Comparison of statistical methods for analysis of small sample sizes for detecting the differences in efficacy between treatments for knee osteoarthritis

Chia-Lung Shih and Te-Yu Hung *

Department of Pediatrics, Chi-Mei Medical Center, Tainan 901, Taiwan

* Correspondence: d97241002@ntu.edu.tw
ABSTRACT

Background A small sample size (n < 30 for each treatment group) is usually enrolled to investigate the differences in efficacy between treatments for knee osteoarthritis (OA). The objective of this study was to use simulation for comparing the power of four statistical methods for analysis of small sample size for detecting the differences in efficacy between two treatments for knee OA.

Methods A total of 10,000 replicates of 5 sample sizes (n=10, 15, 20, 25, and 30 for each group) were generated based on the previous reported measures of treatment efficacy. Four statistical methods were used to compare the differences in efficacy between treatments, including the two-sample t-test (t-test), the Mann-Whitney U-test (M-W test), the Kolmogorov-Smirnov test (K-S test), and the permutation test (perm-test).

Results The bias of simulated parameter means showed a decreased trend with sample size but the CV% of simulated parameter means varied with sample sizes for all parameters. For the largest sample size (n=30), the CV% could achieve a small level (<20%) for almost all parameters but the bias could not. Among the non-parametric tests for analysis of small sample size, the perm-test had the highest statistical power, and its false positive rate was not affected by sample size. However, the power of the perm-test could not achieve a high value (80%) even using the
largest sample size (n=30).

**Conclusion** The perm-test is suggested for analysis of small sample size to compare the differences in efficacy between two treatments for knee OA.

**Keywords:** knee osteoarthritis, Kolmogorov-Smirnov test, Mann-Whitney U-test, permutation test, simulation.
# Background

Osteoarthritis (OA) is a degenerative joint disease and is most commonly located in the knee. Knee OA is a highly prevalent disease afflicting the elderly, and a 44% lifetime risk has been reported in American adults.[1] The degenerative process in the knee joint can lead to joint pain, stiffness, and swelling which could negatively affect physical function and life quality. Several nonsurgical treatments have been developed to ease the symptoms of knee OA such as autologous adipose tissue injections, hyaluronic acid injections, platelet-rich plasma treatment, and laser therapy.[2-4]

A randomized controlled clinical trial is generally designed to compare the differences in efficacy between treatments for knee OA.[5] However, a small sample size (n < 30 patients for each treatment group) is usually enrolled to examine the differences,[6] which might be due to lack of human and financial resources. A small sample size is associated with low statistical power of the study, which could produce inconclusive results. It is important to investigate the effects of statistical methods for analysis of small sample size on the power of detecting clinical differences in efficacy between two treatments for knee OA.

Simulation studies have been adopted to evaluate the accuracy and
appropriateness of various statistical methods under an assumption of the known validity.[7] Due to the invention and progression of computers, simulation studies have been increasingly used in medical research.[8, 9] To compare statistical methods for analysis of small sample size for detecting the differences in efficacy between treatments for knee OA, simulation methods were adopted in this study.

The objective of this study was to compare different statistical methods for analysis of small sample size for detecting the differences in efficacy between two treatments for knee OA. A simulation method was adopted to conduct this research, and the results could suggest a better statistical method for analysis of small sample size to increase statistical power.

Methods

Parameters

To generate data sets based on an assumption of known parameters, these parameters were collected from the previous studies that investigated the differences in efficacy between treatments for knee OA.

Magnetic resonance imaging (MRI)

The changes of cartilage volume, thickness of the synovial membrane, and synovial fluid volume before and after treatments measured by MRI are a relatively new
A technique to assess the efficacy of treatments for knee OA. These parameters were collected from a previous study of assessing the differences in efficacy between treatments for knee OA (ClinicalTrials.gov NCT01354145).[10] It was a randomized double-blind, double-dummy, controlled trail with a maximum follow-up of two years. Two patient groups were treated with chondroitin sulfate and celecoxib respectively. Eight parameters were measured by MRI that presented the mean changes between baseline and 24 months follow-up were adopted in this study, including four parameters that showed a significant difference or a borderline level between treatments and four parameters that showed no significant difference for the use of negative controls in this study (Table 1).

