Doctors and Numbers: An Assessment of the Critical Risk Interpretation Test

Tanner J. Caverly, MD, MPH, Allan V. Prochazka, MD, MSc, Brandon P. Combs, MD, Brian P. Lucas, MD, MS, Shane R. Mueller, MSW, Jean S. Kutner, MD, MSPH, Ingrid Binswanger, MD, MPH, Angela Fagerlin, PhD, Daniel D. Matlock, MD, MPH

Background. *Risk interpretation affects decision making. Yet, there is no valid assessment of how clinicians interpret the risk data that they commonly encounter.* ***Objective.*** *To establish the reliability and validity of a 20-item test of clinicians' risk interpretation.* ***Methods.*** *The Critical Risk Interpretation Test (CRIT) measures clinicians' abilities to 1) modify the interpretation based on meaningful differences in the outcome (e.g., disease specific v. all-cause mortality) and time period (e.g., lifetime v. 10-year mortality), 2) maintain a stable interpretation for different risk framings (e.g., relative v. absolute risk), and 3) correctly interpret how diagnostic testing modifies risk. There were 658 clinicians and medical trainees who participated: 116 nurse practitioners (NPs) at a national conference, 273 medical students at 1 institution, 148 residents in internal medicine at 2 institutions, and 121 internists at 1 institution. Participants completed a self-administered paper test during educational conferences. Seventeen evidence-based medicine experts took the test online and formally assessed content validity. Eighteen second-year medical students were recruited to take the test and a retest 3 weeks later to explore test-retest correlation.* ***Results.*** *Expert review supported test clarity and content validity. Factor analysis supported that the CRIT identifies at least 3 separable areas of clinician knowledge. Test-retest correlation was fair (intraclass correlation coefficient = 0.65; standard error = 0.15). Scores on our test correlated with other tests of related abilities. Mean test scores varied among groups, with differences in prior evidence-based medicine training and experience (93 for NPs, 101 for medical students, 101 for residents, 103 for academic internists, and 110 for physician experts;* $P < 0.001$*).* ***Conclusions.*** *Our results provide supporting evidence for the reliability and validity of the CRIT as an index of critical risk interpretation abilities, which is acceptable and feasible to administer in an educational setting.* ***Key words:*** *medical decision making; risk perception; risk communication; numeracy.* ***(Med Decis Making XXXX;XX:xx–xx)***

When recommending a test or treatment, clinicians should prioritize 2 items: determination of the net benefit and informed patient preference.[1,2] A clinician's ability to appropriately interpret data about harms and benefits (risk data) is essential to achieve both of these goals. Despite the importance of correctly interpreting risk data, both patients and physicians misinterpret this information.[3,4] Patient interpretation of numbers is increasingly recognized as a critical factor in health-related decisions.[5] However, quantitative abilities among clinicians have not been evaluated extensively.

Previous randomized studies have identified that risk presentation influences clinician perceptions of treatment benefit and self-reported decision making,[4] influences that can be broadly termed ''framing effects.''[6] Studies examining framing effects have randomized different physicians to varying presentation formats (assessing between-subject framing effects). For instance, one group of physicians is randomized to view the benefit of a new drug in terms of a relative risk reduction (typically a larger number), while a second group is randomized to view a numerically equivalent benefit in terms of an absolute risk reduction (typically a smaller number). Such studies have demonstrated that relative risk reduction is perceived to be a larger and more persuasive benefit among both patients and physicians.[4]

Although these studies have provided evidence about how presentation affects risk perception on average, they ignore how changes in presentation affect an individual clinician's risk perception. Such within-clinician effects are important because

DOI: 10.1177/0272989X14558423

Table 1 Definition of Terms

Risk	The probability of an event occurring over a certain time period.[38] This definition highlights the 3 components of any risk: the probability, the event, and the time period.
Probability	The relative frequency of an event (e.g., ''the chance of a heart attack is x% in patients who are diabetic v. y% in those who are not'').[38]
Risk interpretation	The bottom line meaning or gist that a person gives to a baseline risk or to the benefit/harm of an intervention.[12] For example, if a patient's precise baseline risk is a 3% lifetime chance of dying from prostate cancer, a clinician's gist might be ''really high''!

some clinicians may be highly influenced by the presentation format while others may be only minimally influenced. Additionally, it is not known whether clinicians can be taught to avoid the influence of the presentation format, an influence that is irrelevant in assessing the extent to which an intervention is net beneficial or net harmful.

We sought therefore to develop a test that has the ability to assess within-subject interpretive abilities that go beyond those assessed by current tests of basic numeracy, which focus on statistical and probabilistic operations and computation.[7–10] We sought, on the other hand, to evaluate the effects of framing on a clinician's highest level risk interpretation (i.e., the meaning or gist that a clinician draws from the risk data). Table 1 presents our definition of risk interpretation and other important terms. We also sought to evaluate the effect that changes to the type of outcome and the time period have on risk interpretation.

The primary focus on gist-level interpretation of risk data also distinguishes our test from the Medical Data Interpretation Test, which assesses specific knowledge and the ability to perform tasks and calculations necessary to make comparisons.[11] Basic numeracy and the specific knowledge and skills assessed with previous tests can be thought of as essential precursors to form an accurate risk interpretation. However, because the gist-level interpretation is thought to be most closely linked with medical decision making,[12] we sought to develop a test that directly assesses this gist-level interpretation.

As evidence accumulates about the important effect of numeracy and data presentation on medical decision making (for both patients and clinicians),[4,5,13] understanding the interpretive abilities of clinicians, and whether these abilities can be improved, is crucial. A valid measure of clinician risk interpretation would allow researchers to explore 2 areas: 1) the effect that risk interpretation has on real-world testing and treatment decisions, and 2) the effect that risk interpretation has on the way that benefits and harms are communicated to patients in actual practice. A valid measure of clinician interpretive abilities would also allow medical educators to assess the impact of educational interventions on these abilities. This paper describes the development of a test of critical risk interpretation and the assessment of its reliability and validity among different groups of clinicians.

