Tohoku J. Exp. Med., 2019, **248**, 253-260

A Medical Record Peer-Review System to Evaluate Residents' Clinical Competence: Criterion Validity Analysis

Junichi Kameoka,[1,2] Makoto Kikukawa,[3] Daiki Kobayashi,[4] Tomoya Okubo,[5] Seiichi Ishii[2] and Yutaka Kagaya[2]

[1]Division of Hematology and Rheumatology, Faculty of Medicine, Tohoku Medical and Pharmaceutical University, Sendai, Miyagi, Japan
[2]Office of Medical Education, Tohoku University Graduate School of Medicine, Sendai, Miyagi, Japan
[3]Department of Medical Education, Faculty of Medical Sciences, Kyushu University, Fukuoka, Fukuoka, Japan
[4]Division of General Internal Medicine, Department of Internal Medicine, St. Luke's International Hospital, Tokyo, Japan
[5]Research Division, The National Center for University Entrance Examinations, Tokyo, Japan

In contrast to input evaluation (education delivered at school) and output evaluation (students' capability at graduation), the methods of outcome evaluation (performance after graduation) of medical education have not been sufficiently established. To establish a method to measure the quality of patient care and conduct outcome evaluation, we have been developing a peer review system of medical records. Here, we undertook a pilot study to evaluate the criterion validity of our system by using "evaluation by program directors (supervisors in the hospitals)" as a criterion standard. We selected 13 senior residents from three teaching hospitals. Five reviewers (general internists working in other hospitals) visited the hospitals independently and evaluated five patients' records for each resident based on the previously established sheet comprising 15 items. Independently, program directors of the senior residents evaluated their clinical performance using an evaluation sheet comprising ten items. Pearson's analysis revealed statistically significant correlation coefficients in three pairs of assessments including clinical reasoning ($r = 0.5848$, $P = 0.0358$). Bootstrap analysis revealed statistically significant correlation coefficients in additional 5 pairs including history taking ($r = 0.509$, 95% confidence interval: 0.034-0.847). In contrast, the correlation coefficients were low in some items: $r = 0.132$ (–0.393-0.639) for physical examination and $r = 0.089$ (–0.847-0.472) for attitude toward patients. To the best of our knowledge, this is the first study, albeit a pilot one, that investigates the criterion validity of medical record evaluations conducted by comparing the assessments of medical records with those by program directors.

Keywords: criterion validity; medical records; outcome evaluation; peer review system; physical examination
Tohoku J. Exp. Med., 2019 August, **248** (4), 253-260.

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, International Public Relations Association 1994). Unlike "input evaluation" and "output evaluation," the methods of "outcome evaluation" have not been sufficiently established (Prystowsky and Bordage 2001). In medical education, outcome evaluation might be best achieved by measuring the clinical competence of residents based on the quality of patient care, which, however, has been hardly used in reforming medical education, at least never in Japan. For example, in 2004, a new postgraduate medical education program including mandatory rotation of various clinical departments, such as pediatrics, obstetrics/gynecology, and psychiatry, was introduced in Japan, and improvement of residents' clinical competency after its introduction has been reported; however, it was an "output" instead of "outcome" evaluation, because it only revealed higher confidence levels among residents than among those who took older programs based on self-administered questionnaires (Nomura et al. 2008). In Japan, the methods of "outcome evaluation" have never been employed in reforming medical education.

To establish a method to measure the quality of patient care and actually conduct outcome evaluation of medical education, we launched a program to develop a peer review system of medical records, which is basically an implicit

Received March 25, 2019; revised and accepted July 23, 2019. Published online August 20, 2019; doi: 10.1620/tjem.248.253.
Correspondence: Junichi Kameoka, Division of Hematology and Rheumatology, Faculty of Medicine, Tohoku Medical and Pharmaceutical University, 1-15-1 Fukumuro, Miyagino-ku, Sendai, Miyagi 983-8536, Japan.
e-mail: j-kame@tohoku-mpu.ac.jp