**Self-report Index**

The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) score is commonly adopted to assess efficacy of knee OA after treatment. WOMAC score contains three kinds of index, including pain index, stiffness index, and function index, and is measured from a patient-reported outcome. These parameters were collected from a randomized, controlled, single-blind study (IRCT2016071513442N11).[11] Two patient groups were treated with plasma rich in growth factor and hyaluronic acid injections respectively, and WOMAC score changes between baseline and six months follow-up were measured (Table 2).
**Statistical methods**

To investigate the effects of statistical methods for analysis of small sample size on statistical power, four methods were considered, including the two-sample t-test (t-test), the Mann-Whitney U-test (M-W test), the Kolmogorov-Smirnov test (K-S test), and the permutation test (perm-test).

**T-test**

The t-test is used to examine whether two population means differ.\[12\] The assumptions of this test are that both populations follow a normal distribution and have equal variances in standard deviation. The formula of the t-test is defined as:

\[
T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}
\]

where \(\bar{x}_1\) and \(\bar{x}_2\) are the sample means, \(s_1\) and \(s_2\) are the sample standard deviations, and \(n_1\) and \(n_2\) are the sample sizes.

The null hypothesis is rejected at the level \(\alpha\):

\[|T| > t_{1-\alpha,\nu}\]

where \(t_{1-\alpha,\nu}\) is the critical value of the \(t\) distribution with \(\nu\) degrees of freedom.

**M-W test**

The M-W test is a non-parametric test used for comparing the difference of two
population medians.[13] This test is often taken as an alternative to the t-test when
the samples do not follow a normal distribution. Test statistic $U$ is defined as:

$$U_1 = \sum R_1 - \frac{n_1(n_1 + 1)}{2}$$

$$U_2 = \sum R_2 - \frac{n_2(n_2 + 1)}{2}$$

where $\sum R$ is the observed rank sum and $n$ is the sample size for each group. If $U_1$ is
smaller than $U_2$ and the null hypothesis is rejected at the level $\alpha$:

$$|Z| = \frac{U_1 - u_u}{s_u} > z_{1-\alpha,v}$$

$$u_u = \frac{n_1n_2}{2}$$

$$s_u = \sqrt{\frac{n_1n_2(n_1 + n_2 + 1)}{2}}$$

where $u_u$ is the mean of all observed ranks, $s$ is the standard deviation of all
observed ranks, and $z_{1-\alpha,v}$ is the critical value of the $z$ distribution with $v$ degrees of
freedom.

**K-S test**

The K-S test is a non-parametric test used for comparing the empirical distribution
functions between two populations. The statistic is defined as:

$$D_{n,m} = \sup_x |F_1(x) - F_2(x)|$$

where $\sup$ is the supremum function, and $F_1$ and $F_2$ are the empirical distribution
functions for each sample group.
The null hypothesis is rejected at the level \( \alpha \):

\[
D_{n,m} > c(\alpha) \sqrt{\frac{n_1 + n_2}{n_1 n_2}}
\]

\[
c(\alpha) = \sqrt{-\frac{1}{2} \ln \alpha}
\]

where \( n_1 \) and \( n_2 \) are the sample sizes for each group.

**Perm-test**

The perm-test is a non-parametric test used for comparing if two populations come from the same distribution. The concept of this test is that the test statistic distribution that is obtained by calculating all possible values of the test statistic under random rearrangements of the labels on the observed sample groups is compared with the real test statistic that is calculated from the observed sample groups.

