Tohoku J. Exp. Med., 2014, **233**, 189-195

Development of a Peer Review System Using Patient Records for Outcome Evaluation of Medical Education: Reliability Analysis

Junichi Kameoka,[1] Tomoya Okubo,[2] Emi Koguma,[1] Fumie Takahashi,[1] Seiichi Ishii[1] and Hiroshi Kanatsuka[1]

[1]Office of Medical Education, Tohoku University Graduate School of Medicine, Sendai, Miyagi, Japan
[2]Research Division, The National Center for University Entrance Examinations, Tokyo, Japan

In addition to input evaluation (education delivered at school) and output evaluation (students' capability at graduation), the methods for outcome evaluation (performance after graduation) of medical education need to be established. One approach is a review of medical records, which, however, has been met with difficulties because of poor inter-rater reliability. Here, we attempted to develop a peer review system of medical records with high inter-rater reliability. We randomly selected 112 patients (and finally selected 110 after removing two ineligible patients) who visited (and were hospitalized in) one of the four general hospitals in the Tohoku region of Japan between 2008 and 2012. Four reviewers, who were well-trained general internists from outside the Tohoku region, visited the hospitals independently and evaluated outpatient medical records based on an evaluation sheet that consisted of 14 items (3-point scale) for record keeping and 15 items (5-point scale) for quality of care. The mean total score was 84.1 ± 7.7. Cronbach's alpha for these items was 0.798. Single measure and average measure intraclass correlations for the reviewers were 0.733 (95% confidence interval: 0.720-0.745) and 0.917 (95% confidence interval: 0.912-0.921), respectively. An exploratory factor analysis revealed six factors: history taking, physical examination, clinical reasoning, management and outcome, rhetoric, and patient relationship. In conclusion, we have developed a peer review system of medical records with high inter-rater reliability, which may enable us, with further validity analysis, to measure quality of patient care as an outcome evaluation of medical education in the future.

Keywords: medical records; outcome evaluation; peer review system; reliability; validity
Tohoku J. Exp. Med., 2014 July, **233** (3), 189-195.

Introduction

The evaluation of education has been divided into three categories: input (education delivered at school), output (students' capability at graduation), and outcome (performance after graduation) evaluations (IPRA Gold Paper No. 11 1994). In medical education, "input evaluation" includes the accreditation of medical schools, such as the Educational Commission for Foreign Medical Graduates (ECFMG) in the United States (Kassebaum 1994), and Japan Accreditation Council for Medical Education (JACME) in Japan. "Output evaluation" includes examinations, both at each medical university and by official institutes, such as the United States Medical Licensing Examination (USMLE) in the United States (Williams 1993), and National Certificate Examination in Japan (Kozu 2006). In contrast, the methods of "outcome evaluation" have not been sufficiently established because of its difficulties (Prystowsky and Bordage 2001). However, considering that the ultimate goal of medical education is to develop good doctors who can provide superior patient care, the development of outcome evaluation methods is mandatory in the field of medical education for the long term.

Outcome evaluation has only been attempted in a few universities such as Thomas Jefferson Medical College, in which the clinical competence of 4,560 graduates between 1975 and 2004 were rated by the program directors of their hospitals (Hojat et al. 2007). Apart from longitudinal analyses, ratings by program directors or other staff members have been investigated for reliability and validity in assessing pediatric trainees' clinical performance (Archer et al. 2010) and physicians' professionalism (Cruess et al. 2006; Tsugawa et al. 2011). These methods mainly assess the "process of clinical performance" rather than "patient outcomes," but the importance of patient outcomes has been increasingly recognized in medical education (Dauphinee 2012; Gonnella and Hojat 2012).