METHODS

We used a previously published approach to evaluating tests of quantitative skills to assess the reliability and validity of the Critical Risk Interpretation Test (CRIT).[14] First, the tool was developed and modified based on a literature review, a comparison with existing tools, feedback from relevant experts, and focus

Received 5 March 2014 from the Veterans Affairs Center for Clinical Management Research, Ann Arbor Veterans Affairs Health System, University of Michigan Medical School, Ann Arbor, MI (TJC, AF); Internal Medicine, University of Colorado School of Medicine, Aurora, CO (AVP, BPC, SRM, JSK, IB, DDM); Ambulatory Care, Denver Veterans Affairs Medical Center, Denver, CO (AVP); and Department of Medicine, John H. Stroger Jr. Hospital of Cook County, Chicago, IL (BPL). Dr. Caverly was supported by an institutional T32HP1006 grant and the 2012 Founder's Award from the Society of General Internal Medicine. Financial support for this study was provided entirely or in part by a grant (Founder's Award) from the Society of General Internal Medicine. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. This research was previously presented in oral and abstract form at the 2013 annual meeting for the Society of General Internal Medicine, May 2013. There are no conflicts of interest. Revision accepted for publication 26 September 2014.

Supplementary material for this article is available on the *Medical Decision Making* website at http://mdm.sagepub.com/supplemental.

Address correspondence to Tanner J. Caverly, MD, MPH, University of Michigan School of Medicine, 2800 Plymouth Road, Building 16, Room 326W, Ann Arbor, MI 48109-2800, USA; telephone: (734) 222-8958; e-mail: tcaverly@med.umich.edu.

groups. Next, we formally tested the final tool for content validity, usability, reliability, and construct validity. This project was approved by the institutional review boards at the University of Colorado and the John H. Stroger Jr. Hospital of Cook County.

Test Development

The literature review identified a number of studies illustrating that many clinicians struggle with appropriately interpreting health statistics.[15–28] In addition, we reviewed items from tests aimed at measuring patient numeracy and data interpretation as well as numerical concepts thought to be influential in patient medical decision making.[5,7,8,11,13,29–32] Existing questions and concepts from this prior work were considered as candidate items in the initial tool development. The final CRIT contains scenarios taken or modified from several resources identified in this literature review.[11,17,26,27,33] Our initial work,[34] a focus group of 8 clinician-researchers with survey and evidence-based medicine expertise, and early pilot testing with residents in internal medicine identified relatively little difficulty with simple calculations such as converting between a decimal and a percentage or converting a relative risk reduction into an absolute risk reduction when directed to do so. On the other hand, we found significant sensitivity to framing effects (e.g., differing interpretations based on presenting an absolute risk v. a relative risk) among clinicians, in agreement with prior research.[4,26] Thus, we narrowed the focus of our test to assess the effects of framing on a clinician's gist interpretation of a risk. The domains of our test of critical risk interpretation were derived from previous work by Gigerenzer and others[25] and Woloshin and others,[32] identified in the literature review, which synthesized important concepts for appropriately interpreting health statistics. We also conducted interviews with experts in the fields of statistics, education, psychology, epidemiology, and evidence-based medicine. The final domains included the following: 1) modify the interpretation based on differences in the type of outcome and the time period of a risk, 2) consistently interpret the same risk when it is formatted in different ways (e.g., absolute risk v. relative risk), and 3) critically interpret how diagnostic testing modifies the risk.

Clinical Scenarios

The CRIT consists of 21 scenarios and takes less than 15 minutes to administer (see the Appendix to view the full test). Scenarios were designed to reflect the type of risk data that clinicians encounter frequently in the medical literature and in clinical practice guidelines. They were revised based on feedback from a focus group of clinician-researchers with survey expertise and pilot testing with medical trainees.

Scored Test Skills

Eighteen of the scored test skills were inferred from pairs of scenarios that were identical except for a single modification to the type of outcome, the time period of the risk, or the risk format (Figure 1). Each scenario in a pair making up a skill was separated by unrelated scenarios to decrease the likelihood of direct comparisons. Although participants were not aware that scenarios would be paired for scoring, they could change responses to earlier scenarios if they noticed a connection. Figure 1 presents 2 different pairs of scenarios and examples of how skills were inferred from responses to each scenario in a pair. Two multiple choice questions on the test were scored directly as individual skills (skills 14 and 20 in Table 3). Responses to all scenarios were closed-ended using either multiple choice options or a 1-to-10 risk perception scale.

Test Evaluation

Initial versions of the test were modified using a focus group of 8 physician-researchers as well as 3 rounds of pilot testing with residents in internal medicine during an outpatient educational conference. Next, 17 experts from www.HealthNewsReview.org—a website on which physician-researchers, health journalists, epidemiologists, and other researchers critically evaluate health news based on specified criteria—were asked to take an online version of the final test to formally evaluate the test's clarity and content validity (7 physicians and 10 nonphysicians). Test scores from the 7 physician experts were also used to assess discriminant validity. Next, 658 clinicians and trainees were recruited to take a paper survey at educational conferences: 116 nurse practitioners (NPs) at a national conference, 273 medical students from a single academic institution, 148 residents in internal medicine at 2 academic institutions, and 121 internists from a single institution. Finally, a small sample of 18 second-year medical students were asked to take the test and repeat the test 3 weeks later to gain preliminary insight into the CRIT's potential test-retest reliability. A larger sample would be necessary to draw more firm

