Validation of the Colorado Psychiatry Evidence-Based Medicine Test

Brian Rothberg, MD
Robert E. Feinstein, MD
Gretchen Guiton, PhD

Abstract

Background Evidence-based medicine (EBM) has become an important part of residency education, yet many EBM curricula lack a valid and standardized tool to identify learners' prior knowledge and assess progress.

Objective We developed an EBM examination in psychiatry to measure our effectiveness in teaching comprehensive EBM to residents.

Methods We developed a psychiatry EBM test using the validated EBM Fresno Test of Competence for family medicine. The test consists of case scenarios with open-ended questions. We also developed a scoring rubric and obtained reliability with multiple raters. Fifty-seven residents provided test data after completing 3, 6, 25, or 31 EBM sessions. The number of sessions for each resident was based on their length of training in our program.

Results The examination had strong interrater reliability, internal reliability, and item discrimination. Many residents showed significant improvement on their examination scores when data were compared from tests taken before and after a sequence of teaching sessions. Also, a threshold for the level of expert on the examination was established using test data from 5 EBM teacher-experts.

Conclusions We successfully developed a valid and reliable EBM examination for use with psychiatry residents to measure essential EBM skills as part of a larger project to encourage EBM practice for residents in routine patient care. The test provides information on residents' knowledge in EBM from entry level concepts through expert performance. It can be used to place incoming residents in appropriate levels of an EBM curriculum and to monitor the effectiveness of EBM instruction.

Introduction

Teaching evidence-based medicine (EBM) skills to physicians in training has garnered support from accreditation groups and is widely incorporated into medical education.[1–3] The content of EBM curriculum and the standard for successful application of EBM are remaining concerns. For instance, Rao and Kanter,[4] summarizing EBM curricula topics for medical students from reports and articles published between 1999 and 2009, found that less than one-third addressed the statistical aspects of research design, analysis, and results. A review of instruments used to evaluate the effectiveness of EBM educational efforts for students, postgraduate trainees, physicians, and nonphysicians concluded that validity instruments were available for evaluating some domains but that instruments were needed to evaluate more recently emphasized skills.[5] Of the 16 instruments meeting the 2 highest levels of validity data in this article, only 2 address the design and statistical skills called for by Rao and Kantor.[4] Moreover, these instruments target different levels of training and specialty groups.

We developed a model of teaching EBM to psychiatric residents that addresses the essential skills identified in the literature.[6–8] Typically, described as the 6 As, these skills include the ability to *assess* a patient, *ask* a clinical question, *acquire* the information, *appraise* the information, *apply* the information to the patient, and *assess* the outcome with a patient. For a copy of the Colorado Psychiatry evidence-based test or scoring rubric, please e-mail the principal investigator of the study (B.R.).

Methods

Our goal was to develop an authentic test of full EBM practice taught in training programs such as the model EBM curriculum adopted by the American Association for Directors of Psychiatric Residency Training (AADPRT).[9,10] Consequently, we chose an open-ended format that begins with clinical conditions and the uncertainty of medical practice. Examination questions require a range of skills: articulating a question, developing a search strategy,

All authors are at the University of Colorado Denver School of Medicine. **Brian Rothberg, MD,** is Assistant Professor in the Department of Psychiatry; **Robert E. Feinstein, MD,** is Professor in the Department of Psychiatry; and **Gretchen Guiton, PhD,** is Associate Professor of Educational Development and Research.

Funding: The authors report no external funding source for this study.

Corresponding author: Brian Rothberg, MD, Department of Psychiatry, University of Colorado Denver School of Medicine, 13001 E. 17th Place, Mail Box F546, Aurora, CO 80045, brian.rothberg@ucdenver.edu

Received July 11, 2012; revisions received October 23, 2012, and December 14, 2012; accepted December 17, 2012.