Received May 8, 2014; revised and accepted June 13, 2014. Published online July 10, 2014; doi: 10.1620/tjem.233.189.
Correspondence: Junichi Kameoka, Office of Medical Education, Tohoku University Graduate School of Medicine, 2-1 Seiryo-machi, Aoba-ku, Sendai, Miyagi 980-8575, Japan.
e-mail: j-kame@med.tohoku.ac.jp

Another approach used to assess clinical competence is a review of medical records, which contain information about "patient outcomes" in addition to the "process of clinical performance." Assessing the quality of patient care by reviewing medical records has been vigorously pursued for many decades, mainly from the viewpoint of health care (Payne 1979; Goldman 1992, 1994; Hayward et al. 1993; Rethans et al. 1994; Smith et al. 1997; Peabody et al. 2000; Hofer et al. 2004; Goulet et al. 2007). However, reviewing medical records, an implicit review in particular, has been met with difficulties because of poor inter-rater reliability (intra-class correlation coefficients (ICCs): 0.16-0.56) (Hayward et al. 1993; Hofer et al. 2004; Goulet et al. 2007). Proposed strategies to achieve adequate reliability include providing structured assessments, higher standards for reviewers, averaging scores from multiple reviewers, adjusting systematic bias resulting from the different backgrounds of individual reviewers, using outcome judgments, and adoption of practice guidelines (Goldman 1992; Smith et al. 1997).

To establish a method to measure quality of patient care and provide outcome evaluation of medical education, we launched a program to develop a peer review system of medical records in 2010. For this purpose, we planned to take two steps: (1) a retrospective study to develop a system with high inter-rater reliability as well as constructive validity, and (2) a prospective study to establish a system with content and criterion validity. Here, we took the first step and, by employing the strategies mentioned above, developed a peer review system of medical records with high inter-rater reliability.

Methods

Study design

For this study, a peer-review system (PRS) committee was constituted at Tohoku University, comprising seven physicians in various fields such as cardiology, gastroenterology, neurology, and hematology. This study was approved by the Tohoku University Research Ethics Board, and the Institutional Review Boards (IRB) of each hospital.

The procedure was as follows: reviewers visited each hospital independently, and evaluated medical records (all medical records of outpatient care and a summary sheet of inpatient care) based on the evaluation sheet described below. Since we wanted to evaluate both the "process" and "outcome" of patient care, we focused on outpatient care because inpatient care is normally performed by teams instead of individual physicians in many Japanese hospitals, making it difficult to evaluate the "process" of patient care by the physician in charge.

To determine the feasibility and examine the appropriateness of an evaluation sheet, we performed a pilot study with 51 cases in February 2012. Having improved the evaluation sheet with the PRS committee members and reviewers in the pilot study, we performed the main study between January and February 2013. After the pilot study, we also developed benchmark case records with varying quality of patient care, which we used to train reviewers in the main study.

Evaluation sheet

The peer review evaluation sheet, an original of ours, was designed by the PRS committee to measure several factors, including those previously reported in the literature (Rethans et al. 1994; Goulet et al. 2007), such as record keeping, gathering of information, clinical assessment, management, as well as factors we developed, such as rhetoric, physician-patient relationship (including "empathy") and overall outcome. "Empathy" here was defined as the ability to understand the feelings and experiences of patients and their family members.

The evaluation sheet consists of two parts: record keeping using a 3-point Likert-type scale: 3 (written), 2 (partially written), and 1 (not written); and quality of care using a 5-point Likert-type scale: 5 (outstanding), 4 (standard), 3 (fair), 2 (poor), and 1 (very poor). After modifications following the pilot study, the final form contained 24 items: 14 items for record keeping and 15 items for quality of care (Table 1). Some of the 15 items for quality of care, such as B8 (appropriate treatment) and B9 (EBM), seemed difficult to assess, but we assumed that excellent reviewers, using their knowledge and experiences, could read between the lines of medical records.

The most controversial issue after the pilot study was whether "NA (not applicable)" in the initial evaluation sheet should be omitted or not, which eventually was left in, in cases of rare, but possible situations such as answering B12 (Is he/she referring other doctors, if necessary?) when "not" necessary, although the presence of NA would hamper the statistical analysis.

