ELSEVIER

European Journal of Cancer 41 (2005) 1735–1738

European Journal of Cancer

www.ejconline.com

Assessment of oral mucositis in clinical trials: Impact of training on evaluators in a multi-centre trial

M.A. Stokman [a,*], S.T. Sonis [b], P.U. Dijkstra [a], J.G.M. Burgerhof [c], F.K.L. Spijkervet [a]

[a] *Department of Oral and Maxillofacial Surgery, University Medical Center Groningen, P.O. Box 30.001, 9700 RB Groningen, The Netherlands*
[b] *Division of Oral Medicine, Brigham and Women's Hospital, Boston, USA*
[c] *Department of Epidemiology and Statistics, University of Groningen, University Medical Center Groningen, The Netherlands*

Received 22 February 2005; received in revised form 14 April 2005; accepted 18 April 2005
Available online 20 July 2005

Abstract

In the assessment of mucositis, the inter-evaluator variability needs to be minimised and would likely to be best accomplished by training. The aim of this study was to evaluate the effect of training on concordance of evaluators in scoring oral mucositis. The evaluators were informed about the pathobiology and clinical appearance of mucositis and were trained in scoring mucositis according the Oral Mucositis Assessment Scale (OMAS). The effect of the training was evaluated by a pre- and post-training test. Each test consisted of 15 slides depicting oral mucositis. The pre- and post-training scores were compared to the reference standard. During 8 months at 6 meetings, 65 evaluators were trained. The mean percentage correctly scored slides according the OMAS increased significantly between the pre- and post-training test ($P < 0.001$). Training evaluators in scoring oral mucositis has a significant improvement on the outcome of mucositis assessment.

Keywords: Inter-observer; Mucositis; Scoring; Training; Multi-centre trial

1. Introduction

Mucositis of the oral mucosa is a frequent cause of morbidity in cancer therapy with a serious burden on patients. Severe mucositis causes considerable pain and discomfort, leading to a higher need of pain medication, parenteral nutrition and length of hospitalisation [1,2]. Many studies in which agents are tested with potential useful outcomes on mucositis have their shortcomings. Studies were underpowered, lacked an adequate control arm, were not investigator or patient blinded or had other major design flaws [3]. To determine the value of new agents aimed at prevention of mucositis, well designed, sufficiently powered and appropriately executed studies are needed. Due to the lack of sufficient patient numbers at single study site, often a multi-centre design is necessary to obtain data within an acceptable time frame. Moreover, time frames for preventive studies on sequelae of cancer therapy are very tight because of new developments or changes in ablative therapies like changes in fractionation schedules in radiotherapy or new combinations of cancer cytotoxic therapies [4].

One of the major concerns in controlled multi-centre trials with mucositis is the establishment of adequate inter-evaluator reliability to reduce outcome variability. In the implementation of the evaluation method in a multi-centre trial, standardisation between the different study sites and evaluators is essential for decreasing error variance and reducing type II error, i.e. failing to detect true differences between active drug and placebo. Furthermore a poor inter-evaluator reliability decreases statistical power, resulting in necessity for larger sample sizes to be able to detect significant differences between drug and placebo [5].

* Corresponding author. Tel.: +31 50 3613840; fax: +31 50 3611136.
E-mail address: m.a.stokman@kchir.umcg.nl (M.A. Stokman).

0959-8049/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ejca.2005.04.030

To increase the inter-evaluator reliability in multicentre trials, start up training meetings are necessary to standardise evaluators to the same method of scoring and baseline knowledge.

According to the regulations of the United States Food and Drug Administration (FDA) selected investigators and evaluators should be qualified by training and experienced to investigate the device (21 Code of Federal Regulations (CFR)812). Industry-sponsored trials, with the intention for FDA approval, will need to have start up training meetings to conform to these regulations.

In testing efficacy of training of evaluators there should be pre- and post-testing conducted to examine on empirical basis whether the training was effective. This testing should evaluate the improvement in: (1) conceptual understanding of the scoring method; (2) accuracy, i.e. how scores by the evaluator agree with the reference standard; and (3) inter-evaluator reliability. Between pre- and post-testing phase a didactic training should be provided [6].

One of the most important issues in research is a well defined endpoint. Regarding mucositis studies, in most instances mucositis will be used as primary endpoint defined as ulcerative or pseudomembranous mucositis. Several mucositis rating scales are clinical observational scores, based on a combination of local mucositis parameters (signs) together with general complaints such as pain and effects on eating [7]. Differences in definition and operationalisation of these general complaints hamper proper comparison of the outcomes using these scoring systems.