The procedure of the perm-test is as follows. In the first step, the mean difference of two groups \( T_{obs} \) is calculated. In the second step, the samples of two groups are pooled, and then randomly divided into two groups. The mean difference of two permutation groups \( T_{perm} \) is calculated. The second step is repeated 1000 times and 1000 values of \( T_{perm} \) are obtained. Finally, the two-sided \( p \)-value is determined by calculating the proportion of sampled permutations where these absolute values of \( T_{perm} \) are greater than or equal to the absolute value of \( T_{obs} \).
Simulation procedure

To investigate how statistical methods for analysis of small sample size affect the power of assessing knee OA treated with two methods, 10,000 replicates of five sample sizes (n=10, 15, 20, 25, and 30 for each group) were generated respectively. The data were generated from a normal distribution of the parameters that had means and standard deviations collected from the previous studies (Tables 1 and 2).[10, 11] The generated data between two treatments for each replicate were tested by the t-test, the M-W test, the K-S test, and the perm-test respectively. The percentage of 10,000 replicates that had a significant difference (p-value < 0.05) was defined as statistical power for p1-p4 and a false positive rate for p5-p8.

A coefficient of variance (CV) of simulated parameter means over 10,000 replicates was calculated as follows:

$$\text{CV} \% = 100 \% \times \left| \frac{sd}{\bar{p}'} \right|$$

where sd is the standard deviation of the simulated parameter means over 10,000 replicates, and $\bar{p}'$ is the mean of simulated parameter means over 10,000 replicates.

The bias of the simulated parameter mean was calculated as the following formula:

$$\text{bias}(\%) = 100 \% \times \left( \frac{p' - p}{p} \right)$$

where $p'$ is the simulated parameter mean from the simulation and p is the
parameter mean from the original study.

Data analysis was conducted using R software version 3.31 (R Development Core Team 2016).

Results

MRI parameters

The 8 MRI parameters were adopted in this simulation study, including four (p1-p4) with significant difference or a borderline level between treatments and four (p5-p8) without significant difference in the original study (Table 1). The effect size of p1-p4 (≥ 0.48) was obviously larger than that of p5-p8 (≤ 0.23) (Table 1). In sensitivity analysis for sample size, 10,000 replicates of larger sample sizes (n=50, 100, 150, 200, 250, and 300 for each group) were simulated, and the difference between treatments was examined by the t-test. For p1, p2, p3, p4, and p6, the mean, median, and variance of p-values decreased obviously with sample size (Figure S1). In contrast, the mean, median, and variance of p-values seemed not to be obviously affected by sample size for p5, p7, and p8.

In analysis for smaller sample sizes (n=10, 15, 20, 25, and 30 for each group), the 95% confidence interval (CI) of bias of simulated parameter means over 10,000 replicates varied with sample sizes, and no decreased trend with sample size was
observed for all parameters (Figure 1). The CV% of simulated parameter means over 10,000 replicates decreased with sample size for the eight parameters (Figure 2). For the largest sample size (n=30), the CV value could achieve a small level (<20%) except for p3, p7, and p8.

The power of the four statistical methods for analysis of parameters p1-p4 showed an increased trend with sample size (Figure 3); however, the power was not over 80% even using the largest sample size (n=30) and varied with parameters. The parameters p1 and p2 had relatively low power. In comparison of power among the four statistical methods, the perm-test and the t-test had the highest power (Figure 3). The power of the M-W test was lower than the perm-test and the t-test but slightly higher than the K-S test. In comparison of false positive rates (p5, p6, p7, and p8) among sample sizes, the false positive rates examined by the four statistical methods would not increase with sample size except for p6.

Self-report Index

WOMAC score (w4) contains pain index (w1), function index (w2), and stiffness index (w3). The effect sizes of these parameters ranged from 0.27 to 0.68 (Table 2). However, the significant differences of these parameters between treatments were not provided by the original study. In sensitivity analysis for sample size, larger sample sizes were simulated for the four parameters (w1-w4), and the differences in
these parameters between treatments was examined by the t-test. For all these parameters, the mean, median, and variance of p-values obviously decreased with sample size (Figure S2), indicating that these parameters had high probability of significant difference between treatments.