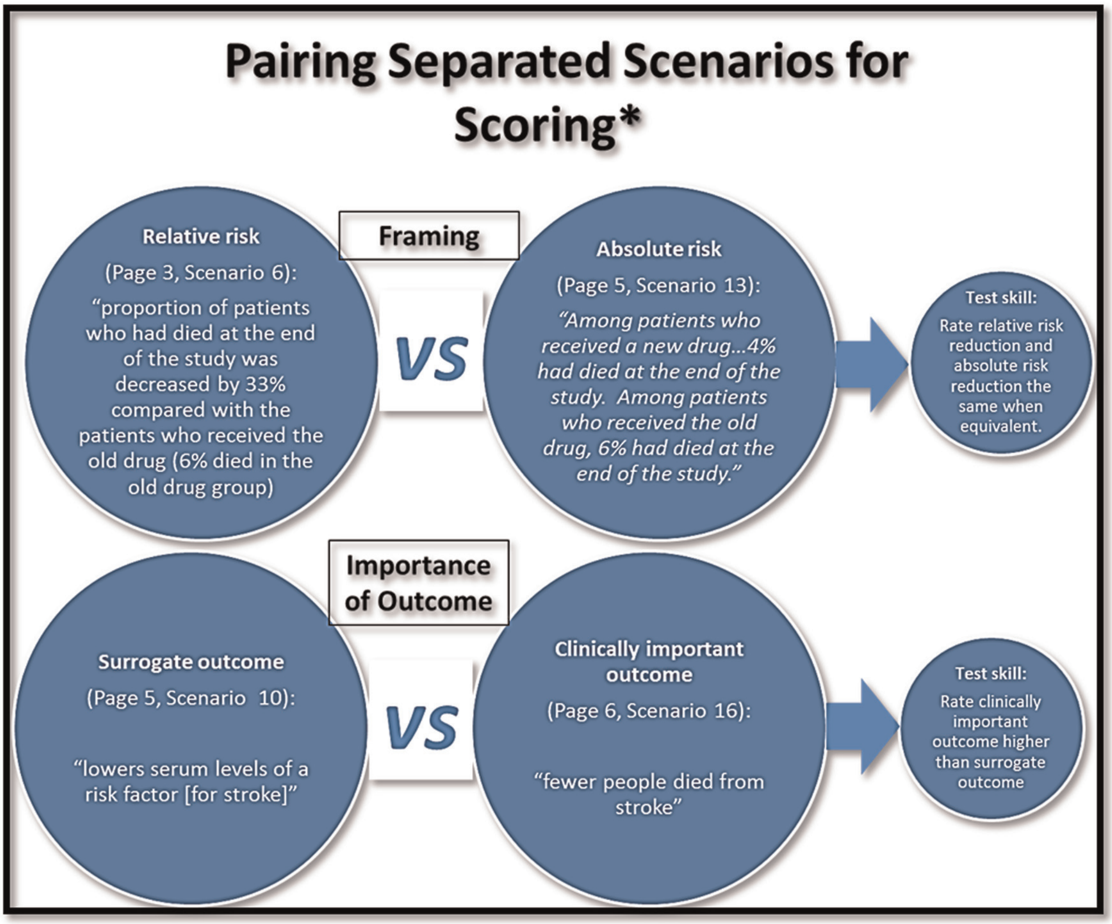

Figure 1 *Two examples of how test skills are inferred from paired scenarios. The first example tests susceptibility to framing effects, and the second example tests sensitivity to the clinical importance of the outcome. On a 1-to-10 scale, respondents were asked to ''rate the extent to which the [benefit] provides proof that the new drug might help people.'' From 1 = no proof to 10 = good proof.*

conclusions about test-retest reliability. However, a larger sample was not possible during the current study. Tests were administered by research staff during educational conferences. All participants were told that their participation was voluntary, that they could stop the test at any time, and that responses would be kept anonymous. The students taking the retest survey and all of the academic internists received gift-card incentives.

All test takers received the exact same order of questions; the question order and the distance between paired scenarios were not varied. It is possible that, given our method of analyzing answers to paired scenarios, order and distance effects could be influential. However, we had several reasons for not testing order effects in this study. Our focus was instead on ensuring that each scenario in a pair was separated as much as possible on the test. Randomizing the order has the downside of potentially bringing scenario pairs closer together. Moreover, the primary aim of this study was to develop a valid and reliable test of critical risk interpretation that could differentiate between individual clinicians based on the overall aggregate score. While we were interested in identifying a set of test skills with a range of difficulty levels, we were less interested in characterizing the difficulty of individual scored skills relative to other scored skills. Thus, randomizing the order was less important for the purposes of this initial work. Finally, we hoped to reduce the number of scenarios in the future to aid additional research using this test in more time-sensitive settings. We felt that testing ordering effects would be more appropriate for future studies because a shorter test could also affect the recall of prior responses apart from any order effects.

Basic Attributes

Scoring. Each of the 20 test skills was assigned a score between 0 (missing or maximally incorrect) and 1 point (ideal response). Direct scoring of the multiple choice scenarios was straightforward (1 if correct and 0 if incorrect). Scoring of skills inferred from responses to paired scenarios was based on simple algorithms, allowing for partial credit (i.e., decimal scores between 0 and 1). Thus, scores for the inferred test skills were kept on a continuous scale, so that we could identify ''more ideal'' and ''less ideal'' responses, allowing us to avoid arbitrary judgments between ''correct'' and ''incorrect'' responses. The Appendix details how all 20 test skills were scored. In addition, to evaluate the robustness of our results to different scoring methods, we also repeated all analyses using a ''dichotomous'' scoring method that graded inferred skills as either correct or incorrect rather than on a continuous scale.

Scenario acceptability and skill difficulty. To measure acceptability, we calculated the proportion of respondents who did not answer one or both of the scenarios making up a skill. Skill difficulty was assessed by calculating the proportion of respondents who gave ideal responses for that skill.

Reliability

Test-retest repeatability. To assess the extent to which an individual's test score remains stable over time, we calculated an intraclass correlation coefficient (ICC) of the 2 continuous CRIT scores for each second-year medical student (n = 18) who completed the test and retest 3 weeks later.

Internal structure. Factor analysis was undertaken to determine the dimensionality of the set of scored skills and the need for scaling correlated skills. Principal factor analysis was used to identify the number of factors with eigenvalues greater than 1. Oblique promax rotation was then used to identify the set of skills that loaded onto each factor with a value greater than 0.3. Nonorthogonal rotation was selected because we hypothesized that even if domains of clinician knowledge/skill are separable, they are still likely to be correlated. Standardized factor scores were created for each participant, and a Cronbach α value was calculated for the set of skills making up each factor. Skills that did not substantively contribute to any of the factors (ie, ''unique'' skills) were also standardized. Because we intended to test a diverse set of skills, we expected that a number of skills were testing unique aspects of clinician ability. Thus, factor scores and unique skill scores were combined to create an overall critical risk interpretation index, with a mean of 100 and a standard deviation of 10.

Validity

Content evidence. In addition to adopting a rigorous development process to support content validity, we also formally evaluated the content validity of the CRIT. The same 17 experts mentioned previously completed the CRIT and then assessed its content validity by rating the following: 1) the clarity of the questions, 2) whether the test covered important domains of risk interpretation, 3) the extent to which it was a valid assessment of the effect of framing on a clinician's interpretation of risk, and 4) the extent to which a poor test score would reflect the inability to interpret basic risk information (similar to other assessments[11]).