Participants

The PRS committee selected five hospitals based on the following criteria: (1) general hospitals in the Tohoku region (northeastern Japan), and (2) approval from the IRB of the hospital was obtained. The average number of beds of the selected hospitals (Ishinomaki Red Cross Hospital, Sendai City Hospital, Yamagata Prefectural Central Hospital, Iwate Prefectural Central Hospital, Osaki Citizen Hospital) was 546 (range: 404-685). Three hospitals were chosen for the pilot study in 2011, and four (including two from the pilot study) were chosen for the main study in 2012. Three hospitals had electronic medical records and one hospital had paper records.

Patients were selected by a representative at each hospital and a member of the PRS committee based on the following criteria: outpatients (1) who visited the hospital for the first time between April 2008 and March 2012 and were eventually hospitalized, (2) who were seen by doctors three to ten years after graduation from medical school, (3) whose final diagnoses did not matter, as long as they were in the field of internal medicine. Patients seen by residents (doctors within two years of graduation) were excluded, because senior doctors always supervised their patient care.

Reviewers were selected by the PRS committee based on the following criteria: general internists (1) who were working in hospitals outside the Tohoku region, and (2) who had reputations for being excellent in a broad field of internal medicine. The selected reviewers came from workplaces all over Japan, from Hokkaido (the most northeastern region of Japan) to Okinawa (the most southwestern region of Japan).

Data analysis

Mean scores and standard deviations of the 29 items were calculated for each hospital. The internal consistency of the items was evaluated using Cronbach's alpha. Inter-rater reliability among the

Table 1. Mean scores (standard deviations) of each item according to hospitals.

