



# Analysis of reporting quality for oral presentations of observational studies at 19<sup>th</sup> National Surgical Congress: Proposal for a national evaluation system

Mustafa Hasbahçeci<sup>1</sup>, Fatih Başak<sup>2</sup>, Aylin Acar<sup>2</sup>, Abdullah Şişik<sup>2</sup>

## ABSTRACT

**Objective:** To compare the quality of oral presentations presented at the 19<sup>th</sup> National Surgical Congress with a national evaluation system with respect to the applicability of systems, and consistency between systems and reviewers.

**Material and Methods:** Fifty randomly selected observational studies, which were blinded for author and institute information, were evaluated by using the Strengthening the Reporting of Observational Studies (STROBE), Timmer Score, and National Evaluation System by two reviewers. Abstract scores, evaluation periods, and compatibility between reviewers were compared for each evaluation system. Abstract scores by three different evaluation systems were regarded as the main outcome. Wilcoxon matched-pairs signed rank and Friedman tests for comparison of scores and times, kappa analysis for compatibility between reviewers, and Spearman correlation for analysis of reviewers based on pairs of evaluation systems were used.

**Results:** There was no significant difference between abstract scores for each system ( $p>0.05$ ). A significant difference for evaluation period of reviewers was detected for each system ( $p<0.05$ ). Compatibility between reviewers was the highest for the Timmer Score (medium,  $\kappa=0.523$ ), and the compatibility for STROBE and National Evaluation System was regarded as acceptable ( $\kappa=0.394$  and  $\kappa=0.354$ , respectively). Assessment of reviewers for pairs of evaluation systems revealed that scores increased in the same direction with each other significantly ( $p<0.05$ ).

**Conclusion:** The National Evaluation System is an appropriate method for evaluation of conference abstracts due to the consistent results between the referees similarly with the current international evaluation systems and ease of applicability with regard to evaluation period.

**Keywords:** Abstract, congress, reporting quality

## INTRODUCTION

Various reporting criteria have been introduced in order to increase both the quality and impact of scientific studies and congress abstracts (1-4). In particular, these criteria were used in order to ensure the standardization of presentations and to increase the reporting quality of conference abstracts (5). There are several studies published in the literature regarding the ease of implementation of these systems during the evaluation and publication process in parallel with conference abstract acceptance (5-8). Nevertheless, application of these systems to each type of randomized-controlled, observational, and experimental studies is a time-consuming part of the selection process. In addition, the subjective features of the parameters constituting the system seem to create an obstacle that precludes these evaluations from wide acceptance (1, 8). Because of these reasons, the search continues in order to find "the most appropriate system" rather than "the best system" (9).

Translation of these systems into different languages from the original creates semantic and conceptual problems. This application problem precludes these systems to be used generally (3, 10). In order to overcome this problem, there are studies conducted on the development of a national congress abstract evaluation system in the Turkish language (11). This recently developed Turkish evaluation system requires further cooperative studies in order to become compliant with the widely used international systems.

The aim of this study was to investigate the oral presentation abstracts of the 19<sup>th</sup> National Surgical Congress (NSC-2014) conducted in 2014 in terms of comparative analysis of international and Turkish congress abstract evaluation systems, the ease of implementation of the system by reviewers, and the determination of harmony between reviewers and system features.

## MATERIAL AND METHODS

Oral abstracts, which were accepted and digitally printed for NSC-2014, were electronically scanned. Abstracts were categorized as "randomized-controlled trials", "observational", "experimental" or "other type" according to study type. *Randomized-controlled trials* were defined as prospective studies whereby participants were randomly allocated to the treatment or control group. *Observational studies* were defined

<sup>1</sup>Department of General Surgery, Bezmialem Vakıf University School of Medicine, Istanbul, Turkey

<sup>2</sup>Clinic of General Surgery, Ümraniye Training and Research Hospital, Istanbul, Turkey

**Address for Correspondence**  
**Mustafa Hasbahçeci**  
e-mail: hasbahceci@yahoo.com

Received: 11.05.2015

Accepted: 19.08.2015

©Copyright 2016  
by Turkish Surgical Association  
Available online at  
www.ulusalcerrahidergisi.org

as prospective descriptive (cohort), retrospective case-control and cross-sectional studies, and descriptive case series. *Experimental studies* were defined as all of the studies performed on any animals or tissues and cells obtained from individuals in the laboratory. *Other type* was defined as cost analysis, and survey reports and case studies that could not be categorized into “randomized-controlled trial”, “observational” or “experimental” groups. The distribution of the oral abstracts accepted in NSC-2014 is given in Table 1.