Relationship across groups. We assessed the validity of our test in 3 additional ways. First, convergent validity was explored by assessing the extent to which the CRIT correlates with validated tests measuring related abilities. Two tests were used to assess convergent validity: 1) the 4-item Berlin Numeracy Test,[9] which assesses statistical numeracy; and 2) the 3-item Cognitive Reflection Test,[35] which tests the ability to suppress an intuitively attractive but incorrect response in favor of a correct response arrived at through deliberation. Second, discriminant validity was explored by evaluating the extent to which subgroups of clinicians with different levels of experience and training in evidence-based medicine scored differently. We used the type of provider and stage of training as a rough identifier of prior evidence-based medicine training and experience. We predicted that scores on the CRIT would be highest among the group of 7 physician-researcher experts, followed by practicing academic internists, residents in internal medicine, medical students, and NPs (who have typically had less robust training on critically appraising evidence for making patient care decisions). Third, we assessed the extent to which answers to individual scenarios

Table 2 Characteristics of the Study Sample

Characteristic	N = 658
Women, n (%)	362 (55)
Nurse practitioners, n (%)	116 (18)
Years since completed training, median (range)	14 (<1–43)
Academic setting, n (%)	24 (21)
Caring for adults, n (%)	104 (90)
Outpatient setting, n (%)	101 (87)
Medical students, n (%)	273 (41)
Third-year students, n (%)	131 (48)
Fourth-year students, n (%)	142 (52)
Internal medicine residents, n (%)	148 (22)
First-year residents, n (%)	72 (49)
Second-year residents, n (%)	38 (26)
Third-year residents, n (%)	38 (26)
Academic faculty in general internal medicine, n (%)	121 (18)
Years since graduated from medical school, median (range)	14 (3–46)
Mostly outpatient, n (%)	73 (60)
Mostly inpatient, n (%)	47 (39)
With degrees in addition to clinical degree (e.g., MPH), n (%)	16 (13)
Physician-researcher, n (%)	19 (16)
With degrees in addition to clinical degree, n (%)	165 (25)
Cognitive Reflection Test score[35] (n = 539; does not include nurse practitioners)	
With 0 correct, n (%)	54 (10)
With 1 correct, n (%)	113 (21)
With 2 correct, n (%)	178 (33)
With 3 correct, n (%)	194 (36)
Berlin Numeracy Test score[9] (n = 655)	
With 0 correct, n (%)	59 (9)
With 1 correct, n (%)	138 (21)
With 2 correct, n (%)	183 (28)
With 3 correct, n (%)	157 (24)
With 4 correct, n (%)	118 (18)

were concordant with previous research regarding the influence of framing effects on risk perception.

Analysis

Linear regression was used to assess the trends of test scores across groups. Paired t tests were used to compare the mean difference between responses on scenarios making up a specific test skill. The mean difference between responses on paired scenarios was divided by their pooled standard deviation to calculate the standardized mean difference (SMD).[36] An ICC was calculated to assess test-retest correlation, and we used Pearson correlation to test for correlations among any factors identified on factor analysis. A 2-tailed P value of <0.05 was considered statistically significant. All analyses were performed using Stata version 13.1 (StataCorp, College Station, TX).

RESULTS

The characteristics of the study sample are described in Table 2. Fifty-five percent of the participants were female, and 25% had additional degrees such as an MPH or a PhD. Scores on 2 previously validated tests assessing statistical numeracy (Berlin Numeracy Test[9]) and the ability to suppress an intuitively appealing but incorrect answer (Cognitive Reflection Test[34]) varied substantially, indicating a range of abilities among the participating clinicians.

Scored Test Skills

The proportion of test skills with missing responses was low (<2%), indicating good acceptability (Tables 3 and 4). Test skills were difficult overall, with fewer than 60% of clinicians scoring optimally on most skills (Table 3). The easiest skill assessed whether clinicians interpreted a reduction in mortality as better proof that a screening test was beneficial over increased disease detection rates alone (78% scored optimally). The most difficult skill assessed understanding that a very low pretest probability limits the predictive value of even a highly sensitive and specific test (10% scored correctly).

Test-Retest Repeatability

The CRIT had fair test-retest reliability (ICC = 0.65; standard error = 0.15). Scores on repeat testing were slightly lower than on initial testing, suggesting minimal or no learning effects between tests.

Internal Structure

Three factors were identified on factor analysis. The 3 factors were made up of 11 scored test skills, while 9 other skills were unique and did not load onto additional factors. The first factor was named the ''Susceptibility to Framing Scale'' because it comprised the 6 skills that tested how interpretation changed when numerically equivalent drug benefits were presented as a relative risk reduction (with baseline risk given), an absolute mortality benefit, an absolute survival increase, or a number needed to treat (skills 3–8 in Table 3; Cronbach $\alpha = 0.76$). The second factor was named the ''Valuing All-cause Mortality Scale'' because it comprised the 3 skills that tested how a participant rated an improvement in all-cause mortality compared with improvements in a surrogate outcome, a surrogate-enriched composite outcome, and an improvement in stroke-specific mortality (skills 16, 18, and 19; Cronbach $\alpha = 0.82$). The third factor was named the ''Valuing Stroke-specific Mortality Scale,'' which contained 1 skill in common with the ''Valuing All-cause Mortality Scale'' (skill 19) as well as 2 additional skills that tested how a participant rated an improvement in stroke-specific mortality compared with a surrogate and a surrogate-enriched composite outcome (skills 15 and 17; Cronbach $\alpha = 0.48$).

The 3 factor scores were all positively correlated with each other: the ''Susceptibility to Framing Scale'' score was significantly correlated with both ''Valuing All-cause Mortality Scale'' and ''Valuing Stroke-specific Mortality Scale'' scores (Pearson r = 0.46 and 0.24, respectively; $P < 0.001$ after the Sidak adjustment to significance levels); ''Valuing All-cause Mortality Scale'' scores were also positively correlated with ''Valuing Stroke-specific Mortality Scale'' scores ($r = 0.32$; $P < 0.001$).

Index Scores

The mean index score among all participants was set at 100 with a standard deviation of 10. Index scores ranged from a low of 44 to a high of 123 and were normally distributed (Table 4). The bottom 25% of participants scored below 95, while the top 25% of participants scored above 106.

Validity

The expert health news reviewers rated the clarity and content validity of the CRIT favorably (Table 4). Seventy-five percent of the 17 experts thought that the CRIT did an excellent or good job at covering the important concepts inherent in the phrase ''basic risk interpretation,'' and 69% rated the clarity of the test as excellent or good. However, there was disagreement about the extent to which scoring poorly on the test indicated an inability to interpret basic risk information. The distribution of CRIT scores among subgroups supported the convergent and discriminant validity of the test (Figure 2). Clinicians with better statistical numeracy, better cognitive reflection scores, and more training and experience with evidence-based medicine concepts scored significantly better on the test (r = 0.36, 0.26, and 0.35, respectively; $P < 0.001$ for all trends on bivariable linear regression).

