Development, implementation and reliability assessment of an emergency physician performance evaluation tool

Jeremy Etherington, MD; Grant Innes, MD; James Christenson, MD; Jonathan Berkowitz, PhD; Robert Chamberlain, MD; Ross Berringer, MD; Cosmas Leung, MD

ABSTRACT

Evaluation of physician practice is necessary, both to provide feedback for self-improvement and to guide department heads during yearly evaluations.

Objective: To develop and implement a peer-based performance evaluation tool and to measure reliability and physician satisfaction.

Methods: Each emergency physician in an urban emergency department evaluated their peers by completing a survey consisting of 21 questions on effectiveness in 4 categories: clinical practice, interaction with coworkers and the public, nonclinical departmental responsibilities, and academic activities. A sample of emergency nurses evaluated each emergency physician on a subset of 5 of the questions. Factor analysis was used to assess the reliability of the questions and categories. Intra-class correlation coefficients were calculated to determine inter-rater reliability. After receiving their peer evaluations, each physician rated the process's usefulness to the individual and the department.

Results: 225 surveys were completed on 16 physicians. Factor analysis did not distinguish the nonclinical and academic categories as distinct; therefore, the survey questions fell into 3 domains, rather than the 4 hypothesized. The overall intra-class correlation coefficient was 0.43 for emergency physicians, indicating moderate, but far from perfect, agreement. This suggests that variability exists between physician evaluators, and that multiple reviewers are probably required to provide a balanced physician evaluation. The intra-class correlation coefficient for emergency nurses was 0.11, suggesting poor reliability. Overall, 11 of 15 physicians reported the process valuable or mostly valuable, 3 of 15 were unsure and 1 of 15 reported that the process was definitely not valuable.

Conclusion: Physician evaluation by a single individual is probably unreliable. A useful physician peer evaluation tool can be developed. Most physicians view a personalized, broad-based, confidential peer review as valuable.

RÉSUMÉ

L'évaluation des médecins dans leurs fonctions est nécessaire tant pour leur offrir des conseils d'amélioration personnelle que pour guider les chefs de départements lors des évaluations annuelles.

Objectif : Développer et mettre en place un outil d'évaluation par les pairs de la compétence et de mesurer sa fiabilité et le taux de satisfaction des médecins face à celui-ci.

Méthodes : Chaque urgentologue du département d'urgence d'un hôpital urbain évalua ses collègues en répondant à un sondage composé de 21 questions sur l'efficacité dans 4 catégories : pratique clinique, interactions avec les collègues de travail et le public, responsabilités autres que cli-

From the Department of Emergency Medicine, Providence Health Care, St. Paul's Hospital Site, Vancouver, BC

Received: Dec. 29, 1999; final submission received: June 15, 2000; accepted: July 3, 2000

This article has been peer reviewed.

niques au département et activités académiques. Un échantillon d'infirmières à l'urgence évalua chaque urgentologue pour un sous-groupe de 5 des questions. On eut recours à l'analyse factorielle pour évaluer la fiabilité des questions et des catégories. Les coefficients de corrélation intra-classes furent calculés afin de déterminer la fiabilité inter-évaluateurs. Après avoir reçu leurs évaluations par leurs pairs, chaque médecin évalua l'utilité du processus pour un individu et pour le département.
Résultats : 225 sondages évaluant 16 médecins furent remplis. L'analyse factorielle ne faisait pas de distinction entre les catégories non cliniques et académiques; par conséquent, les questions du sondage s'inséraient dans 3 catégories plutôt que dans les 4 de l'hypothèse de départ. Le coefficient de corrélation intra-classes global était de 0,34 pour les urgentologues, indiquant une concordance modérée, mais loin d'être parfaite. Ces résultats évoquent l'existence d'une variabilité entre les évaluateurs médecins et la nécessité de recourir à de multiples évaluateurs pour arriver à une évaluation équilibrée. Le coefficient de corrélation intra-classes pour les infirmières était de 0,11, indiquant une fiabilité faible. Globalement, 11 des 15 médecins considérèrent le processus valable ou très valable, 3 sur 15 étaient incertains et 1 sur 15 jugea que le processus n'était définitivement pas valable.
Conclusion : L'évaluation des médecins par une seule personne n'est probablement pas fiable. Il est possible d'élaborer un outil d'évaluation du travail du médecin par les pairs. La plupart des médecins considèrent comme valable une revue par les pairs confidentielle et réunissant plusieurs personnes.

