APS
Archives of Plastic Surgery

Review Article

Trends in statistical methods in articles published in *Archives of Plastic Surgery* between 2012 and 2017

Kyunghwa Han[1], Inkyung Jung[2]

[1]*Department of Radiology, Research Institute of Radiological Science, Center for Clinical Imaging Data Science and* [2]*Division of Biostatistics, Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Korea*

This review article presents an assessment of trends in statistical methods and an evaluation of their appropriateness in articles published in the *Archives of Plastic Surgery* (*APS*) from 2012 to 2017. We reviewed 388 original articles published in *APS* between 2012 and 2017. We categorized the articles that used statistical methods according to the type of statistical method, the number of statistical methods, and the type of statistical software used. We checked whether there were errors in the description of statistical methods and results. A total of 230 articles (59.3%) published in *APS* between 2012 and 2017 used one or more statistical method. Within these articles, there were 261 applications of statistical methods with continuous or ordinal outcomes, and 139 applications of statistical methods with categorical outcome. The Pearson chi-square test (17.4%) and the Mann-Whitney U test (14.4%) were the most frequently used methods. Errors in describing statistical methods and results were found in 133 of the 230 articles (57.8%). Inadequate description of P-values was the most common error (39.1%). Among the 230 articles that used statistical methods, 71.7% provided details about the statistical software programs used for the analyses. SPSS was predominantly used in the articles that presented statistical analyses. We found that the use of statistical methods in *APS* has increased over the last 6 years. It seems that researchers have been paying more attention to the proper use of statistics in recent years. It is expected that these positive trends will continue in *APS*.

Keywords Statistics / Parametric methods / Nonparametric methods / Statistical software

Correspondence: Inkyung Jung
Division of Biostatistics, Department of Biomedical Systems Informatics, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea
Tel: +82-2-2228-2494
E-mail: ijung@yuhs.ac

Received: 2 Jan 2018 • Revised: 24 Apr 2018 • Accepted: 2 May 2018
pISSN: 2234-6163 • eISSN: 2234-6171 • https://doi.org/10.5999/aps.2018.00010 • Arch Plast Surg 2018;45:207-213

INTRODUCTION

Evidence-based medicine (EBM) is an approach to medical practice intended to optimize decision-making by emphasizing the use of evidence from well-designed and well-conducted research [1]. The ability to understand sources of bias in the medical literature is required to practice EBM properly. Bias can occur from study design and/or the inappropriate use of statistical tests. Study design affects the choice of statistical methods. Moreover, statistics is deeply involved in all stages of research, from data collection to analysis and interpretation. An improper choice of study design and statistical techniques may lead to improper results and conclusions. Therefore, statistics plays a crucial role in evaluating the evidence of medical research.

Archives of Plastic Surgery (*APS*) adheres to the guidelines and best practices of the International Committee of Medical Journal Editors (ICMJE). The ICMJE recommendations for statistics state that authors should describe statistical analyses with enough detail to enable a reader to verify the reported results and that authors need to provide appropriate indicators of measurement errors or uncertainty, such as confidence intervals, beyond the P-value [2,3]. Furthermore, they recommend specifying the statistical software program(s) and versions used.

The aim of this article was to assess trends in statistical methods, and to evaluate their appropriateness, in papers published in *APS* from 2012 to 2017. *APS* is the official journal of the Korean Society of Plastic and Reconstructive Surgeons and is published 6 times per year. Since 2012, it continues the *Journal of the Korean Society of Plastic and Reconstructive Surgeons*, which was launched in 1974. This review article provides an overview of recent trends in the statistical methodology used in *APS*.

METHODS

This study is a retrospective literature analysis, neither approval from the Institutional Review Board nor informed consent was required.

We collected 388 original articles published in *APS* from 2012 to 2017. Case reports, ideas and innovations, review articles, and letters were excluded. Of these articles, 230 (59.3%) used statistical methods to analyze data and to report results. We classified them according to the types of statistical methods and software used, and checked whether there were errors in the description of statistical methods and results. We counted the number of statistical methods applied. When multiple statistical analyses were used in a study, each method was counted separately. The Cochran-Armitage trend test was used to assess the presence of linear trends in the percentages of statistical methods and statistical software used in the published articles by year from 2012 to 2017. R version 3.4.2 (R Foundation for Statistical Computing, Vienna, Austria) was used to perform the statistical tests, and P-values < 0.05 were considered to indicate statistical significance.

