Teaching and evaluating first and second year medical students' practice of evidence-based medicine

Robert Holloway,[1,2] Kathryn Nesbit,[3] Donald Bordley[4] & Katia Noyes[2]

PURPOSE To implement an evidence-based medicine (EBM) curriculum for Year 1 and 2 medical students, and to develop a method to evaluate their practice of EBM in discrete and relevant worksteps.

METHODS For the 100 students entering Year 1 of their medical education in 2000, we implemented a curriculum with 25–30 student contact hours of EBM instruction which used a variety of teaching formats and spanned the first and second years of their training. We developed an evaluation module that assessed the following 5 steps in the practice of EBM: generating well built questions; searching for evidence; critical appraisal; applying the evidence, and self-evaluation. We tested 2 different versions of the test module 3-months apart with the same cohort of second year students, and correlated their scores on the second module with examination components of a comprehensive assessment. We obtained feedback from the students regarding the EBM curriculum and evaluation method.

RESULTS Each test module took 2–4 hours to complete and 5–8 minutes to grade. There was moderate test–retest reliability for the total test scores ($r = 0.35$, $P < 0.001$). Step 1 scores correlated with the mock board examination scores ($r = 0.23$, $P = 0.05$). Step 2 scores correlated with the peer assessment factor 'work habits' ($r = 0.24$, $P = 0.02$), and Step 3 scores correlated with clinical reasoning exercises ($r = 0.31$, $P = 0.002$). Step 4 scores lacked test–retest reliability and did not correlate with components of the comprehensive assessment. The majority of students felt there was too much focus on EBM during the first 2 years of the curriculum and they rated the EBM test module the lowest rated component of the comprehensive assessment.

CONCLUSIONS Although we have demonstrated preliminary reliability and validity of a new evaluation instrument that assess the domains of scientific knowledge, work habits and reasoning skills required in the practice of EBM, many of the correlations were weak, and we remain in the very early stages of determining if, when and how EBM instruction should occur in medical education.

KEYWORDS education, medical, continuing/*methods; evidence-based medicine/*education; curriculum; students, medical; attitude of health personnel.

*Medical Education 2004; **38**: 868–878*
doi:10.1046/j.1365-2929.2004.01817.x

INTRODUCTION

Evidence-based medicine (EBM) integrates individual clinical expertise with the best available evidence from systematic research.[1] It has been promoted as a means to reduce clinical practice variation and to improve the quality of patient care. The practice of EBM demands a set of skills to help clinicians retrieve, appraise and apply current best evidence. Despite the ample literature on the formats and approaches to teaching EBM, there is little evidence that teaching EBM leads to sustained changes in behaviour, let alone changes in quality of care or patient outcomes.

[1]Department of Neurology, University of Rochester, Rochester, New York, USA
[2]Department of Community and Preventive Medicine, University of Rochester, Rochester, New York, USA
[3]Department of Medical Informatics, University of Rochester, Rochester, New York, USA (deceased)
[4]Department of Medicine, University of Rochester, Rochester, New York, USA

Correspondence: Robert Holloway MD, MPH, Department of Neurology, University of Rochester, 1351 Mount Hope Avenue, Suite 220, Rochester, New York 14620, USA. Tel: 00 1 585 275 1018; Fax: 00 1 585 461 3554; E-mail: Robert.Holloway@ctcc.rochester.edu

Key learning points

We implemented an evidence-based medicine (EBM) curriculum for Year 1 and 2 medical students, and developed a method to evaluate their practice of EBM in discrete and relevant worksteps.

We used a modified script concordance methodology to assess critical appraisal skills. Preliminary test–retest reliability and validity have been demonstrated.

Many students felt there was too much emphasis on EBM instruction during the first 2 years in education and were distrustful of the methods used in the evaluation. Teaching and evaluation of EBM must be sensitive to learner needs and must evolve along with the experience of the student, resident or practising clinician.

We remain in the very early stages of determining if, when and how EBM instruction should occur in medical education.

It is not surprising, therefore, that there is little consensus as to the proper timing and content of EBM instruction across the spectrum of medical education.

There are many reasons why we lack sound evidence regarding effective methods of teaching EBM.[2] The complexities of the educational system, the difficulty in keeping an adequate study sample, the difficulty in controlling for co-interventions, and the difficulty in changing doctor behaviour are all contributing factors. An additional important factor is the lack of validated outcomes.[3,4] Frequently, reported outcomes refer to subjective variables such as satisfaction or self-reported changes in attitude or knowledge. No educational outcomes have yet been developed to capture the required skills that encompass the practice of EBM.