For assessment of the mucosal changes related to anti-cancer therapies the Oral Mucositis Assessment Scale (OMAS) has been developed [8]. The OMAS is a simple, quantitative and accurate mucositis score especially validated for research application in multicentre clinical trials. In this score, a clear definition of mucositis symptoms, erythema and ulceration, are established. The OMAS has been shown to be highly reproducible between observers ($r > 0.8$), responsive over time ($r > 0.9$) and accurately records the anatomic elements considered being associated with mucositis [8].

The aim of this study was to evaluate the effect of training on concordance of evaluators in scoring oral mucositis according to the OMAS score.

2. Patients and methods

Training meetings were organised as start up of a phase III multi-centre clinical trial. During these meetings, evaluators were trained in scoring mucositis according to OMAS and informed about the intention of the study, the pathobiology and clinical appearance of mucositis. All meetings were conducted by the same trainer (F.S.).

The training meeting consisted of a pre-testing phase, didactic training and post-testing phase, all of which was performed on the same day. The pre-testing phase consisted of a slideshow of 15 slides with clinical pictures of different regions of the mouth with or without different stages of mucositis. For each slide the evaluator had to fill in on a form the mucositis score, according the OMAS of the depicted region, based on visible ulceration and/or erythema, and the size of the lesion (ulceration) or intention (erythema) (Table 1). The score forms were collected after the slideshow. The didactic training consisted of a thorough review of the pathobiology of mucositis, the scoring method of mucositis according to OMAS, and an explanation of the different clinical aspects of mucositis. The post-testing phase consisted of a retest of the same 15 mucositis slides of which the evaluator again needed to fill in the score forms.

The scores of the evaluators were compared to a reference standard score. The reference standard was developed as follows. All 15 slides were scored for ulceration and erythema according the OMAS and rated for visible ulceration and/or erythema, independently by three experienced evaluators (S.S., F.S., M.S.) in the field of mucositis scoring prior to the start up meetings. A consensus meeting was held to discuss discrepancies in their scores. Consensus was reached by means of discussion. The outcomes of these scores were used as the reference standard score for each slide.

2.1. Statistical analysis

The scores of the evaluators according to OMAS and presence or absence of ulcerations and/or erythema, were dichotomised in either correct or incorrect in comparison to the reference standard. The Student's *t*-test

Table 1
Scoring mucositis according to the Oral Mucositis Assessment Scale (OMAS)

Location	Erythema[a]			Ulcerations/pseudomembranes[b]			
Upper lip	0	1	2	0	1	2	3
Lower lip	0	1	2	0	1	2	3
Right cheek	0	1	2	0	1	2	3
Left cheek	0	1	2	0	1	2	3
Right ventral and lateral tongue	0	1	2	0	1	2	3
Left ventral and lateral tongue	0	1	2	0	1	2	3
Floor of the mouth	0	1	2	0	1	2	3
Soft palate	0	1	2	0	1	2	3
Hard palate	0	1	2	0	1	2	3

[a] Erythema: 0 = none; 1 = not severe; 2 = severe.
[b] Ulcerations/pseudomembranes: 0 = no lesion; 1 = <1 cm^2; 2 = 1–3 cm^2; 3 = >3 cm^2.

for dependent samples was used to analyse the mean performance (expressed as the percentages of correct scores) of the evaluators in two ways, one ignoring the missing and secondly considering the missing as incorrect scores. McNemar's test was used to analyse the percentages correct scores of the different slides separately.

A P-value of <0.05 was considered significant.

3. Results

In the course of 8 months, 6 start-up training meetings were organised and a total of 65 evaluators were trained. The average group size was 11 evaluators (range 8–15). The professional background of the evaluators varied from dentists, physicians, nurses and research assistants. The results of the dichotomised scores on the 15 slides of the pre- and post-test training are summarised in Table 2.

Comparing the mean performance of the evaluators, ignoring the missing, the mean evaluator's percentage correctly scored slides according to the OMAS increased significantly between pre- and post-test training for both ulceration from 47% to 69% (95% CI: 17–26) and erythema from 63% to 77% (95% CI: 11–17). Considering the missing as incorrect scores, the mean percentages increased less pronounced but still significant at the 0.001 level. The mean percentage correctly scored ulceration increased between pre- and post-test training from 45% to 62% (95% CI: 12–23) and for erythema from 59% to 69% (95% CI: 5–14).