evaluation based on the assumption that good reviewers can read between the lines. Assessing the quality of patient care by reviewing medical records has been vigorously pursued, mainly from the perspective of health care (Hayward et al. 1993; Goldman 1994; Rethans et al. 1994; Smith et al. 1997; Peabody et al. 2000). However, reviews of medical records—implicit reviews in particular—have been met with difficulties because of poor inter-rater reliability (Hayward et al. 1993; Hofer et al. 2004; Goulet et al. 2007). In the previous study, by selecting good reviewers who are proficient in broad fields of medicine, providing detailed criteria obtained from the pilot study, and using the summary sheet of inpatient care to evaluate the "outcome" of outpatient care, we established high inter-rater reliability of the peer-review system of medical records (average measure intraclass correlations for the reviewers: 0.917) in addition to construct validity (Kameoka et al. 2014). Here, we attempt to undertake a pilot study to evaluate criterion validity, one of the three traditional types of validity (content validity, criterion validity, and construct validity). To evaluate the criterion validity of our peer-review system of medical records, we selected "evaluation by program directors" as a criterion standard, because it represents a situation similar to our medical record evaluation; both methods are workplace assessments (Etheridge and Boursicot 2013), by which the performance of doctors can be evaluated in clinical instead of test situations (Al-Wassia et al. 2015).

Although many studies have investigated the reliability and construct validity of various tools to assess physician performance or medical professionalism (Hojat et al. 2007; Archer et al. 2010; Tsugawa et al. 2011), studies that investigate criterion validity have been lacking. A meta-analysis of 35 studies on the multi-source feedback process to assess physician performance between 1975 and 2012 revealed that only four studies reported data on physician/surgeon performance in comparison with other criterion measures (e.g., objective structured clinical examination (OSCE)) (Al Ansari et al. 2014). Another meta-analysis of 28 studies relating to the utility of a mini-Clinical Evaluation exercise (mini-CEX) between 1995 and 2013 revealed that only one study analyzed the criterion validity of the mini-CEX among specialty trainees, in which a good correlation between mini-CEX scores and outcome in the Canadian version of the Membership of the Royal College of Physicians examination was demonstrated among medicine trainees (n = 22) (Yates 2013). When Li et al. (2017) performed a systematic review of the instruments assessing medical professionalism between 1990 and 2015, they concluded that only a limited number of studies were methodologically sound, with criterion validity being either unreported or having negative ratings in most studies. According to the systematic review of the measurement of physician-patient communication, although reliability and structural validity were rated mainly of fair quality, criterion validity was not investigated (Zill et al. 2014).

As for medical record evaluations (also called medical chart audits, particularly when explicit criteria are employed), only a few studies have attempted to investigate criterion validity. McDermott et al. (2006) compared the chart review and direct observations among 51 patients with asthma; however, this study dealt with one specific disease with an explicit evaluation, such as "whether peak flow was obtained within one hour of arrival." Stange et al. (1998) compared the medical record evaluations and direct observation of patients visits; however, the medical record review was an explicit one, performed by eight trained research nurses. Ramsey et al. (1989) investigated the predictive validity of the American Board of Internal Medicine (AMIM) certification by comparing the ratings of clinical skills by professional associates and those of medical record evaluations between certified and noncertified physicians; however, the medical records were reviewed using explicit criteria focusing on acute infections (respiratory and urinary infections) and some chronic diseases (hypertension, diabetes mellitus, and coronary diseases). Moreover, their study compared certified and noncertified physicians as groups, not individually (Ramsey et al. 1989). Thus, to the best of our knowledge, this is the first study, albeit a pilot one, that investigates the criterion validity of medical record evaluations of various diseases conducted by comparing the assessments of medical records with those by program directors.

Methods

Ethics approval and consent to participate

This study was designed by the peer-review system (PRS) committee, as described previously (Kameoka et al. 2014). It was approved by the Tohoku University Research Ethics Board (2014-1-651), St. Luke's International Hospital Ethics Committee Institutional Review Board (IRB) (15-R043), Kawakita General Hospital IRB (2015-0007), and Seirei Hamamatsu General Hospital IRB (No. 1787). All senior residents were informed of the study in detail, and were given the option to opt out of participation. Informed consent from the patients was not required for this retrospective study.

Evaluation of medical records

The procedure of the peer review of medical records was the same as the one we used in the reliability study (Kameoka et al. 2014). Briefly, reviewers visited each hospital and evaluated medical records (all outpatient care medical records and an inpatient care summary sheet) based on the evaluation sheet, which comprised 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). Among the fifteen items for evaluating quality of care, thirteen items mainly evaluate the "process" of patient care, whereas the fourteenth item evaluates the "outcome" of patient care; the fifteenth item is an overall evaluation. "Outcome" is defined as the morbidity and mortality of the patient, regardless of the process: for example, if the diagnosis and treatment of multiple myeloma was delayed and a pathological fracture occurred with prolonged morbidity, then its score should be low; but, if the diagnosis and treatment of multiple myeloma was slightly delayed but complete remis-

sion was eventually obtained without any morbidity, then its score should not be low.