Examination of the response distribution for individual scenarios identified patterns consistent with previous research findings[4] (distributions displayed in the Appendix). For instance, the mean response when a drug benefit was presented as a relative risk reduction was 7.0 on a scale from 1 (no proof of benefit) to 10 (good proof of benefit). The mean response was 5.1 when the same information was presented as an absolute risk reduction (mean difference = 1.9; $P < 0.001$). The SMD was 0.8, indicating a ''large'' difference.[36] This finding is consistent with multiple previous studies that have demonstrated framing effects on patient and clinician perceptions.[4] Similarly, when the same risk was presented as an absolute number (1 million) as opposed to a proportion (0.3%; i.e., 1 million out of 300 million), responses also varied significantly in the expected direction, that is, with the absolute number appearing more impressive than the equivalent proportion[37] (SMD = 0.5; $P < 0.001$). The extent that responses on our test align with previous research findings further

Table 3 Proportion of Ideal Responses and Missing Answers to the 20 Skills on the Critical Risk Interpretation Test

	N = 658		
	Ideal Response, %	**Left Blank, %**	**Test Skill[a]**
Modify interpretation based on differences in the type of outcome			
For a screening test, answer that mortality reduction is a better endpoint than disease detection rates.[b]	78	0.3	13
Answer that a patient-oriented endpoint (stroke mortality) is a better endpoint than a surrogate endpoint (improved risk factor).[b]	63	0.9	15
For a screening test, answer that mortality reduction is a better endpoint than improved 5-year survival rates.[b]	59	0.5	12
Select a larger risk estimate for deaths from all causes than death from a specific disease.[c]	48	0.8	1
Answer that a patient-oriented endpoint (all-cause mortality) is a better endpoint than a surrogate endpoint.[b]	43	1.2	16
Rate the risk of dying from a disease as more important than the risk of getting the disease.[d]	39	0.6	9
Answer that a patient-oriented endpoint (stroke mortality) is a better endpoint than a composite containing surrogate endpoints.[b]	29	0.8	17
Answer that a patient-oriented endpoint (all-cause mortality) is a better endpoint than a composite containing surrogate endpoints.[b]	18	1.2	18
Answer that all-cause mortality is a better endpoint than disease-specific mortality.[b]	17	1.1	19
Modify interpretation based on meaningful differences in the time period			
Rate the risk of dying over the next 10 years as more important than the same risk over a lifetime for a 40-year-old patient.[d]	58	0.9	10
Select a larger risk estimate for a 20-year than a 10-year risk.[c]	53	0.9	2
Maintain a stable interpretation when the same risk is presented in different ways			
Rate absolute survival increase and absolute mortality reduction the same when they are numerically equivalent.[b]	37	0.5	8
Rate an absolute number the same as a proportion when they are equivalent.[d]	31	0.5	11
Rate the number needed to treat and relative risk reduction the same when they are numerically equivalent.[b]	28	1.7	3
Rate a relative risk reduction and absolute mortality reduction the same when they are numerically equivalent.[b]	26	0.6	7
Rate the number needed to treat and absolute mortality reduction the same when they are numerically equivalent.[b]	21	1.7	5
Rate a relative risk reduction and absolute survival increase the same when they are numerically equivalent.[b]	20	0.5	6
Rate the number needed to treat and absolute survival increase the same when they are numerically equivalent.[b]	17	1.5	4
Critically interpret how testing modifies risk			
Answer that changing a test cutoff affects the number of false-positives and false-negatives.[c]	67	2.1	20
Answer that a very low pretest probability limits the predictive value of even a highly sensitive and specific test.[c]	10	0.3	14

a. Provided for reference when looking at the full test in the Appendix.
b. Based on a 10-point response scale that ranged from 1 to 10 with anchors at 1 (‘‘no proof’’) and 10 (‘‘good proof’’).
c. Based on multiple choice questions with 4 options.
d. Based on a 10-point response scale that ranged from 1 to 10 with anchors at 1 (‘‘not important’’) and 10 (‘‘very important’’).

Table 4 Psychometric Properties for the CRIT

Property	Scale Performance		
Basic attributes			
Individual scenarios			
Proportion not answering skill (goal is to minimize skill nonresponse), %	0.3–2.1 (mean = 0.9)		
Proportion answering skill correctly (goal is a broad range of skill difficulty), %	10 (skill 14)–78 (skill 13)		
Aggregate index score			
25th percentile, 75th percentile	95, 106 (range = 44–123)		
Reliability			
Test-retest repeatability (whether answers are the same over time)			
Correlation of scores at test and retest 3 weeks later (goal is ICC > 0.6)	ICC = 0.65		
Internal structure (3 factors): Cronbach α (goal is 0.7–0.8; lower suggests lack of coherence; higher suggests skills are redundant)			
"Susceptibility to Framing Scale" (skills 3–8)	α = 0.76		
"Valuing All-cause Mortality Scale" (skills 16, 18, 19)	α = 0.82		
"Valuing Stroke-specific Mortality Scale" (skills 15, 17, 19)	α = 0.48		
Validity			
Content evidence (are relevant domains of "basic risk interpretation" covered?): scenarios developed and revised based on feedback from experts in evidence-based medicine, statistics, and epidemiology			
Expert health news reviewer responses (n = 17)	*Excellent/good, %*	*Fair, %*	*Poor, %*
How well would you say the test covers the important concepts inherent in the phrase "basic risk interpretation"?	75	25	0
How would you rate the clarity of the test?	69	25	6
How well does this test assess the influence of framing on a clinician's interpretation of a risk?	76	19	6
If a clinician scored poorly on our test, this would indicate that their ability to interpret basic risk information is limited.	50% agree (suggests that the CRIT may also assess more difficult interpretive abilities)		
Relationship across groups (scale discriminates among groups with different attitudes or abilities)[a]: comparison of mean scores with the following hypothesized relationships (mean score group 1 > mean score group 2)			
Clinicians with the highest > lowest statistical numeracy	105 v. 94 ($P < 0.001$)		
Clinicians with the highest > lowest cognitive reflection scores	104 v. 96 ($P < 0.001$)		
Expert physician > nurse practitioner	110 v. 93 ($P < 0.001$)		

Note: CRIT = Critical Risk Interpretation Test; ICC = intraclass correlation coefficient.
a. Each comparison is between those with the highest versus lowest level of the characteristic (e.g., statistical numeracy score of 4 v. score of 0). Significance levels are given using marginal differences by group on bivariate linear regression, adjusting for multiple comparisons using the Scheffe method.