Key words: peer review, physician assessment, performance evaluation, physician competence

Introduction

Like much of health care practice, the evaluation of physician performance is undergoing rapid change. Physicians have traditionally accepted evaluation as part of the process of obtaining a medical degree, postgraduate training and licensure; but, once in practice, they rarely participate in formal performance evaluation unless issues of incompetence or impairment are registered with provincial licensing bodies. Recently, however, a renewed international interest in maintenance of professional knowledge and skills has flourished.[1–5] This has led to the development of performance-evaluation programs by national certifying bodies, provincial and state colleges, professional associations, hospitals and their respective medical departments.[6–10]

Comprehensive databases are a rarity in Canadian emergency departments (EDs), and it is difficult for medical directors (department heads) to complete a systematic review of physician performance compared with that of a peer cohort.[11] As a result, physician re-appointment is often based on anecdotal or incomplete information and is, therefore, potentially biased. In our department, like many others, emergency physicians (EPs) frequently work together on double- or triple-covered shifts. This gives us the opportunity to observe the practice of our peers.

After establishing that our EPs wished to be evaluated comprehensively by more than one individual (the department head acting alone), we postulated that peer review by EPs and nurse coworkers could provide a reliable and valid evaluation of physician performance. Our objective was to develop a tool to evaluate physician performance based on peer and coworker assessment of practice. The steps in this process are: tool development, pilot testing, tool revision, reliability assessment, and validation. This article addresses development, pilot testing, revision, and reliability assessment of the performance evaluation tool (PET). For the purpose of this work, peer review is defined as "a formal or informal assessment of the medical knowledge, technical skills, and interpersonal skills of a physician as determined by a professional colleague."

Methods

Setting

This project was conducted at St. Paul's Hospital (SPH), an inner-city tertiary-level Vancouver teaching centre. SPH has an annual census of 54,000 patient visits, is staffed by full-time EPs and is the base site for a CCFP-EM residency program.

PET development

At a departmental retreat in May 1995 we established, by consensus, minimal practice standards for EPs who work in our department (Table 1). These standards, which outline physician expectations for clinical and academic performance, were subsequently ratified by all members of the department. This process ensured that physicians understood and subscribed to the standards of practice from

which evaluation criteria were derived. We then identified 4 key attributes (domains) that broadly defined these performance standards: clinical effectiveness, interaction effectiveness (relationships with colleagues, coworkers and the public), effectiveness in discharging nonclinical departmental responsibilities, and academic effectiveness (Table 2).

Next, also by consensus, we developed a PET consisting of 21 questions designed to elicit information about the 4 domains. Each answer was scored from 1 to 5 using the following modified Likert scale: "your peer is 1) among the least proficient in the group, 2) below average, 3) average, 4) above average, 5) among the most proficient in the group." In addition, each question encouraged respondents to provide detailed, constructive written feedback. To enhance survey reliability, we defined specific behaviours that would warrant a high score for the parameter being evaluated (Table 2).

Peer practice exposure

In St. Paul's ED, clinicians have substantial exposure to their peers' practice. In addition to informal "curbside consultation" over interesting cases and difficult x-rays, there are several established mechanisms for formal practice exposure. Twice during every clinical shift, the EPs on duty make rounds and briefly review the status of all patients in the department. At monthly morbidity and mortality rounds, we review all deaths that occurred in the department, in addition to cases that were interesting or associated with patient morbidity. Several times annually, we each perform formal quality audits on our peers' ED charts. Based on this relatively high level of exposure, we felt peer review was viable.

Pilot study and PET refinement

In June 1996, the draft PET was pilot tested to assure clarity and to identify problems with the question format and scoring system. All full-time EPs at St. Paul's Hospital completed an evaluation for each of their peers. Based on physician feedback, the PET was revised and several questions were modified. A summary statement was added, requiring evaluators to mark one of the following responses on each PET: a) "this physician's performance meets or exceeds the baseline clinical and academic expectations for an emergency physician working at St. Paul's Hospital" or b) "this physician requires formal review by St. Paul's Emergency Associates in conjunction with the department head." The physicians agreed that, until the PET was refined and validated, these summary responses would be the only information provided to the department head. They further agreed that if 2/3 of the group recommended that a physician undergo formal review, the complete performance evaluation for that physician would be made available to the department head and all physicians would assist with the subsequent review process.

At this time, we derived an abbreviated PET comprised of a subset of 5 questions physicians felt nurses could best answer (Table 3). Our purpose was to determine whether nurse assessment of physician performance could provide meaningful information not available from physician assessment alone.