Items		hospitals				total	% of NA
		1	2	3	4		
I. Record keeping							
<general>							
A1	style of medical records (3: electronic, 2: paper-based, 1: combined)					1.59 (0.68)	2.3
A2	legibility (1: easily legible, 2: hard to read, 3: illegible)					1.01 (0.10)	6.5
<specific> (3: written, 2: partially written, 1: not written)							
A3	chief complaint	2.87 (0.48)	2.71 (0.66)	2.64 (0.71)	2.86 (0.47)	2.79 (0.58)	0.2
A4	past history	2.93 (0.32)	2.80 (0.57)	2.93 (0.33)	2.94 (0.24)	2.90 (0.38)	0.2
A5	family history	1.52 (0.86)	1.44 (0.82)	1.39 (0.78)	1.24 (0.65)	1.42 (0.80)	1.6
A6	social history	1.96 (0.91)	2.18 (0.96)	1.89 (0.95)	1.89 (0.95)	1.98 (0.94)	2.8
A7	history of allergies	2.24 (0.97)	1.38 (0.78)	1.85 (1.00)	2.03 (1.00)	1.93 (1.00)	1.4
A8	present illness	2.93 (0.25)	2.97 (0.18)	2.91 (0.29)	2.92 (0.26)	2.93 (0.25)	1.4
A9	physical examination	2.66 (0.58)	2.64 (0.61)	2.64 (0.58)	2.68 (0.61)	2.66 (0.59)	3
A10	medication	2.03 (0.94)	2.56 (0.81)	2.53 (0.83)	2.54 (0.79)	2.36 (0.89)	0
A11	diagnosis	2.81 (0.51)	2.68 (0.62)	2.85 (0.42)	2.86 (0.45)	2.80 (0.51)	0.5
A12	assessment and plans	2.86 (0.40)	2.76 (0.50)	2.64 (0.57)	2.84 (0.42)	2.79 (0.47)	0
A13	explanation to the patient and family members	1.67 (0.86)	1.85 (0.82)	2.22 (0.84)	2.25 (0.87)	1.95 (0.89)	0.5
A14	signature of the doctor	2.84 (0.55)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	2.94 (0.34)	0
	Average of A3 through A14	2.45 (0.47)	2.41 (0.29)	2.48 (0.49)	2.51 (0.43)	2.46 (0.26)	
II. Quality of care (5: outstanding, 4: standard, 3: fair, 2: poor, 1: very poor)							
B1	Is he/she taking a history related to the chief complaint?	3.71 (0.53)	3.35 (0.56)	3.47 (0.64)	3.44 (0.59)	3.52 (0.59)	0.2
B2	Is he/she taking a history unrelated to the chief complaint?	3.22 (0.73)	2.77 (0.72)	3.00 (0.73)	3.19 (0.76)	3.07 (0.75)	0.9
B3	Is he/she performing a CC-focused physical examination?	3.42 (0.84)	3.20 (0.82)	3.28 (0.87)	3.29 (0.82)	3.32 (0.84)	0.5
B4	Is he/she performing a systemic physical examination?	2.96 (0.90)	2.61 (0.75)	2.85 (0.97)	2.80 (0.80)	2.83 (0.87)	0.2
B5	Is he/she ordering diagnostic tests appropriately?	3.81 (0.46)	3.72 (0.50)	3.83 (0.51)	3.78 (0.46)	3.79 (0.48)	0.2
B6	Is he/she interpreting the results of examinations appropriately?	3.86 (0.44)	3.80 (0.45)	3.75 (0.55)	3.81 (0.57)	3.82 (0.50)	0.5
B7	Is he/she adequately listing differential diagnoses?	3.69 (0.62)	3.44 (0.58)	3.62 (0.69)	3.68 (0.70)	3.62 (0.65)	0.5
B8	Is he/she treating the patient appropriately?	3.89 (0.33)	3.77 (0.47)	3.80 (0.57)	3.74 (0.59)	3.81 (0.49)	6.8
B9	Is he/she following EBM?	3.78 (0.45)	3.71 (0.52)	3.81 (0.52)	3.75 (0.48)	3.76 (0.49)	0.2
B10	Are the medical records well-written?	3.85 (0.44)	3.88 (0.35)	3.91 (0.47)	3.77 (0.47)	3.85 (0.44)	1.1
B11	Is he/she making referrals to other doctors, if necessary?	3.94 (0.50)	3.95 (0.46)	3.84 (0.59)	3.95 (0.65)	3.92 (0.55)	41.2
B12	Does he/she have empathy towards the patient?	3.60 (0.54)	3.58 (0.61)	3.65 (0.53)	3.55 (0.62)	3.60 (0.57)	1.1
B13	Is the explanation to the patient and family members enough?	3.11 (1.14)	3.11 (1.08)	3.47 (0.95)	3.25 (1.13)	3.21 (1.09)	0.5
B14	Outcome assessment of the patient	3.92 (0.34)	3.92 (0.31)	3.89 (0.35)	3.86 (0.52)	3.90 (0.38)	0
B15	Overall assessment of patient care	3.80 (0.57)	3.61 (0.60)	3.59 (0.80)	3.75 (0.65)	3.71 (0.65)	0
	Average of B1 through B15	3.67 (0.52)	3.49 (0.31)	3.60 (0.49)	3.58 (0.57)	3.57 (0.34)	2.3

scores by the four reviewers was examined by calculating ICCs. In addition, an exploratory factor analysis was performed in order to investigate the construct validity of the evaluation sheet. Parameters were estimated by maximum-likelihood estimation, and Promax rotation was employed for the rotation of the estimated factors. SPSS (version 15.0) was used for the statistical analysis.

Results

Time

It took three to four days for the reviewers to visit the

hospitals and complete the review. The total time required for an evaluation ranged from 1,170-1,405 minutes (mean 1,260 minutes, 11.3 minutes per patient).

Scores

Among the 112 cases reviewed, two cases were excluded from the analyses below because of incomplete evaluation sheets. The diagnoses of 110 cases included 30 gastrointestinal diseases, 28 cardiovascular diseases, 12 respiratory diseases, and 40 other diseases.

