ARTICLE

Development, Validation, and Implementation of a Medical Judgment Metric

Rami A. Ahmed, DO, MHPE, Michele L. McCarroll, PhD, Alan Schwartz, PhD, M. David Gothard, MS, S. Scott Atkinson, Patrick G. Hughes, DO, Jose Ramon Cepeda Brito, MD, Lori Assad, MPH, Jerry G. Myers, PhD, Richard L. George, MD, MSPH, FACS

Background: *Medical decision making is a critical, yet understudied, aspect of medical education.* ***Aims:*** *To develop the Medical Judgment Metric (MJM), a numerical rubric to quantify good decisions in practice in simulated environments; and to obtain initial preliminary evidence of reliability and validity of the tool.* ***Methods:*** *The individual MJM items, domains, and sections of the MJM were built based on existing standardized frameworks. Content validity was determined by a convenient sample of eight experts. The MJM instrument was pilot tested in four medical simulations with a team of three medical raters assessing 40 participants with four levels of medical experience and skill.* ***Results:*** *Raters were highly consistent in their MJM scores in each scenario (intraclass correlation coefficient 0.965 to 0.987) as well as their evaluation of the expected patient outcome (Fleiss's Kappa 0.791 to 0.906). For each simulation scenario, average rater cut-scores significantly predicted expected loss of life or stabilization (Cohen's Kappa 0.851 to 0.880).* ***Discussion****: The MJM demonstrated preliminary evidence of reliability and validity.* ***Key words:*** *clinical judgment; medical judgment; decision making; simulation.* ***(MDM Policy & Practice XXXX;XX:1–8)***

Medical decision making is a critical, yet understudied, aspect of medical education.[1] The ability to demonstrate sound medical decision making is among the most highly sought characteristics of medical professionals,[2] such that the American College of Graduate Medical Education (ACGME) has included medical decision making as a core competency in several specialties.[3-5] For example, a poorly conducted initial history and physical examination can lead to poor medical decision making in acute care situations and can result in patient death.[6] The role of decision making in the academic sense focuses on the rational and methodical approach to diagnostic reasoning versus the underlying subconscious nature of decision making that generally takes over in uncertain medical situations.[7]

As medical students graduate from a wide variety of training programs, there remains significant variability in medical decision making at the initiation of their residency training.[8,9] Unfortunately, measuring and quantifying medical decision making is difficult in a consistent form, especially in varying medical conditions, and the training and assessment of medical decision making is still underdeveloped.[10] As medical training continues to evolve, valid assessments will be required to ensure that trainees have achieved and maintained required proficiencies as well as good clinical decision–making skills.

Medical simulation, despite its popularity in critical care training and mainstream assimilation into academia, provides a yet underutilized methodology to train and assess medical decision making in a variety of environments. Importantly, simulation labs provide

Received 7 March 2016 from Summa Health, Akron, Ohio (RAA, MLM, SSA, PGH, JRCB, LA, RLG); Northeast Ohio Medical University, Rootstown, Ohio (RAA, MLM, RLG); University of Illinois at Chicago, Chicago, Illinois (AS); Biostats, Inc., East Canton, Ohio (MDG); and NASA Glenn Research Center, Cleveland, Ohio (JGM). Financial support for this study was provided entirely by a subcontract with Zin Technologies, Inc. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. Revision accepted for publication 23 February 2017.

Address correspondence to Michele L. McCarroll, PhD Summa Center for Women's Health Research, Summa Health, 525 East Market Street, Medical Building 2nd Floor, Akron, OH 44304, USA; telephone: (330) 375-4880; e-mail: mccarrollm@summahealth.org.

Reprints and permission:
http://www.sagepub.com/journalsPermissions.nav
DOI: 10.1177/2381468317715262

a standardized environment in which high-fidelity simulations allow a comprehensive evaluation of the entire process from data assimilation, diagnostic planning and interpretation, as well as integrated management actions. The development of a medical decision-making assessment with evidence for validity could support future evaluation, training, and monitoring. The purpose of this study was twofold: 1) to develop the Medical Judgment Metric (MJM), a numerical rubric to quantify good decisions in practice in simulated environments; and 2) to obtain initial preliminary evidence of reliability and validity of the tool.