Randomized-controlled studies, experimental studies, and other type studies were excluded from the analysis. The sample group was selected as 50 abstracts by using a computer-assisted random number system among 404 observational studies presented in NSC-2014. This enabled the ability to identify a 10-15% difference with a predicted accuracy of 90%. The relevant parts of the reports in the sample groups were transformed into an invisible digital form and transferred into the image files to ensure blindness to the authors and clinics. The Turkish versions of STROBE (Strengthening the Reporting of Observational Studies), Timmer Score, and the previously proposed National Evaluation System (NES) were loaded up to a web-based abstract assessment system and were made available to reviewers in order to aid in the evaluation of the observational studies (<http://ulusaldegerlendirmesistemi.blogspot.com.tr/>).

After a personal assessment authority was given on the system, two researchers that were currently working as general surgeons in an education and research hospital (AS, AA) performed their evaluations independent of one another.

The system also noted the total time spent by each reviewer according to each type of evaluation. Abstracts in the sample group were scored by the users through the evaluation system using STROBE, Timmer, and NES instruments.

Each parameter developed under the system was assessed as “0” or “1” according to abstracts whether containing that feature in STROBE, Timmer and NES instruments or not. Declared scores were recorded as total scores. Three parameters among 16 parameters in NES and four parameters among 19 parameters in Timmer Score were excluded from the evaluation due to not being applicable to observational studies. Thus, the range of assessment tools was defined as 0-11 for STROBE (Appendix 1), 0-15 for Timmer (Appendix 2), and 0-13 for NES (Appendix 3).

### Statistical Analysis

Descriptive statistical methods (mean, standard deviation, rate, and range of values) were used for assessment of the study data. Wilcoxon matched-pairs, signed rank, and Friedman tests were used for comparison of the scores and times spent for

evaluation by reviewers according to each assessment system. Spearman correlation was used for analysis of reviewers based on pairs of the evaluation systems. Kappa analysis was used to assess compatibility between the two reviewers in terms of the results of each evaluation system. Kappa coefficient ( $\kappa$ ) less than 0.20 was defined as weak, 0.21-0.40 was defined as acceptable, 0.41-0.60 was defined as moderate, 0.61-0.80 was defined as good, and a value greater than 0.80 was defined as very good when determining the degree of compliance. A p-value of less than 0.05 with a confidence level of 95% was considered as significant.

### RESULTS

Total scores and times spent for evaluation for each reviewer according to study type were given in Table 2 and Table 3. The highest score was 9 according to STROBE (maximum score 11), 14 according to Timmer (maximum score 15), and 12 according to NES (maximum score 13) when both reviewers were considered together. There was no statistically significant difference in terms of scores given by the reviewers for each evaluation system (Table 2).

There was a statistically significant difference in terms of duration of log-in to the system during the assessment of reviewers (Table 3). The duration of log-in to the STROBE system was significantly shorter in reviewer 1 ( $p=0.001$ ) and significantly shorter for reviewer 2 ( $p=0.019$ ) in the NES system when compared to the other two evaluation systems (Table 3).

Kappa analysis showed that the two reviewers were compatible in every evaluation system; however, the degree of compatibility was best in Timmer evaluation system (moderate,  $\kappa=0.523$ ) (Table 4). The Kappa values for STROBE and NES were  $\kappa=0.394$  (acceptable) and  $\kappa=0.354$  (acceptable), respectively. All evaluations of each reviewer between systems (Strobe-Timmer, Strobe-NES, and Timmer-NES) resulted in a significantly positive correlation with an increase in scores in correlation analysis (For reviewer 1:  $p=0.003$ , 0.001, and 0.001, and for reviewer 2:  $p=0.014$ , 0.001, and 0.001, respectively) (Table 5).