The PRS committee selected five reviewers—the same reviewers who were used in the reliability analysis because they were well trained and well aware of the criteria for evaluation. They were all males, aged between 45 and 56 years, working as teaching doctors in general hospitals (400 to 1,166 beds), and had the reputation for being experts in broad fields of internal medicine.

Three hospitals were selected by the PRS committee based on the following criteria: (1) general hospitals outside the Tohoku region (northeastern Japan), (2) hospitals where senior residents—here defined as doctors with three years' experience after graduation from medical school—see outpatients independently (without the supervision of senior doctors) and (3) hospitals with IRB approval for the study. The selected hospitals (St. Luke's International Hospital, Tokyo, Japan; Kawakita General Hospital, Tokyo, Japan; Seirei Hamamatsu General Hospital, Hamamatsu, Japan) were tertiary-level community teaching hospitals, which had 520, 331, and 750 beds, respectively. All of these hospitals maintained electronic medical records.

Patients (outpatients) were selected by a representative at each hospital and a member of the PRS committee based on the following criteria: (1) they visited the hospital for the first time between April 2014 and March 2015 and were eventually hospitalized, (2) they were independently seen by senior residents, and (3) their final diagnoses were related to the field of internal medicine, regardless of the specific diagnosis.

Evaluation of residents by program directors

Program directors, all of whom were males, evaluated residents' clinical performances independent of the peer review of medical records by using an evaluation sheet comprising ten items. This evaluation sheet was designed by the PRS committee, based on the American Board of Internal Medicine (ABIM) rating form (Haber and Avins 1994) (Table 1). Among the items, "medical knowledge" was excluded because it could not be evaluated in the peer review system of the medical records; however, two new items were added: "medical record keeping" and "outcome of patients." The former was added not only because it was included in some previous reports (Rethans et al. 1994; Goulet et al. 2007) but also because we wanted to determine whether this system does more than just evaluate the quality of medical record keeping. The latter was added because it was one of the exact characteristics we wanted to measure, as described later.

Data analysis

The mean scores and standard deviations of the evaluation sheets were calculated for each item. Pearson's correlation coefficients were calculated between the mean scores of the medical record assessments (C1-15) and the program directors' assessments (P1-10) for each resident. P values < 0.05 were considered statistically significant. Because the sample number was small, bootstrap resampling method was applied to measure the accuracy of correlation coefficients, by calculating 95% confidence intervals. Correlations were considered statistically meaningful when confidence intervals did not contain 0. Bootstrapping is a highly computer-intensive statistical procedure for estimating the sampling distribution of an estimator by sampling a replacement of the original sample (Kisielinska 2013). STATA 11 was used for the statistical analysis.

Results

Evaluation of medical records

We selected 65 patients who were seen by 13 senior residents—9 males and 4 females—in the 3 hospitals (5 patients per resident). The diagnoses of 65 cases included 16 cardiovascular diseases, 12 respiratory diseases, 11 gastrointestinal diseases, 9 neurological diseases, and 17 other diseases.

The total time required for an evaluation of both record keeping (14 items) and quality of care (15 items) ranged from 500 to 855 minutes (mean: 710 minutes, 10.9 minutes per patient). The mean time required per patient was comparable to that of the previous reliability study (Johnson et al. 2011) (11.3 minutes per patient). Since the purpose of the present study was to compare the quality of care between the two assessments, only quality of care data (15 items) were used for the analyses described below.

The mean scores and standard deviations of the 15 items assessed by the peer-review system are shown in Table 2. The average score (standard deviation) of items

Table 1. Correspondence between assessment items: program director assessments vs. ABIM rating form.

Items of assessments by program directors		Items of ABIM rating form
P1	History taking	History taking skills
P2	Physical examination	Physical examination skills
P3	Clinical reasoning	Clinical judgment
P4	Treatment	Medical care
P5	Clinical skills	Procedural skills
P6	Medical record keeping	-
P7	Attitudes towards patients and family members	Attitudes and professionalism
P8	Cooperation with other members	Interpersonal skills
P9	Outcome of patients	-
P10	Overall evaluations	Overall competence
	-	Medical knowledge

ABIM, the American Board of Internal Medicine.