DOI: http://dx.doi.org/10.4300/JGME-D-12-00193.1

determining criteria for critically appraising a research article, calculating and interpreting relevant quantitative measures, and interpreting findings in relation to a patient. We developed a scoring rubric to aid determination of whether performance criteria were met and to provide feedback to examinees. Interrater reliability was estimated to ensure that the rubric could be applied consistently. Finally, we presented evidence for the validity of the examination, including instrument development to ensure content validity, its ability to discriminate experts from novices, and its ability to differentiate skills at different levels of training.

Setting and Participants

The instructional series involves sets of 3, 6, and 25 sessions that progress over the 3-year residency. First-year residents receive EBM instruction in 3 2-hour sessions designed to create interest and introduce them to asking a question related to their patient and acquiring information. Second-year residents participate in 6 1-hour sessions which focus on the basics of critical appraisal, EBM math, patient preferences and values, and outcomes. Third-year residents attend twenty-five 75-minute sessions that require them to apply the 6As, use EBM math, and appraise different types of research designs as they present their own cases.

Data come from all psychiatric residents participating in EBM sessions over 3 academic years (2005–2008) in 1 residency program. Data come from 56 residents, 50 of whom completed the examination after completing an instructional series. At entry to EBM, course residents included 17 postgraduate year (PGY)-1, 18 PGY-2, 13 PGY-3, and 8 PGY-4 residents. Some residents (n = 33) completed multiple instructional series so that examination performance after 1 series functions as a pretest to examination performance following the next instructional series.

Test Development, Test Outcomes, and Analysis of Outcomes

The Colorado Psychiatry EBM (CP-EBM) examination tests knowledge of the 6 As of EBM through 14 open-ended questions. In question 1, residents must first *assess* and *ask* a focused clinical treatment question following the patient population or problem, intervention, comparison group, and outcome (PICO) format for each of 2 clinical cases involving depression and schizophrenia (BOX). In questions 2 to 4, residents are to name resources and identify a search strategy to *acquire* the information needed to answer the clinical questions. In questions 5 to 9, the resident's *appraisal* skills are evaluated by testing the concepts of validity, relevance, magnitude, and significance of research findings. Also, the residents are asked to calculate relative risk, relative risk reductions, attributable risk, and number needed to treat to explain the meaning of each value calculated in relation to a clinical scenario. Questions 10 and 11 test a resident's basic knowledge of diagnostic and prognostic research designs. Question 12 tests the resident's ability to *apply* outcome measures in clinical practice. Finally, questions 13 and 14 ask about *assessing* a patient with an outcome measure over time. The examination is administered with an open time frame, generally requiring 1 to 2 hours to complete, and is scored using a rubric with a maximum total of 226 points.

What was known

Evidence-based medicine (EBM) is an important component of medical knowledge and practice-based learning and improvement, yet there is a dearth of valid tools to assess EBM skills acquisition in learners.

What is new

A psychiatry EBM assessment adapted from another validated tool had strong interrater reliability, internal reliability, and item discrimination.

Limitations

Single-site study, small sample with reduced statistical power to effects, and a focus on testing knowledge, not its application in clinical practice.

Bottom line

The psychiatry EBM tool facilitated an assessment of residents' developing EBM knowledge after a series of teaching sessions.

Case scenarios represent relatively common conditions seen in psychiatry. The depression and schizophrenia conditions offer a rich context for question formulation and investigation as they relate to a large and evolving literature. Content specifications were based on the instructional model and specific question types modeled after the Fresno Test of Competence in EBM, a validated instrument.[11] Short answers and calculations congruent with EBM in practice are elicited by the questions.

We developed a standard scoring rubric with explicit grading criteria for each question.[12] For example, item 1 asks the candidate to formulate an EBM-focused question for the 2 clinical scenarios. To score answers, each element

BOX **Test Question Example**

Case 1 Depression: You have just seen David who recently became sad, suicidal, and depressed. He plans to seek treatment with both medication and cognitive behavioral therapy. To begin treatment you generally prefer to recommend the combination of both a selective serotonin reuptake inhibitor (SSRI) and crisis intervention first, followed by an SSRI and psychotherapy, if needed. You have been told that it may not be necessary to recommend both medications and psychotherapy simultaneously.