Table 1. Distribution of verbal presentations with regard to types of studies accepted for 19<sup>th</sup> National Surgical Congress

Type of study	n (%)
Randomized controlled	9 (1.9)
Observational	404 (85.4)
Experimental	34 (7.2)
Other	26 (5.5)
Total	473

Table 2. Total scores of 19. National Surgical Congress with regard to evaluation systems and reviewers

Evaluation System	Maximum score	Score for reviewer 1 <sup>§</sup>	Range for reviewer 1	Score for reviewer 2 <sup>§</sup>	Range for reviewer 2	p <sup>¶</sup>
STROBE	11	5.5±1.4 (5)	2-8	5.4±1.9 (5)	1-9	0.708
Timmer	15	7.1±2.8 (7)	0-14	7.5±2.2 (7)	3-13	0.206
NES	13	7.6±1.8 (8)	1-12	7.4±2.1 (8)	2-11	0.386

<sup>§</sup>:mean±standard deviation (median); <sup>¶</sup>: Wilcoxon matched-pairs signed rank test; STROBE: Strengthening the Reporting of Observational Studies; NES: National Evaluation System

Table 3. Total length of evaluation time of 19. National Surgical Congress with regard to evaluation systems and reviewers

Evaluation System	Maximum score	Total time for reviewer 1 (sec) <sup>¶</sup>	Range for reviewer 1 (sec)	Total time for reviewer 1 (sec) <sup>¶</sup>	Range for reviewer 1 (sec)	p <sup>¶</sup>
STROBE	11	54.5±17.6	31-107	111.3±48.2	28-272	0.001
Timmer	15	75.8±20.7	44-133	123.0±44.2	39-257	0.001
NES	13	82.8±33.8	40-225	103.2±38.4	38-201	0.009
p <sup>§</sup>		0.001		0.019		

<sup>¶</sup>:mean±standard deviation; <sup>¶</sup>:Wilcoxon matched-pairs signed rank test; <sup>§</sup>:Friedman test; STROBE: Strengthening the Reporting of Observational Studies; NES: National Evaluation System

Table 4. Compatibility analysis of each evaluation system between reviewers

	Kappa value	p
STROBE	0.393	0.001
Timmer	0.523	0.001
NES	0.354	0.001

STROBE: Strengthening the Reporting of Observational Studies; NES: National Evaluation System

Table 5. Correlation analysis of reviewers between evaluation systems

	STROBE x Timmer p (correlation coefficient) <sup>¶</sup>	STROBE x NES p (correlation coefficient) <sup>¶</sup>	Timmer x NES p (correlation coefficient) <sup>¶</sup>
Reviewer 1	0.003 (0.408)	0.001 (0.503)	0.001 (0.599)
Reviewer 2	0.014 (0.346)	0.001 (0.501)	0.001 (0.444)

<sup>¶</sup>:Spearman correlation; STROBE: Strengthening the Reporting of Observational Studies; NES: National Evaluation System

**DISCUSSION**

Evaluation and scoring of congress abstracts with any evaluation system clearly enhances writing quality and standardization. Worldwide accepted evaluation systems can be used for this purpose. However, development and dissemination of a national based system seem to be useful. Besides, parameters for the quality of research method can also be evaluated with that of reporting (8). However, it must be always kept in mind that there are no gold standards during arbitration and evaluation processes (9).

Observational studies constitute most of the oral and poster abstracts in scientific congresses, while randomized controlled trials and experimental studies play a minor role (1, 5, 6). Consistent with this data, observational studies constituted 85.4% of oral abstracts in NSC-2014. In addition, there has been a tendency of accepting most of the poster abstracts evaluated for being published in congresses (12). Thus, only observational studies containing oral abstracts were evaluated in this study.

Evaluation systems such as Consolidated Standards of Reporting Trials (CONSORT) and STROBE are widely used in the scientific field. It seems that these systems should be used separately for randomized-controlled trials and observational studies. In addition, the determination of separate criteria for each type of study has been suggested (3, 4, 10, 13). On the contrary, evaluation systems designed by Timmer et al. (8) seem to be applicable for more various type of studies, includ-

ing experimental studies. However, it has been commonly observed that some changes can be made according to the type of study (5, 8, 11). Although NES is designed for evaluation of abstracts from both observational and randomized controlled studies, some parameters require distinction according to study type (11). Thus, implementation of two different evaluation systems for evaluation of observational and randomized-controlled studies seems to be more appropriate.