Table 2. Mean scores for medical record evaluation items.

	Items	mean scores (standard deviations)
C1	Is he/she taking a history related to the chief complaint?	3.06 (0.83)
C2	Is he/she taking a history unrelated to the chief complaint?	2.78 (0.79)
C3	Is he/she performing a CC-focused physical examination?	2.96 (0.85)
C4	Is he/she performing a systemic physical examination?	2.82 (0.87)
C5	Is he/she ordering diagnostic tests appropriately?	3.58 (0.71)
C6	Is he/she interpreting the results of examinations appropriately?	3.64 (0.81)
C7	Is he/she adequately listing differential diagnoses?	3.38 (0.90)
C8	Is he/she treating the patient appropriately?	3.65 (0.82)
C9	Is he/she following EBM?	3.56 (0.88)
C10	Are the medical records well-written?	3.57 (0.75)
C11	Is he/she making referrals to other doctors, if necessary?	3.54 (0.83)
C12	Does he/she have empathy towards the patient?	2.98 (0.87)
C13	Is the explanation to the patient and family members enough?	2.44 (1.19)
C14	Outcome assessment of the patient	3.67 (0.75)
C15	Overall assessment of patient care	3.30 (0.78)
	Average of C1 through C15.	3.26 (0.93)

Table 3. Mean scores for program director assessment items.

	Items	mean scores (standard deviations)
P1	History taking	3.54 (0.97)
P2	Physical examination	3.46 (0.88)
P3	Clinical reasoning	3.62 (1.04)
P4	Treatment	3.85 (0.69)
P5	Clinical skills	3.85 (0.80)
P6	Medical record keeping	3.85 (1.21)
P7	Attitude towards patients and family members	4.15 (0.99)
P8	Cooperation with other members	4.08 (0.95)
P9	Patient outcome	3.38 (0.77)
P10	Overall evaluation	3.69 (1.03)

C1 through C15 (quality of care) for the 65 cases was 3.26 (0.93). The general tendency of the scores was similar to that of the previous study (Kameoka et al. 2014): the total mean score for item C14 (outcome) was high despite the relatively low scores for items C1 through C4 (history taking and physical examination). The average scores (standard deviation) of C1 through C15 of the three hospitals were 3.15 (0.88), 3.22 (0.98), and 3.50 (0.84).

Evaluation of residents by program directors

The mean scores and standard deviations of the 10 items assessed by the program directors are shown in Table 3. The average scores (standard deviation) of P1 through P10 of the three hospitals were 3.42 (1.01), 4.33 (0.30), and 3.63 (0.49), indicating that the assessment by the program director of the second hospital was somewhat lenient.

Correlation coefficients between the scores of the two methods

The correlation coefficients between the scores of medical record assessments (C1-C15) and program directors' assessments (P1-10) are shown in Table 4. The scattergrams of the representatives are presented in Fig. 1.

With regard to Pearson's analysis, the correlation coefficients were statistically significant in three pairs of medical record and program directors' assessments including C9 (following EBM) versus P3 (clinical reasoning). With regard to bootstrap analysis, the correlation coefficients were statistically significant in an additional five pairs including C1 (history taking related to chief complaint) and P1 (history taking). In contrast, the correlation coefficient between C3 (chief complaint-focused physical examination) or C4 (systemic physical examination) and P2 (physical examination) was low ($r = 0.132$, -0.091), and P7 (attitude toward patients and family members) had no positive correlations with C12 (empathy toward the patient) ($r = -0.089$) or C13 (explanation given to the patient and family members) ($r = -0.089$). Although statistically insignificant, P4 (treatment) was mostly correlated with C4 (treatment) ($r = 0.299$) and, conversely, C4 was mostly correlated with P4, which was unexpected because treatment was, in most

Table 4. Correlation coefficients between medical record evaluations and assessments by program directors.