The focus of most EBM educational offerings has targeted practising doctors, graduate trainees, and Year 3 and 4 medical students. Recently, many medical schools have introduced EBM during the preclinical years, in an effort to promote the lifelong, self-directed behaviours required to practise EBM during the clinical years.[5] The feasibility of introducing an 8-hour, problem-based EBM course to Year 1 medical students has been recently described.[6]

Here we report our experience with a curriculum that introduces and reinforces EBM throughout the first 2 years of medical school. We also describe our method of evaluating the practice of EBM.

METHODS

Setting and sample

The study took place at the University of Rochester School of Medicine and Dentistry (URSMD) and involved the 100 students (mean age 24 years, 60% female) who entered Year 1 of a 4-year medical school education in August 2000. We followed this cohort through Years 1 and 2 of medical school until March 2002. During this time, the students were exposed to a structured EBM curriculum, and participated in the development and testing of an EBM evaluation instrument.

EBM curriculum

In 1999, URSMD implemented a new design for medical student education (the 'double-helix' curriculum) that integrates the clinical and basic science strands through all 4 years.[7] During the first 2 years, 30% of curricular time was devoted to clinically relevant experiences. An important component of clinical integration was an early introduction and reinforcement of EBM (Table 1). There were 25–30 student contact hours of EBM instruction, employing a variety of teaching formats including lectures, laboratory sessions, small group sessions, skill-based workshops and individualised instruction. A 43-page syllabus was created to support this instruction.

Students were introduced to EBM during the first 4 weeks of the curriculum. This introduction included 8–10 student contact hours during the course Mastering Medical Information. Topics included an EBM overview, medical informatics, EBM databases, framing a well built searchable question, and critical appraisal of the literature. There were a series of EBM reinforcements during the remainder of Year 1 as part of an ambulatory clerkship course. These 5 student contact hours included an advanced skills workshop, a patient-centred EBM exercise with individualised feedback, a peer comparison report, and an EBM review. Evidence-based medicine reinforcements continued in the second year with a series of small group EBM tutorials, led by local experts in EBM. These tutorials consisted of 7 monthly, 2-hour sessions to review clinical cases and practise building clinical questions, and searching and appraising the literature.[8]

Table 1 Components of the evidence-based medicine curriculum

	Year 01 2000–01	Year 02 2001–02
Introduction (8–10 hours)	**Reinforcement**	
Lectures Laboratories Small groups	Skills workshop (2 hours) Patient-centred exercises with individualised instruction (2 hours) EBM review (1 hour) Peer comparison report	EBM tutorials (12–15 hours)

EBM evaluation module development

As the conceptual framework for developing an evaluation instrument we used the 5 steps outlined in the practice of EBM:

1 converting knowledge gaps into focused answerable questions;
2 searching for the best external evidence;
3 critically appraising that evidence for its validity and importance;
4 applying the evidence in clinical practice, and
5 self-evaluation.[1]

The evaluation exercise centred on a clinical case, and provided individual scores for steps 1–4, as well as a total score.

We administered an EBM pilot module to the student study cohort after they completed their first class, Mastering Medical Information, in September 2000. We administered the same pilot EBM module to 46 Year 3 medical students during their medicine clerkship in the winter/spring of 2001. The 46 Year 3 students had not taken part in a structured EBM curriculum during their first 2 years of medical school. From this pilot, we learned that the clinical case had to be standardised, each student had to appraise the same article, and that multiple-choice questions were insufficient to evaluate critical appraisal skills.

Informed by the pilot experience, we developed 2 EBM test modules and administered these modules to the student study cohort during their second year of medical school. The modules were administered 3 months apart, in November 2001 and March 2002, under slightly different conditions. The March 2002 examination took place within the comprehensive assessment, a 2-week assessment system that assessed the habits of competence in those students participating in the double-helix curriculum.[9] The comprehensive assessment was designed to assess professional competencies by using a series of exercises that linked basic science knowledge with clinical and interpersonal skills.

Test modules

In both test EBM modules, the students were presented with a clinical case. In the module administered in November 2001, a written clinical vignette describing a patient with headaches was used. In the module administered in March 2002, a standardised patient (SP) with lower back pain was used. After obtaining the clinical information from the written clinical vignette or the SP, the student completed a 5-step EBM evaluation module (Table 2). Each test module was contained as a 15–20-page evaluation packet (available from the author on request) using multiple testing strategies described below.