METHODS

Development

The individual judgment items, clinical domains, and competency sections of the initial MJM were created based on the approach of Weber and others,[11] existing ACGME Clinical Competency Committees,[12] and the Association of American Medical Colleges[13] conceptual frameworks for clinical judgment. Over the course of 4 months, a team of medical simulation, emergency medicine, medical decision making, and trauma experts devised a list of medical judgment items not specific to any particular medical condition. The MJM development team defined medical decision making as how people routinely perceive, understand, and judge things to make medical decisions for others measured by a score to quantify good decisions in practice.[14] The MJM medical judgment items were developed with three to six internal anchors (core observed medical behaviors). This methodology is modeled after other validated teamwork/nontechnical skill checklists employed in surgery[15] and trauma[16] that provide several internal anchors in an attempt to improve interrater reliability. Subsequently, the MJM judgment items were stratified into clinical domains that were defined as history and physical, diagnostic, interpretation, and management, with a maximum score of four in each domain on a 0.5 interval scale (see Figure 1a–d). Once the list of medical judgment items and domains were reviewed, items were stratified into competency levels: Novice, Intermediate, Proficient, and Advanced. The intended use of the MJM is for raters to select the items (with core internal anchors) in each domain (history and physical, diagnostic, interpretation, and management) where the maximum score for each domain is 4. The domain scores represented the observed performance during simulation scenarios, thus adding up to an overall performance MJM maximum score of 16.

Content Validity

Using the method by Wynd and others,[17] eight experts from cardiology, emergency medicine, family medicine, and medical simulation backgrounds reviewed the MJM for content validity. Experts were recruited in two waves. In the first wave, $n = 4$ experts were from a single American College of Surgeons (ACS) verified level I trauma institution in the United States (a Level I Trauma Center is a comprehensive regional resource capable of providing total care for every aspect of injury—from prevention through rehabilitation).[18] In the second wave, $n = 4$ experts were from other ACS level I trauma institutions in the United States. No expert participated in both waves, and the authors were not eligible to be recruited as experts in the second wave.

In each wave, the experts were asked to rate whether the MJM judgment items, clinical domains, and competency levels were ‘‘not relevant,’’ ‘‘somewhat relevant,’’ ‘‘quite relevant,’’ or ‘‘very relevant’’ content. Also, the content experts were asked to determine if content items should be added, deleted, or modified in each of the clinical domains and competency levels. The MJM was revised after each wave based on the expert agreement for each item, domain, and competency level. The last draft of the MJM was distributed among the MJM development team for final review of flow, ease of use, and formatting for preliminary field-testing.

Internal Structure

Preliminary field-testing for clarity and feasibility of the MJM instrument was conducted by administering the instrument in four groups of 10 participants each. Due to the pilot nature of the study, this sample size was chosen to establish feasibility and provide preliminary interrater reliability and validity figures to conduct a future power analysis for further interrater reliability and validity. The groups were recruited to target a variety of experience and skill levels in the medical field from highly educated engineers with no medical background or experience (layperson), physicians completing their postgraduate medical training, and seasoned physicians with mastery level experience, in addition to spanning a wide spectrum of scientific disciplines. Participants undertook four medical simulations: biliary colic, pneumothorax, ST elevation myocardial infarction, and renal colic, which fall into the two most common reasons for emergency