1) Ask and then construct a focused treatment question for the clinical situations. Ask a question that will help you organize an evidence-based medicine search of the world literature's database and help you to acquire an article for an answer.

TABLE 1 **Paired *t* Test Results for Residents Completing 6 or 31 Sessions**[a]

Number of Sessions	n	Mean Pretest	Pretest SD	Mean Posttest	Posttest SD	*t*	*P*	Cohen *d*
6	11	116.27	44.63	141.09	51.04	2.222	.05	0.52
31	8	91.00	34.94	161.50	24.14	7.654	.001	2.33

[a] Only 2 of the residents had pre-post data for 3 sessions, and only 2 of the residents with 25 sessions had pre-post data. Due to the small n for these groups, no analyses were conducted.

in the PICO format receives points according to the following criteria: not evident, 0 points; minimal, 1 point; strong, 2 points; and excellent, 3 points. The sum of points for all elements is the total score for the item. Items are scored separately for each case, resulting in a maximum of 24 points per case.

Reliable application of the scoring rubric was developed using data from a pilot sample of 8 fourth-year psychiatric residents who had received EBM instruction from the senior investigator. Four examinations were selected at random for scoring, and 4 EBM experts in our program independently scored 1 examination and compared results on all questions. The raters reviewed and discussed all items when scores differed until all agreed on the scoring of an examination. The process was repeated for each of the remaining 3 tests until all 4 raters agreed on the scores. The rubric was clarified as needed during this process.

To estimate scoring reliability, these same 4 raters, blinded to resident identity and year, independently scored 4 additional examinations from the larger study sample. Interrater reliability is reported for these data. Internal consistency of the test is estimated using a sample of all residents who completed 6 instructional sessions. All examinations were scored by 1 of the 4 raters who participated in establishing the reliability of the scoring system.

Initially, content validity was emphasized in the development process by including questions that address all components of the EBM guidelines. Next, we examined the criterion validity by comparing the scores of EBM experts and novices. We examined the number of sessions needed for a resident to reach the level of an expert using the lowest expert score as the criterion. Finally, we examine whether the pattern of scores relates to the amount of instruction. We compared posttest mean scores after 3, 6, 25, and 31 sessions. Data analyses were conducted using SPSS version 18 software (IBM Corp, Armonk, NY).

The Colorado Multiple Institutional Review Board granted approval for the study.

Results

Reliability

Mean ratings across the 4 raters for the 4 examinations ranged from 144 to 205, representing the variety of responses expected in practice. The average correlation across all 4 raters, estimated by Cronbach's alpha (r = 0.987), indicates a high level of consistency among raters (eg, raters order examinees similarly). Because instructional use of the examination was a primary goal, we also estimated the intraclass correlation coefficient (ICC) by using the formula for absolute agreement (r = 0.93, 95% confidence interval [CI] 0.724–0.995). This ICC is the equivalent of the dependability index (ie, Phi for an absolute decision) in a single-facet, fully crossed G-study. Finally, we conducted a 2-factor, random effects analysis of variance and estimated the generalizability coefficient for a single rater as 0.88.[13]

Expert Performance

After we standardized the examination scoring, psychiatry experts were identified using the AADPRT list-serve. Ten psychiatric residency faculty members from 10 different US programs identified themselves as expert EBM teachers, and 5 members completed and returned the examination via the honor system following the same test conditions as those offered to residents. Two blinded investigators scored the faculty examinations. Expert faculty members scored in the range of 178 to 219 (mean = 206.4, SD = 16.35). Experts performed significantly higher ($t = 8.5$; $P < .001$; 95% CI 77.27–131.28) than 8 novices who completed the examination prior to instruction (range = 70–136; mean = 102.13; SD = 27.87).

