Evaluation of the Sample Size and Power Analysis of the Research Articles Published in Turkey Centered SCI-E Indexed Journals of Orthopedics and Traumatology

Türkiye Merkezli Ortopedi ve Travmatoloji Alanındaki SCI-E Kapsamındaki Dergilerde Güç Analizi ve Örneklem Büyüklüğü Değerlendirmesi

Onur KOCADAL,[a]
Cem Nuri AKTEKİN,[a]
Murad PEPE,[a]
Ertuğrul AKŞAHİN,[a]
Bilgehan TAĞRIKULU,[a]
Asiye UĞRAŞ DİKMEN[b]

[a]1st Clinic of Orthopedic and Traumatology, Ankara Training and Research Hospital, [b]Department of Public Health, Gazi University Faculty of Medicine, Ankara

Geliş Tarihi/*Received:* 31.01.2015
Kabul Tarihi/*Accepted:* 03.06.2015

This study was presented as a verbal announcement at 24th National Turkish Orthopedics and Traumatology Congress, 12-16 November 2014, Antalya, Turkey.

Yazışma Adresi/*Correspondence:*
Onur KOCADAL
Ankara Training and Research Hospital, 1st Clinic of Orthopedic and Traumatology, Ankara,
TÜRKİYE/TURKEY
onurkocadal@gmail.com

ABSTRACT Objective: The aim of this study was to conduct a power analysis and assess the adequacy of the sample size of the original research articles published in the Turkey-centered SCIE (Science Citation Index Expanded) indexed orthopedics and traumatology journals. **Material and Methods:** 1036 articles published between January 2010 and May 2014 were evaluated. 190 original research articles were included in the study. The power analysis was performed and the adequacy of the sample size were measured with Cohen's effect size and the α values. The alpha error level was set at 5%. **Results:** Only 5 trials (2.6%) of the 190 articles examined had sufficient power. The targeted power was achieved in 3 of the 5 trials. One article in the small effect size (1.5%) had adequate power (β <0.2). It was determined that 76 articles did not fulfill the basal power (β <0.2) criteria even at the large effect size. **Conclusion:** As the number of prospective randomized trials increase in the field of orthopedics and traumatology, more attention should be paid to power analysis in the trial design. Carrying out a trial following power analysis would help improve the its quality.

Key Words: Sample size; statistics; orthopedics

ÖZET Amaç: Bu çalışmada SCIE (Science Citation Index Expanded) kapsamında bulunan, Türkiye merkezli, ortopedi ve travmatoloji alanındaki orjinal araştırma makalelerinin güç analizi ve örneklem büyüklüğü değerlendirilmesi amaçlandı. **Gereç ve Yöntemler:** Ocak 2010-Mayıs 2014 tarihleri arasında basılmış olan 1036 yayın tarandı. 190 orjinal araştırma makalesi çalışmaya dahil edildi. Güç analizi ve örneklem sayısı değerlendirmesinde α değeri ve Cohen'in etki büyüklüğü değerleri kullanıldı. Alfa hata düzeyi 0,05 olarak belirlendi. **Bulgular:** Değerlendirilen makalelerin sadece 5 tanesinde güç analizi yapılmıştı (%2,6). Bunların 3 tanesinde amaçlanan güce ulaşılmıştı. Sadece 1 makale (%1,5) küçük etki düzeyinde yeterli güce (β<0,2) sahipti. 76 makalenin geniş etki büyüklüğü düzeyinde, bazal çalışma gücü (β<0,2) kriterini sağlamadığı saptandı. **Sonuç:** Ortopedi ve travmatoloji alanında özellikle prospektif randomize çalışmaların sayısının artması nedeniyle çalışma tasarımında güç analizinin daha dikkatli yapılması gerekmektedir. Çalışmaların güç analizi yapılarak gerçekleştirilmesi çalışma niteliğinin artmasına yardımcı olacaktır.

Anahtar Kelimeler: Örneklem büyüklüğü; istatistikler; ortopedi

Turkiye Klinikleri J Med Sci 2015;35(3):146-51

The calculation of the sample size, is an essential step in designing a clinical trial. The quantity of cases in a study is expressed as the sample size. Trials with excessive sample size may lead to a loss of resources.[1] Trials with inadequate sample size might have controversial

doi: 10.5336/medsci.2015-43953

results. To calculate the required sample size in trials, terms such as power analysis, Type 1 and Type 2 errors, and effect size should be taken into consideration.