Step 1

Based on the clinical case, the student had to write 3 searchable questions: 1 each pertaining to diagnosis, therapy and prognosis. The maximum total score for Step 1 was 11 points. Diagnosis and therapy questions were worth 4 points, with 1 point each being given for clear mention of the patient, intervention, comparison intervention and outcome. Prognosis questions were worth 3 points, with 1 point each being given for clear mention of the patient, prognostic risk factor or time, and outcome.

Step 2

After completing Step 1, students were asked to complete a Medline search. The Medline search

Table 2 Components of the evidence-based medicine modules

	EBM test module 1 November 2001	EBM test module 2 March 2002
Method to launch EBM module	Clinical case vignette	Standardised patient
Clinical content area	Headache	Lower back pain
Administration	3-day take home	In-class, open book
Student task	**Evaluation method**	
Step 1: Generating PICO question	Open-ended questions	
Step 2: Medline searching	Printed Medline search tactics	
Step 3: Critical appraisal	Modified script concordance questions Therapeutic terms exercise	
Step 4: Apply results of appraisal to patient	Modified script concordance questions Short answer questions	
Step 5: Self-assessment (not graded)	Multiple-choice questions	

PICO = patient, intervention, comparison, outcome.

was based on a standardised search question developed by the faculty, relevant to the clinical case and contained in the evaluation packet (e.g. 'How effective are anti-inflammatory drugs in treating lower back pain in adults over the age of 25?'). The search question was designed to evaluate the student's ability to complete search tactics taught in the EBM curriculum. The students had to search Medline through the University of Rochester's Ovid Web search service, and create a search strategy to answer the search question. The students submitted a printed copy of their search strategy for grading.

The students earned points based on their ability to complete search tactics (Table 3). Eleven tactics were defined, based on a search of the literature and a survey of librarians. Tactics were assigned point values with a total of 100 points for each search, and the more important tactics were assigned higher point values.[10] Five librarians graded 25 student searches. Nine of the 11 test items had weighted kappa scores between 0.73

Table 3 Criteria for grading Medline search tactics

	Medline search evaluation criteria	Points earned
1	Break down the question into components and search for each	15
2	Select appropriate MeSH terms	15
3	Select appropriate explodes for MeSH terms	15
4	Select appropriate MeSH subheading(s)	15
5	Use appropriate MeSH age terms	10
6	Try finding the evidence using EBM reviews, publication types or filters	10
7	Combine using Boolean operators	5
8	Combine all concepts in one search set	5
9	Modify strategy effectively by language, date, focus or text words	5
10	Find a sufficient number of citations for 'best set'	5
	Total points:	100
11	Try text words for missed MeSH terms in 2, 4, or 5 (if applicable)	(+ 10)

and 1.0. Item 8 had a weighted kappa of 0.61, and item 11 a weighted kappa of 0.56.

Step 3

Step 3 assessed student ability to critically appraise a research article for its validity and study results. The article to be appraised was provided to the students within the evaluation packet. We used a modified script concordance format to assess student ability to appraise a study's validity.[11,12] We created a series of 'what if' questions changing the study design or conduct to alter the study validity (Table 4A). Students had to determine the effect of the alteration on the study validity using a 5-point Likert scale. A final question had the students rate the overall validity of the study on a scale of 1–9, where 1 = definitely not valid and 9 = definitely valid.

To create a validity grading grid, 13 URSMD faculty experts in EBM read the research article and answered the same questions. For each question, the experts' answers were assigned a weight corresponding to the proportion of experts who selected it. Credits for each answer were transformed proportionally to obtain a maximum score of 1 point for modal experts' choice(s) on each item, other experts' choices receiving partial credit. Answers not chosen by the experts' received 0 points. Each 'expert' received $20 for completing the exercise.

To assess the students' understanding of the study results, we created a therapeutic terms exercise to evaluate student understanding of relative risk reduction, absolute risk reduction, numbers needed to treat, numbers needed to harm, and confidence intervals. Step 3 had a total of 20 points, 13 points for the script concordance questions and 7 points for the therapeutic terms exercise.