It was determined that 225 seconds was spent per each abstract using the evaluation system developed by Timmer. In addition, a handbook that has been prepared pertaining to relevant parameters was given to the reviewers in advance (8). Three hours was spent for 63 abstracts in the evaluation system that consisted of five parameters by Bydder (7). This system was solely designed for their studies. In that system, it was observed that a period of 180 seconds was necessary for each abstract. In our study, the time required for evaluation by the Timmer system was shorter (75.8-123.0 seconds). The time required for evaluation by STROBE and NES were different, 54.5-111.3 and 82.8-103.2 seconds, respectively. The effects of spending more time by reviewers on the evaluation remain unclear. Thus, no assessment could be performed regarding the effect of this relationship. However, it can be stated that the newly developed NES has been evaluated within acceptable periods as in other systems.

Another important problem during the evaluation of congress abstracts was the compatibility between reviewers (7). Incompatibility becomes more obvious in subjective parameters of the evaluation system such as importance, originality, triggering property of academic and scientific contributions (9, 14). Reviewer contribution from different centers also affects this relationship (14). Moreover, some studies in the literature suggest that the 'luck factor' and coefficients are very close in the harmony assessment of several reviewers among themselves shown by kappa analysis during the evaluation process of congress abstracts (12). Kappa coefficient values ranging from 0.11 and 0.60 have been reported related to the compatibility of reviewers in evaluation. As an interesting issue, a value higher than 0.40 has been rarely seen (9, 12). Taking these into account, it can be stated that there has been only a moderate compatibility between reviewers at best. In this study, kappa values showing compatibility between reviewers were 0.393, 0.523, and 0.354 for STROBE, Timmer, and NES, respectively. These values are acceptable and consistent when compared with similar studies in the literature. A high compatibility degree between reviewers indicates a reliable system. However, this can also indicate a bias where a group of abstracts are affected either positively or negatively in a systematic way (9).

Taking these factors into account, a moderate compatibility degree seems to be the most eligible degree.

It is emphasized that a valuable scientific arbitration requires an age younger than 40 and education on epidemiology or statistics (14). Reviewers participating in this study were younger than 40; however, they were not educated on epidemiology or statistics. There are some approaches to increase the compatibility degree such as preparation of a user manual handbook, publication of articles explaining each parameter by samples, and giving less space on subjective parameters as possible (8, 10, 13, 14). Despite that, significant differences are observed among accepted and rejected abstracts for conferences (7, 9, 12). Thus, evaluation systems that are currently used seem to be an efficient way of assessing abstracts despite the incompatibility between reviewers (7, 14).

Well-attended harmony meetings are organized in order to determine parameters related to quality in current instruments during the preparation of a control list while evaluating an article or an abstract (2, 4, 9). Agreed parameters are revised at regular intervals. Moreover, STROBE recommends prior coordination with their centers before translation of evaluation system used for observational studies into another language. Turkish translations of STROBE and CONSORT criteria have been published as a result of mutual adaptation studies (15, 16). Besides, it has been turned into a system where public criticism and contributions can be received via a web page (3, 10). Additional review and development compliance meetings are required for ensuring the widespread use of a national based congress abstract assessment system.

Evaluation instruments specially designed for specific conferences can also be found in the literature along with CONSORT and STROBE evaluation systems, which are well described and prepared with the implementation of specific processes (7, 14). Regardless of being evaluated with any of these evaluation systems, it is more important just to evaluate the congress abstracts with one of these evaluation systems.

There are articles in the literature containing detailed descriptions of all criteria related to CONSORT and STROBE evaluation systems (10, 13). Evaluation systems have been recommended for use with explanatory articles (4). This approach enhances the compatibility between reviewers, and ensures the resulting score to be more accurate.

### Study Limitations

There are several limitations of this study. Due to a lack in the number of randomized-controlled trials, observational studies were investigated solely during the study. Reviewers who evaluated the abstracts did not have any particular training on statistics or epidemiology. Correlation analysis between systems could not be assessed and interpreted properly due to different maximum scores of the systems.