PET implementation

In October 1997, each EP in active practice at St. Paul's Hospital completed the PET for every other physician in the group. All data were collated and analyzed by an independent third party (Wilson Banwell and Associates, Vancouver) to ensure confidentiality and respondent anonymity. Based on these evaluations, domain-specific and overall mean scores were determined for each physician. Mean scores in each domain and the overall mean score were con-

Table 1. Minimal standards for emergency practice

Clinical

1. All patients should be approached with dignity
2. All care delivered within the department should be current and evidence-based
3. Standards of care must be acceptable to peer physicians, consulting staff and nursing staff
4. Physicians must be prepared to defend any variance in their patterns of practice
5. Physicians must participate in CME in order to maintain competency
6. Physicians should attend monthly M&M and Grand Rounds
7. Physicians must support ongoing quality improvement projects
8. Physicians must attempt to expedite care and facilitate patient flow when necessary
9. Relationships with peers should be professional and courteous
10. Physicians should be punctual for change-of-shift

Research

1. Physicians should have a working knowledge of research design and statistical methods
2. Physicians should actively participate in research projects
3. Physicians must consistently enter patients in ongoing research trials

Teaching

1. Physicians have responsibility for teaching housestaff at all levels of training
2. Physicians on duty must provide bedside teaching for all housestaff
3. Physicians will conduct a 30-minute teaching seminar on weekdays prior to their 10–6 shift
4. Physicians will present M&M and Grand Rounds on a rotational basis
5. Physicians should present a three-hour EM resident seminar annually

verted to an EP-based rank from 1 to 16. Ties were dealt with by assigning the average of the ranks to the tied scores. The ranks were combined across raters to provide an overall ranking of the physicians.

Concurrently, a sample of 22 full-time emergency nurses (RNs), selected alternately from a list of 43 full-time RNs, evaluated the same physicians by completing the abbreviated PET (Table 3). RN responses were based on the same modified Likert scale. Mean scores for these evaluations were similarly converted to an RN-based rank from 1 to 15.

Scores (mean, range and standard deviation) as well as compiled written comments, question by question, were provided confidentially to each physician. To put individual results in context, we provided the group mean score for each question. EPs were not given access to other physicians' individual scores or rankings.

Validity and reliability

The PET's 21 questions and 4 domains are consensus-based; therefore, they require validation. To determine whether the questions actually measure the intended attributes (domains), factor analysis was performed, using an orthogonal rotation. Factor analysis is a statistical technique that examines the structure of the relationship among variables to determine whether they actually cluster together to explain a smaller number of underlying constructs.[12] In this case, the variables are the 21 questions, and the constructs are the 4 domains to which these questions were initially ascribed. This factor analysis examined all pair-wise correlations between responses (given by the study physicians about their peers) and extracted factors, by matching questions that were most inter-correlated. This process — having several peers evaluate the same physician — may pro-

Table 2. Performance evaluation questions and descriptors

Clinical Domain (CD): Effectiveness as a practitioner of emergency medicine

CD1: Completes appropriate and thorough patient evaluations and treatment plans. The high-scoring physician treats patients expeditiously. He/she is neither precipitous nor slow to treat, performs sufficient rather than exhaustive workups, excludes serious causes of illness, and arranges for continuity of care.

CD2: Determines priority of patient care and manages multiple patients effectively. The high-scoring physician can triage patients appropriately and treat them in order of severity. He/she manages multiple patients in varying stages of assessment and treatment and is able to maintain control of the department despite the pressure and stress produced by simultaneous ongoing activity.

CD3: Performs procedural skills correctly and appropriately. The high-scoring physician can skillfully perform a variety of procedures from simple suturing to placement of chest tubes or central lines, and recognizes the indications, contraindications, complications and options for procedural intervention.

CD4: Demonstrates clinical acumen. The high-scoring physician can readily identify non-urgent, urgent and emergent problems. He/she makes correct clinical diagnoses, prescribes care according to current standards and guidelines, and is neither too slow nor too quick to refer to consultants.

CD5: Practices effectively in all areas of the department: Trauma, Acute Care, Fast Track. The high-scoring physician is a capable practitioner, whether dealing with resuscitations, acute medical and surgical problems, or minor illnesses and injuries.