The mean scores (standard deviations) of each item according to the hospitals are shown in Table 1. The average score (standard deviation) of items B1 through B15 (quality of care) for the 110 cases was 3.57 (0.34). The average scores (standard deviations) of items B1 through B15 for the four reviewers were 3.73 (0.51), 3.46 (0.44), 3.60 (0.51), and 3.55 (0.41) (data not shown). The average scores (standard deviations) of items B1 through B15 of gastrointestinal diseases, cardiovascular diseases, respiratory diseases, and other diseases were 3.57 (0.37), 3.54 (0.26), 3.62 (0.34), and 3.59 (0.35), respectively (data not shown).

The percentages of "NA" were very high in item B11 (referral to other doctors, 41.2%), and high in item B8 (treatment, 6.8%), probably because we focused on outpatient care, in which patients were sometimes hospitalized quickly before receiving any treatment or being referred to other doctors. Among the record keeping items, the percentage of "NA" was high in A6 (social history, 2.8%), possibly because reviewers may have decided this information was unnecessary in some cases.

Although preliminary, several observations can be made from this table. First, in regards to record keeping, each hospital had weak items, such as A13 (explanation to the patient) in hospital 1 (1.67), and A7 (history of allergies) in hospital 2 (1.38). These weak items appeared to be correlated with the forms used by the hospitals; the chart in hospital 1 had no form for "explanation to the patient" and the chart in hospital 2 had no form for "history of allergies." Second, the total mean score for item B14 (outcome) was high (3.90) despite the relatively low scores for items B1 through B4 (history taking and physical examination). Hospital 2 presented a typical case whose mean score for item B14 (3.92) was the highest, while mean scores for items B1 through B4 were the lowest among the four hospitals.

Reliability and validity

Cronbach's alpha was approximately 0.8 for all 29 items, indicating sufficient internal consistency among the items (Table 2I). ICCs for reviewers revealed high correlations, 0.733 for the single measure and 0.917 for the average measure, indicating a high inter-rater reliability among the scores by the four reviewers (Table 2II).

An exploratory factor analysis revealed six factors involving "history taking," "physical examination," "clinical reasoning," "management and outcome," "rhetoric," and "patient relationship" (Table 3). We removed the following 16 items from the factor analysis: A1 through A13 since they were objective facts, B11 because of the high rates of "NA" (41%), and B15 because "overall assessment" was not suitable for factor analysis.

Table 2. Reliability analyses.

I. Cronbach's alpha for items

	all 29 items	13 items selected for factor analysis
with all items	0.798	0.769
With one item deleted		
A-1	0.807	
A-2	0.799	
A-3	0.792	
A-4	0.795	
A-5	0.8	
A-6	0.796	
A-7	0.795	
A-8	0.796	
A-9	0.791	
A-10	0.806	
A-11	0.795	
A-12	0.792	
A-13	0.799	
A-14	0.801	
B-1	0.781	0.74
B-2	0.783	0.744
B-3	0.78	0.75
B-4	0.778	0.747
B-5	0.79	0.75
B-6	0.787	0.746
B-7	0.783	0.732
B-8	0.789	0.748
B-9	0.793	0.756
B-10	0.792	0.769
B-11	0.795	
B-12	0.791	0.76
B-13	0.796	0.8
B-14	0.793	0.759
B-15	0.777	

II. Intraclass correlations (95% confidence interval) for reviewers

single measure	0.733 (0.720-0.745)
average measure	0.917 (0.912-0.921)

Discussion

In the present study, we have developed a peer review system of medical records with high inter-rater reliability (exhibiting one of the highest ICCs ever reported). We have also shown some construct validity of the evaluation sheet by factor analysis. What we need to do next is to per-

Table 3. Factor analysis.