Statistical methods according to the objective of the analysis

Table 1 lists statistical methods according to the objective of the analysis and whether they involve continuous or ordinal, categorical, or time-to-event outcomes. We classified the statistical methods into 1 of 3 commonly used objectives: comparisons, correlations, and regression analyses.

Comparisons can be performed using paired or independent samples. Paired data arise from the same individual at different points in time or from different regions of the body, while unpaired (or independent) data arise from distinct individuals. In paired data, the variables to be compared are correlated with each other, so that correlation should be considered in the analysis. In plastic surgery research, clinical assessments before and after surgery or from multiple regions within the same subject can be treated as paired (or clustered) data.

The association between two variables can be assessed through either a correlation or regression analysis. Correlation analyses quantify the relationship between the variables, while regression analyses model the relationship between an outcome variable and one or more explanatory variables. Regression can be used to predict an outcome based on one or more predictors.

Other statistical analyses that were not listed in Table 1 but were used in articles published in *APS* include the normality test, power analysis, multivariate analysis, and reliability analysis using such as the intraclass correlation coefficient, Bland-Altman plots, the Cronbach alpha, and the kappa statistic.

Table 1. Statistical methods applied in the articles published in *Archives of Plastic Surgery*

Objective	Continuous or ordinal outcomes		Categorical data	Time-to-event data
	Parametric methods	Nonparametric methods	Statistical methods	
Comparison of paired samples	Paired t-test Repeated-measures ANOVA	Wilcoxon signed-rank test Friedman test	McNemar test	-
Comparison of independent samples	Two-sample t-test ANOVA	Mann-Whitney U test (Wilcoxon rank-sum test) Kruskal-Wallis test	Pearson chi-square test Fisher exact test	Kaplan-Meier curve and log-rank test
Correlation	Pearson correlation coefficient	Spearman correlation coefficient	-	-
Regression analysis	Linear regression model Linear mixed model	-	Logistic regression model	Cox proportional hazards model

ANOVA, analysis of variance.

Parametric or nonparametric methods

The statistical methods for continuous or ordinal outcomes can be further classified as parametric or nonparametric (Table 1). Parametric statistical methods assume a specific parametric form of the distribution for the underlying population, while nonparametric methods do not assume any parametric form. For example, the t-test assumes that the variables are normally distributed. When comparing the central tendency between groups, one should check whether the data can be assumed to be normally distributed before applying parametric tests. Nonparametric methods need fewer assumptions about the underlying distribution. In many cases, nonparametric methods are more appropriate when the sample size is not very large. Nonparametric methods based on ranks are especially useful for testing ordinal scale variables such as the visual analogue scale, which is widely used in the plastic surgery field.

Errors in reporting statistical methods and results

We assessed whether there were errors in reporting statistical methods and results in the articles published in *APS*. Errors in presenting P-values were observed, such as writing "P = 0.00" or "P = 1.00" instead of indicating that the P-value was very small or large (e.g., P < 0.001 or P > 0.999) and insufficient descriptions of the P-value, such as mentioning only the significance of the results without an exact P-value. Errors in describing the statistical methods were evaluated in terms of whether the applied statistical methods were described in the Methods section and whether the description of the applied statistical methods was complete and correct.

Statistical software

The frequency of the use of various statistical software packages was counted, as well as the type of statistical software package used.

RESULTS

Frequency and types of statistical methods

Of the 388 articles published in *APS* between 2012 and 2017, 230 (59.3%) used one or more statistical method. Fig. 1 shows a statistically significant increase in the number of articles that used statistical methods over 6 years (P for trend = 0.023). In 2012 and 2013, the percentage of articles using statistics was around 50%. In 2017, 64.7% of the articles published in *APS* used statistical methods. The number of statistical methods used per article in *APS* was 1.87 ± 1.06 (mean ± standard deviation). Almost half of the articles (47.1%) using statistics employed one method (Table 2). One article used six statistical methods.

Table 3 shows the frequencies of the statistical methods applied in the articles published in *APS* by year. There were 261 applications of statistical methods for continuous or ordinal outcomes, and 139 applications of statistical methods for cate-

Fig. 1. Frequency of statistical methods used in *APS*

The number of articles that used statistical methods and the proportion thereof among all articles published in *Archives of Plastic Surgery* (*APS*) by year.