Step 4

Step 4 assessed the students' ability to apply the evidence from the article to the clinical case. We again used a modified script concordance format (Table 4B). We created a series of 'what if' scenarios that altered the disease characteristics, comorbidities, setting and preferences of the patient in the clinical case. Students determined the effect of the modifications on applying the evidence to the patient in the clinical case using a 5-point Likert scale. An applicability grading grid was developed in a similar manner as the validity grading grid (available from the author on request). Step 4 had a total of 7 points, 5 points

Table 4(A) Template used for 'critical appraisal' (Step 3) grading grid

If.....[alter characteristics of the study]	**...the validity of the study becomes**				
Randomisation	− 2	− 1	0	+ 1	+ 2
Allocation concealment	− 2	− 1	0	+ 1	+ 2
Sample size (random error)	− 2	− 1	0	+ 1	+ 2
Follow-up: length	− 2	− 1	0	+ 1	+ 2
Follow-up: loss to follow-up (count)	− 2	− 1	0	+ 1	+ 2
Follow-up: contamination	− 2	− 1	0	+ 1	+ 2
Follow-up: co-intervention	− 2	− 1	0	+ 1	+ 2
Follow-up: compliance	− 2	− 1	0	+ 1	+ 2
Follow-up: cross-over	− 2	− 1	0	+ 1	+ 2
Blinding: subjects	− 2	− 1	0	+ 1	+ 2
Blinding: investigators	− 2	− 1	0	+ 1	+ 2
Analysis: intention to treat	− 2	− 1	0	+ 1	+ 2
Analysis: balance of groups and adjustments if necessary	− 2	− 1	0	+ 1	+ 2
Analysis: subgroup analysis	− 2	− 1	0	+ 1	+ 2
Publication issues: funding source	− 2	− 1	0	+ 1	+ 2
Publication issues: conflict of interest	− 2	− 1	0	+ 1	+ 2

− 2 = much less valid; − 1 = less valid; 0 = neither more nor less valid; + 1 = more valid; + 2 = much more valid.

Table 4(B) Template used for 'applicability' (Step 4) grading grid

If.....[alter characteristic of the patient]		...the results of this study become				
Change:	characteristics of the patient pathophysiological differences in the illness differences in patient compliance comorbid conditions	− 2	− 1	0	+ 1	+ 2
Change:	setting feasibility issues provider compliance or skill	− 2	− 1	0	+ 1	+ 2
Change risk/benefit profile		− 2	− 1	0	+ 1	+ 2
Change values of the patient		− 2	− 1	0	+ 1	+ 2

− 2 = much less applicable to patient; − 1 = less applicable to patient; 0 = neither more or less applicable to patient; + 1 = more applicable to patient; + 2 = much more applicable to patient.

for the script concordance questions and 2 points for a short answer response giving an overall assessment of whether the trial results applied to the patient.

Step 5

Step 5 consisted of 17 self-assessment questions, which were modified from an EBM textbook.[8] These questions asked students to rate how often they engaged in each step-related activity, using a 5-point Likert scale. An example of a question was 'I am critically appraising external evidence', with the following standard response options: 0%, 25%, 50%, 75%, or 100% of the time. Of the 17 self-assessment questions, 7 questions addressed step 1, 4 questions addressed step 2, 3 questions addressed step 3, and 3 questions addressed step 4. Step 5 results were not graded but were used for peer comparison feedback and analysis purposes.

Student satisfaction

Following the comprehensive assessment in March 2002, we obtained student feedback on the EBM curriculum and the EBM test modules from a survey and written comments.

Analysis

All available data were used in all analyses and no imputation was carried out for missing data. All analyses were conducted using SAS Version 8 for Windows (SAS Institute Inc., Cary, North Carolina, USA) and STATA Release 6 (College Station, Texas, USA).

Total score and weighting of steps

Scores for each step were created by transforming raw scores to a 0–100 measurement scale by summing the items within each step, dividing by the maximum possible total score for that step, and multiplying by 100. A total score for the test modules was computed by taking a weighted average of the individual step scores. We calculated a total score based on the following weightings: Step 1 = 10%, Step 2 = 15%, Step 3 = 60%, and Step 4 = 15% (weightings were determined by EBM faculty). Distributions of scores were examined using mean scores, standard deviations and ranges.

Test–test reliability

We determined the correlation using Pearson correlation coefficients between each step and the total scores for the 2 EBM test modules administered 3 months apart, by individual student.

Validity

Firstly, we determined the correlation between each step within test module 1, as well as within test module 2. Secondly, we determined correlations with test module 2 scores and components of the comprehensive assessment. The following prespecified comprehensive assessment components were used:

1 mock written boards;
2 peer assessment evaluations, and
3 standardised patient post-encounter exercises.