a

Health & Physical

	Level 1 (Novice)	Level 2 (intermediate)	Level 3 (Proficient)	Level 4 (Advanced)
Health & Physical	□ Does not collect accurate historical data	□ Limited ability to acquire accurate historical information in an organized fashion	□ Demonstrates ability to acquire accurate and relevant histories from patient	□ Proficient at acquiring accurate histories from the patients in an efficient prioritized and hypothesis driven fashion
	□ Does not use Physical exam to confirm history	□ Limited thorough physical exam or misses key physical exam findings	□ Demonstrates ability to perform accurate and appropriately thorough physical exams	□ Proficient at performing accurate physical exams that are targeted to the patients complaints
	□ Does not recognize potentially life threatening presentations	□ Limited ability to seek secondary data (e.g. medical records)	□ Demonstrates ability to seek and obtain data from secondary sources (e.g. medical records) when needed	□ Proficient at synthesizing history and physical exam findings to generate a prioritized differential diagnosis and problem list
	□ Does not recognize patients central clinical problem	□ Limited ability to recognize patient's central clinical problem or develop limited differential diagnosis	□ Demonstrates ability to use history and exam findings to define patient's central clinical problems	□ Proficient at using history and physical exam effectively to minimize the need for further diagnostic

Overall Score □ 1 □ 1.5 □ 2 □ 2.5 □ 3 □ 3.5 □ 4

b

Interpretation

	Level 1 (Novice)	Level 2 (intermediate)	Level 3 (Proficient)	Level 4 (Advanced)
Interpretation	□ Does not interpret basic diagnostic test accurately	□ Limited ability to interpret basic diagnostic test accurately	□ Demonstrates ability to interpret most basic and some complex diagnostic testing accurately	□ Proficient at interpretation of basic and complex findings accurately
	□ Does not understand the limitations of diagnostic studies	□ Limited ability to understand the limitations of diagnostic studies	□ Demonstrates ability to understand the limitations of diagnostic studies	□ Proficient at recognizing limitations of diagnostic tests to identify deviations from common patterns
	□ Does not recognize or seek guidance in the interpretation of diagnostic studies when	□ Limited ability to recognize or seek guidance in the interpretation of diagnostic studies when needed	□ Demonstrates ability to mostly recognize or seek guidance in the interpretation of diagnostic studies when needed	□ Proficient at recognizing and seeking guidance in the interpretation of diagnostic studies when needed
	□ Does not recognize abnormal test results warranting immediate intervention	□ Limited ability to recognize abnormal test results warranting immediate intervention	□ Demonstrates ability to recognize most abnormal test results warranting immediate intervention	□ Promptly proficient to recognize abnormal findings warranting immediate intervention

Overall Score □ 1 □ 1.5 □ 2 □ 2.5 □ 3 □ 3.5 □ 4

c

Diagnostic

	Level 1 (Novice)	Level 2 (intermediate)	Level 3 (Proficient)	Level 4 (Advanced)
Diagnostic	□ Does not apply foundational knowledge to diagnostic testing and procedures to patient care	□ Limited understanding of basic necessity for acquisition	□ Demonstrates appropriate and correct diagnostic testing according to priority	□ Proficient in orders and correctly prioritizes appropriate diagnostic testing taking into account subtle and/or conflicting history and physical findings
	□ Does not determine the necessity for diagnostic studies	□ Limited understanding of the concept of pre-test probability and test performance characteristics	□ Demonstrates ability to use a diagnostic testing based on the pre-test probability of disease and likelihood of test results altering management	□ Proficient in anticipating and accounting for pitfalls and biases when ordering or performing diagnostic testing
	□ Does not perform bedside diagnostic studies	□ Limited basic competency for performing simple or non-complex bedside diagnostic studies	□ Demonstrates basic competency for performing most complex bedside diagnostic testing	□ Proficient an advanced competency for performing all complex bedside diagnostic testing
	□ Is not aware of the risks, benefits, and contradictions to a diagnostic study	□ Limited understanding of the risks or complications for common diagnostic testing	□ Demonstrates understanding of the implications of false positives and negatives of diagnostic testing prior to initiation of such studies	□ Proficient in using diagnostic testing for alternative and/or off-label use in crisis situations or unusual circumstances