Interaction Domain (ID): Effectiveness of interaction with colleagues, coworkers and the public

ID1: Interacts professionally and cordially with emergency physicians and medical staff. The high-scoring physician has good interpersonal and communication skills with other emergency physicians and medical staff. He/she confronts interpersonal conflict directly, constructively and in a timely fashion. This physician is able to engage the appropriate parties in conflict resolution, and does not rely on others to solve interpersonal problems.

ID2: Arrives punctually for professional obligations. The high-scoring physician arrives and departs in a timely manner with respect to shifts, rounds, meetings and professional obligations.

ID3: Accepts responsibility at beginning-of-shift. The high-scoring physician is willing to "pitch in" at the beginning of each shift to help clear any backlog of patients waiting to be seen, or to facilitate his/her colleagues' ability to complete their patient-care responsibilities.

ID4: Discharges obligations at end-of-shift. The high-scoring physician does a thorough work-up and reliable disposition for each patient under his/her care. This physician is willing to stay to the end of his/her shift (i.e., 10 hours) to complete patient care, and can account for all patients in the emergency department before hand over.

ID5: Functions as a "team player" with nurses, technologists and clerical workers. The high-scoring physician has good interpersonal relations and communicates effectively with ancillary staff. His/her behaviour demonstrates recognition that the delivery of effective emergency care requires a team effort. This physician encourages and supports coworkers and does not belittle their role in the provision of care.

ID6: Respects patients and families and involves them in clinical decision-making. The high-scoring physician treats patients professionally and with empathy, respects their wishes and privacy, and assists them and their families to make informed treatment decisions.

ID7: Interacts professionally with housestaff and medical students. The high-scoring physician has professional interpersonal relations and communicates effectively with medical students and residents. He/she is able to guide and direct housestaff in their management of patients, and in the development of their skills related to emergency medicine. This physician recognizes that medical education is sequential, and maintains appropriate expectations for the level of housestaff training.

duce an exaggerated level of correlation between the evaluations. Orthogonal rotation is a statistical method that helps reduce inappropriate intra-individual correlation.

Inter-rater reliability for each question was determined by calculating intra-class correlation coefficients (ICC), based on physician scores. Intra-class correlation coefficients are analogous to kappa values; therefore, high coefficients suggest less variability between raters, which implies that fewer peers are necessary to perform an adequate evaluation. Low coefficients suggest that there is substantial variability between different observers responding to the same question. Low coefficients suggest either that many peers are necessary to evaluate the parameter or that the question may not be useful. We considered an ICC of =0.7 to indicate high reliability, 0.3–0.7 to indicate moderate reliability, and <0.3 to indicate poor reliability.

Outcomes

Primary outcomes included EP intra-class correlation (reliability) for each question and overall EP intra-class correlation. Secondary outcomes included RN intra-class correlation (reliability) and EP satisfaction with the performance evaluation process. Physician satisfaction with the process was determined by asking 4 questions: 1) Is participation in peer review a worthwhile process for professional growth? 2) Is the peer review process worthwhile for the improvement of our physician group as a whole? 3) What aspects of the performance evaluation process did you find personally useful? and 4) Should we continue the peer review process?

Sample size and analyses

EP sample size was defined by the number of full-time ED physicians (n = 16) in the group. The RN sample (n = 22)

Table 2. continued

Departmental domain (DD): Effectiveness in discharging non-clinical departmental responsibilities

DD1: Regularly attends and contributes to required rounds and meetings. The high-scoring physician attends rounds and meetings within the department, as well as those required by the Medical Staff Bylaws of the hospital (e.g., quarterly staff meetings). This physician is actively involved in departmental decision-making and is knowledgeable about programs, initiatives and undertakings by the department, and by the hospital insofar as they impact the practice of Emergency Medicine.

DD2: Completes charts and related paperwork legibly, thoroughly and punctually. The high-scoring physician is timely, thorough and clear in the completion of all medical records. He/she discharges obligations to the Medical Records Department as requested, and comprehensively completes documentation required by the department as a matter of business (e.g., WCB forms).

DD3: Assists with departmental responsibilities in the hospital: meetings, committees, planning sessions. The high-scoring physician recognizes the need for Emergency Department representation throughout the hospital, and devotes time to committee work as necessary. This physician accepts such obligations as a matter of medical-staff responsibility, and undertakes to represent the interests of the department through active, regular participation in any areas he/she undertakes.