		P1 History taking	P2 Physical examination	P3 Clinical reasoning	P4 Treatment	P5 Clinical skills	P6 Medical records	P7 Attitude[1)]	P8 Cooperation[2)]	P9 Outcome of patients	P10 Overall evaluation
C1	Is he/she taking a history related to the chief complaint?	**0.509***	**0.530**	0.269	0.156	**0.564****	**0.515***	0.126	0.244	0.243	0.299
C2	Is he/she taking a history unrelated to the chief complaint?	0.205	0.243	0.129	−0.096	**0.342**	**0.322**	0.050	0.200	0.171	0.079
C3	Is he/she performing a CC-focused physical examination?	0.121	0.132	0.065	−0.042	0.227	**0.315**	0.078	0.247	0.171	0.010
C4	Is he/she performing a systemic physical examination?	−0.150	−0.091	−0.161	−0.334	0.125	0.030	−0.195	−0.012	−0.101	−0.239
C5	Is he/she ordering diagnostic tests appropriately?	−0.044	0.023	0.151	−0.244	0.154	0.179	−0.328	0.018	0.054	−0.124
C6	Is he/she interpreting the results of examinations appropriately	0.042	−0.034	0.240	0.095	0.190	0.281	0.124	**0.346**	**0.300**	0.160
C7	Is he/she adequately listing differential diagnoses?	0.299	0.246	0.102	0.025	**0.440***	**0.325**	0.135	0.121	0.098	0.196
C8	Is he/she treating the patient appropriately?	**0.308**	0.158	0.262	**0.307**	0.242	0.116	0.043	0.197	0.028	0.292
C9	Is he/she following EBM?	**0.564****	**0.445**	**0.585****	0.263	**0.394**	**0.443**	0.177	**0.512***	**0.365**	**0.460**
C10	Are the medical records well-written?	0.189	0.252	**0.331**	−0.106	**0.369**	**0.410**	−0.044	0.241	**0.317**	0.184
C11	Is he/she making referrals to other doctors, if necessary?	−0.153	−0.056	0.039	−0.426	0.255	0.007	−0.139	0.043	0.042	−0.044
C12	Does he/she have empathy towards the patient?	−0.262	−0.151	−0.078	−0.300	0.250	−0.124	−0.089	−0.017	−0.036	−0.101
C13	Is the explanation to the patient and family members enough?	−0.216	−0.202	−0.148	−0.215	−0.048	−0.306	−0.281	−0.142	−0.274	−0.223
C14	Outcome assessment of the patient	**0.302**	0.192	**0.331**	0.239	0.205	**0.359**	0.025	**0.316**	0.150	0.159
C15	Overall assessment of patient care	**0.424**	**0.316**	0.277	0.279	**0.477***	**0.410**	0.257	**0.339**	0.255	**0.360**

Items in bold indicate correlation coefficients higher than 0.3.
*Confidence intervals did not contain 0 in Bootstrap analysis.
**P values were less than 0.05 in Pearson's analysis.
[1)]Attitude towards patients and family members.
[2)]Cooperation with other members.