	Factor loading					
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
1. Gathering of information						
B1	**0.574**	0.130	0.208	0.022	0.181	0.129
B2	**0.905**	0.030	0.094	0.017	0.046	−0.036
B3	**0.407**	0.058	0.114	−0.029	**0.900**	0.044
B4	**0.660**	0.013	0.059	0.035	**0.430**	−0.032
2. Clinical assessment						
B5	0.103	0.183	**0.623**	0.025	0.119	0.239
B6	0.154	**0.302**	**0.719**	0.049	0.023	0.148
B7	**0.387**	**0.326**	**0.496**	0.234	0.017	−0.061
3. Management						
B8	0.124	**0.934**	0.257	0.027	0.035	0.198
B9	0.017	**0.608**	**0.335**	−0.017	0.036	0.230
4. Rhetoric						
B10	−0.017	0.164	0.124	0.085	0.008	**0.450**
5. Physician-patient relationship						
B12	0.021	0.052	0.049	**0.926**	0.016	**0.362**
B13	0.034	−0.008	0.052	**0.650**	−0.018	−0.070
6. Outcome						
B14	0.102	**0.346**	0.268	0.033	0.038	**0.374**

Bold values indicate factor loadings higher than 0.3.
Factor 1: history taking, Factor 2: management and outcome, Factor 3: clinical reasoning, Factor 4: patient relationship, Factor 5: physical examination, Factor 6: rhetoric.

form a prospective study to determine content and criterion validity, as well as further construct validity.

The present system, in which medical records are reviewed by visiting each hospital, proved feasible with no practical problems. However, considering the time and cost of visiting hospitals, a system by which reviewers can review records in their own workplace, similar to the current peer review system of academic papers, may be a preferable alternative in the future, if the security of the patients' information can be guaranteed.

High inter-rater reliability was obtained, probably because (1) we selected reviewers who had a reputation as good internists in a broad field of medicine, (2) we provided criteria to the reviewers by presenting benchmark medical records obtained from the pilot study, (3) the evaluation sheet was modified after the pilot study by reviewers as well as members of the PRS committee, and (4) reviewers were able to read the summary sheet of inpatient care to evaluate the "outcome" of outpatient care. These structured conditions were among the previously proposed strategies to achieve adequate reliability, as described in the Introduction section (Goldman 1992; Smith et al. 1997).

Construct validity was supported by exploratory factor analysis, indicating that our evaluation sheet measured various skill domains, including those previously emphasized, such as history taking, physical examination, clinical reasoning, and management (Rethans et al. 1994; Goulet et al. 2007). In addition to these established skill domains, we attempted to measure the physician-patient relationship, mainly using items B12 and B13. Whether we can measure the empathy of doctors by reviewing medical records remains to be determined, despite the internal consistency obtained in the current study: Cronbach's alpha was lower when the item B12 (empathy) was deleted than when all items were included. The measurement of empathy has been receiving international attention, and a study using the Japanese version (Kataoka et al. 2009) of the Jefferson Scale of Physician Empathy (JSPE) (Hojat et al. 2002), composed of 20 items answered on a seven-point Likert-type scale, implicated cultural differences on empathic behaviors. We hope we can examine the correlation between empathy measured by our system and that by these established systems in the future.

Several issues from the data are worth mentioning, although we recognize that they are preliminary. First, some weak items of record keeping appeared to correlate with the format of the charts used in hospitals, confirming the theory that the quality of patient care depended on the "structure" as well as "process" and "outcome" (Donabedian 1988). Second, the low mean scores of B1 through B4 (history taking and physical examination) indicated that Japanese physicians, at least those in the current study, may not be good at systematic history taking and physical examinations, as has been pointed out by Western-trained

Japanese physicians (Shimahara 2002). Third, despite the low scores on items B1 through B4, the mean score of item B14 (outcome) was high, perhaps because many physicians in the current study quickly resorted to laboratory examinations such as CT scans. Both the number of CT scanners per million population and the estimated number of radiation-induced cases of cancer per year were the highest in Japan (Berrington de González and Darby 2004; Hall and Brenner 2008); however, the advantages and disadvantages of these findings need further discussion.

Our study has several limitations. First, the number of reviewers and record samples were both low; therefore, a generalizability analysis was not performed. An extension study, including a greater number of hospitals outside the Tohoku region and more reviewers from various backgrounds, is underway for generalizability analysis. Second, validity analysis was insufficient. To further examine validity, particularly content and criteria validity, a prospect study investigating the correlation between the assessments in the current system and those by persons familiar with the doctors' performance (program directors, comedical staff members, and patients) is being planned. Third, we only focused on outpatient care, which was performed by individual physicians. Whether we can evaluate inpatient care, which is normally performed by teams instead of individual physicians, remains to be investigated.