DD4: Promotes a positive image of the department and hospital through extracurricular professional activities. The high-scoring physician engages in academic or other professional activities that promote the reputation of the department and the hospital. This might include providing medical education to lay groups, paramedics, students or peers at the local, provincial, national or international level. This physician might also participate in academic or advisory committees to such groups as UBC, the BCMA section of Emergency Medicine, the College of Family Physicians or the Royal College of Physicians and Surgeons of Canada, CAEP, ACEP, SAEM, ATLS, ACLS, etc. Non-academic activities might include such things as assisting the hospital foundation with fundraising.

DD5: Understands the hospital mission and works towards its fulfillment. The high-scoring physician provides comprehensive, compassionate care to all patients without regard to race, creed, colour, religion, sex, sexual orientation or lifestyle.

Academic Domain (AD): Effectiveness as an academic emergency physician

AD1: Participates in sufficient CME to maintain a current, comprehensive knowledge base of emergency medicine. The high-scoring physician demonstrates knowledge of and familiarity with current medical literature in Emergency Medicine and current practice guidelines. His/her practice is evidence-based. This physician studies and applies new knowledge and techniques as they are proven to enhance patient outcomes.

AD2: Presents comprehensive, well-prepared academic rounds. The high-scoring physician is articulate in highlighting clinically relevant topics. This physician demonstrates careful preparation, clear understanding and effective organization when presenting rounds. He/she uses a variety of techniques and teaching aids to enhance the quality of rounds.

AD3: Teaches emergency medicine effectively. The high-scoring physician is committed to teaching Emergency Medicine and participates in a wide variety of activities, including bedside teaching, case reviews, housestaff rounds and daily performance evaluation of residents. This physician endeavours to be an effective teacher, acts as an educational resource for housestaff, and endeavours to answer their questions or direct them to appropriate resources.

AD4: Demonstrates a commitment to academic emergency medicine through research, or works consistently to support research activities on the part of colleagues. The high-scoring physician is actively engaged in research. He/she holds an academic appointment in a department or division of the University Faculty of Medicine, and is regularly promoted on the basis of performance. The high-scoring physician who is not actively involved in research modifies his/her practice to assist peers with their work (e.g., agrees to chart pertinent information or provide standardized diagnoses such as ICD 9) and consistently enters patients in approved studies.

represents half of the full-time ED nurses (n = 43). Each physician's scores, including the mean, range and standard deviation for each question, were compared with those for the entire group. An independent statistician (J.B.) computed physician scores, performed factor analysis, and determined intra-class correlations using SPSS Windows V.7.0.

Results

Sixteen full-time ED physicians were eligible to complete the performance evaluation, but one physician who was on medical leave at the time of the study did not evaluate his peers. The remaining 15 evaluated all eligible physicians except themselves (16 – 1); therefore 225 performance evaluations were completed. Each evaluation consisted of 21 responses, for a total of 4,725 evaluable responses. The 22 RNs each provided 5 responses for the available 15 EPs — a total of 1,650 evaluable responses.

Table 3. Performance evaluation questions (nurses' assessment)

CD1:	Completes appropriate and thorough patient evaluations and treatment plans
CD5:	Practices effectively in all areas of the department (Trauma, Acute Care, Fast Track)
ID5:	Functions as a "team player" with nurses, technologists and clerical workers
ID6:	Respects patients and families and involves them in clinical decision-making
DD5:	Understands the hospital mission and works toward its fulfillment

Table 4 lists the 16 physicians (A to P) and shows how physician raters ranked each EP by domain and overall, as well as how RN raters ranked each EP overall. The "SD" column shows standard deviations computed from the ranks in the "Overall rank (EP ranked)" column. Higher SD values indicate greater variability in the assessment of rank by fellow physicians.

Table 5 shows intra-class correlation (reliability) for each question, as well as overall ICC for ranking by physician and nurse evaluators. Correlation coefficients for RN responses range from 0.09–0.13. Coefficients for EP responses in the "academic" and "departmental" domains averaged 0.19, while coefficients for EP responses in the "clinical" and "interaction" domains averaged 0.36. In assessment of overall ranking, the ICC is 0.43 for EPs and 0.11 for RNs.

Factor analysis extracted only 3 factors (domains), rather than the 4 hypothesized; however, the items comprising the factors corresponded extremely closely with those initially hypothesized. Factor 1 (clinical) consists of questions 1 through 5, as hypothesized; factor 2 (interaction) consists of questions 6 through 12, as hypothesized; factor 3 consists of questions 13 through 21, but does not separate into the two hypothesized domains ("departmental" and "academic"). These 3 extracted factors explain 74.3% of the total variability.