The mock written boards assessed basic science knowledge, and only 73 of the 97 students completed them. The peer assessment exercise comprised a 12-item rating scale on professionalism, communication and teamwork. Each student was rated by 15 random peers. Factor analysis supported a 2-factor solution – 'interpersonal sensitivity' and 'work habits'.[13] After 8 SP exercises, the students completed 2 exercises, a 'post-encounter probe' and a 'long exercise', that assessed clinical reasoning skills. The first exercise component, the post-encounter probe, consisted of an intentional mix of short answer questions that tapped into multiple domains of knowledge application including basic, clinical and the social sciences. The second component, the long exercise, included short answers and essays that again assessed clinical integration of knowledge across the disciplines of medical sciences.

Correlation between self-assessment and performance

For each student, we averaged the self-assessment responses (i.e. 0%, 25%, 50%, 75% and 100%) within each step: 7 questions for Step 1, 4 questions for Step 2, 3 questions for Step 3, and 3 questions for Step 4. We then correlated student performance on test EBM module 2 step scores with their self-assessment step scores (e.g. Step 1 performance on EBM module 2 compared with mean response to the 7 Step 1 self-assessment questions).

RESULTS

EBM pilot module: comparison of students on 'new' versus 'old' curricula

Steps 1 and 2 of the pilot module were similar to Steps 1 and 2 of the test modules. The 2 student groups scored similarly on Step 1 of the pilot module, with mean scores of 93.6 (SD 6.7) and 91.2 (SD 8.5) for Year 1 and 3 students, respectively. In contrast, Year 1 students performed better than Year 3 students on Step 2 of the developmental module, with mean scores of 86.2 (SD 14.0) versus 62.3 (SD 13.2) for Year 1 and 3 students, respectively ($P < 0.01$). The Step 3 and 4 components of the pilot module were based on 20 multiple-choice questions. These scores are not reported due to the differences in testing and grading methodologies compared to the test modules.

EBM test modules: examination procedures and scores

The test modules took 2–4 hours to complete and the students handed in the EBM evaluation packet and a printed copy of their Medline searches for grading purposes. The medical librarian graded the searches; each search took approximately 2–3 minutes to grade. Faculty graded Step 1 and the short answer components of Steps 3 and 4, which took approximately 3–5 minutes per examination. After grading was complete, the students were given their total scores and their scores for each step. They were also provided with the distribution of the expert responses for the script concordance questions and a written explanation of the answers.

The total EBM test scores showed similar means and standard deviations (Table 5). Approximately 50% of the students scored full credit (100%) for the Step 1 exercise on both test modules, whereas only 10–15% of students received full credit (100%) for the Step 2 exercise.

Test–retest reliability

When comparing EBM test modules 1 and 2, there were weak to moderate correlations between Step 1 scores ($r = 0.28$, $P = 0.006$), Step 3 scores ($r = 0.26$, $P = 0.009$), and the total scores ($r = 0.35$, $P < 0.001$). There were no correlations between Step 2 scores ($r = 0.12$, $P = 0.26$) and Step 4 scores ($r = 0.08$, $P = 0.43$). When only the script concordance questions were analysed, the following correlations were obtained: Step 3: r = 0.32, $P = 0.001$; Step 4: r = 0.02, $P = 0.83$.

Validity

There were no statistically significant correlations between the steps within either test module. However, there were weak to moderate correlations with the step and total scores of EBM test module 2 and components of the comprehensive assessment (Table 6). For example, the Step 1 scores showed a weak correlation with the mock board examination scores ($r = 0.23$, $P = 0.05$), the Step 2 scores showed a weak correlation with the peer assessment factor 'work habits' ($r = 0.24$, $P = 0.02$), and the Step 3 scores showed a moderate correlation with the SP post-encounter probe questions ($r = 0.31$, $P = 0.002$). Step 4 scores did not correlate with any component of the comprehensive assessment. Finally, the total EBM test scores correlated with the SP exercises ($r = 0.30$, $P = 0.002$), peer assessment 'work habits' ($r = 0.22$, $P = 0.03$), and the mock board examination scores ($r = 0.23$, $P = 0.05$).