Power analysis is the statistical method used to detect the required sample size in a planned trial or to check if a completed trial had the adequate sample size or not. Power analysis is based on Type 1 and Type 2 errors and the effect size.[2] Type I error (α error) is the probability of detecting a difference between 2 applications at the end of the trial when there is no difference. Type II error (β error) defines the non-finding of a difference between 2 applications when there is a difference.[3] Type 2 error might be decreased by increasing the sample size. In scientific trials, the aim is to keep the α error at 0.05, and the minimum '1- β' value at 0.80 levels.[4,5]

Effect size can be defined as the size of the expected difference between 2 mean values or 2 results.[6] It is calculated as the difference between the two means scores divided by the pooled standard deviation.[5,7] In his different statistical measurement methods, Cohen classified effect size values as small, medium or large.[8] Effect size determines the sample size that will be included in the trial. While a smaller sample size is required to perform a trial with an expectation of high intergroup difference (large effect size), a large sample size is needed to detect the difference between the values in a trial with an expectation of a lower intergroup difference (small effect size).[9,10]

The number of articles performing power analysis and assessing sample size is increasing in the literature.[2,10-12] The present article reports that an important percentage of the trials have been performed with inadequate sample sizes.[13,14] While there have been trials concerning sample size and power analysis in different disciplines in Turkey in recent years, to the best of our knowledge, no trials have been conducted in the field of Orthopedics and Traumatology.[15-17]

In the current study, the main aim was to assess the original articles in the Turkey-centered SCIE (Science Citation Index Expanded) indexed orthopedics and traumatology journals published between January 2010 and May 2014 to determine whether or not they had adequate power ($\beta < 0.20$) at small, medium or large effect size levels and whether or not they were performed with adequate sample size according to their effect size.

MATERIAL AND METHODS

The articles published between January 2010 and May 2014 in 'Acta Orthopedica et Traumatologica Turcica', the 'Turkish Journal of Trauma and Emergency Surgery' and the 'Joint Diseases and Related Surgery', which are Turkey-centered SCIE (Science Citation Index Expanded) indexed orthopedics and traumatology journals, were reviewed retrospectively. The case reports, the case series, the review articles, studies with only descriptive statistics (the trials that did not include any comparison with reported data such as the mean value or median value), the technical notes, the letters to the editor, the basic science trials, the experimental trials and the trials outside the orthopedics and traumatology discipline in the Turkish Journal of Trauma and Emergency Surgery were excluded from the study. The retrospective cohort, prospective cohort and the prospective randomized articles with these criteria were included in the study. The number of publication in the journals and their distribution are presented in Table 1.

The total sample size for each trial given the material and method sections of the original articles, the number of groups and the sample size in each group if there were any intergroup comparisons and the statistical methods used were recorded. Whether or not a power analysis had been performed was noted. Outcomes were defined as 'positive' when significant differences were found or 'negative' if no statistically differences occur. In the trials examining more than one hypothesis and using more than one statistical method, the statistical method that examined the main hypothesis of the trial was taken into consideration.

The power analysis was performed using the G*Power 3.1 statistical program.[18] The Cohen's effect size value, α value and the sample sizes in the

TABLE 1: The distribution of articles in journals.

		Acta Orthopedica et Traumatologica Turcica	Joint Diseases and Related Surgery	Turkish Journal of Trauma and Emergency Surgery	Total
Included in the study	Prospective randomized	12	6	1	19
	Prospective cohort	38	16	1	55
	Retrospective cohort	72	36	8	116
	Total	122	58	10	190
	Case reports	85	33	13	131
	Case series	12	10	2	24
	Review articles	1	19	0	20
Excluded in the study	Studies with only descriptive statistics	75	15	14	104
	Technical notes	2	0	0	2
	Letters to the editor	5	4	0	9
	Basic science trials	57	28	2	87
	Trials out of the orthopedics and traumatology discipline	0	0	470	470
	Total	237	108	501	846
	TOTAL	359	166	511	1036

trials were used for the power analysis. For the calculation of the power analysis, the small, medium and large effect size limit values as defined by Cohen were used (Table 2).[2,8,18] The alpha (type 1) error level was set at 5%.