Table 4. Performance ranking of emergency physicians (EPs), based on assessments by coworkers (other EPs and registered nurses)

	Categories, assessed by EPs					
					Overall assessment (and SD)	
EP	Clinical practice	Interaction with coworkers and the public	Nonclinical departmental responsibilities	Academic activities	EPs	RNs
A	1	1	4	5	1 (1.8)	1
B	6	3	2	1	2 (2.5)	10
C	7	6	1	3	3 (2.5)	8
D	5	2	6	4	4 (1.6)	9
E	9	5	9	7	5 (3.0)	3
F	3	7	8	9	6 (2.9)	4
G	2	11	10	6	7 (2.4)	12
H	10	14	3	2	8 (3.4)	11
I	11	4	5	12	9 (2.5)	2
J	4	12	7	8	10 (2.6)	13
K	8	8	11	10	11 (2.3)	5
L	13	9	12	13	12 (1.9)	6
M	12	10	14	14	13 (2.1)	14
N	14	15	13	11	14 (2.3)	7
O	15	16	15	15	15 (1.9)	15
P	16	13	16	16	16 (0.8)	–

In response to the question "Is participation in peer review a worthwhile process for professional growth?" 11 physicians felt it was definitely or mostly valuable, 3 were unsure, and 1 felt it was definitely not valuable. In response to the question "Is the peer review process worthwhile for the improvement of our physician group as a whole?" 12 physicians felt it was definitely or mostly valuable, 1 was unsure, and 2 felt it was mostly or definitely not valuable.

Of 15 physicians, 13 appreciated receiving peer feedback, 11 appreciated feedback from the RNs, 12 appreciated the chance to provide peer feedback, and 11 felt this process was better than leaving performance evaluation to the medical director alone. Of the 2 physicians who felt performance

evaluation was not personally useful, both felt their evaluation did not reflect their actual performance, and both felt that this instrument was not the best method of physician assessment. Because of study anonymity, it was impossible to determine whether these physicians scored well or poorly on the PET.

Five physicians felt we should continue the peer review process "as is," 8 felt it should be continued with "minor changes," 1 said it should be discontinued, and 1 failed to respond to the question.

Discussion

In British Columbia, physician performance review is mandated by the Medical Practitioners Act and the Hospital Act.[10] The former charges the College of Physicians and Surgeons of British Columbia to "establish, monitor and enforce standards of practice."[13] The latter directs the hospital medical staff to "discipline its own members in such manner as it thinks fit and, if circumstances warrant, recommend to the board the suspension, restriction, cancellation or non-renewal of the permit of a member to practice within the hospital."[14] A joint policy document developed by the British Columbia Medical Association and the British Columbia Hospitals Association states that criteria for evaluating physicians should be "explicit, objective and developed in conjunction with the medical staff."[15] Nash and coworkers[3] suggest that the evaluation process and criteria should be incorporated into medical staff bylaws. These recommendations imply a duty to evaluate physician performance through the utilization of specific parameters that are developed collaboratively and defined clearly in hospital regulations governing medical staff practice.

According to the Canadian Medical Association, hospital department heads "are responsible for setting standards of care, ensuring that department members are aware of these, and providing evidence to the hospital administration and board that care has been reviewed and is within established standards."[16] Medical staff reappointment is expected to reflect satisfactory review of clinical performance. Although Canadian hospitals have always reviewed some aspects of physician performance, few have developed mechanisms to ensure that a valid and comprehensive process occurs;[2,3] nor has such a process been clearly defined, even in recently developed medical staff bylaws.[17]

Legislated requirements notwithstanding, there is a need for physician evaluation. Physicians benefit from constructive feedback, patients need to know that their doctors are competent, and hospital administrators worry about quality of care, public perception of care, and public protection.[4,7] As well, rising health costs have led to increased monitoring of physician-initiated resource utilization and cost-effectiveness.[3]

Although there are forces driving physician evaluation, there are also forces opposing it. Provincial governments mandate the collection of only rudimentary outpatient data; therefore, few hospitals collect detailed ED information, and valid data regarding physician practice patterns, interactive skills, resource utilization and patient outcomes are seldom available. Currently, in Canada, there is insufficient incentive to justify the expense of more comprehensive data capture and analysis.[5] To complicate matters, physicians may resist evaluation because they fear loss of respect, feelings of inadequacy and failure, or loss of livelihood. In addition, "turf" issues may exist inasmuch as physicians expect their professional associations, not hospitals, to evaluate them. Some suspect that hospital-driven evaluation is more concerned with good business than good medicine. Physicians may also eschew evaluation of their peers be-