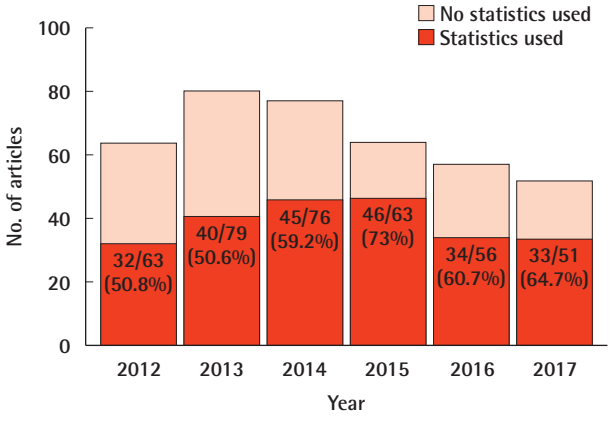


Table 2. The number of statistical methods used in the articles published in *Archives of Plastic Surgery* by year

No. of statistical methods used	Year						Total
	2012	2013	2014	2015	2016	2017	
1	18 (56.3)	19 (47.5)	21 (46.7)	20 (43.5)	15 (46.9)	14 (43.8)	107 (47.1)
2	7 (21.9)	9 (22.5)	15 (33.3)	13 (28.3)	7 (21.9)	8 (25.0)	60 (26.4)
3	5 (15.6)	8 (20.0)	5 (11.1)	9 (19.6)	6 (18.8)	8 (25.0)	40 (17.6)
≥ 4	2 (6.3)	4 (10.0)	4 (8.9)	4 (8.7)	4 (12.5)	2 (6.3)	20 (8.8)
Total	32	40	45	46	32	32	227[a]

Values are presented as number (%).
[a]Although 230 articles used statistical methods according to Table 1, three articles were not included in this table because they presented statistical results without a description of the statistical methods.

Table 3. Frequencies of the statistical methods applied in the articles published in *Archives of Plastic Surgery* by year

Objective	Year						Total
	2012	2013	2014	2015	2016	2017	
For continuous or ordinal outcomes							
Comparison of paired samples	14 (25.5)	14 (18.4)	4 (4.8)	15 (16.7)	9 (14.3)	11 (17.7)	67 (15.6)
Comparison of independent samples	13 (23.6)	28 (36.8)	37 (44.1)	36 (40.0)	23 (36.5)	24 (38.7)	161 (37.4)
Correlation and regression	4 (7.3)	5 (6.6)	6 (7.1)	4 (4.4)	7 (11.1)	7 (11.3)	33 (7.7)
For categorical outcomes							
Comparison of paired samples	0	1 (1.3)	2 (2.4)	1 (1.1)	0	0	4 (0.9)
Comparison of independent samples	20 (36.4)	19 (25.0)	28 (33.3)	26 (28.9)	14 (22.2)	10 (16.1)	117 (27.2)
Regression	1 (1.8)	4 (5.3)	3 (3.6)	3 (3.3)	1 (1.6)	6 (9.7)	18 (4.2)
For other outcomes and/or objectives							
Descriptive statistics	0	1 (1.3)	0	1 (1.1)	0	0	2 (0.5)
Normality test	1 (1.8)	1 (1.3)	1 (1.2)	1 (1.1)	2 (3.2)	1 (1.6)	7 (1.6)
Reliability analysis	1 (1.8)	1 (1.3)	1 (1.2)	1 (1.1)	4 (6.4)	2 (3.2)	10 (2.3)
Power analysis	0	1 (1.3)	0	0	3 (4.8)	1 (1.6)	5 (1.2)
Multivariate analysis	0	0	0	0	1 (1.6)	0	1 (0.2)
Survival analysis	0	0	2 (2.38)	2 (2.22)	1 (1.6)	0	5 (1.2)
Total	55	76	84	90	63	62	430

Values are presented as number (%).
Multiple counting was applied for articles that used multiple statistical methods.