Overall Score □ 1 □ 1.5 □ 2 □ 2.5 □ 3 □ 3.5 □ 4

d

Management

	Level 1 (Novice)	Level 2 (intermediate)	Level 3 (Proficient)	Level 4 (Advanced)
Management	□ Develops no differential diagnosis	□ Develops limited and narrow differential diagnosis	□ Develops an appropriate differential diagnosis by using most available medical information	□ Develops an accurate differential diagnosis by using all available medical information including those that pose the greatest potential for morbidity or mortality
	□ Does not prioritize critical initial stabilization/ actions in the resuscitation of a critically ill or injured patient	□ Limited prioritization of critical initial stabilization/ actions in the resuscitation of critically ill or injured patient	□ Demonstrates prioritization of critical initial stabilization/ actions in the resuscitation of most critically ill or injured patients	□ Proficient at prioritizing critical initial stabilization/ actions in the resuscitation of a critically ill or injured patient
	□ Does not recognize disease presentations that deviate from common patterns	□ Limited recognition of disease presentations that deviate from common patterns	□ Demonstrates ability to recognize most disease presentations that deviate from common patterns and avoiding premature assessments	□ Proficient at recognizing disease presentations that deviate from common patterns requiring complex decision-making
	□ Does not revise differential diagnoses in response to changes in a patient course over time	□ Limited revisions of the differential diagnosis in response to changes in a patient course over time	□ Demonstrates ability to use history and exam findings to define patient's central clinical problems	□ Proficient at revising the differential diagnoses in response to changes in a patient course over time
	□ Does not coordinate activities with other team members to optimize care	□ Limited coordination of activities with other team members to optimize care	□ Demonstrates coordination of activities with most other members to optimize care	□ Proficiently integrates all members of the team that each is able to maximize their skills to optimize care
	□ Does not handle multiple competing demands and synthesize all of the data in a timely and effective manner	□ Limited ability to handle multiple competing demands and synthesize all of the available data in a timely and effective manner	□ Demonstrates prioritization of multiple competing demands and synthesizes most of the available data in a timely and effective manner	□ Proficiently prioritizes multiple competing demands and synthesizes all of the available data in a timely and effective manner

Overall Score □ 1 □ 1.5 □ 2 □ 2.5 □ 3 □ 3.5 □ 4

Figure 1 Medical Judgment Metric: (a) History and Physical domain; (b) Diagnostic domain; (c) Interpretation domain; (d) Management domain. The Medical Judgment Metric (MJM) is a tool that can be applied to anyone interested in assessing medical decision-making capacity. The raters of the MJM should be well experienced physicians.

intervention: chest pain and abdominal pain. The participants were enrolled on an individual basis in each simulation scenario and each participant completed all four cases. Prior to starting these medical simulations, an unrelated practice simulation was conducted to ensure the comfort of each participant and answer questions about their ability to perform in the medical simulation scenarios. The participants were assessed in a strictly summative manner during all scenarios. The participants did not receive feedback, education, or a debriefing.

In each tested medical simulation, participants were informed that they were in a moderate-sized community hospital emergency department in the United States and were asked to care for the patient as best as possible by identifying the appropriate screening, testing, treatment, and diagnosis. Each participant had access to standard medical equipment and a nurse was available to use the equipment at the command of the participant; however, no training was provided to the participant on how to use the equipment. The full body simulation mannequin was capable of receiving any of the tests and maneuvers as directed by the participant. Verbal and/or visual feedback was provided for all inquiries as well as requested and available tests were provided in a scaled time-delay fashion. All laboratory values, radiographs, electrocardiograms, and representative ultrasound images were provided without interpretation beyond reference laboratory values. Participants were informed if certain tests, treatments, or specialists were not available. An agreed-upon safe word (MUSKRAT) was established prior to starting the medical simulations in the event that the participant felt ill (e.g., due to anxiety) or was injured (e.g., accidental needle stick) during the medical simulation. If any participant in the scenarios said the word “MUSKRAT,” the medical simulation team would immediately end the scenario and attend to the individual in need.

A team of four medical raters with backgrounds in emergency medicine, trauma, and medical simulation scored the participants using the MJM either live or using a video recording of the medical simulation, as all scenarios were recorded. A minimum of three raters was required for each medical simulation. The raters were blinded to the participant’s name and skill level as each participant was assigned a study identification number on scheduling. Each rater (either live or watching the video) was in a separate room and viewing area when they observed the simulation and completed the MJM and critical action evaluation. The raters handed their results to the research coordinator and she was the only one that reviewed and entered the final interpretation of the simulation. For each rater and medical simulation scenario, the numerical MJM total score (across the four domains of the MJM) was examined for interrater reliability using a two-way mixed average score intraclass correlation coefficient (ICC). Since the four expert raters were recruited to serve as our entire population of raters their effects were considered to be fixed.