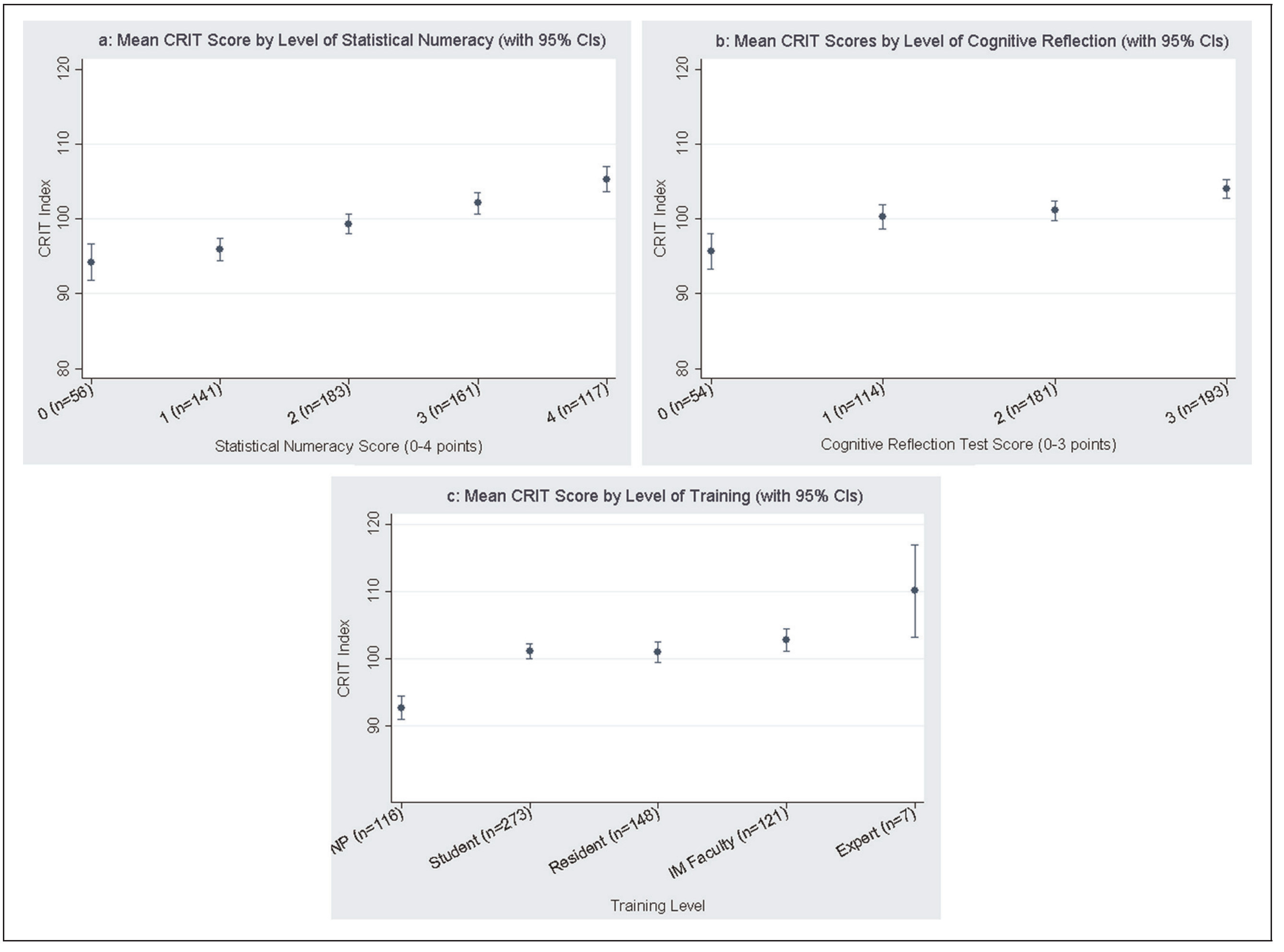

Figure 2 *Mean Critical Risk Interpretation Test (CRIT) scores for different subgroups with 95% confidence intervals. Differences in scores across all subgroups were statistically significant at* $P < 0.001$ *on linear regression.*

supports the content and construct validity of the CRIT. For comparison, changes in the type of outcome (getting cancer v. dying from cancer) and changes in the time period (10-year risk v. lifetime risk) were both associated with medium effect sizes (SMD = 0.6).

Finally, secondary analysis demonstrated similar psychometric properties for the CRIT when changing the scoring methodology to a dichotomous correct/incorrect answer rather than keeping inferred skills on a continuous scale. In addition, we conducted post hoc bivariate linear regression to assess for the presence of any significant distance effects (i.e., effects due to the number of intervening scenarios between pairs making up a test skill). With the mean score on each test skill as the dependent variable and the distance between scenarios as the independent variable, there was no statistically significant relationship ($P = 0.08$). Given the small sample of only 18 observations (i.e., the 18 test skills), we also looked at the 95% confidence interval (CI) for the regression coefficient to evaluate whether large distance effects could be excluded (β coefficient = −0.01; 95% CI = −0.03 to 0.002). The lower limit of the 95% CI for the coefficient of −0.03 indicates the potential for only small distance effects and excludes large or even moderate effects. We repeated these post hoc analyses using the proportion with an ''ideal response'' as the dependent variable (rather than mean scores). Again, we found no significant relationship between distance and the proportion with an ''ideal response,'' and moderate to large effects due to distance between scenarios could be excluded.