Japanese medical education has undergone significant changes since 1990, such as the introduction of problem-based learning tutorials, objective structured clinical examination (OSCE), and clinical clerkships (Kozu 2006; Teo 2007). In 2004, a new postgraduate medical education program including mandatory rotations in various clinical departments, such as pediatrics, obstetrics/gynecology, and psychiatry, was introduced (Nomura et al. 2008). To recruit students from various backgrounds such as the Humanities/Social Science track to medical schools, introducing a new medical school system has also been proposed (Tokuda et al. 2008). However, these reforms have been conducted and are being discussed without any measures of outcome evaluation. We hope the current system will enable us to contribute to the measurement of outcome evaluation in the future.

Acknowledgments

This work was supported in part by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology of Japan (22590448). We thank Drs. Yutaka Kagaya, Yoshiyuki Ueno, Akira Imatani, Atsushi Takeda, and Masaki Kanemura (Tohoku University Graduate School of Medicine) for cooperating as members of the PRS committee, Dr. Mitsunori Miyashita (Tohoku University Graduate School of Medicine, Department of Health Sciences) for statistical analysis in the pilot study, and Dr. Yasumichi Kinoshita (Ishinomaki Red Cross Hospital), Dr. Masao Hiwatari (Sendai City Hospital), Dr. Hiroaki Takahashi (Iwate Prefectural Central Hospital), Dr. Makio Gamo (Osaki Citizen Hospital), and Dr. Toshikazu Goto (Yamagata Prefectural Central Hospital) for their support and cooperation in reviewing patients' medical records. We also thank all the reviewers for reviewing the medical records of patients, Dr. Makoto Kikukawa (Kyushu University) and Dr. Junya Iwazaki (Tohoku University) for critical reading of the manuscript, and Mr. Yutaro Arata, Mr. Katsunori Tanaka, Mr. Shinya Otsuki, Ms. Naoko Chiba, and Ms. Ayaka Arata (Office of Medical Education, Tohoku University) for their technical assistance.

Conflict of Interest

All authors declare no conflict of interest.

References

Archer, J., McGraw, M. & Davies, H. (2010) Assuring validity of multisource feedback in a national programme. *Arch. Dis. Child.*, **95**, 330-335.

Berrington de González, A. & Darby, S. (2004) Risk of cancer from diagnostic X-rays: estimates for the UK and 14 other countries. *Lancet*, **363**, 345-351.

Cruess, R., McIlroy, J.H., Cruess, S., Ginsburg, S. & Steinert, Y. (2006) The Professionalism Mini-evaluation Exercise: a preliminary investigation. *Acad. Med.*, **81**, S74-S78.

Dauphinee, W.D. (2012) Educators must consider patient outcomes when assessing the impact of clinical training. *Med. Educ.*, **46**, 13-20.

Donabedian, A. (1988) The quality of care. How can it be assessed? *JAMA*, **260**, 1743-1748.

Goldman, R.L. (1992) The reliability of peer assessments of quality of care. *JAMA*, **267**, 958-960.

Goldman, R.L. (1994) The reliability of peer assessments: a meta-analysis. *Eval. Health Prof.*, **17**, 3-21.

Gonnella, J.S. & Hojat, M. (2012) Medical education, social accountability and patient outcomes. *Med. Educ.*, **46**, 3-4.

Goulet, F., Jacques, A., Gagnon, R., Racette, P. & Sieber, W. (2007) Assessment of family physicians' performance using patient charts: interrater reliability and concordance with chart-stimulated recall interview. *Eval. Health Prof.*, **30**, 376-392.

Hall, E.J. & Brenner, D.J. (2008) Cancer risks from diagnostic radiology. *Br. J. Radiol.*, **81**, 362-378.