Participants (ratees) were volunteers who served as representatives of their medical and expertise skillset and therefore their effects were considered to be random. The only other person available to assist the participant was a confederate nurse who only executed requested orders. Each participant did all four cases individually. The two-way mixed model ICC was chosen to account for variation of the rater scores in two ways, the fixed rater effect and the random ratee effect. Since the collective average score of the four raters for each ratee was used in subsequent analyses the average agreement of the score from the raters was selected rather than the individual agreement score. The final model for assessing the interrater reliability between the raters for each medical simulation scenario was the ICC (three-raters or four-raters) model. The ICC (three-raters or four-raters) average measure score then provide a metric for the interrater reliability of the raters by determining the proportion of the variation in average rater scores due to the different ratees. An ICC value approaching 1 indicates a higher reliability.

Relationships With Other Variables

In addition to completing the MJM for each medical simulation, raters were asked to complete a simulation-specific critical action checklist (reporting whether or not the participant had performed actions determined a priori to be critical to management of the patient) and to make a prediction of the patient’s condition at the conclusion of the simulation (loss of life, loss of function, or stabilized). The agreement of the raters in assessing each ratee’s simulation outcome was determined using Fleiss’s Kappa, an extension of Cohen’s Kappa. Only the three raters who completed evaluations for all patient outcomes were included in the determination of the Fleiss’ Kappa values. Since there are more than two raters assigning categorical ratings,

Table 1 Content Validity Index Average Percentage of Agreement on Medical Judgment Metric Items in the Clinical Domains Needing Team Agreement

	History and Physical	Diagnostic Evaluation	Interpretation of Diagnostic	Management	Other category
1. Lacks foundational knowledge to apply to diagnostic testing and procedures to patient care.	—	85%	5%	10%	—
2. Understands basic necessity for acquisition of diagnostic studies.	5%	85%	10%	—	—
3. Recognizes disease presentations that deviate from common patterns and require complex decision making as a result of the interpretation of diagnostic testing.	—	—	87.5%	12.5%	—
4. Orders and correctly prioritizes appropriate diagnostic testing.	7.5%	85%	2.5%	5%	—
5. Orders and correctly prioritizes appropriate diagnostic testing taking into account subtle and/or conflicting history and physical findings.	5%	81.25%	8.75%	5%	—
6. Inability to recognize patients' central clinical problem or develops very limited differential diagnosis.	6.25%	—	6.25%	62.5%	25%
7. Inconsistently recognizes patient's central clinical problem or develops limited differential diagnosis.	6.25%	—	6.25%	62.5%	25%
8. Uses all available medical information to develop an appropriate differential diagnosis.	6.25%	—	6.25%	62.5%	25%
9. Synthesizes all of the available data and narrows and prioritizes the list of weighted differential diagnoses to determine appropriate management including those that are the greatest potential for morbidity or mortality.	8.75%	2.5%	8.75%	80%	—
10. Does not understand the concept of pretest probability and test performance characteristics.	—	48.75%	15%	11.25%	—
11. Understands the concept of pretest probability and test performance characteristics and uses the diagnostic testing based on the pretest probability of disease and the likelihood of test results altering management.	—	48.75%	15%	11.25%	—
12. Fails to recognize patient's central clinical problem.	25%	—	—	50%	25%
13. Does not seek or is overly reliant on secondary data.	25%	—	—	50%	25%
14. Synthesizes data to generate a prioritized differential diagnosis and problem list.	10%	2.5%	5%	37.5%	25%
15. Ability to utilize diagnostic testing in alternative and/or off-label use in crisis situations or unusual circumstances.	—	59.75%	13.75%	27.5%	—
16. Inability to recognize disease presentations that deviate from common patterns.	—	—	6.25%	43.7%	50%
17. Consistently recognizes disease presentations that deviate from common patterns.	—	—	6.25%	43.7%	50%
18. Inability to recognize disease presentations that deviate from common patterns.	—	—	6.25%	43.7%	50%
19. Fails to recognize potentially life-threatening problems.	31.25%	6.25%	6.25%	56.25%	—
20. Inconsistently recognizes patient's central clinical problem or develops limited differential diagnosis.	31.25%	6.25%	6.25%	31.25%	25%
21. Unable to recognize critical/severely abnormal findings warranting immediate intervention.	35%	—	35%	30%	—
22. Inconsistently able to recognize critical/severely abnormal findings warranting immediate intervention.	35%	—	35%	30%	—
23. Consistently able to recognize critical/severely abnormal findings warranting immediate intervention.	35%	—	35%	30%	—