In analysis for smaller sample sizes, the 95% CI of bias of simulated parameter means over 10,000 replicates varied with sample sizes for all parameters (Figure 4). The CV% of simulated parameter means over 10,000 replicates decreased with sample size for all of the parameters (Figure 5). For the largest sample size (n=30), the CV value could reach a small level (< 20%) except for w3 of the PRGF group.

The power of the four statistical methods for examining the parameters w1-w4 showed an increased trend with sample size (Figure 6); however, the power was not over 80% even using the largest sample size (n=30) and it varied with parameters. The parameter w3 had relatively lower power than the other three parameters. In comparison of power among statistical methods, the perm-test and the t-test had similar power and their power had the highest value (Figure 6). The power of the M-W test was lower than the perm-test and the t-test but slightly higher than the K-S test. Interestingly, w1 examined by the K-S test had obviously higher power than the other 3 statistical methods as sample size ≥ 15.

Discussion
Randomized controlled study designs are high level of evidence to determine if one treatment is superior to any other; however, a small sample size is usually enrolled in such design to examine the differences in efficacy between treatments for knee OA. To suggest a better statistical method for analyzing small sample size, we used simulation to compare four statistical methods for analysis of small sample sizes for detecting the differences in efficacy between treatments. The 95% CI of bias of simulated parameter means decreased with sample size; however, CV% of simulated parameter means over 10,000 replicates did not show an obviously decreased trend with sample size. In general, the perm-test and the t-test had the highest power. The power of the M-W test was lower than the perm-test and the t-test but slightly higher than the K-S test. However, the perm-test and the t-test could not raise the power to a high level (80%), even using the largest sample size (n=30).

An acceptable bias (within±10%) has been suggested for parameters estimated.[14] However, the 95% CI of bias of simulated parameter means over 10,000 replicates seriously exceeded this acceptable level even using the largest sample size (n=30). The variation of these parameters was large (CV > 45%) (Tables 1 and 2), and thus these parameters were seriously over- or underestimated using small sample sizes (n≤30). On the other hand, a CV% value is used to assess the precision of a replication experiment.[15] The CV% value of estimated parameters is
expected to decrease with sample size. However, the CV% value varied with sample
size in this study, which might have resulted from the inherent variation in these
parameters. A CV of 20% has been suggested as an accepted level.[15] The CV% of
simulated parameter means could achieve this accepted level using the largest
sample size (n=30) except for p3, p7, p8, and w3 of the RGFA group. A large variation
was inherent in these parameters (Tables 1 and 2), and thus the CV% of simulated
parameter means over 10,000 replicates could not achieve this accepted level as
sample size <30.

Parameter p6 had no significant difference between treatments in an original
study [10] and was taken as a negative control in this study. In sensitivity analysis for
larger sample sizes (n=50, 100, 150, 200, 250, and 300 for each group), the p-values
of p6 showed a decreased trend with sample size (Figure S1). In addition, the effect
size of p6 (0.23) was obviously larger than that of p5 (0.10), p7 (0.04), and p8 (0.05).
These indicate that p6 should have significant difference between treatments and it
could not be taken as a negative control. This could explain why a false positive rate
of p6 examined by the four statistical methods showed a slightly increased trend with
sample size (Figure 3).

WOMAC score is a traditional measure to assess the efficacy of knee OA
patients after receiving treatments. It is a self-reported score and is considered as a
subjective method. WOMAC score is still used in recently researches [16, 17] because it has been validated for assessing the efficacy of knee OA after treatment [18]. In contrast, the knee OA structural changes measured by MRI are a relatively new and objective method for assessing efficacy of treatments. Therefore, the two types of parameters were simultaneously considered in this study, and this could provide more comprehensive information about statistical methods for analysis of small sample size for detecting the differences in efficacy of treatments for knee OA.