Table 4. Errors in describing statistical methods and results by year

Error	2012	2013	2014	2015	2016	2017	Total
Inadequate description of P-values	11 (34.4)	19 (47.5)	22 (48.9)	17 (37.0)	9 (26.5)	12 (36.4)	90 (39.1)
No mention of statistical methods	4 (12.5)	2 (5.0)	2 (4.4)	5 (10.9)	7 (20.6)	1 (3.0)	21 (9.1)
Incomplete or wrong descriptions	6 (18.8)	6 (15.0)	1 (2.2)	5 (10.9)	8 (23.5)	6 (18.2)	32 (13.9)
Total[a]	32	40	45	46	34	33	230

Values are presented as number (%).
[a]Total indicates the number of published articles using statistical methods, not the sum of the above 3 cells.

gorical outcomes. Statistical methods for comparisons of independent samples were most commonly used. Among the methods of comparison, the Pearson chi-square test (17.4%) and the Fisher exact test (11.3%) for categorical outcomes and the Mann-Whitney U test (14.4%) and independent t-test (13.7%) for continuous or ordinal outcomes were the most frequently used methods. The Wilcoxon signed-rank test, paired t-test, Kruskal-Wallis test, and analysis of variance (ANOVA) were also widely used, accounting for more than 7% of the published articles using statistics. Within the category of regression analysis, logistic regression was used almost twice as much as linear regression. More complicated methods, such as repeated-measures ANOVA or linear mixed models, were applied in very few articles.

There were 30 applications of other statistical methods for other outcomes and objectives. Eight articles evaluated questionnaires or inter-rater agreement using reliability statistics. Seven articles checked the assumption of normality using the Kolmogorov-Smirnov test or the Shapiro-Wilk test. Five articles performed a power analysis prior to beginning a study or retrospectively. Survival analyses were applied in three articles.

Errors in reporting statistical methods and results

We found errors in describing statistical methods and results in 133 of the 230 articles (57.8%). The frequency of various types of errors is presented in Table 4. Errors in P-values were found in 90 articles, with 25 instances of presenting inadequate description of P-values as equal to 0 or 1, and 67 instances of not presenting the exact P-values. For example, reporting "P = NS (not significant)", or "P < 0.05" or "P < 0.01" instead of an exact P-value was the most common error in presenting P-values. Although such errors are not critical, they are worth mentioning and can be easily corrected.

Twenty-one articles did not state which statistical methods were applied, and 32 articles presented incomplete or wrong descriptions and applications. The statistical methods used in the article should be described in the Methods section, but some articles only reported P-values, along with the statistical meth-

ods used, in the Results section. For correlation analyses, for instance, the method for estimating a correlation coefficient (e.g., Pearson or Spearman) should be described. Another example is just mentioning the "t-test" without providing further details. Whether the independent or paired t-test was used should be explicitly stated, because that choice depends on the study design and the data structure. An example of an incorrect description of statistical methods was the use of the Wilcoxon signed-rank test for the comparison of independent observations. The Wilcoxon signed-rank test is used for paired comparisons, while the Wilcoxon rank-sum test (the same as the Mann-Whitney U test) is used for independent comparisons. These two methods were sometimes misused or misstated, due to confusion arising from the similar names.

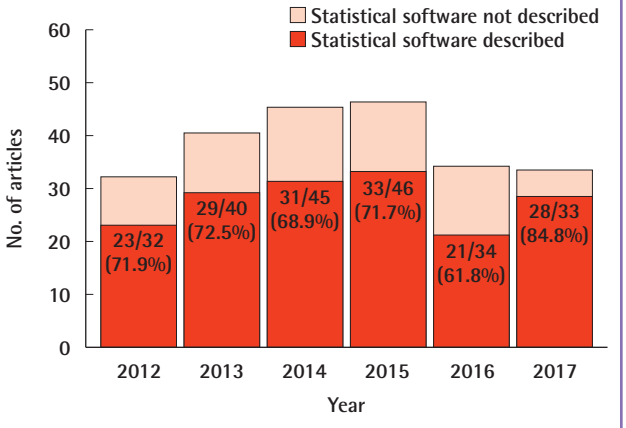
Statistical software packages

Among the 230 articles published in *APS* that used statistical methods, 165 (71.7%) provided details about the statistical software programs used for analyses. Seventy-five articles did not provide any such information. The percentage of articles presenting information about the statistical software used has increased by over 10%, from 71.9% in 2012 to 84.8% in 2017, although a statistically significant increasing trend was not observed (P for trend = 0.597) (Fig. 2).