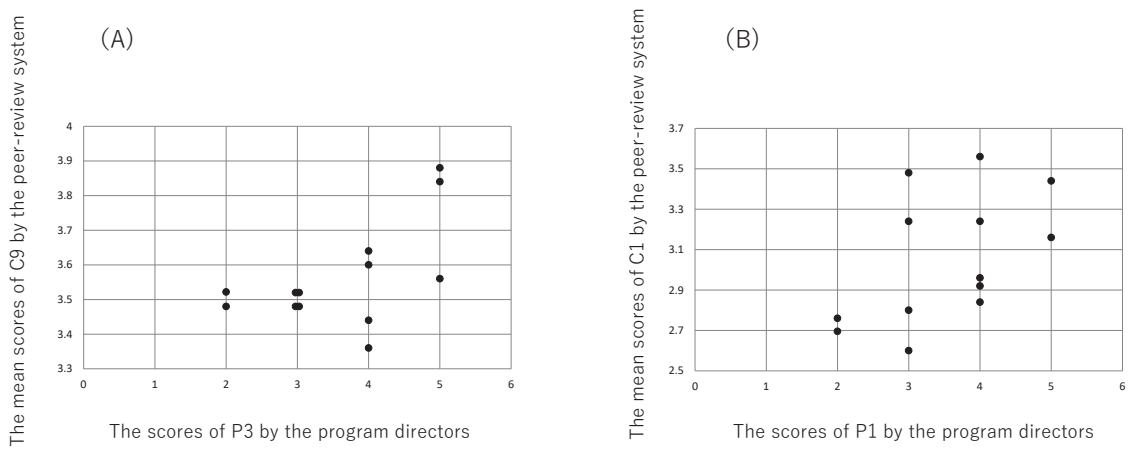

Fig. 1. Scattergram depicting the scores by the peer-review system versus the scores by the program directors.
(A): C9 (following EBM) versus P3 (clinical reasoning), which was significant both in Pearson's analysis (r = 0.585, P = 0.0358) and in bootstrap analysis (95% confidence interval: 0.013-0.928). (B): C1 (history taking related to chief complaint) versus P1 (history taking), which was not significant in Pearson's analysis (r = 0.509, P = 0.0753), but was significant in bootstrap analysis (95% confidence interval: 0.034-0.847).

cases, not completed as part of the initial outpatient care. C10 (well-written medical records) was mostly correlated with P6 (medical records) (r = 0.410), and P10 (overall evaluation) showed the second highest correlation coefficient with C15 (overall evaluation) (r = 0.360).

Discussion

In the current pilot study, we attempted to evaluate the criterion validity of our peer-review system by using evaluations by program directors as a criterion standard of evaluating the clinical competence of residents. The results revealed that we were able to obtain statistically significant correlations among some, but not many, items. The reasons why we could not obtain significant correlations among many items could be attributed to the following factors: (1) the number of residents was too small to show statistical significance, (2) the two scales were measured in very different ways to be closely correlated and, thus, (3) the evaluation by program directors may not serve as a golden standard of measuring the residents' competence. Given these considerations, the results of the current study, in which significant correlations were obtained among some major items including history taking, are promising for the future development and utilization of this system.

The traditional concept of validity, which comprises content, criterion, and construct validity, has been challenged since Messick (1994) proposed six aspects (content, substantive, structural, generalizability, external, and consequential), and Kane proposed four aspects (scoring, generalization, extrapolation, and implications) (Kane 2013; Cook et al. 2015). However, these arguments stem from a practical point of view and are relevant in cases such as in deciding whether one passes or fails a certain test. For example, "extrapolation" in Kane's framework refers to real-world instead of test-world performance, which is actually the concept of "criterion validity." Therefore, we have used the term "criterion validity" in this paper.

The present study suggested that some items such as

"physical examination" and "attitude toward patients" might be difficult to evaluate by merely reading medical records. Assessing the ability to write down physical findings may not be difficult, but assessing the ability to detect abnormal physical findings may be challenging. Measuring physician-patient relationships, such as empathy toward patients, by reviewing medical records may also be challenging. Velez and others' comparison of the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) scores measuring patient satisfaction with the Four Habits Coding Scheme (4HCS) scores measuring physicians' bedside communication skills including empathy showed no correlations (Velez et al. 2017). Although, the Leicester Assessment Package—a tool to assess consultation competence with established reliability—has been developed, this system uses simulated patients, not a workplace-based evaluation (Fraser et al. 1994). In another study, although the factors relating to the communication skills of physicians were reviewed, the analyses were based on videotaped primary care visits, not on medical record evaluations (Tallman et al. 2007). Hemmerdinger et al. (2007) conducted a systematic review of tests of empathy in medicine and reported that, among 36 identified instruments of empathy assessments, 8 demonstrated evidence of reliability and validity, of which 6 were self-rated measures, 1 was a patient-rated measure, and 1 was an observer-rated measure. No instruments using medical record evaluations have succeeded in demonstrating reliability and validity.

Methods for evaluating the clinical competence of residents include (1) performance examinations such as OSCE and CEX, (2) ratings by program directors or other staff members, and (3) medical record evaluations with either implicit or explicit criteria (Holmboe and Hawkins 1998). The advantages and disadvantages of these methods are shown in Table 5. Despite some disadvantages, such as poor inter-rater reliability particularly when implicit criteria are used (Hayward et al. 1993; Hofer et al. 2004; Goulet et al. 2007), and time-consuming procedures (Holmboe and Hawkins 1998), medical record evaluations have many advantages over other methods including the following: (1) they can be conducted in a longitudinal manner because medical records are usually available over time; (2) detailed analyses of performances are possible, for example, physicians detecting specific physical findings, such as lymphadenopathy; and, most of all, (3) this is the only method of assessing "patient outcomes," the importance of which is being increasingly recognized in medical education (Dauphinee 2012; Gonnella and Hojat 2012). Thus, overcoming the disadvantages described above and establishing a system of medical record evaluations with high reliability and validity is important.