Table 2 Reliability of Summed Rater Evaluations by Scenario

Scenario	ICC Value (95% CI)	*P* Value
Biliary colic	0.986 (0.976-0.992)	<0.001
Pneumothorax	0.980 (0.965-0.989)	<0.001
STEMI	0.965 (0.941-0.980)	<0.001
Renal colic	0.987 (0.978-0.993)	<0.001

Note: ICC = intraclass correlation coefficient; CI = confidence interval; STEMI = ST elevation myocardial infarction. ICC values are two-way mixed average values. *P* value from *F* test of ICC value equal to 0.

Fleiss' Kappa with 95% confidence interval was calculated to determine the consistency of the raters in evaluating the expected patient outcome. The relationship between MJM scores and outcomes was examined through three-way ROC (receiver operating characteristic curve) analyses of the association between mean (over raters) MJM scores and the expected patient outcome selected by the majority of raters for each video.[19] For each ratee a majority outcome was identified among all rater outcome scores; there was never a split decision in the raters' outcomes. A partitioning strategy was used to guide the cutoff determination since the outcomes were trichotomous and one partition value may influence another partition value. The ordinality of the outcomes was considered by giving priority to loss of life using a one-versus-all grouped comparison and then by giving priority to stabilization. The volume under the surface generated by the specific cutoffs was determined with chance performance equal to 0.17 and perfect performance equal to 1.00. Finally, an overall agreement between the cutoffs and the outcome predicted directly by the rater for each video was determined using Cohen's Kappa. Cohen's Kappa was utilized since the agreement between two categorical measures was assessed for each simulation outcome. The two measures consisted of the majority outcome of the raters and the categorical outcome determined in reference to the cutoff of the mean rater score. For example, if the majority of the raters determined the outcome to be "loss of life" for that subject's patient and the average rater score was below the cutoff for "loss of life" then those measures were determined to be in agreement. Using this approach, significant Cohen's Kappa values would then be interpreted as evidence of concurrent validity, that the average rater scores numerically could be used as valid differentiators of simulation outcomes.

Response Process

Finally, a project team member recorded any questions/comments about clarity of the MJM judgment items, clinical domains, competency sections, grammar, syntax, organization, appropriateness, acceptability to the clinical raters, ease of use, and logical flow. All statistical analyses were completed using STATA version 14, SPSS version 23.0, and Microsoft Excel version 2007. Institutional review board approval was obtained for conducting human subjects research.

RESULTS

Content Validity

Content validity (CV) data from the waves of expert reviews were reviewed and items with a 100% agreement rate among raters were retained for the final metric. There were n = 20 statements, items, or domains that did not have 100% or even 50% complete agreement among reviewers (Table 1). Percentage agreements from the reviewers are detailed and voted on by the authors for group categorization as very relevant (75% to 100% relevance to domain), quite relevant (50% to 74% relevance to domain), somewhat relevant (25% to 49% relevance to domain), and not relevant (0% to 24% relevance to domain). Categories of relevance were then collapsed and identified as "relevant" versus "not relevant." The grey highlight in Table 1 indicates the final category the authors agreed upon for the final draft of the MJM.

Internal Structure

Figure 1a through d display the developed MJM domains that the raters used to evaluate in the medical simulations. The ICC values ranged from 0.965 to 0.987, supporting the consistency of the raters across the different participants for each scenario (Table 2).