Comparing the results of the evaluators of scoring the absence or presence of ulceration and/or erythema, the correctly scored slides increased significantly between pre- and post-test training for both ulceration from 83% to 92% (95% CI: 2–16) and erythema from 79% to 85% (95% CI: 1–12).

The proportion improvements for the different slides are shown in Table 3. Analysing the scores for each of the 15 slides separately, 14 improvements were found for assessment of ulceration and of these improvements, 11 were significant. Also on erythema, on 14 slides improvements were found and 11 were significant.

Table 2
Summary of the correct, incorrect and missing scores of the 65 evaluators for the 15 slides according to OMAS: T1 is pre-test training scores; T2 is post-test training scores

	T1 (%)	n	T2 (%)	n
Ulceration				
Correct	45	435	62	608
Incorrect	50	491	28	269
Missing	5	49	10	98
Erythema				
Correct	59	567	69	669
Incorrect	35	344	21	207
Missing	6	55	10	99

Table 3
Proportion of scoring improvement of each slide after training of the evaluators

Slide	Mucositis	Proportion of improvement	95% CI
1	Ulceration	0.00	(−0.05 to 0.05)
	Erythema	0.12	(−0.01 to 0.25)
2	Ulceration	0.37*	(0.24 to 0.50)
	Erythema	0.25*	(0.14 to 0.37)
3	Ulceration	0.03	(−0.08 to 0.15)
	Erythema	−0.03	(−0.21 to 0.14)
4	Ulceration	−0.04	(−0.13 to 0.05)
	Erythema	0.06	(−0.02 to 0.14)
5	Ulceration	0.40*	(0.28 to 0.52)
	Erythema	0.02	(−0.06 to 0.09)
6	Ulceration	0.63*	(0.50 to 0.75)
	Erythema	0.20*	(0.07 to 0.33)
7	Ulceration	0.22*	(0.09 to 0.34)
	Erythema	0.31*	(0.17 to 0.45)
8	Ulceration	0.33*	(0.19 to 0.47)
	Erythema	0.42*	(0.27 to 0.57)
9	Ulceration	0.11*	(0.03 to 0.19)
	Erythema	0.29*	(1.13 to 0.45)
10	Ulceration	0.30*	(0.12 to 0.48)
	Erythema	0.02	(−0.02 to 0.06)
11	Ulceration	0.18*	(0.06 to 0.30)
	Erythema	0.27*	(0.10 to 0.44)
12	Ulceration	0.25*	(0.09 to 0.41)
	Erythema	0.07	(−0.08 to 0.22)
13	Ulceration	0.00	(−0.12 to 0.12)
	Erythema	0.08*	(0.00 to 0.15)
14	Ulceration	0.52*	(0.38 to 0.66)
	Erythema	0.14	(−0.02 to 0.30)
15	Ulceration	0.19*	(0.04 to 0.35)
	Erythema	0.07	(−0.01 to 0.15)

* Represents a significant improvement ($P < 0.05$).

4. Discussion

This study has shown that training of evaluators has a positive significant influence on scoring oral mucositis.

Analysing the slides separately, three slides were poorly scored. More than 87% of the evaluators scored these three slides wrong in the pre-test and post-test after training according to OMAS score. These slides were not easy to evaluate with respect to the size of the lesion. A possible explanation might be a poor slide exposure or poor depiction of the anatomic site. In the future, it would probably be better to keep these three slides out of the training meeting. In a post hoc analysis,

leaving these three slides out of the training meeting, the mean evaluator's percentage correctly scored slides according to OMAS increased for ulceration from 52% to 81% and for erythema from 62% to 78%. However, scoring oral mucositis in the clinical situation should be easier because it allows inspection of the whole mouth and all anatomic sites are visible and the evaluator can change the position of the patient in obtaining a better view.