DISCUSSION

The CRIT was developed to assess a set of critical thinking skills that we believe are essential for high-value medical decision making and optimal risk communication with patients. The domains of the CRIT were structured around the fact that there are 3 components to any risk: the probability (e.g., 3%), the type of outcome (e.g., 3% chance of dying from cancer), and the time period (e.g., 3% chance of dying from cancer over the next 10 years). Specifically, the CRIT was designed to evaluate a clinician's ability to 1) modify the risk interpretation based on meaningful differences in the type of outcome and the time period of a risk, 2) consistently interpret the same risk when it is framed in different ways (e.g., absolute risk framing v. relative risk framing), and 3) critically interpret how testing modifies risk. Experts in evidence-based medicine rated the content validity of the CRIT favorably. While our literature review indicated that we were testing core concepts regarding the interpretation of risk information, there was disagreement among the experts regarding the extent to which scoring poorly on the test indicated an inability to interpret basic risk information. Given this disagreement and the high difficulty level of the test, it may be that these skills are both essential and difficult to attain in practice. Convergent validity was supported by the fact that performance on the CRIT was correlated with performance on 2 previously validated measures assessing related constructs. Discriminant validity was supported by the fact that the CRIT was able to discriminate among groups of clinicians with different backgrounds in evidence-based medicine training and experience. Finally, the results of this study also support the acceptability and feasibility of administering the CRIT in an educational setting.

Factor analysis indicated that the CRIT is multidimensional and best viewed as an index containing 3 scales and a number of unique skills. This is an unsurprising finding as we intended to assess a diverse set of skills and these skills were measured in different ways (i.e., scales with different anchors, several multiple choice questions, some skills inferred from 2 separated scenarios while 2 skills assessed using a single multiple choice question). When multiple skills measured the same concept in the same way, we did find several scales within the overall index. Two scales demonstrated good reliability: the ''Susceptibility to Framing Scale'' and the ''Valuing All-cause Mortality Scale,'' (α = 0.76 and 0.82, respectively). A third scale, the ''Valuing Stroke-specific Mortality Scale,'' had suboptimal reliability (α = 0.48).

Response distributions for individual scenarios were consistent with findings from previous research,[4] further contributing to the CRIT's construct validity. Importantly, by separating scenarios and inferring interpretation skills, we were able to demonstrate these framing effects within the same individual (within-subject effects) as opposed to the between-subject effects identified in previous studies. This distinction is critically important and represents a major objective of developing and validating the CRIT. A valid tool that assesses within-subject effects is necessary to explore variation in interpretive abilities, the extent to which these abilities are modifiable, the role that framing effects have on test and treatment practices, and the role that framing effects have on how clinicians communicate the benefits and harms of medical interventions. Clinicians would likely have scored better if the same interpretive abilities were directly assessed rather than inferred. However, risk data are often encountered without comparison to other presentations in day-to-day practice. Given this fact, our inferred assessment approach may be representative of how clinicians often view and interpret risk data. In addition, the subjective response scales are a particular strength of our measure, given that a person's gist of risk data is likely to have the most impact on medical decision making.[12]

To our knowledge, some skills on the CRIT have not been previously assessed. For example, we found that irrelevant changes to the presentation format had at least as big an effect on clinician risk perception as significant changes to the type of outcome or the time period of a risk. This suggests that clinicians may be overly sensitive to irrelevant framing effects and not sensitive enough to components of a risk that are crucial to accurate risk interpretation (i.e., the type of outcome or time period of a risk).

This study has several limitations. We used a convenience sample of clinicians who were predominantly from a single academic institution. This limits the generalizability of our findings. However, our goal was not to generalize the results to a large population but to assess the validity of the CRIT in typical users. We aimed to develop a coherent set of skills to assess the construct of ''critical risk interpretation.'' However, some may disagree with the content of our test or feel that we left important skills out. Nonetheless, our development process utilized an extensive literature review and interviews of experts in different disciplines; the validity of the final content was supported by the formal assessment of a group with content expertise. Training level and professional degree were used as rough proxies for prior training in and experience with evidence-based medicine principles. A more accurate assessment of prior evidence-based medicine experiences would have been helpful but was not possible in this study. Use of the CRIT in research studies beyond the educational setting may require a shortened test. Finally, the CRIT should not be used for high-stake evaluations of individual clinicians, which would require a higher level of reliability. The CRIT can be used, however, to assess correlations between CRIT scores and clinical performance among groups of clinicians or to evaluate the efficacy of educational interventions aimed at improving interpretive abilities.

Further assessment of the predictive validity of the CRIT or its subscales, that is, how individual clinicians' CRIT scores correlate with the quality of risk communication practices and the quality of testing and treatment decisions, would usefully extend previous findings about the effects of risk presentation on interpretation and theoretical decision making in vignette-based studies. Such assessments would allow insight into the role that risk interpretation might play in real-world clinical performance. For instance, with this new and useful measure, future research could investigate whether more skillful risk interpretation is associated with higher patient ratings of shared decision making or a higher quality of shared decision making in observations of routine clinical encounters. Moreover, simply administering the CRIT and providing feedback on performance may be a useful educational tool. Test takers at all levels of medical training may be surprised by how their interpretation varies based on the way that a risk is presented and thus motivated to interpret risk information with a more critical eye in the future.

CONCLUSIONS

The results of this single-site measurement development study provide support for the reliability and validity of using the CRIT as an index of critical risk interpretation. We found that clinicians have significant room for improvement when it comes to critically interpreting the risk data found in health news, medical literature, and clinical practice guidelines. As has been done with patient tests of numeracy[5,13] and data interpretation skills,[11] we hope that the CRIT can be used to inform educational, risk communication, and policy interventions to ensure that clinicians are optimally contributing to the shared decision-making process, thereby promoting high-value care.

ACKNOWLEDGMENTS

The authors acknowledge Lisa Schwartz, MD, MS, and Steven Woloshin, MD, MS, for their early feedback on the domains of our test, sharing previous work developing and validating the Medical Data Interpretation Test, and highlighting the importance of understanding health statistics, which motivated and informed our study.

REFERENCES

1. Eddy DM. Clinical Decision Making: From Theory to Practice. A Collection of Essays from the Journal of the American Medical Association. 1st ed. Boston: Jones & Bartlett; 1996.

2. Mulley AG, Trimble C, Elwyn G. Stop the silent misdiagnosis: patients' preferences matter. BMJ. 2012;345:e6572.

3. Akl EA, Oxman AD, Herrin J, et al. Framing of health information messages. Cochrane Database Syst Rev. 2011;(12):CD006777.

4. Akl EA, Oxman AD, Herrin J, et al. Using alternative statistical formats for presenting risks and risk reductions. Cochrane Database Syst Rev. 2011;(3):CD006776.

5. Reyna VF, Nelson WL, Han PK, Dieckmann NF. How numeracy influences risk comprehension and medical decision making. Psychol Bull. 2009;135(6):943–73.