Hayward, R.A., McMahon, L.F. Jr. & Bernard, A.M. (1993) Evaluating the care of general medicine inpatients: how good is implicit review? *Ann. Intern. Med.*, **118**, 550-556.

Hofer, T.P., Asch, S.M., Hayward, R.A., Rubenstein, L.V., Hogan, M.M., Adams, J. & Kerr, E.A. (2004) Profiling quality of care: is there a role for peer review? *BMC Health Serv. Res.*, **4**, 9.

Hojat, M., Gonnella, J.S., Nasca, T.J., Mangione, S., Veloksi, J.J. & Magee, M. (2002) The Jefferson Scale of Physician Empathy: further psychometric data and differences by gender and specialty at item level. *Acad. Med.*, **77**, S58-S60.

Hojat, M., Paskin, D.L., Callahan, C.A., Nasca, T.J., Louis, D.Z., Veloski, J., Erdmann, J.B. & Gonnella, J.S. (2007) Components of postgraduate competence: analyses of thirty years of longitudinal data. *Med. Educ.*, **41**, 982-989.

IPRA Gold Paper No. 11 (1994) Public Relations Evaluation: Professional Accountability.

Kassebaum, D.G. (1994) LCME accreditation standards for management of the medical school curriculum: a clarification. Liaison Committee on Medical Education. *Acad. Med.*, **69**, 37-38.

Kataoka, H.U., Koide, N., Ochi, K., Hojat, M. & Gonnella, J.S. (2009) Measurement of empathy among Japanese medical students: psychometrics and score differences by gender and level of medical education. *Acad. Med.*, **84**, 1192-1197.

Kozu, T. (2006) Medical education in Japan. *Acad. Med.*, **81**, 1069-1075.

Nomura, K., Yano, E., Aoki, M., Kawaminami, K., Endo, H. & Fukui, T. (2008) Improvement of residents' clinical compe-

tency after the introduction of new postgraduate medical education program in Japan. *Med. Teach.*, **30**, e161-e169.
Payne, B.C. (1979) The medical record as a basis for assessing physician competence. *Ann. Intern. Med.*, **91**, 623-629.
Peabody, J.W., Luck, J., Glassman, P., Dresselhaus, T.R. & Lee, M. (2000) Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality. *JAMA*, **283**, 1715-1722.
Prystowsky, J.B. & Bordage, G. (2001) An outcomes research perspective on medical education: the predominance of trainee assessment and satisfaction. *Med. Educ.*, **35**, 331-336.
Rethans, J.J., Martin, E. & Metsemakers, J. (1994) To what extent do clinical notes by general practitioners reflect actual medical performance? A study using simulated patients. *Br. J. Gen. Pract.*, **44**, 153-156.
Shimahara, N.K. (2002) *Teaching in Japan: a cultural perspective*, Routledge, New York, NY.
Smith, M.A., Atherly, A.J., Kane, R.L. & Pacala, J.T. (1997) Peer review of the quality of care. Reliability and sources of variability for outcome and process assessments. *JAMA*, **278**, 1573-1578.
Teo, A. (2007) The current state of medical education in Japan: a system under reform. *Med. Educ.*, **41**, 302-308.
Tokuda, Y., Hinohara, S. & Fukui, T. (2008) Introducing a new medical school system into Japan. *Ann. Acad. Med. Singapore*, **37**, 800-802.
Tsugawa, Y., Ohbu, S., Cruess, R., Cruess, S., Okubo, T., Takahashi, O., Tokuda, Y., Heist, B.S., Bito, S., Itoh, T., Aoki, A., Chiba, T. & Fukui, T. (2011) Introducing the Professionalism Mini-Evaluation Exercise (P-MEX) in Japan: results from a multicenter, cross-sectional study. *Acad. Med.*, **86**, 1026-1031.
Williams, R.G. (1993) Use of NBME and USMLE examinations to evaluate medical education programs. *Acad. Med.*, **68**, 748-752.