The t-test and the M-W test are commonly used to examine the clinical differences in efficacy between two treatments for knee OA [10, 11]. To our best knowledge, the two non-parametric tests (the perm-test and the K-S test) have not been applied to test these clinical differences. The assumptions when conducting the t-test are that both populations follow a normal distribution and have equal variances in standard deviation. When sample size is large (n≥30), the data set is considered to follow a normal distribution and the t-test can be applied. When sample size is small, the data may violate the normality assumption [19], and a non-parametric test is applied instead of the t-test. However, our results discovered that the perm-test could produce the highest power among the three non-parametric methods. A power of 80% is usually expected from statistical tests [11]. However, using the perm-test for analysis of small sample size could not achieve
this level even using the largest sample size (n=30).

There were two research limitations in this study that might affect our conclusion. First, parameter p3 collected from the previous study had large level of effect size (0.83), but it could not achieve the acceptable power (80%) even using the largest sample size (n=30). If parameters between treatments have larger effect size, the power could achieve this acceptable level examined by small sample size. Second, the parameter w1 examined by the K-S test had abnormally higher power than the other three statistical methods as sample size ≥ 15. In comparison with the parameters, w1 had no special effect size or CV% (Tables 1 and 2). To our best knowledge, we could not explain why w1 examined by the K-S test had abnormally high power.

Conclusion

This report used simulation to compare statistical methods for analysis of small sample size for detecting the clinical differences in efficacy between treatments for knee OA. The parameters were seriously over- or underestimated using small sample sizes (n≤30). Two types of measures of treatment efficacy were considered in this study, including MRI measurements and WOMAC score. The simulation produced similar results between the two types of measures. Among the non-parametric tests for analysis of small sample size, we discovered that the perm-test had the highest
power, and its false positive rate was not affected by sample size. However, the power of the perm-test could not achieve a high value (80%) even using the largest sample size (n=30) in this study.

**Abbreviations**

CV: coefficient of variance; HA: hyaluronic acid; K-S test: Kolmogorov-Smirnov test; M-W test: Mann-Whitney U-test; MRI: magnetic resonance imaging; OA: osteoarthritis; perm-test: permutation test; PRGF: plasma-rich in growth factor; t-test: two-sample t-test; WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index;

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Availability of data and material**

Not applicable.

**Competing interests**

The author claims no conflict of interest.

**Funding**

Not applicable.

**Authors’ contributions**

CLS did the data analysis. TYH wrote and revised the manuscript. All authors read and approved the final manuscript.
Acknowledgements

We appreciate the assistance of Jie-Huei Wang, PhD with r code editing.

References

1. Murphy L, Schwartz TA, Helmick CG, Renner JB, Tudor G, Koch G, Dragomir A, Kalsbeek WD, Luta G, Jordan JM. Lifetime risk of symptomatic knee osteoarthritis. *Arthritis Rheum-Arthr* 2008;59(9): 1207-13.

2. Jones IA, Wilson M, Togashi R, Han B, Mircheff AK, Thomas Vangsness C, Jr. A randomized, controlled study to evaluate the efficacy of intra-articular, autologous adipose tissue injections for the treatment of mild-to-moderate knee osteoarthritis compared to hyaluronic acid: a study protocol. *BMC musculoskeletal disorders* 2018;19(1): 383.

3. Ahmad HS, Farrag SE, Okasha AE, Kadry AO, Ata TB, Monir AA, Shady I. Clinical outcomes are associated with changes in ultrasonographic structural appearance after platelet-rich plasma treatment for knee osteoarthritis. *International journal of rheumatic diseases* 2018;21(5): 960-66.

4. Alqualo-Costa R, Thome GR, Perracini MR, Liebano RE. Low-level laser therapy and interferential current in patients with knee osteoarthritis: a randomized controlled trial protocol. *Pain management* 2018;8(3): 157-66.