Relationships With Other Variables

For the expected patient outcomes derived from the critical action checklists, Fleiss' Kappa values ranged from 0.791 to 0.906, demonstrating the consistency of the raters in evaluating the outcomes of each simulation (Table 3). Cohen's Kappa between

Table 3 Reliability of Rater Evaluations by Scenario Outcome

Scenario	Fleiss' Kappa Value (95% CI)	*P* Value
Biliary colic	0.804 (0.670-0.938)	<0.001
Pneumothorax	0.874 (0.727-1.000)	<0.001
STEMI	0.906 (0.751-1.000)	<0.001
Renal colic	0.791 (0.656-0.926)	<0.001

Note: CI = confidence interval; STEMI = ST elevation myocardial infarction. *P* value from test of Fleiss Kappa value equal to 0.

the outcome predicted by the average MJM score and the outcome expected by the majority of raters ranged from the 0.851 to 0.880, demonstrating the ability of the average rater score to consistently differentiate between the three outcome levels (stabilization, loss of function, or loss of life) for each medical simulation scenario (Table 4).

DISCUSSION

The pilot study reports the development of the MJM tool and its ability to quantify the quality of the medical decision capacity of an individual in four distinct time-critical medical simulations. Specifically, a strong correlation was found of the student's MJM score (below or above the MJM cutoff) with rater evaluations of "loss of life" or "stabilization;" perhaps the small sample size impaired the ability to discern "loss of function." Ultimately, the results of this pilot study suggest that quantifying medical decision making is possible with high interrater reliability and close associations with scenario outcome in a medical simulation environment.

High-fidelity medical simulation provides an operational environment in which the MJM can be used to measure medical decision making. Several recent studies have demonstrated the importance and effectiveness of medical simulation complemented by high-quality debriefing enabling the transfer of knowledge, skills, and attitudes to the clinical arena.[20-22] Thus, due to its high interrater reliability and evidence suggesting validation, the MJM is a promising evaluation tool as part of a medical education curriculum to measure and subsequently modify the quality of medical decision making. The long-term goal of research into medical decision-making tools is to potentially predict patient outcomes. At this time, there is still little evidence to suggest such tools can predict these outcomes.[23]

Several considerations warrant further discussion. As a pilot study, a limited number of participants and scenarios were assessed. All four raters were from the same institution with three having clinical backgrounds in emergency medicine and

Table 4 Concurrent Validation of Categorized Rater Scores to Simulation Outcomes

	Simulation Outcome				
Scenario/Outcome Average Score Cutoff	Loss of Life	Loss of Function	Stabilization	Cohen's Kappa	Volume Under Surface
Biliary colic				0.880	0.86
Loss of Life: Score <22.33	13				
Loss of Function: 22.33 ≤ Score ≤ 27.83		6	3		
Stabilization: Score > 27.83			18		
Pneumothorax				0.855	0.25
Loss of Life: Score < 24.83	12	3			
Loss of Function: 24.83 ≤ Score ≤ 26.00		1			
Stabilization: Score > 26.00			23		
STEMI				0.868	0.6
Loss of Life: Score < 24.5	19	1			
Loss of Function: 24.50 ≤ Score ≤ 28.16	2	2			
Stabilization: Score > 28.16			16		
Renal colic				0.851	0.25
Loss of Life: Score < 19.67	12				
Loss of Function: 19.67 ≤ Score ≤ 23.00		1			
Stabilization: Score > 23.00		3	24		

Note: STEMI = ST elevation myocardial infarction. Average score cutoff is determined by providing priority to loss of life followed by stabilization. The volume under the surface (VUS) is calculated at the fixed cutoff scores with chance performance equal to a VUS value of 0.17 and perfection equal to a VUS value of 1.00.

fellowship training in medical simulation. The fourth rater was a general surgeon. In addition, both the MJM ratings and the critical action checklists with expected patient outcomes were determined by the same raters while viewing each simulation, thus potentially contributing to scoring predictability bias of the outcomes and inflated MJM score associations. Also, the raters' clinical judgment may be inaccurate in an uncertain clinical environment. While this pilot study was not powered to undertake an analysis of the performance of subsets of the sample, the study design included the evaluation of common critical simulations undertaken by study participants with a wide range of skill and vocational discipline. Also, due to the pilot nature of the study, a ''loss of function'' outcome in the MJM was not robust (Cohen Kappa) enough to detect significant differences. Additionally, the classification of ''loss of life'' and ''stabilization'' using Cohen Kappa demonstrated the magnitude of agreement. Future studies should look at larger sample sizes of students, include a variety of raters from different specialties, and to include a wider variety of scenarios.