SPSS (SPSS Inc., Chicago, IL, USA or IBM Corp., Armonk, NY, USA) was predominantly used in the articles that presented statistical analyses (141 of 168 cases, 83.9%). SAS (SAS Institute Inc., Cary, NC, USA) was used in roughly 4% of the articles. Other programs, such as GraphPad PRISM (GraphPad Software Inc., La Jolla, CA, USA), STATA (StataCorp LCC, Lakeway, TX, USA), and SigmaPlot (Systat Software Inc., Point Richmond, CA, USA), were only employed in one to four articles.

Fig. 2. Frequency of articles in *APS* describing statistical software

The number of articles that indicated which statistical software packages were applied and the proportion thereof among all articles using statistical methods by year. *APS*, *Archives of Plastic Surgery*.



DISCUSSION

In this review, we analyzed articles published in *APS* from 2012 to 2017 with respect to the use and type of statistical methods and statistical software packages. The results showed an increasing trend in the application of statistical methods and the use of statistical software packages.

Two relevant articles—one review and one editorial—regarding statistics have been published in specialized plastic surgery journals [4,5]. Januszyk and Gurtner [4] presented a practical overview of statistics in medicine, ranging from basic principles in statistics to descriptive and inferential statistical methods, with detailed guidelines for the interpretations of statistical tests. Freshwater [5] published a letter pleading for improvements in statistical analyses in plastic surgery. Some medical journals have published articles like this one, providing a systematic review and/or analysis of trends in the statistical methods applied [6-9]. However, we did not find such articles in the field of plastic surgery. To our knowledge, based on PubMed and KoreaMed (http://koreamed.org), this is the first article to report and summarize trends in the application of statistical methods in a plastic surgery journal.

Altman [10] reviewed the statistical contents of medical research published in the journal *Statistics in Medicine*. He found a considerable increase in the use of statistics and reported that a much greater use of complex statistical methodology in medical research was detected. The review articles regarding the use of statistics in medical journals [6-9] reflect Altman's findings. Altman [10] also said as a final comment, "Reviewing medical papers is difficult, time-consuming, occasionally frustrating, and educational. Many journals are desperate for expert statistical help." *APS* invited a statistical editor to join the editorial team in 2012, and started having statistical reviewers assess the submitted articles to improve the quality of statistical applications.

Despite the increasing use of statistics in *APS*, there were some statistical errors in the articles, including the presentation of P-values and the description of statistical methods and/or statistical software used. Some authors stated whether the results were statistically significant without providing exact P-values, especially for non-significant results; frequently presented as "P = NS." Moreover, some authors did not report the P-values throughout the article even for significant results, only stating whether the results were statistically significant. The exact P-values are useful information for interpreting the statistical results

of hypothesis testing. A very small P-value indicates that the null hypothesis is very incompatible with the data that have been collected [11-13]. Some software packages output results with the P-value listed as 0.000 or 1.000. Researchers usually copy and paste the P-value into the paper as is; however, such values should be presented as "P < 0.001" or "P > 0.999." "P = 0.000" means that there is absolutely zero chance of getting the results (and more extreme results) if the null hypothesis is true. However, there is always some chance of such an outcome, and we cannot definitively say that the probability is either 0 or 1. Some authors reported P-values without details regarding the data (e.g., summary estimates such as mean ± standard deviation, number [%], or odds ratio). The P-value has nothing to do with the magnitude or the importance of an observed effect [11,12]. For example, a difference in the visual analogue scale for pain assessment before and after surgery of 0.1 with a P-value of 0.2 would be interpreted as a non-significant difference, while a difference of 0.01 with a P-value of 0.003 would be presented as significant. As argued by Wasserstein and Lazar [13], statistical significance is not equivalent to scientific, human, or economic significance. Recently, some statements about the misuse of P-values were announced by a statistical society [13] and presented in a major medical journal [14]. To provide a broad and appropriate interpretation of the results of research, authors should report not only P-values with summary estimates, but also uncertainty measures such as the 95% confidence interval and/or standard error of estimates.

No indication of which statistical software package was used was provided in 75 of the 230 articles using statistical methods (28.3%). Different statistical programs could present different statistical results. For example, the median values computed in SPSS and R are not the same, because different algorithms are employed to calculate the median in the default settings. Another salient difference is the default setup of the event probability of the binary dependent variable in logistic regression, for which SAS uses a smaller value as a default, while SPSS uses a higher value. Data interpretation can be influenced by these defaults, so authors should understand the statistical software they use in detail and indicate which statistical software was used in the article.