### CONCLUSION

National evaluation system, which is open to further development and is expected to gain national access, seems to be an appropriate system for evaluation of congress abstracts. Evalu-

ation meetings with broad participation and further studies evaluating the system are required in order to receive criticism and contributions to the recommended system. A guideline should be published in order to increase the harmony between reviewers and ease of application. The regulation according to different study types is thought to lead to the ease of implementation.

**Peer-review:** Externally peer-reviewed.

**Author Contributions:** Concept – M.H., F.B.; Design – M.H., F.B.; Supervision – M.H., F.B.; Resources – A.A., A.Ş.; Materials – A.Ş., A.A.; Data Collection and/or Processing – M.H., F.B., A.A., A.Ş.; Analysis and/or Interpretation – M.H., F.B.; Literature Search – M.H., F.B., A.Ş., A.A.; Writing Manuscript – M.H., F.B.; Critical Review – M.H., F.B., A.A., A.Ş.; Other – A.A., A.Ş.

**Conflict of Interest:** No conflict of interest was declared by the authors.

**Financial Disclosure:** The authors declared that this study has received no financial support.

### REFERENCES

1. Can MF, Öztaş M, Yağcı G, Öztürk E, Yıldız R, Peker Y, ve ark. Uluşal Cerrahi Kongreleri'nde sunulan randomize kontrollü çalışma özetlerinin raporlama kalitesi: CONSORT kılavuzuna dayalı değerlendirme. *Uluşal Cerrahi Derg* 2011; 27: 67-73.
2. Vandembroucke JP, STREGA, STROBE, STARD, SQUIRE, MOOSE, PRISMA, GNOSIS, TREND, ORION, COREQ, QUOROM, REMARK and CONSORT: for whom does the guideline toll? *J Clin Epidemiol* 2009; 62: 594-596. [\[CrossRef\]](#)
3. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandembroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol* 2008; 61: 344-349. [\[CrossRef\]](#)
4. Hopewell S, Clarke M, Moher D, Wager E, Middleton P, Altman DG, et al. CONSORT for reporting randomised trials in journal and conference abstracts. *Lancet* 2008; 371: 281-283. [\[CrossRef\]](#)
5. Knobloch K, Yoon U, Rennekampff HO, Vogt PM. Quality of reporting according to the CONSORT, STROBE and Timmer instrument at the American Burn Association (ABA) annual meetings 2000 and 2008. *BMC Med Res Methodol* 2011; 11: 161. [\[CrossRef\]](#)
6. Yoon U, Knobloch K. Assessment of reporting quality of conference abstracts in sports injury prevention according to CONSORT and STROBE criteria and their subsequent publication rate as full papers. *BMC Med Res Methodol* 2012; 12: 47. [\[CrossRef\]](#)
7. Bydder S, Marion K, Taylor M, Semmens J. Assessment of abstracts submitted to the annual scientific meeting of the Royal Australian and New Zealand College of Radiologists. *Australas Radiol* 2006; 50: 355-359. [\[CrossRef\]](#)
8. Timmer A, Sutherland LR, Hilsden RJ. Development and evaluation of a quality score for abstracts. *BMC Med Res Methodol* 2003; 3: 2. [\[CrossRef\]](#)
9. Rowe BH, Strome TL, Spooner C, Blitz S, Grafstein E, Worster A. Reviewer agreement trends from four years of electronic submissions of conference abstract. *BMC Med Res Methodol* 2006; 6: 14. [\[CrossRef\]](#)
10. Vandembroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Int J Surg* 2014; 12: 1500-1524. [\[CrossRef\]](#)
11. Hasbahçeci M, Başak F, Uysal Ö. Evaluation of reporting quality of the 2010 and 2012 National Surgical Congress oral presentations by CONSORT, STROBE and Timmer criteria. *Uluşal Cerrahi Derg* 2014; 30: 138-146. [\[CrossRef\]](#)