Table 5 Component and total scores of EBM test modules 1 and 2

	EBM test module 1				EBM test module 2			
	Mean	**SD**	**Range**	**% receiving maximum (100%)**	**Mean**	**SD**	**Range**	**% receiving maximum (100%)**
Step 1: Generating PICO question	90.1	13.8	27.3–100	47.2%	91.7	10.2	63.6–100	51.2%
Step 2: Medline searching	81.0	12.8	45–100	14.4%	83.5	11.7	50–100	11.5%
Step 3: Critical appraisal	72.8	10.9	42.2–93.2	0	69.5	12.4	31.8–93.0	0
Step 4: Apply results of appraisal to patient	78.2	12.6	43.3–100	1.0%	76.6	14.2	38.6–100	2.1%
Total score*	76.6	7.6	54.6–91.4	0	74.8	8.3	46.9–90.7	0

* Total score = Step 1 (10%) + Step 2 (15%) + Step 3 (60%) + Step 4 (15%).
PICO = patient, intervention, comparison, outcome.

Table 6 Correlation of EBM module step and total scores with elements of the comprehensive assessment

Component of the comprehensive assessment	**Students *n***	**Step 1**	**Step 2**	**Step 3**	**Step 4**	**Total**
Basic science examination – mock boards	72	**0.23***	0.02	0.17	0.14	**0.23**
		0.05	0.89	0.14	0.24	**0.05**
Peer assessment	97	**0.21**	0.17	0.09	− 0.06	0.12
		0.04	0.10	0.39	0.55	0.22
Interpersonal sensitivity		0.18	0.048	< − 0.01	− 0.12	< − 0.01
		0.07	0.64	0.99	0.23	0.99
Work habits		0.18	**0.24**	0.16	0.03	**0.22**
		0.08	**0.02**	0.12	0.79	**0.03**
SPs – Post-encounter exercises	97	0.14	0.13	**0.29**	0.02	**0.30**
		0.16	0.21	**0.004**	0.83	**0.002**
Post-encounter probe		0.12	0.11	**0.31**	0.02	**0.32**
		0.25	0.26	**0.002**	0.85	**0.002**
Long exercise		0.14	0.11	0.15	0.019	0.18
		0.18	0.30	0.14	0.86	0.08

* Correlation coefficients and associated *P*-values. Item pairs are shown in bold when the *P*-value is ≤ 0.05.
SP = standardised patient.

Correlation between self-assessment and performance

There were no statistically significant correlations between the mean self-assessment scores and student performance on any of the 4 steps.

Student satisfaction

Student satisfaction for EBM test module 2 is shown in Table 7. The EBM take-home examination was the lowest rated component of the comprehensive assessment. Fifty of the 95 student

Table 7 Student ratings of EBM test module 2

	Strongly disagree or disagree (1 or 2)	Mixed feelings (3)	Agree or strongly agree (4 or 5)	Mean	(SD)
Exercise directly related to SP case	11*	31	54	3.63	(0.94)
Addressed important issues	16	36	44	3.27	(1.00)
Good test of Medline searching	17	22	57	3.47	(0.98)
Good test of critical appraisal skills	20	29	47	3.23	(0.97)
Reinforced value of good EBM skills	24	26	45	3.17	(1.12)

* Counts (not percentages); SP = standardised patient.

respondents had mixed feelings or disagreed that the examination reinforced the value of good EBM skills.

A total of 67 students provided written feedback on their experience with the EBM teaching and evaluation over the prior 19 months. Nine of the 67 comments were positive statements about EBM teaching and the evaluation modules. The remainder of the comments, however, were negative. The following themes emerged:

1 there was too much curricular emphasis on EBM within the first 2 years;
2 the EBM test modules had little perceived value;
3 distrust of the script concordance questions;
4 frustration over the method of Medline grading, and
5 the EBM test took too long to complete.

DISCUSSION

We developed a test to evaluate the knowledge base, work habits and reasoning skills required in the practice of EBM, and have demonstrated that the assessment can be administered using written clinical vignettes or standardised patients. The correlation of the total and step scores to different areas of the comprehensive assessment (basic science knowledge, work habits and reasoning skills), and the lack of correlation among steps within a test module suggests that we may be evaluating different domains of competence. The test–retest reliability of the evaluation method is supported by the moderate correlation between test modules 1 and 2 for the total scores, as well as the Step 1 and Step 3 component scores. These correlations persisted despite different methods of administration, different clinical cases, and different articles to appraise suggesting short-term stability in the domains being assessed.

The correlation of the Step 1 component to the mock board scores for the 72 students who completed the mock boards suggests that a student's ability to generate a searchable question is partly dependent on their scientific knowledge base. The Step 1 ceiling effect also suggests that it may be a better measure of obtaining a minimum standard of performance, rather than distinguishing between degrees of excellence.