5. Lee S, Kim SJ. Efficacy and safety of ChondroT on knee-osteoarthritis: Protocol for a 8-week, randomized, double-blind, placebo-controlled, multicenter therapeutic exploratory clinical trial. *Medicine* 2018;97(12): e0170.

6. Kim SH, Ha CW, Park YB, Nam E, Lee JE, Lee HJ. Intra-articular injection of mesenchymal stem cells for clinical outcomes and cartilage repair in osteoarthritis of the knee: a meta-analysis of randomized controlled trials. *Archives of orthopaedic and trauma surgery* 2019.

7. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med* 2006;25(24): 4279-92.

8. Kristman V, Manno M, Cote P. Loss to follow-up in cohort studies: how much is too much? *Eur J Epidemiol* 2004;19(8): 751-60.

9. Vaeth M, Skovlund E. A simple approach to power and sample size calculations in logistic regression and Cox regression models. *Stat Med* 2004;23(11): 1781-92.

10. Pelletier JP, Raynauld JP, Beaulieu AD, Bessette L, Morin F, de Brum-Fernandes AJ, Delorme P, Dorais M, Paiement P, Abram F, Martel-Pelletier J. Chondroitin sulfate efficacy versus celecoxib on knee osteoarthritis structural changes using magnetic resonance imaging: a 2-year multicentre exploratory study.
11. Raeissadat SA, Rayegani SM, Ahangar AG, Abadi PH, Mojgani P, Ahangar OG. Efficacy of Intra-articular Injection of a Newly Developed Plasma Rich in Growth Factor (PRGF) Versus Hyaluronic Acid on Pain and Function of Patients with Knee Osteoarthritis: A Single-Blinded Randomized Clinical Trial. *Clinical medicine insights Arthritis and musculoskeletal disorders* 2017;10:1179544117733452.

12. Snedecor GW, Cochran WG. Statistical Methods. *Eighth Edition, Iowa State University Press* 1989.

13. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 1947;18(1): 50-60.

14. Forero CG, Maydeu-Olivares A. Estimation of IRT Graded Response Models: Limited Versus Full Information Methods. *Psychol Methods* 2009;14(3): 275-99.

15. European_commission. Method validation and quality control procedure for pesticide residues analysis in food and feed. *Document No SANTE 11945/2015* 2015.

16. Su K, Bai YM, Wang J, Zhang HS, Liu H, Ma SY. Comparison of hyaluronic acid and PRP intra-articular injection with combined intra-articular and intraosseous PRP injections to treat patients with knee osteoarthritis. *Clin Rheumatol* 2018;37(5): 1341-50.

17. Duymus TM, Mutlu S, Dernek B, Komur B, Aydogmus S, Kesiktas FN. Choice of intra-articular injection in treatment of knee osteoarthritis: platelet-rich plasma, hyaluronic acid or ozone options. *Knee Surg Sport Tr A* 2017;25(2): 485-92.

18. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation-Study of Womac - a Health-Status Instrument for Measuring Clinically Important Patient Relevant Outcomes to Antirheumatic Drug-Therapy in Patients with Osteo-Arthritis of the Hip or Knee. *J Rheumatol* 1988;15(12): 1833-40.

19. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism* 2012;10(2): 486-89.
**Figure legends**

Figure 1 The 95% confidence interval of bias of simulated parameters over 10,000 replicates for p1-p8.

Figure 2 Coefficient of variation (CV%) of simulated parameters over 10,000 replicates for p1 (1), p2 (2), p3 (3), p4 (4), p5 (5), p6 (6), p7 (7), and p8 (8).

Figure 3 Power of parameters p1 (1), p2 (2), p3 (3), and p4 (4) and false positive rates of parameters p5 (5), p6 (6), p7 (7), and p8 (8) between groups over 10,000 replicates examined by the two-sample t-test (t-test), the Mann-Whitney U-test (M-W test), the Kolmogorov-Smirnov test (K-S test), and permutation test (perm-test).