Table 5. Inter-rater reliability based on intra-class correlation coefficients (ICC)

	ICC	
Domain*	EPs	RNs
Clinical (CD)		
CD1	0.31	0.11
CD2	0.41	–
CD3	0.38	–
CD4	0.34	–
CD5	0.37	0.09
Interaction (ID)		
ID1	0.27	–
ID2	0.45	–
ID3	0.36	–
ID4	0.35	–
ID5	0.33	0.09
ID6	0.45	0.11
ID7	0.33	–
Departmental (DD)		
DD1	0.14	–
DD2	0.24	–
DD3	0.05	–
DD4	0.17	–
DD5	0.33	0.13
Academic (AD)		
AD1	0.23	–
AD2	0.25	–
AD3	0.23	–
AD4	0.06	–
Overall rank	0.43	0.11

*See Tables 1 and 2 for descriptions.

cause of a natural reluctance to "rock the boat" or because they fear potential litigation by disgruntled colleagues. Nonetheless, physicians participate in performance evaluation for several reasons: chiefly to obtain or maintain qualifications, and to secure or ensure continued employment.[2] From the perspective of personal development, performance evaluation may be seen as a guide to professional growth and a tool to enhance self-esteem.

Faced with the expectation of physician performance evaluation and the lack of data to accomplish this, we developed, implemented, and tested a PET based on anonymous peer review. Our data show that peer review is well accepted by EPs and that broad-based peer evaluation by physicians is probably more reliable (particularly in the "interaction" and "clinical" domains) than evaluation by a single individual.

Not unexpectedly, physicians differ somewhat in their observations of each other. The overall intra-class correlation of 0.43 indicates moderate variability among physician peer reviewers; however, ICC values are sufficiently high to indicate some commonality of opinion. These data suggest that an adequate evaluation is unlikely when only one or two raters assess a peer, and supports our hypothesis that a department head acting alone can not provide optimal physician evaluation.

In our setting, the PET was most reliable in the "interaction" and "clinical" domains, and factor analysis suggests that these are valid domains. Factor analysis did not distinguish between the "academic" and "departmental" domains, suggesting that the behaviours by which we judge physician effectiveness in these areas may, in fact, be similar. These attributes should be combined into a single domain in future iterations of the PET. Furthermore, because reliability was poor in the "academic" and "departmental" domains, other methods of evaluating these parameters should be explored.

Since we work closely with our nurses, we expected their evaluations would provide useful information to help measure EP performance. We were surprised to find that RN-based ratings were inconsistent and that inter-rater reliability for nurse evaluations was uniformly poor (Table 5) on all individual survey items and for overall rating. This renders their scores difficult to interpret. Whereas some may conclude that nurses cannot evaluate physician performance, it is more likely that the questions we asked them may not reflect the knowledge they are best able to provide. Further peer review participation by our nurses will require collaborative work to develop questions that focus on aspects of physician performance they can best evaluate.

We chose to define physician competencies relative to other group members (e.g., "your peer is among the most proficient in the group") rather than taking a more traditional approach, with rankings such as "poor," "fair," "good," "very good" or "excellent." Our experience using the latter system to evaluate residents shows that almost all residents are ranked "good" or "very good" regardless of performance; therefore, such an evaluation system fails to discriminate. In addition, being compared to one's peers provides an implicitly understood benchmark.

Limitations of the study

An important goal of the PET is to provide the department head with useful information for performance evaluation, mentoring and, ultimately, reappointment. In our setting, however, concerns about the use of PET data led to an agreement that the data remain confidential unless 2/3 of the group felt a physician required formal review. It is possible that, in a situation where PET data is provided freely to the department head, survey responses may become less frank or reliable. Peer review will also be less reliable in settings where other "practice exposure mechanisms" (e.g., shift rounds, morbidity/mortality rounds, chart audits) are not in place.

Another goal of the PET is to provide feedback that physicians can use to guide personal and professional development. However, we did not demonstrate that peer feedback leads to measurable changes in practice or behaviour. This is an area for future study.

Questionnaires such as the PET may be limited by several factors: interpersonal conflict among respondents, the respondent's frame of mind during questionnaire completion, exaggerated responses generated by the juxtaposition of questions that have positive or negative emotional impact on the respondent, and respondent fatigue from answering multiple questions.