The correlation of the Step 2 scores with the 'work habits' factor within the peer assessment, although weak, is not surprising, given the set of routine tasks students were asked to perform as part of creating and printing their Medline search strategy. However, the lack of test–retest reliability suggests that it was difficult to distinguish between student skill levels. This lack of reliability may be due to student factors, including inconsistent diligence and effort in performing searches, or technical factors in that different search questions led to unstable or unpredictable search outputs that were not amenable to structured assessments. It is less likely to be due to grader factors, given the high levels of agreement demonstrated. More research is needed to broaden the scope of search assessment to include other search engines and evidence sources, to reward search 'outcomes' – the articles – rather than the process, and to reward effective shortcuts taken by advanced searchers of evidence.[14]

The Step 3 scores showed moderate test–retest reliability and moderate correlation with clinical reasoning skills as assessed by the post-encounter exercises that followed SP cases. We adapted the methods in developing script concordance questions for assessing clinical reasoning skills, and applied them to critical appraisal skills, which is also characterised by organisation of knowledge into networks or 'scripts' that allow for efficient critical appraisal tasks.[11,12] Script concordance questions have several advantages, including straightforward scoring, use when there is no consensus among experts, and re-usability in test–retest situations. In fact, one can imagine the creation 'off-the-shelf' grading grids for a set of articles that can be used across time and settings.[15] A disadvantage of the script concordance methodology is that creating the grading grids is fairly resource-intensive.

More research is needed to improve the methods of assessing the application of evidence to enhance patient care, as the Step 4 component lacked test–retest reliability and did not correlate with other components of the comprehensive assessment. In addition, the lack of correlation between self-reported insight of knowledge and skill, and its objective assessment, reinforces the findings that self-assessment techniques to evaluate EBM educational offerings may not be valid.[16]

One of the most important findings of our study, however, was the feedback we obtained from students. Students felt that there was too much emphasis on EBM during the first 2 years of the curriculum. The introduction of EBM during the first 2 years may be possible and even desirable,[6] but continued reinforcement of EBM principles without the clinical experience to support its application may not, and more research is needed to establish the timing and amount of EBM instruction within the medical school curriculum that will meet the needs of the learner.

Students also expressed significant dissatisfaction with the evaluation methods. In particular, they voiced frustration over the grading of the Medline search tactics, and, as 1 student stated, 'At this point, we feel the end result is more important than the process.' In addition, students expressed considerable dissatisfaction and mistrust with the use of the 'expert panel' to develop the grading grid. Our attempt to apply objective assessment methods to subjective appraisal questions appeared to be unsettling to many students. Our study and the reaction of the students emphasise that more attention and research is needed to develop distinct evaluation methods along the continuum of EBM competence, with reference to novice, advanced beginner, competent, proficient and expert practitioners.[17]

We have approached the development of an EBM assessment instrument as a staged validation process. The first stage focused on establishing that competencies embodied in the practice of EBM can be reliably measured. The next stage will focus on establishing whether the reliably measured competencies correlate with or predict behaviours that ultimately lead to higher quality care. Focusing too much on patient outcomes may make this task too difficult and distracting, and a more proximal measure of clinical reasoning skills, patient-centredness, and professionalism may need to suffice. The final stage of instrument development will be to establish whether these measured competencies respond to different educational interventions.

Our research indicates that various EBM competencies can be reliably measured and do correlate with other competencies deemed necessary for the practice of contemporary medicine.[18] However, many of these correlations were weak and more research is needed to determine whether these EBM competencies anchor with other domains of clinical practice and whether these associations persist, weaken or strengthen over time. We plan on following the study cohort to see if their EBM behaviours as measured during Years 1 and 2 predict their performance in Years 3 and 4.

In conclusion, our experience has been revealing. Firstly, serious consideration needs to be given to the amount of EBM instruction provided during the first 2 years of medical school: an introduction may be valuable and well received; continual reinforcement may not. Secondly, EBM teaching and evaluation must be sensitive to learner needs and must evolve along with the experience of the student, resident and practising clinician. Finally, although we have demonstrated preliminary reliability and validity of a new evaluation instrument that assesses the domains of scientific knowledge, work habits and reasoning skills required in the practice of EBM, we remain in the very early stages of determining if, when and how EBM instruction should occur in medical education.

CONTRIBUTORS

RH contributed to study design, study implementation, data analysis and the first draft of the manu-

script. KNe contributed to study design, study implementation, data analysis and critical revision of the manuscript. DB contributed to study design, data analysis and critical revision of the manuscript. KNo contributed to data analysis and critical revision of the manuscript.