The required sample sizes in the trials were calculated one by one for small, medium and large effect sizes. In addition to Cohen's effect size value, the α value and sample sizes were used for the assessment of the adequacy of the sample size. The alpha (type 1) error level was set at 5%. The number of the articles with adequate sample size within small, medium and large effect size was assessed. The term inadequate sample size was defined as the arithmetical difference between the minimum required sample size – for each of the effect size levels – and the actual sample size used in that trial.[2]

RESULTS

Following a retrospective search of a total of 1036 articles, 190 articles were included in the study. One hundred and twenty-two articles (64.3%) were published in the Acta Orthopedica et Traumatologica Turcica Journal, 58 articles (30.5%) in the Joint Diseases and Related Surgery and 10 articles (5.2%) in the Turkish Journal of Trauma and Emergency Surgery. There were 116 retrospective cohort, 55 prospective cohort and 19 randomized prospective trials. Of the 190 primary outcomes, 127 were 'positive' with statistically significant differences and 63 were 'negative' with no statistically significant differences. Only 5 of the trials in the 190 articles (2.6 %) had included power analysis. The targeted powers of 4 trials were 0.8. Although the targeted power was achieved in 2 of them, the

TABLE 2: Cohen's effect size limit values used in the calculations.

Group	Outcome and group characteristics	Small Effect Size	Medium Effect Size	Large Effect Size	Formulae
t-test	Two grouped continuous outcome	0,2	0,5	0,8	$d = u_a - u_b / \sigma$
F Test	Three or more grouped continuous outcome	0,1	0,25	0,4	σ of means/pooled σ
chi-square [d]	Categorical outcome	0,1	0,5	0,8	$\sqrt{\sum_{i=1}^{k} \frac{(p_{1i} - p_{0i})^2}{p_{0i}}}$

d: effect size d; u_a: mean of sample a; u_b: mean of sample b; σ: standard deviation; $p(p_i)$: probability of success (in group i)

remaining 2 trials were unable to achieve the targeted power due to patient censoring or data being analyzed with unsuitable statistical methods. The targeted power of the last trial was 0.95 and this was achieved.

STUDIES WITH POSITIVE RESULTS

None of the articles in the study had adequate power (β <0.20) in the small effect size. However, 35 articles in the medium effect size and 76 articles in the large effect size had adequate power (β <0.20). Fifty-one articles demonstrated inadequate power in the large effect size value (β >0.20), as can be seen in Table 3.

STUDIES WITH NEGATIVE RESULTS

One article in the small effect size, 19 articles in the medium effect size and 38 articles in the large effect size had adequate power (β <0.20). Twenty-five articles demonstrated inadequate power in the large effect size value (β >0.20) (Table 4).

The mean sample size in the trials with 'negative' results was 69.5. The mean inadequate sample size in 63 articles that could not provide a small effect size was 633.5. The mean inadequate sample size in 44 articles that could not provide a medium effect size was 71.5. The mean inadequate sample size in 24 articles that could not provide a large effect size was 20.2.

DISCUSSION

An important indicator for the quality of a clinical trial is its adequate sample size and its statistical power. The statistical results in the clinical trials are controversial if the sample size is inadequate. Although the number of sample sizes in trials might be inadequate, it is still possible to achieve results with statistical significance. Furthermore, the existence of statistically significant difference might be reported as 'not statistically significant' due to the inadequacy of the study samples.[13] In our study, the adequacy of the sample sizes in Turkey-centered articles was examined. For the negative studies, the required minimum power (β <0.20) at the large effect size,[2,19] which is the basic proficiency level, was determined to be 60.3%. These results show that greater attention should be paid to power analysis and sample size calculation in future trials.

Terms such as 'p value' or 'statistically significant difference' are frequently used in the statistical assessment of clinical trials. Before the assessment of these terms, the sample size and the power analysis of the trial should also be taken into consideration.[3] To perform power analysis, some parameters should be known. These can be counted as effect size, alpha and beta type error levels and the rate of the sample distribution in the groups.[9] In our study, a power analysis was conducted and the

TABLE 3: Power distribution of the articles with positive studies according to small, medium and large effect size. Articles have sufficient power at the level of the basal force necessity are denoted by *.