In addition, medical record evaluations will enable us to compare the differences between clinical performances by the same individuals in various situations. For example, in the present study, the mean scores of the overall evaluations of five patients treated by one resident were 3.4, 3.2, 3.0, 3.0, and 3.0 (mean: 3.12), showing consistent scoring, whereas those treated by another resident were 4.2, 4.0, 3.8, 2.8, and 2.6 (mean: 3.48), showing a large variation (data not shown). Although these variations may be due to the strength and weakness of each resident, they could also perhaps be because the efforts they make may not be consistent. The extent of the efforts they make can only be measured by unannounced evaluations, which can be determined only by medical record evaluations.

Our study has several limitations. First, as described previously, the sample size, particularly the number of residents, was too small to show statistical significance in some items; therefore, we attempted to overcome this limitation by utilizing the bootstrap method. Second, the reliability of the evaluations by the program directors has not been as vigorously pursued in the current study as in the study on

Table 5. Methods of evaluating clinical competence of residents.

	Performance examinations		Ratings by program directors	Medical record audits	
	OSCE	CEX		explicit criteria	implicit criteria
Time needed	moderate		short	long	
Reliability	high	moderate	high	high	low
Validity	moderate	high	moderate	low	high
Patient reality	artificial	real	real	real	
Physician-based or case-based	both		physician-based only	both	
Process measures	high		moderate	low〜high	
Outcome measures	low		moderate	high	
Detailed analyses	possible		impossible	possible	
Unannounced analyses	impossible		possible	possible	
Longitudinal analyses	difficult		difficult	easy	

OSCE, objective structured clinical examinations; CEX, clinical evaluation exercises.

the medical record peer-review system. Third, as described above, we assumed that an evaluation by the program directors should serve as a criterion standard for evaluating the clinical competence of residents; however, a golden standard of criterion validity of the evaluation may not have been established.

In conclusion, in the current pilot study, we were able to obtain promising results of the criterion analysis of our peer-review system. From the point of view of time consumption, our system seems feasible, because the time required for the review (11.3 minutes per patient) was acceptable, as compared with other methods such as case-based discussions, which took 25 ± 16 minutes in one study (Brittlebank et al. 2013). Although we need to further improve the procedure, such as by establishing an efficient training system for general reviewers or amending some items in our evaluation sheet, we hope our peer-review system will enable us to evaluate the quality of patient care and use it as an outcome evaluation of medical education in the future.

Acknowledgments

This project was supported in part by Grants-in-aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan (25460609). We thank Drs. Hiroshi Kanatsuka, Yoshiyuki Ueno, Akira Imatani, Atsushi Takeda, and Masaki Kanemura (Tohoku University Graduate School of Medicine) for cooperating as members of the PRS committee, and Dr. Osamu Takahashi (St. Luke's International Hospital), Dr. Naoki Asazuma (Kawakita General Hospital), and Dr. Satoru Nakayama (Seirei Hamamatsu General 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. Mitsunori Miyashita (Tohoku University) and Dr. Kiyoshi Kinjo (Okinawa Chubu Hospital) for the critical reading of the manuscript, and Ms. Kinue Utsumi, Ms. Saori Sato, Ms. Emi Koguma, Mr. Yutaro Arata, Mr. Shinya Otsuki, Ms. Nozomi Chubachi, and Ms. Fumie Takahashi (Office of Medical Education, Tohoku University) for their technical assistance.

Conflict of Interest

The authors declare no conflict of interest.

References

Al Ansari, A., Donnon, T., Al Khalifa, K., Darwish, A. & Violato, C. (2014) The construct and criterion validity of the multi-source feedback process to assess physician performance: a meta-analysis. *Adv. Med. Educ. Pract.*, **5**, 39-51.

Al-Wassia, H., Al-Wassia, R., Shihata, S., Park, Y.S. & Tekian, A. (2015) Using patients' charts to assess medical trainees in the workplace: a systematic review. *Med. Teach.*, **37** Suppl 1, S82-87.

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

Brittlebank, A., Archer, J., Longson, D., Malik, A. & Bhugra, D.K. (2013) Workplace-based assessments in psychiatry: evaluation of a whole assessment system. *Acad. Psychiatry*, **37**, 301-307.