ACKNOWLEDGEMENTS

We would like to thank the following individuals for curricular support: Edward Hundert, Elaine Dannefer, Ron Epstein, Anne Nofziger, Julia Sollenberger, Kathryn Markakis and Scott Trippler. We also thank Elaine Dannefer for her careful review of the manuscript, Keith Skelton and Sean Meldrum for SAS support, and Carolynn O'Connell for project management and for preparation of the manuscript.

FUNDING

This project was funded (in part) by a National Board of Medical Examiners' (NBME Stemmler Medical Education Research Fund grant. The project does not necessarily reflect NBME policy, and NBME support provides no official endorsement.

RH was supported in part by a K24 NS42098–01 from the National Institute of Neurological Disorders and Stroke. KNo was supported in part by a K01 AG 20980–01 from the National Institute of Ageing.

ETHICAL APPROVAL

We obtained an exemption from the University of Rochester's Research Subject's Review Board.

REFERENCES

1 Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based Medicine. How to Practise and Teach EBM*. 2nd edn. Edinburgh: Churchill Livingstone 2000.
2 Hatala R, Guyatt G. Evaluating the teaching of evidence-based medicine. *JAMA* 2002;**288**:1110–2.
3 Taylor R, Reeves B, Mears R *et al.* Development and validation of a questionnaire to evaluate the effectiveness of evidence-based practice teaching. *Med Educ* 2001;**35**:544–7.
4 Fliegel JE, Frohna JG, Mangrulkar RS. A computer-based OSCE station to measure competence in evidence-based medicine skills in medical students. *Acad Med* 2002;**77**:1157–8.
5 Barnett SH, Kaiser S, Kasner-Morgan L *et al.* An integrated programme for evidence-based medicine in medical school. *Mt Sinai J Med* 2000;**67**:163–8.
6 Srinivasan M, Weiner M, Breitfeld PP, Brahmi F, Dickerson KL, Weiner G. Early introduction of an evidence-based medicine course to preclinical medical students. *J Gen Intern Med* 2002;**17**:58–65.
7 Hundert EM, Dannefer EF. University of Rochester School of Medicine and Dentistry. *Acad Med* 2000;**75** (9):252–255.
8 Cox J. Evidence-based Medicine Tutorials for Second Year Medical Students. *Early Observations of a New Curriculum Component*. Paper presented at the Undergraduate Medical Education for the 21st Century National Symposium. Sponsored by the Health Resources and Services Administration. Baltimore, Maryland 2002.
9 Epstein RM, Dannefer E, Nofziger AC *et al.* Comprehensive assessment of professional competence: the Rochester experiment. *Teach Learn Med*. in press.
10 Holloway R, Nesbit K, Bordley D. *Evaluating Medical Students' Ability to Practise Evidence-based Medicine*. Paper presented at the 40th Annual Conference on Research in Medical Education (RIME)/Annual Meeting of the Association of American Medical Colleges. Washington DC 2001.
11 Charlin B, Roy L, Brailovsky C, Goulet F, Van der Vleuten C. The script concordance test: a tool to assess the reflective clinician. *Teach Learn Med* 2000;**12**:189–95.
12 Brailovsky C, Charlin B, Beausoleil S, Cote S, Van der Vleuten C. Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: an experimental study on the script concordance test. *Med Educ* 2001;**35**:430–6.
13 Dannefer EF. Peer and Self-Assessment of Professionalism for Medical Students. Presented at Ottawa Conference on Medical Education and Assessment. Ottawa 2002.
14 Haynes RB. Of studies, syntheses, synopses and systems: the '4S' evolution of services for finding current best evidence. *ACP J Club* 2001;**134**:11–13.
15 Sibert L, Charlin B, Corcos J, Gagnon R, Grise P, Van der Vleuten C. Stability of clinical reasoning assessment results with the script concordance test across two different linguistic, cultural and learning environments. *Med Teach* 2002;**24**:522–7.
16 Khan KS, Awonuga AO, Dwarakanath LS, Taylor R. Assessments in evidence-based medicine workshops: loose connection between perception of knowledge and its objective assessment. *Med Teach* 2001;**23**:92–4.
17 Dryefus HL. *On the Internet (Thinking in Action)*. New York: Routledge 2001.
18 Medical School Objectives Project Writing Group. Learning objectives for medical student education – guidelines for medical schools. Report 1 of the Medical School Objectives Project. *Acad Med* 1999;**74**:13–8.

Received 7 May 2003; editorial comments to authors 20 June 2003; accepted for publication 1 July 2003