12. Rothwell PM, Martyn CN. Reproducibility of peer review in clinical neuroscience. Is agreement between reviewers any greater than would be expected by chance alone? *Brain* 2000; 123: 1964-1969. [CrossRef]
13. Hopewell S, Clarke M, Moher D, Wager E, Middleton P, Altman DG, et al. CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. *PLoS Med* 2008; 5: e20. [CrossRef]
14. Montgomery AA, Graham A, Evans PH, Fahey T. Inter-rater agreement in the scoring of abstracts submitted to a primary care research conference. *BMC Health Serv Res* 2002; 2: 8. [CrossRef]
15. Karaçam Z. STROBE bildiri: Epidemiolojide gözlemsel araştırma raporu yazımının güçlendirilmesi için bir rehber. *Anadolu Hemşirelik ve Sağlık Bilimleri Dergisi* 2014; 17: 1.
16. Available from: <http://www.consort-statement.org/downloads/translations>.

Appendix 1. STROBE Statement		
Number	Item	Recommendation
1	Title	Indicate the study's design with a commonly used term in the title
2	Authors	Contact details for the corresponding author
3	Study design	Description of the study design
4	Objective Methods	Specific objectives or hypothesis
5	Setting	Description of setting, follow-up dates or dates at which the outcome events occurred or at which the outcomes were present, as well as any points or ranges on other time scales for the outcomes
6	Participants	<p>Cohort study—Give the most important eligibility criteria, and the most important sources and methods of selection of participants. Describe briefly the methods of follow-up</p> <p>Case-control study—Give the major eligibility criteria, and the major sources and methods of case ascertainment and control selection</p> <p>Cross-sectional study—Give the eligibility criteria, and the major sources and methods of selection of participants</p> <p>Cohort study—For matched studies, give matching and number of exposed and unexposed</p> <p>Case-control study—For matched studies, give matching criteria and the number of controls per case</p>
7	Variables	Clearly define primary outcome for this report.
8	Statistical methods	Describe statistical methods, including those used to control for confounding
9	Participants	Report Number of participants at the beginning and end of the study
10	Main results	<p>Report estimates of associations. If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period</p> <p>Report appropriate measures of variability and uncertainty</p>
11	Conclusions	General interpretation of study results

Appendix 2. Timmer abstract evaluation tool	
Number	Parameters
1	Question / objective sufficiently described?
2	Design evident and appropriate to answer study question?
3	Subject characteristics sufficiently described?
4	Subjects appropriate to the study question?
5	Controls used and appropriate? (if no control, check no)
6	Method of subject selection described and appropriate?
7	If random allocation to treatment groups was possible, is it described? (if not possible, check n/a)
8	If blinding of investigators to intervention was possible, is it reported? (If not possible, n/a)
9	If blinding of subjects to intervention was possible, is it reported? (If not possible, n/a)
10	Outcome measure well defined and robust to measurement bias? Means of assessment reported?
11	Confounding accounted for?
12	Sample size adequate?
13	Post hoc power calculations or confidence intervals reported for statistically non significant results?
14	Statistical analyses appropriate?
15	Statistical tests stated?
16	Exact p-values or confidence intervals stated?
17	Attrition of subjects and reason for attrition recorded?
18	Results reported in sufficient detail?
19	Do the results support the conclusions?
	Excluded parameters

Appendix 3. National evaluation system		
No	Group	Parameter
1	Title	Identification of study design in the title with common terms (randomized controlled, cohort, case-control, cross-sectional, case series, case report)
2	Author	Identification of contact information of corresponding author including e-mail address and exclusion of the Institution/Hospital name within the abstract
3	Aim	Identification of problem/aim/hypothesis
4	Method	Relevancy of the study design with the problem/aim/hypothesis of the study
5		Appropriate inclusion criteria, identification of selection, resource and methods
6		Identification of control group-if applicable
7		Identification of randomized allocation and management-if applicable
8		Identification of blinding the investigators, participants and reviewers, if applicable
9		Planned interventions for treatment groups
10		Specification of follow-up and result periods
11		Identification of variables and main outcome
12		Statement of participant number at the beginning and end of the study and reasons for exclusion
13	Statistics	Statistical methods, p value and confidence interval identification
14	Results	Reporting results in detail and according to aim/hypothesis
15	Conclusion	Interpretation of results in relevance to the aim/hypothesis
16		Relevancy of conclusions with results Excluded parameters