MRI-guided transurethral ultrasound ablation (TULSA) for benign prostatic obstruction: 12-month prospective clinical outcomes

Detailed statistical methods

Contents

Statistical methods

Here, we give a more detailed discussion and rationale for statistical tests and methods chosen for statistical analysis in this study.

As a preliminary phase I study for MRI-guided transurethral ultrasound ablation (TULSA) in treatment of benign prostate obstruction (BPO), our aim was to investigate the safety and feasibility of the method with a small sample of subjects in a follow-up study of 12 months. The following issues, common in early phase clinical studies, guided our choices in designing the statistical analyses:

1. Small sample size
2. Unknown distribution of measured variables
3. The measured variables are not independent (e.g. some questionnaires measure partly similar things, different flow parameters are also not independent etc.)

Choosing statistical test

Student's t-test is probably the most commonly used statistical test in medical literature when comparing the difference between two groups. The principles of Student's t-test could be used in time series data using a paired test. However, issues 1 and 2 in our data conflict with the assumptions of Student's t-test. Thus, we chose to use Wilcoxon signed rank test as a non-parametric paired test suitable for time series data.

The null hypothesis in our study is that TULSA does not have any effect on the parameters measured. Comparison between baseline and 12-month follow-up (the endpoint of follow-up period) was considered most relevant to show possible deviations from the null hypothesis, and thus the statistical tests were evaluated between these two timepoints.

In theory, Wilcoxon signed rank test assumes the variables to be continuous. This assumption may be violated in some questionnaire data, thus we used continuity correction (parameter "correct = TRUE" in R function wilcox.test) for questionnaire data, but not for physiological measurements such as flow parameters, prostate volume or PSA ("correct = FALSE").

Adjusting p-values for multiple comparisons

A well-known issue of making multiple statistical tests is the increasing risk of false positives when the number of tests increases. The issue is most serious in omics-type of data, often demanding thousands of tests. As the traditional significance level is defined to be at $p < 0.05$, this leads to on average every 20 statistical test to give false positives finding in case of no real differences in the measured variables. In our study, 13 tests were made and in order to minimize false positives (or type I errors), we conducted a p-value adjustment for multiple comparisons.

Common procedures used for multiple comparison adjustment of significance levels include a rather conservative Bonferroni's method (Bonferroni 1936), a slightly improved Šidàk's method (Šidàk 1967) which both focus on controlling the probability of observing *at least one significant difference* in case of true null hypothesis, or algorithms focusing on controlling false discovery rate, such as Benjamini-Hochberg correction (Benjamini & Hochberg 1995).

All the abovementioned methods have assumptions on independence of variables. In our case, assuming independent variables would be rather brave, as many of questionnaire scores come from partly similar questions, and e.g. flow parameters most likely correlate with each other. In an attempt to take the supposedly complex correlation structure of our data into consideration, we decided to build a permutation-based correction for multiple comparisons that is sometimes considered a golden standard for multiple test adjustments with correlating set of tests (Conneely & Boehnke 2007). Our approach is somewhat similar to what Westfall & Young introduced as the Algorithm 2.5 in their textbook on resampling-based multiple comparison (Westfall & Young 1993).

Permutation-based method for multiple comparison adjustment

The statistical comparisons are calculated between baseline and 12 month follow-up measurements. Eight patients completed the follow-up. In permutation-based (sometimes referred to as resampling) method, the goal is to create a *null distribution* of p-values similar to what would be created from a data that is generated from similar distribution to actual measurements. To respect the "exact" distribution of the measurements, this can be achieved by shuffling timepoint-labels of measurements from each patient and running the same statistical tests as are done in the actual data analysis. As the variables are considered to be dependent from each other, the correlation structure is maintained by not shuffling the time labels randomly for each variable, but for each patient only.

More formally: For a follow-up study of t timepoints, let us have datasets $X_1, X_2, ..., X_t$. Each dataset consists of n patients and m measurements, yielding a matrix sized $m \times n$. Wilcoxon signed rank test may be applied to a pair of these datasets (e.g. X_1, X_t). In each of $k \leq ((t(t-1))^n$ rounds, following steps are made:

1. Generate a random dataset pair $X_{a,b}$ by choosing a random pair of timepoints for each patient independently. The pair of timepoints are not required to be in chronological order, but duplicate pairs ($X_{a,a}$) are not allowed
2. Calculate statistical tests and p-values for each variable using $X_{a,b}$
3. Save p-values as a vector p_r (where $r \in [1, k]$) in a $k \times m$ matrix collecting all p-values generated during the process

This random process will generate a null distribution of p-values from a symmetric data. In other words, there will be as many (or, approximately as many, if $k < ((t(t-1))^n$, such as when working with large datasets that require optimization of computing time) "changes up" as "changes down" due to no restrictions on the chronological order of the randomized timepoints. To control type 1 error, next step will be to calculate at which significance level α no more than 5% of the k rounds find *at least one* significant finding at random. This may be calculated by finding the minima of each p-value vector p_r and finding the 5th percentile of these minima. Formally, if $Q_{min(p_r)}$ is the quantile function of minima of p-value vectors, then $\alpha = Q_{min(p_r)}(0.05)$.

In our study, the computational time was not a challenge for going through all possible permutations ($k = 2^8 = 256$ in total). Other constants in our analyses, as defined above, were $n = 8$, $m = 18$ and $t = 2$.

R code used in analysis

All R codes used in statistical analysis and building figures for this article are open access in GitHub:

`https://github.com/topihovinen/tulsastudy/`

References

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B 1995;57(1):289â300.

Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilitÃ . *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 1936.

Conneely KN, Boehnke M. So many correlated tests, so little time! Rapid adjustment of p-values for multiple correlated tests. The American Journal of Human Genetics. 2007;81(6):1158-68.

ŠidÃ¡k ZK. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. Journal of the American Statistical Association 1967; 62(318):626â633.

Westfall PH, Young SS. 2.3. Single-step Methods. In *Resampling-based multiple testing: Examples and methods for p-value adjustment*. 1993:43-53.