Power	Small Effect Size	Medium Effect Size	Large Effect Size
≥0.80*	0	35*	76*
0.60≤x<0.80	2	16	23
0.40≤x<0.60	5	26	18
0.20≤x<0.40	28	39	7
<0.20	92	11	3

TABLE 4: Power distribution of the articles with negative studies according to small, medium and large effect size. Articles have sufficient power at the level of the basal force necessity are denoted by *.

Power	Small Effect Size	Medium Effect Size	Large Effect Size
≥0.80*	1*	19	38*
0.60≤x<0.80	0	7	14
0.40≤x<0.60	0	15	9
0.20≤x<0.40	15	19	2
<0.20	47	3	0

adequacy of the sample sizes were assessed using the data in published articles.

'The effect size' term and the sample size were considered to a lower extent than the statistical significance by researchers. This topic has been studied frequently in recent years in the literature and it has been determined that the sample size is often not adequately taken into consideration in a significant portion of existing articles. In the study of Moher et al., where they evaluated 383 randomized articles that had been published in selected medical journals, it was found that 68% of the articles had inadequate power.[14] The orthopedic literature seems to follow a course parallel to this. In his study, where he evaluated 117 articles related to orthopedic trauma, Lochner found a type 2 error rate of 90.5%.[10] Our study shows a similar result in Turkey.

The effect size is the difference level that is sought in a trial. There are different methods described for effect size calculation in the literature[7,20,21] due to the variations of standard deviation calculations. There are also various limit values for effect size.[7,21] Cohen classified the effect size as large, medium and small levels.[8,21] We preferred to make use of Cohen's effect size limit values due to its general utilization in the medical literature.[2,3,10,13,21] To calculate the small differences between the groups with high similarity rates, the trials should be performed at small effect size. In the case of visible difference levels, however, the trials should be performed at large effect size. In the trials, detecting the small differences requires a large number of samples. This situation causes difficulties in the cost and period of the study.[1] Having adequate power and not including more samples than necessary in the trial can save time and be more cost-effective. It is not always mandatory to perform all the scientific studies with large numbers of samples and to provide small effect size.[22]This situation may result in ethical issues. On the other hand, carrying out a study with adequate power and the minimum sample size is an ethical requirement. The scientific quality of the "significant" results of a trial with an inadequate sample size is also questionable.

Besides inadequate sample size selection, there are other important factors that can lead to inadequate power analysis. The selection of inappropriate statistical methods such as nonparametric methods rather than parametric ones or vice versa is common.[23] Another typical problem is neglecting to consider the confounding factors that can affect outcomes.[24] This might under- or overestimate the statistical relation and miss categorization of the study variables, which might increase the false positivity rates.[25] The power analysis of the trials can be performed before or after the trial.[26] The ideal situation is before the trial; yet, this is usually not the case.[27] Researchers usually perform the power analysis at the end of the trial to test the power of the trial when they cannot reach significant results.[3] We also determined that the power analysis was performed before the trial in only 5 of 190 articles. Power analysis prior to study was reported to be carried out in 4-9% of the studies.[10,13] Performing the statistical power analysis before the study will increase the trial's scientific quality. Some selected medical journals today question the criteria used to determine the sample size when the study is assessed.[16] Therefore, in the future, this situation (inadequate sample size) is expected to occur less frequently in scientific literature.

This study has some limitations. First of all, prospective randomized trials played a small part in the articles included in the study. The trials with relatively low levels of evidence, such as retrospective cohort studies, may be predicted to have lower power. Another weak point of this study was that only the main hypothesis or the conclusion was assessed and the secondary results were not taken into consideration.

CONCLUSION

As a result, it is mandatory to calculate the minimum sample size during the design period of a scientific trial, to consider the effect size and to keep the alpha and beta type errors within acceptable ranges. Greater attention should be paid to power analysis as it is one of the most important steps in producing quality articles.