A potential problem with the PET was its length; future research will address item reduction. By eliminating questions that perform poorly, we can improve overall PET reliability and reduce respondent fatigue. Since questions do, in fact, cluster well to the defined domains, fewer questions may be required to measure that domain adequately in future PET iterations.

We did not evaluate intra-observer reliability in this study, nor was external validity formally tested. Because the survey results are based on rankings within a group, it is possible that good physicians could score poorly in a strong department and that low-functioning physicians could score well in a weak department.

We did not show that peer evaluations correlate with true

"gold standard" measures of physician performance, including clinical outcomes, patient satisfaction, and resource utilization — all difficult and expensive to monitor on an ongoing basis. Future studies should focus on achieving this level of validation. If PET scores correlate with clinical gold standards, then the PET could be considered a valid surrogate marker of physician performance.

Conclusions

Peer performance evaluation is feasible and is accepted by physicians. Broad-based peer evaluations are probably more reliable than evaluations provided by single individuals (e.g., department heads). Peer evaluation by physicians provides moderate reliability for overall ranking and for "interaction" and "clinical" domains. Peers cannot reliably evaluate "departmental" and "academic" performance; therefore, other mechanisms for evaluating these parameters should be explored. RN-based evaluations are not reliable in our setting.

Acknowledgments: We thank the emergency physicians and nursing staff of St. Paul's Hospital for endorsing the concept of peer review and committing the time necessary to complete the performance evaluations. We also acknowledge Wilson Banwell and Associates for their invaluable advice as we developed the performance evaluation tool, and for their assistance in data analysis and reliability assessment of the instrument.

References

1. Davis DA, Norman GR, Painvin A, Lindsay E, Ragbeer MS, Rath D. Attempting to ensure physician competence. JAMA 1990;263:2041-2.

2. Langsley DG. Medical competence and performance assessment: a new era. JAMA 1991;266:977-80.

3. Nash DB, Markson LE, Howell S, Hildreth EA. Evaluating the competence of physicians in practice: from peer review to performance assessment. Acad Med 1993;68:519-22.

4. Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. JAMA 1993;269:1655-60.

5. Norman GR, Davis DA, Lamb S, Hanna E, Caulford P, Kaigas T. Competency assessment of primary care physicians as part of a peer review program. JAMA 1993;270:1046-51.

6. McAuley RG, Henderson HW. Results of the peer assessment program of the College of Physicians and Surgeons of Ontario. CMAJ 1984;131:557-61.

7. Gilbert G. Relating quality assurance to credentials and privileges. QRB 1984;May:130-5.

8. Page GG, Bates J, Dyer SM, Vincent DR, Bordage G, Jacques A, et al. Physician assessment and physician enhancement programs in Canada. CMAJ 1995;153:1723-8.

9. Bates J, Page G, Vincent D, Dyer S. The assessment and enhancement of clinical competence in British Columbia, Canada. Proceedings of the 6th Ottawa Conference on Medical Education, Certification, Licensure, and CME; 1995.

10. College of Physicians and Surgeons of British Columbia. 1st College Quarterly. The Association; Winter 1995.

11. Eliasoph H, Ashdown C. Development, testing and implementation of an emergency services methodology in Alberta. Healthcare Manage Forum 1995;8:31-7.

12. Norman GR, Streiner DL. Biostatistics: the bare essentials. St. Louis: Mosby–Year Book, Inc; 1994. p.129-42.

13. Medical Practitioners Act, British Columbia. Revised Statutes of British Columbia, 1979: chap. 254. Available: http://bbs.qp.gov.bc.ca/bcstats

14. Hospital Act, British Columbia. Revised Statutes of British Columbia, 1995. chap. 200. Available: http://bbs.qp.gov.bc.ca/bcstats

15. Granting of Privileges–Guidelines for Hospital Trustees, Administration and Medical Staff: Part III, Appointment/Reappointment Process. BC Medical Association, College of Physicians and Surgeons of BC and BC Hospitals Association: joint publication; June 1991.

16. Cruess SR. Medical department heads. In: DD Gelman, editor. The CMA guide to medical administration in Canadian hospitals. Ottawa, 1996. E3.3-3.4.

Correspondence to: Dr. J.M. Etherington, Department of Emergency Medicine, St. Paul's Hospital, 1081 Burrard St., Vancouver BC V6Z 1Y6; 604 806-8480, fax 604 806-8488, jetherington@stpaulshosp.bc.ca