Which statistical methods were used should be presented in the Methods section of the article. We noticed that some articles presented the results without mentioning statistical methods. We included these articles in the category of articles that used statistics. However, which statistical methods/software were used and how the significance level was set cannot be known. The instructions for authors in *APS* state that "methods of statistical analysis and criteria for statistical significance should be described" in the Methods section. Not only the names of the statistical analyses, but also the objectives of the study for using statistical methods should be described in detail in the Methods section.

The inclusion of a small number of subjects could limit the use of statistical analysis. Plastic surgery is a predominantly clinical field, so many plastic surgeons have focused their efforts on improving clinical results, and particularly on improving surgical techniques [15]. Assessments of newly updated surgical techniques or preliminary studies to generate an idea based on an animal experiment generally have small sample sizes. In some cases, neither statistical tests nor regression analysis might be necessary. Indeed, sophisticated statistical techniques are not always needed. Nonetheless, good data summarization using appropriate descriptive statistics can be very helpful for understanding the data. If statistical tests are required for a study with a small sample size, nonparametric statistical methods may be useful.

Less familiar statistical methods, such as reliability analyses and power analysis, were infrequently but consistently applied in the articles published in *APS*. Reliability analyses for evaluating internal consistency, test-retest repeatability, or inter-rater agreement are performed to assess reproducibility or repeatability among techniques/modalities or human readers. Power analysis is needed when planning a prospective study to achieve an adequate number of subjects. One may want to perform power analysis if non-significant results are obtained due to a small sample size.

This article can serve as the first step for obtaining a better understanding of the statistical methods frequently used in *APS*. In conclusion, the use of statistical methods has increased in *APS* over the last 6 years. Although there is room for improvement, researchers have been paying more attention to the proper use of statistics in recent years. These positive trends in *APS* are expected to continue in the future.

NOTES

Conflict of interest

No potential conflict of interest relevant to this article was reported.

REFERENCES

1. Wikipedia. Evidence-based medicine [Internet]. Wikipedia, The Free Encyclopedia [cited 2017 Dec 4]. Available from: https://en.wikipedia.org/wiki/Evidence-based_medicine.
2. International Committee of Medical Journal Editors. Uni-

form requirements for manuscripts submitted to biomedical journals. N Engl J Med 1997;336:309-15.
3. International Committee of Medical Journal Editors. Recommendations for the conduct, reporting, editing and publication of scholarly work in medical journals [cited 2017 Nov 26]. Available from: http://www.icmje.org/recommendations/browse/manuscript-preparation/preparing-for-submission.html.
4. Januszyk M, Gurtner GC. Statistics in medicine. Plast Reconstr Surg 2011;127:437-44.
5. Freshwater MF. A plea to improve statistical analyses and methods in plastic surgery. J Plast Reconstr Aesthet Surg 2013;66:1447-8.
6. Kim J, Yoon S, Kang JJ, et al. Research designs and statistical methods trends in the annals of rehabilitation medicine. Ann Rehabil Med 2017;41:475-82.
7. Sato Y, Gosho M, Nagashima K, et al. Statistical methods in the journal: an update. N Engl J Med 2017;376:1086-7.
8. Arnold LD, Braganza M, Salih R, et al. Statistical trends in the Journal of the American Medical Association and implications for training across the continuum of medical education. PLoS One 2013;8:e77301.
9. Qualls M, Pallin DJ, Schuur JD. Parametric versus nonparametric statistical tests: the length of stay example. Acad Emerg Med 2010;17:1113-21.
10. Altman DG. Statistical reviewing for medical journals. Stat Med 1998;17:2661-74.
11. Jung I. Some facts that you might be unaware of about the p-value. Arch Plast Surg 2017;44:93-4.
12. van Rijn MH, Bech A, Bouyer J, et al. Statistical significance versus clinical relevance. Nephrol Dial Transplant 2017;32 (suppl_2):ii6-ii12.
13. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Stat 2016;70:129-33.
14. Kyriacou DN. The enduring evolution of the p value. JAMA 2016;315:1113-5.
15. Lee WJ. Research, plastic surgery, and archives of plastic surgery. Arch Plast Surg 2017;44:359-60.