Cook, D.A., Brydges, R., Ginsburg, S. & Hatala, R. (2015) A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med. Educ.*, **49**, 560-575.

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

Etheridge, L. & Boursicot, K. (2013) Performance and workplace assessment. In *A practical guide for medical teachers*, 4th ed., edited by Dent, J.A. & Harden, R.M. Churchill Livingstone, London, pp. 307-313.

Fraser, R.C., McKinley, R.K. & Mulholland, H. (1994) Consultation competence in general practice: testing the reliability of the Leicester assessment package. *Br. J. Gen. Pract.*, **44**, 293-296.

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.

Haber, R.J. & Avins, A.L. (1994) Do ratings on the American Board of Internal Medicine Resident Evaluation Form detect differences in clinical competence? *J. Gen. Intern. Med.*, **9**, 140-145.

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.

Hemmerdinger, J.M., Stoddart, S.D. & Lilford, R.J. (2007) A systematic review of tests of empathy in medicine. *BMC Med. Educ.*, **7**, 24.

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., 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.

Holmboe, E.S. & Hawkins, R.E. (1998) Methods for evaluating the clinical competence of residents in internal medicine: a review. *Ann. Intern. Med.*, **129**, 42-48.

International Public Relations Association (IPRA) (1994) Public relations evaluation: professional accountability. IPRA Gold Paper, 11.

Johnson, G., Booth, J., Crossley, J. & Wade, W. (2011) Assessing trainees in the workplace: results of a pilot study. *Clin. Med. (Lond)*, **11**, 48-53.

Kameoka, J., Okubo, T., Koguma, E., Takahashi, F., Ishii, S. & Kanatsuka, H. (2014) Development of a peer review system using patient records for outcome evaluation of medical education: reliability analysis. *Tohoku J. Exp. Med.*, **233**, 189-195.

Kane, M.T. (2013) Validating the interpretations and uses of test scores. *J. Educ. Meas.*, **50**, 1-73.

Kisielinska, J. (2013) The exact bootstrap method shown on the example of the mean and variance estimation. *Comput. Stat.*, **28**, 1061-1077.

Li, H., Ding, N., Zhang, Y., Liu, Y. & Wen, D. (2017) Assessing medical professionalism: a systematic review of instruments and their measurement properties. *PLoS One*, **12**, e0177321.

McDermott, M.F., Lenhardt, R.O., Catrambone, C.D., Walter, J. & Weiss, K.B. (2006) Adequacy of medical chart review to characterize emergency care for asthma: findings from the Illinois Emergency Department Asthma Collaborative. *Acad. Emerg. Med.*, **13**, 345-348.

Messick, S. (1994) The interplay of evidence and consequences in the validation of performance assessments. *Educ. Res.*, **23**, 13-23.

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**, e161-169.
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.
Ramsey, P.G., Carline, J.D., Inui, T.S., Larson, E.B., LoGerfo, J.P. & Wenrich, M.D. (1989) Predictive validity of certification by the American Board of Internal Medicine. *Ann. Intern. Med.*, **110**, 719-726.
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.
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.
Stange, K.C., Zyzanski, S.J., Smith, T.F., Kelly, R., Langa, D.M., Flocke, S.A. & Jaén, C.R. (1998) How valid are medical records and patient questionnaires for physician profiling and health services research? A comparison with direct observation of patients visits. *Med. Care*, **36**, 851-867.
Tallman, K., Janisse, T., Frankel, R.M., Sung, S.H., Krupat, E. & Hsu, J.T. (2007) Communication practices of physicians with high patient-satisfaction ratings. *Perm. J.*, **11**, 19-29.
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.
Velez, V.J., Kaw, R., Hu, B., Frankel, R.M., Windover, A.K., Bokar, D., Rish, J.M. & Rothberg, M.B. (2017) Do HCAHPS doctor communication scores reflect the communication skills of the attending on record? A cautionary tale from a tertiary-care medical service. *J. Hosp. Med.*, **12**, 421-427.
Yates, P.J. (2013) The Mini-CEX is not valid or reliable in assessing the clinical competence of higher surgical trainees. *Ann. R. Coll. Surg. Engl.*, **95**, 1-4.
Zill, J.M., Christalle, E., Muller, E., Harter, M., Dirmaier, J. & Scholl, I. (2014) Measurement of physician-patient communication: a systematic review. *PLoS One*, **9**, e112637.