6. Kahneman D. Thinking, Fast and Slow. 1st ed. New York: Farrar, Straus and Giroux; 2011.

7. Schwartz LM, Woloshin S, Black WC, Welch HG. The role of numeracy in understanding the benefit of screening mammography. Ann Intern Med. 1997;127(11):966.

8. Lipkus IM, Samsa G, Rimer BK. General performance on a numeracy scale among highly educated samples. Med Decis Making. 2001;21(1):37–44.

9. Cokely ET, Galesic M, Schulz E, Ghazal S, Garcia-Retamero R. Measuring risk literacy: the Berlin Numeracy Test. Judgm Decis Mak. 2012;7(1):25–47.

10. Schapira MM, Fletcher KE, Gilligan MA, et al. A framework for health numeracy: how patients use quantitative skills in health care. J Health Commun. 2008;13(5):501–17.

11. Schwartz LM, Woloshin S, Welch HG. Can patients interpret health information? An assessment of the Medical Data Interpretation Test. Med Decis Making. 2005;25(3):290–300.

12. Reyna VF. A theory of medical decision making and health: fuzzy trace theory. Med Decis Making. 2008;28(6):850–65.

13. Nelson W, Reyna VF, Fagerlin A, Lipkus I, Peters E. Clinical implications of numeracy: theory and practice. Ann Behav Med. 2008;35(3):261–74.

14. Wallace D, Rheinlander K, Woloshin S, Schwartz L. Quantitative literacy assessments: an introduction to testing tests. Numeracy. 2009;2(2). Available from: http://scholarcommons.usf.edu/numeracy/vol2/iss2/art3/

15. Berwick DM, Fineberg HV, Weinstein MC. When doctors meet numbers. Am J Med. 1981;71(6):991–8.

16. Wulff HR, Andersen B, Brandenhoff P, Guttler F. What do doctors know about statistics? Stat Med. 1987;6(1):3–10.

17. Abernathy CM, Hamm RM. Surgical Intuition: What It Is and How to Get It. 1st ed. Philadelphia: Hanley & Belfus; 1995.

18. Fahey T, Griffiths S, Peters TJ. Evidence based purchasing: understanding results of clinical trials and systematic reviews. BMJ. 1995;311(7012):1056–60.

19. Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. Acad Med J Assoc Am Med Coll. 1998;73(5):538–40.

20. Gigerenzer G, Hoffrage U, Ebert A. AIDS counselling for low-risk clients. AIDS Care. 1998;10(2):197–211.

21. Fritsche L, Greenhalgh T, Falck-Ytter Y, Neumayer H-H, Kunz R. Do short courses in evidence based medicine improve knowledge and skills? Validation of Berlin questionnaire and before and after study of courses in evidence based medicine. BMJ. 2002;325(7376):1338–41.

22. Sheridan SL, Pignone M. Numeracy and the medical student's ability to interpret data. Eff Clin Pract ECP. 2002;5(1):35–40.

23. Windish DM, Huot SJ, Green ML. Medicine residents' understanding of the biostatistics and results in the medical literature. JAMA. 2007;298(9):1010–22.

24. Shaneyfelt T, Baum KD, Bell D, et al. Instruments for evaluating education in evidence-based practice. JAMA. 2006;296(9): 1116–27.

25. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping doctors and patients make sense of health statistics. Psychol Sci Public Interest. 2007;8(2):53.

26. Perneger TV, Agoritsas T. Doctors and patients' susceptibility to framing bias: a randomized trial. J Gen Intern Med. 2011;26(12): 1411–7.

27. Wegwarth O, Schwartz LM, Woloshin S, Gaissmaier W, Gigerenzer G. Do physicians understand cancer screening statistics? A national survey of primary care physicians in the United States. Ann Intern Med. 2012;156(5):340–9.

28. Moyer VA. What we don't know can hurt our patients: physician innumeracy and overuse of screening tests. Ann Intern Med. 2012;156(5):392–3.

29. Fagerlin A, Zikmund-Fisher BJ, Ubel PA, Jankovic A, Derry HA, Smith DM. Measuring numeracy without a math test: development of the Subjective Numeracy Scale. Med Decis Making. 2007; 27(5):672–80.

30. Peters E, Hibbard J, Slovic P, Dieckmann N. Numeracy skill and the communication, comprehension, and use of risk-benefit information. Health Aff (Millwood). 2007;26(3):741–8.

31. Peters E, Västfjäll D, Slovic P, Mertz CK, Mazzocco K, Dickert S. Numeracy and decision making. Psychol Sci. 2006;17(5): 407–13.

32. Woloshin S, Schwartz LM, Welch HG. Know Your Chances: Understanding Health Statistics. 1st ed. Berkeley: University of California Press; 2008.

33. Windish DM, Price EG, Clever SL, Magaziner JL, Thomas PA. Teaching medical students the important connection between communication and clinical reasoning. J Gen Intern Med. 2005; 20(12):1108–13.

34. Caverly TJ, Prochazka AV, Binswanger IA, Kutner JS, Matlock DD. Confusing relative risk with absolute risk is associated with more enthusiastic beliefs about the value of cancer screening. Med Decis Making. 2014;34(5):686–92.

35. Frederick S. Cognitive reflection and decision making. J Econ Perspect. 2005;19(4):25–42.

36. Guyatt G. JAMA's Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice. New York: McGraw-Hill; 2008.

37. Huff D. How to Lie with Statistics. London: Penguin Books; 1991.

38. Rothman KJ, Lash TL, Greenland S. Modern Epidemiology. 3rd rev. Philadelphia: Lippincott Williams & Wilkins; 2012.

Corrigendum: Doctors and Numbers: An Assessment of the Critical Risk Interpretation Test

Caverly TJ, Prochazka AV, Combs BP, Lucas BP, Mueller SR, Kutner JS, Binswanger I, Fagerlin A, McCormick J, Pfister S, Matlock DD. Doctors and numbers: An assessment of the critical risk interpretation test. Med Decis Making. 2015;35(4):512-524. (Original DOI: 10.1177/0272989X14558423)

In the OnlineFirst version of this article, authors Jacqueline McCormick and Shirley Pfister were inadvertently left out of the author list. Their names have been included in the version of the article that appears in this issue. All authors have confirmed their contributions as authors.

DOI: 10.1177/0272989X15585799

For information about MDM manuscript requirements and submission, please see http://mdm.sagepub.com/ and click the Submit a Manuscript navigation button.