Discrimination

Using the lowest score of the 5 experts in the initial validity study, a threshold of 178 was set for determining expertise. No residents achieved this level of proficiency after 3 sessions. Overall, 18% of those participating in the program reached this level of proficiency (none after 3 sessions, 14% after 6 sessions, 24% after 25 sessions, and 25% after 31 sessions).

TABLE 2 **Summary of Validity Evidence for Colorado Psychiatric EBM Examination**

Validity Evidence	Measure Used	Acceptable Results	Test Performance
Content			
The test covers the main aspects of EBM called for in the literature and requires open-ended responses	Expert opinion development process	Match to literature	Revisions based on expert suggestions; cases limited to psychiatric context
Internal structure			
Interrater reliability (degree to which 2 scorers rate a single performance consistently and absolutely and estimated reliability for single rater)	Interrater correlation (multirater study)	> 0.80	Cronbach's alpha = 0.95
	Intraclass correlation (ICC) Estimating G-study (Phi) G coefficient for single rater (D)	> 0.80	ICC = 0.93 D = 0.88
Internal reliability (degree to which all test questions measure a single construct)	Cronbach's alpha (postexamination after 6 sessions, n = 32)	> 0.8 (for use with individuals)	Cronbach's alpha = 0.84
Item difficulty (relative difficulty of each item)	% of candidates answering question correctly	Variability allows assessment of expert and novice groups	Ranged from 0.09 to 0.79; mean of 0.52
Item discrimination	Item discrimination index ranges from −1.0 to 1.0	> 0.30	Ranged from 0.24 to 0.78; mean of 0.52
Relationship to other variables			
Discrimination	Mean scores of experts and novices compared by *t* test	Significant difference with experts outperforming novice groups	On a 226-point examination, expert mean was 206.5 and novice mean was 106.22, a difference that is significant ($P = .001$; Cohen's $d = 3.109$)
Gain in knowledge and skills	Mean scores for residents at different levels of training	Increasing performance as instruction increases	Comparison of pre-post scores prior to instruction, after 6 sessions, and after 31 sessions indicate significant gains after 31 sessions ($P = .001$) and gains nearing significance after 6 sessions ($P = .05$). Gain scores after 6 sessions (mean = 24.82, SD = 37.04) significantly less than gain scores after 25 sessions (mean = 70.5; SD =34.52; $P = .01$).

Abbreviation: EBM, evidence-based medicine.

Impact of Instruction

Resident performance earns progress from 124.33 after 3 sessions to 136.81 (6 sessions) to 158.67 (25 sessions) to 162.25 (31 sessions), although differences do not reach statistical significance. To further examine the impact of instruction, we compared pre-post gains in performance after 6 and 31 sessions. As shown in TABLE 1, scores increased for both groups.

Summary of Validity Evidence

In TABLE 2, we summarize the test properties of the CP-EBM examination in relation to content, internal structure of the examination, and relationship of examination performance to other variables.

Discussion

The CP-EBM examination provides a reliable and valid assessment of a trainee's ability to identify the information needed and to formulate a clinical question, conduct a search for information, understand and critically appraise the evidence including design and statistics, and integrate the evidence into clinical decisions for a given patient. Numerous standards and authors call for a range of validity evidence to support interpretation of test scores.[14] The development process supports the content of the examination and its authenticity for use with psychiatric residents. A high level of interrater reliability was achieved with brief training and guided by the scoring rubric. Analyses indicate that the examination can be reliably scored by a single rater and that results can be used to distinguish individual performance. Evidence of validity is provided by the test's ability to measure change in knowledge after progressive instruction in EBM and its ability to discriminate experts and novices. No floor or ceiling effect is evident. Although the EBM skills tested apply across disciplines, the use of psychiatric cases provides authenticity for residents in this

specialty. As new residents enter with various facility in EBM, the instrument can be used to identify skill levels. It could be used to set a standard for basic competency in psychiatric EBM practice. As EBM practice in both medical schools and psychiatry residency programs becomes widespread, this basic standard may need to be adjusted.