Acknowledgment

We thank Dr. Anita L. Akkas (PhD in English Literature) for contributing to the English editing.

REFERENCES

1. Streiner DL. Sample size and power in psychiatric research. Can J Psychiatry 1990;35 (7):616-20.
2. Freedman KB, Bernstein J. Sample size and statistical power in clinical orthopaedic research. J Bone Joint Surg Am 1999;81(10): 1454-60.
3. Biau DJ, Kernéis S, Porcher R. Statistics in brief: the importance of sample size in the planning and interpretation of medical research. Clin Orthop Relat Res 2008;466(9): 2282-8.
4. Çakır B. [BIAS in medical research: types, classification, sources and control measures: review (1)]. Turkiye Klinikleri J Med Sci 2005; 25(1):100-10.
5. Bertrand C, Van Riet R, Verstreken F, Michielsen J. A basic introduction to statistics for the orthopaedic surgeon. Acta Orthop Belg 2012;78(1):1-5.
6. Sullivan GM, Feinn R. Using Effect Size-or Why the P Value Is Not Enough. J Grad Med Educ 2012;4(3):279-82.
7. Cohen L, Manion L, Morrison K. Quantitative data analysis. Research Methods in Education. 6th ed. USA: Taylor & Francis; 2007. p.501-58.
8. Cohen J. The analysis of variance. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. L. Hillsdale, NJ: L. Erlbaum Associates; 1988. p.273-406.
9. Kirby A, Gebski V, Keech AC. Determining the sample size in a clinical trial. Med J Aust 2002;177:256-7.
10. Lochner HV, Bhandari M, Tornetta P 3rd. Type-II error rates (beta errors) of randomized trials in orthopaedic trauma. J Bone Joint Surg Am 2001;83-A(11):1650-5.
11. Bhandari M, Lochner H, Tornetta P 3rd. Effect of continuous versus dichotomous outcome variables on study power when sample sizes of orthopaedic randomized trials are small. Arch Orthop Trauma Surg 2002;122(2):96-; 8.
12. Chung KC, Kalliainen LK, Hayward RA. Type II (beta) errors in the hand literature: the importance of power. J Hand Surg Am 1998;23 (1):20-5.
13. Freedman KB, Back S, Bernstein J. Sample size and statistical power of randomised, controlled trials in orthopaedics. J Bone Joint Surg Br 2001;83(3):397-402.
14. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272(2): 122-4.
15. Demirel N, Gürler S. [Practical approaches to the determination of the sample size in clinical studies]. Kafkas Univ Vet Fak Derg 2010; 16(2):205-11.
16. Süt N. [Sample size determination and power analysis in clinical trials]. RAED Dergisi 2011;3(1-2):29-33.
17. Kılıç S. Effect size. JMOOD 2014;4(1):44-6.
18. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39(2):175-91.
19. Malavolta EA, Demange MK, Gobbi RG, Imamura M, Fregni F. Randomized controlled clinical trials in orthopedics: difficulties and limitations. Rev Bras Ortop 2011;46(4):452-9.
20. Breaugh JA. Effect size estimation: factors to consider and mistakes to avoid. Journal of Management 2003;29(1):79-97.
21. Fritz CO, Morris PE, Richler JJ. Effect size estimates: current use, calculations, and interpretation. J Exp Psychol Gen 2012;141(1): 2-18.
22. Karlsson J, Engebretsen L, Dainty K; ISAKOS Scientific Committee. Considerations on sample size and power calculations in randomized clinical trials. Arthroscopy 2003;19(9):997-9.
23. Fagerland MW. t-tests, non-parametric tests, and large studies--a paradox of statistical practice? BMC Med Res Methodol 2012;12: 78.
24. Kılıç S. A major obstacle in correct interpretation of causal association: confounder. Journal of Mood Disorders 2015;5(1):44-6.
25. Taşdelen B, Kanık EA. [The correct usage and reporting of biostatistical techniques in health researches]. Sağlık Bilimleri Enstitüsü Dergisi 2009;2(1):1-13.
26. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med 1994;121(3): 200-6.
27. Bailey CS, Fisher CG, Dvorak MF. Type II error in the spine surgical literature. Spine (Phila Pa 1976) 2004;29(10):1146-9.