Scoring only the absence or presence of mucositis signs erythema and/or ulceration gave a higher mean percentage of correctly scored slides than when the size or the intention of the lesion had to be taken into account both at pre- and post-testing. Scoring of dimensions on a slide is very difficult and can only be done if a reference, like a tooth, is visible on the slide. Some of the slides did not have such a reference. This could be the explanation of the difference in mean percentages correctly scored slides between both scoring methods. These outcomes can be an argument for the use of a scoring method, which only evaluates the absence or presence of the mucositis signs erythema and/or ulceration. Clinically, the ulcerative stage of mucositis is the most considerable stage and responsible for the pain and loss of function. In several studies aimed at prevention of mucositis, the primary endpoint is prevention of ulcerations. The World Health Organization Oral Toxicity Scale (WHO) score measures anatomic, symptomatic, and functional components of oral mucositis [9]. The WHO score is easy to use, to learn and measures no lesion dimensions. In contrast, the OMAS is focused on the anatomic compound of mucositis alone, and is more precisely related to the mucosal changes and dimensions of the lesions due to cancer therapies. In clinical research with mucositis as primary endpoint, it would probably be best to use both, a method measuring mucositis in a subjective way and a method measuring only in an objective way.

In this present study, the professional background of the evaluators varied and unfortunately the distribution was unknown. This drawback could be interpreted as flaw of this study, however it is known from others studies that the type of medical professional background does not influence the scoring outcome [8].

Post-training monitoring of the evaluators and calibration during a (multi-centre) study is necessary to determine reliability and to prevent evaluator drift [6]. An evaluator can have high levels of competence and reliability before the study starts but this can fade away in time. It is necessary to involve a positive feedback loop post-training, which could measure the degree of the evaluators' comprehension. The phase III study, for which these trainings were accomplished, was stopped prematurely due to the instability of the intervention agent. Therefore, no information is available from this study about the evaluators reliability and drift with time.

Training is essential to gain standardisation of the evaluators and to increase intra- and inter-evaluator reliability. Not only for standardising of scoring procedures but also to provide every study site and evaluator similar information. Moreover in multi-centre trials, concordance between the evaluators at different study sites is a prerequisite.

Based on this study and previous research, the following training prerequisites are recommended for future multi-centre trails on mucositis prevention [6,10]. Training needs to consist of: (1) training of the scoring method used as endpoint of the study; (2) didactic training comprising a review of pathobiology of mucositis and of the scoring method; (3) training in scoring and processing of data; (4) testing efficacy of training intervention by pre- and post-testing and (5) post-training monitoring of the quality of scoring.

In conclusion, training evaluators in scoring oral mucositis has a significant improvement on the outcome of mucositis evaluation.

Conflict of interest statement

None declared.

References

1. Elting LS, Cooksley C, Chambers M *et al*. The burdens of cancer therapy – clinical and economic outcomes of chemotherapy-induced mucositis. *Cancer* 2003, **98**, 1531–1539.
2. Sonis ST, Oster G, Fuchs H *et al*. Oral mucositis and the clinical and economic outcomes of hematopoietic stem-cell transplantation. *J Clin Oncol* 2001, **19**, 2201–2205.
3. Rubenstein EB, Peterson DE, Schubert M *et al*. Clinical practice guidelines for the prevention and treatment of cancer therapy-induced oral and gastrointestinal mucositis. *Cancer* 2004, **100**, 2026–2046.
4. Cooper JS, Pajak TF, Forastiere AA *et al*. Postoperative concurrent radiotherapy and chemotherapy for high-risk squamous-cell carcinoma of the head and neck. *N Engl J Med* 2004, **350**, 1937–1944.
5. Muller MJ, Szegedi A. Effects of interrater reliability of psychopathologic assessment on power and sample size calculations in clinical trials. *J Clin Psychopharmacol* 2002, **22**, 318–325.
6. Kobak KA, Engelhardt N, Williams JBW *et al*. Rater training in multicenter clinical trials: issues and recommendations. *J Clin Psychopharmacol* 2004, **24**, 113–117.
7. Parulekar W, Mackenzie R, Bjarnason G *et al*. Scoring oral mucositis. *Oral Oncol* 1998, **34**, 63–71.
8. Sonis ST, Eilers JP, Epstein JB *et al*. Validation of a new scoring system for the assessment of clinical trial research of oral mucositis induced by radiation or chemotherapy. Mucositis Study Group. *Cancer* 1999, **85**, 2103–2113.
9. Anon. Handbook for reporting results of cancer treatment. WHO Offset Publ, Vol. 48; 1979. p. 15–22.
10. Kobak KA, Lipsitz JD, Feiger A. Development of a standardized training program for the Hamilton Depression Scale using internet-based technologies: results from a pilot study. *J Psychiatr Res* 2003, **37**, 509–515.