A number of limitations should be considered when interpreting our results. The study was conducted in a single institution with an EBM program designed to enhance the specific skills assessed. The limited sample size reduced the statistical power to identify all program effects. As efforts to increase students' knowledge of biostatistics on entry to medical school and EBM are increasingly included in the medical school curriculum, the sensitivity of the instrument to knowledge change may be affected. Furthermore, although this examination is a knowledge and skills examination, it does not address whether classroom EBM learning ultimately leads to improved resident attitudes toward the use of EBM. This examination also does not address whether classroom-acquired EBM knowledge and skills will lead to use in a clinical setting or whether classroom EBM practice improves patient outcomes. As noted in Coomarasamy,[15] standalone teaching and integrated teaching are both effective in improving the knowledge base, but it is clinically integrated teaching with live patients that is more likely to bring about changes in clinician behaviors with real patients.

Conclusion

The CP-EBM test is part of a larger project described in the University of Colorado Department of Psychiatry EBM project,[9] which includes curricular content, resident attitudes, and clinically integrated teaching activities. The validation of our test is an important step to examine how best to teach and measure EBM knowledge and skills, apply these skills to clinical practice, and improve patient outcomes.

References

1 Guyatt G, Rennie D, eds. *Users' Guides to the Medical Literature: A Manual for Evidence-Based Medical Practice*. Chicago, IL: AMA Press; 2002.
2 Lurie SJ, Mooney CJ, Lyness JM. Measurement of the general competencies of Accreditation Council for Graduate Medical Education: a systematic review. *Acad Med*. 2009;84(3):301–309.
3 Barzansky B, Etzel SI. Educational programs in US medical schools, 2002–2003. *JAMA*. 2003;290(9):1190–1196.
4 Rao G, Kanter SL. Physician numeracy as the basis for an evidence-based medicine curriculum. *Acad Med*. 2010;85(11):1794–1799.
5 Shaneyfelt T, Baum KD, Bell D, Feldstein D, Houston TK, Kaatz S, et al. Instruments for evaluating education in evidence-based practice: a systematic review. *JAMA*. 2006;296(9):1116–1127.
6 Sackett DL. *Evidence-Based Medicine: How to Practice and Teach EBM*. 2nd ed. Edinburgh, Scotland: Churchill Livingstone; 2000.
7 Feinstein RE. Evidence-based medicine. In: Blumenfield M, Strain JJ, eds. *Psychosomatic Medicine*. Philadelphia, PA: Lippincott Williams & Wilkins; 2006:881–897.
8 Sackett DL, Rosenberg WMC. The need for evidence-based medicine. *J R Soc Med*. 1995;88(11):620–624.
9 Feinstein RE, Rothberg B, Weiner N, Savin DM. University of Colorado department of psychiatry evidence-based medicine educational project. *Acad Psychiatry*. 2008;32(6):525–530.
10 Feinstein, R. (2011) Evidence-based Medicine Educational Project Model Curriculum. American Association of Directors of Psychiatric Training Website. www.aadprt.org (membership required to access). Accessed February 2, 2012.
11 Ramos KD, Schafer S, Tracz SM. Validation of the Fresno test of competence in evidence based medicine. *BMJ*. 2003;326(7384):319–321.
12 Feinstein R. *University of Colorado Department of Psychiatry Evidence-Based Medicine Test Scoring Rubric*. Denver: University of Colorado; 2005:1–13.
13 Crocker L, Algina J. *Introduction to Classical and Modern Test Theory*. New York, NY: Harcourt Brace Jovanovich; 1986.
14 Downing M, Haladyna TH. *Validity and its threats*. In: Downing SM, Yudkowsky R, eds. *Assessment in Health Professions Education*. New York, NY: Routledge; 2009.
15 Coomarasamy A, Khan KS. What is the evidence that postgraduate teaching in evidence based medicine changes anything? A systematic review. *BMJ*. 2004;329(7473):1017.