleaning.** No data were missing, but one observation of a diastolic blood pressure of 66 was considered unlikely and this patient was removed from the dataset. To ensure that all investigated patients were sufficiently exposed to the treatment, only participants with an adherence of at least 80% were selected for analysis.[[1]](#footnote-1) This led to a further removal of 29 patients from the dataset.

**Power comparison.** Two variables, the diastolic blood pressure measured at the last visit (denoted *dbp\_end*) and the change in blood pressure from baseline (denoted *dbp\_dif)*, were investigated for the first treatment arm and the placebo arm, which consisted of 39 and 44 patients. Although the normal QQ-plots of the variables revealed deviations from normality, the sample sizes were considered large enough for the Central Limit Theorem to apply and to assume that the sample means are normally distributed. The F-ratio test was only suggestive of unequal variances, but we chose the Welch two-sample t-test to make sure all assumptions were fulfilled in the simulations. The test was applied one-sided, with alternative hypothesis , where is the mean of treatment group and that of the placebo group.

For each of the two variables, the power of the t-test was estimated via 10 000 Monte Carlo simulations. The treatment and placebo groups were simulated by random samples from a normal distribution with the mean and standard deviation of the respective group and were then compared via the Welch t-test. The power was determined as the percentage of p-values below 0.05 in 10 000 tests. Because simulation with the original means yielded a power of 1, the simulations were repeated with delta () simulated in two ways. **A:** as a percentage of the original difference in mean in the data for each variable. **B:** as the same difference in means for both variables.

|  |  |
| --- | --- |
| ***A.***  *Figure 1: power of the Welch t-test when applied to the variables* dbp\_dif *(blue) and* dbp\_end *(red) for multiple differences in means (as a percentage of the observed difference in the data) between the two treatment arms.* | ***B.***    *Figure 2: power of the Welch t-test when applied to the variables* dbp\_dif *(blue) and* dbp\_end *(red) for multiple differences in means.* |

**A.** Figure 1 was created by 41 series of simulations, with the difference in means of each variable starting at the observed difference and decreasing by 2.5% each series. The plot shows that both variables have a similar power close to 1 as long as the difference in means is more than 75% of its original value. They converge to a power of 0.05, corresponding to the significance level of 0.05, when there is no difference in means. However, when the means differ 20% - 75% of their observed differences, the test using the variable *dbp\_end* has more power than the test applied to *dbp\_dif*.

**B.** Figure 2 was created by 60 series of simulations, with the difference in means starting at minus 12 and converging to 0 with 0.2 each series. This figure shows that *dbp\_dif* has more power in detecting a difference for a certain range of Delta values. We will explain this in the following.

**Analytical approach to the observed power findings.** To explain the discrepancy in power of the Welch T-test the following power function was used:

Where , and the non-centrality parameter is defined as:

For **A** this function was applied to a difference in means () decreased to 42.5% of the observed difference, as simulating with this led to the largest difference in power. The power is 70% for *dbp\_dif* and 74% for *dbb\_end*.

As , and are fixed, the difference in power can either be due to a different or different values of and for the variables. As a higher increases the power, the variable with the largest and smallest variances should give the highest power. In this case, *dbp\_dif* has smaller values for and , but *dpb\_end* has a larger . The larger here has outweighed the smaller variances, leading to a higher power for *dbp\_end*.

Following the previous statement, we would expect to find a higher power for *dbp\_dif* when comparing absolute differences (**B**), because of the smaller variance. When using the same method as above mentioned, this is exactly what we found.

**Testing strategy using the lowest p-value.** A testing strategy was assessed which considers the treatment’s effect to be significant whenever either the test on *dbp\_end* or the test *dbp\_dif* yields a p-value below 5%. To calculate the Type 1 error rate for this strategy, 10 000 Monte Carlo simulations were performed for both variables simultaneously. A baseline blood pressure variable *dbp\_base* and the final blood pressure variable *dbp\_end* were each simulated by taking samples from two normal distributions representing the treatment and the placebo group. In order to simulate the null hypothesis, the means of these groups were set equal, while they retained their original standard deviation. The variable *dbp\_dif* was then created as the difference between *dbp\_base* and *dbp\_end* to simulate dependency, as these variables are related in real patients. The difference between the treatment and placebo group was then tested for both variables via the one-sided Welch two-sample t-test.

The Welch t-test is Type 1 error robust, meaning that it controls the Type 1 error rate at 5 %. As a result, the when evaluating the two variables separately, around 5% of the simulations under H0 had resulted in a p-value smaller than 0.05. However, with the assessed testing strategy, a Type 1 error is made whenever either measurement under H0 yields a p-value below 0.05, resulting in an increased Type 1 error rate of approximately 0.077

**Correcting for multiple comparisons.** When confronted with the multiple testing problem, the question can be asked if it is possible to solve the inflation of the type I error. One possible option is to control the family-wise error rate (FWER), the risk of a Type I error when all null hypotheses hold, by using the Bonferroni correction. This method relies on Boole’s inequality:

\*m

With as the significance level (0.05 in this case) and m as the number of simultaneously performed tests (2 in this case). To avoid that the Type I error rate exceeds , the p-value of each test can be multiplied by 2.

The simulation of the testing strategy using the lowest p-value under the null hypothesis was repeated, but this time the lowest p-value of the two was multiplied by 2. This correction reduced the Type I error rate to approximately 0.039. The alternative testing strategy with Bonferroni correction was also applied to the original data to investigate the power of this strategy. The strategy had a power of 0.8045, far lower than the power of practically 1 for the original tests. This could partly be explained by the fact that the Bonferroni correction is conservative, reducing the Type I error rate more than strictly necessary and thereby also reducing the power. Probably more important in this case is that the two variables are related, while the Bonferroni correction assumes independence between the tests. Because our variables and therefore the corresponding p-values are related, a Type I error in one variable increases the likelihood of a Type I error in the other. As double Type I errors are not counted twice when only the lowest p-value is used, the Type I error rate is lower than expected for independent variables.

1. 1 Valgimigli M, Garcia Garcia H, Vrijens B *et al*. Standardized classification and framework for reporting, interpreting, and analysing medication non-adherence in cardiovascular clinical trials: a consensus report from the Non-adherence Academic Research Consortium (NARC). *European Heart Journal* 2018; **00**: 1-16. [↑](#footnote-ref-1)