Figure 4 The 95% confidence interval of bias of simulated parameters (w1-w4) over 10,000 replicates. There were two treatment groups, including plasma rich in growth factor (PRGF) and hyaluronic acid (HA) groups.

Figure 5 Coefficient of variation (CV) of simulated parameters over 10,000 replicates for parameters w1 (1), w2 (2), w3 (3), and w4 (4). There were two treatment groups, including plasma-rich in growth factor (PRGF) and hyaluronic acid (HA) groups.

Figure 6 Power of parameters w1 (1), w2 (2), w3 (3), and w4 (4) between groups over 10,000 replicates examined by the two-sample t-test (t-test), the Mann-Whitney U-test (M-W test), the Kolmogorov-Smirnov test (K-S test), and the permutation test (perm-test).
Table 1 MRI Changes in patients treated for knee OA between baseline and 24 months follow-up collected from the literature [10].

| Parameter | Position | Chondroitin sulfate group | Celecoxib group | p-value | Effect size$^*$ |
|-----------|----------|---------------------------|-----------------|---------|-----------------|
|           |          | mean (SD) | CV% | mean (SD) | CV% |              |
| Cartilage volume loss | Medial compartment (%) | $-6.3$ ($\pm3.2$) | 51 | $-8.1$ ($\pm4.2$) | 52 | 0.018* | 0.48 |
| Cartilage volume loss | Condyle (%) | $-5.5$ ($\pm3.9$) | 71 | $-7.7$ ($\pm4.7$) | 61 | 0.008* | 0.51 |
| Cartilage volume loss | Medial suprapatellar bursa Synovial membrane thickness loss (mm) | $-0.66$ ($\pm22.72$) | 3432 | $17.96$ ($\pm33.73$) | 188 | 0.076 | 0.84 |
| Cartilage volume loss | Medial compartment | $-6.8$ ($\pm3.5$) | 51 | $-9.4$ ($\pm4.7$) | 48 | 0.045* | 0.63 |
| Cartilage volume loss | Latera compartment (%) | $-4.1$ ($\pm3.1$) | 76 | $-4.4$ ($\pm3.0$) | 68 | 0.814 | 0.10 |
| Synovial membrane thickness (mm) | Plateau (%) | $-7.6$ ($\pm4$) | 53 | $-8.6$ ($\pm4.8$) | 56 | 0.276 | 0.23 |
| Synovial fluid volume (ml) | $0.13$ ($\pm0.25$) | 192 | 0.14 ($\pm0.24$) | 171 | 0.948 | 0.04 |
| Synovial fluid volume (ml) | $-2.6$ ($\pm14.5$) | 558 | $-2.0$ ($\pm11.8$) | 590 | 0.776 | 0.05 |

*: p-value < 0.05. CV: coefficient of variation. $^*$: effect size calculated based on equal sample sizes.
Table 2 WOMAC index changes in patients treated for knee OA between baseline and 6 months follow-up collected from the literature [11].

| Parameter | WOMAC index | PRGF group | HA group | p-value | Effect size$^*$ |
|-----------|-------------|------------|----------|---------|----------------|
|           | mean (SD)   | CV%        | mean (SD)| CV%     |                |
| W1        | Pain        | -3.9 (±4.06) | 104       | -2.8 (±1.75) | 63  | NA | 0.35 |
| W2        | Function    | -12.9 (±9.40) | 73       | -7.6 (±5.81) | 76  | NA | 0.68 |
| W3        | Stiffness   | -1.6 (±1.98) | 124       | -1.0 (±1.76) | 176 | NA | 0.27 |
| W4        | Total       | -18.5 (±14.08) | 76       | -11.5 (±7.66) | 67  | NA | 0.62 |

HA: hyaluronic acid; PRGF: plasma rich in growth factor; CV: coefficient of variation. NA: not available. $^*$: effect size calculated based on equal sample sizes.
Figure 1.
Figure 1 Continued.
Figure 2.
Figure 3.
Figure 4.
Figure 5.
Figure 6.