CONCLUSION

Additional research using the MJM is still needed to validate further the ability to measure medical decision making in medical simulation. Despite the strength of these pilot data, greater evidence for validity and interrater reliability will be required from the analysis of a larger sample of participants, medical cases, and a range of complex medical cases for greater generalizability. As of right now, there is still a need to support clinical faculty with additional evaluation metrics, outside of typical skill-based metrics, to assess training progress in medical students, and resident physicians. The MJM provides a promising tool to assess medical decision making.

ACKNOWLEDGMENTS

We wish to thank the following participating experts for helping the study team with content and face validity reviews of the MJM: Erik L. Antonsen, MD, PhD; Jeffery D. Kerby, MD, PhD; Lynn M. Hamrich, MD, FAAFP; Tiffany Marchand, MD; Janyce M. Sanford, MD, MBA, FACEP; and Jordan A. Weinberg, MD, FACS.

REFERENCES

1. Schwartz A. Medical decision making and medical education: challenges and opportunities. Perspect Biol Med. 2011;54(1):68–74.

2. Allen GD, Rubenfeld MG, Scheffer BK. Reliability of assessment of critical thinking. J Prof Nurs. 2004;20:15–22.

3. Lyss-Lerman P, Teherani A, Aagaard E, Loeser H, Cooke M, Harper GM. What training is needed in the fourth year of medical school? Views of residency program directors. Acad Med. 2009; 84(7):823–9.

4. Frischknecht AC, Boehler ML, Schwind CJ, et al. How prepared are your interns to take calls? Results of a multi-institutional study of simulated pages to prepare medical students for surgery internship. Am J Surg. 2014;208(2):307–15.

5. Chamberland M, Mamede S, St-Onge C, Setrakian J, Schmidt HG. Does medical students' diagnostic performance improve by observing examples of self-explanation provided by peers or experts? Adv Health Sci Educ Theory Pract. 2015;20(4): 981–93.

6. Azagury D, Liu RC, Morgan A, Spain DA. Small bowel obstruction: a practical step-by-step evidence-based approach to evaluation, decision making, and management. J Trauma Acute Care Surg. 2015;79(4):661–8. doi:10.1097/TA.0000000000000824.

7. Lighthall GK, Vazquez-Guillamet C. Understanding decision making in critical care. Clin Med Res. 2015;13(3–4):156–68. doi: 10.3121/cmr.2015.1289.

8. Croskerry P, Petrie D, Reilly J, Tait G. Deciding about fast and slow decisions. Acad Med. 2014;89(2):197–200.

9. Zendehrouh S. A new computational account of cognitive control over reinforcement-based decision making: modeling of a probabilistic learning task. Neural Netw. 2015;71:112–23.

10. Murray DJ, Freeman BD, Boulet JR, Woodhouse J, Fehr JJ, Klingensmith ME. Decision making in trauma settings: simulation to improve diagnostic skills. Simul Healthc. 2015;10(3):139–45.

11. Weber JM, Kopelman S, Messick DM. A conceptual review of decision making in social dilemmas: applying a logic of appropriateness. Pers Soc Psychol Rev. 2004;8(3):281–307.

12. Andolsek K, Padmore J, Hauer K, Holmboe E. Clinical Competency Committees. A Guidebook for Programs. Available at: https://www.acgme.org/acgmeweb/Portals/0/ACGMEClinicalCompetencyCommitteeGuidebook.pdf.

13. Association of American Colleges of Medicine. Situational judgment. Available at: https://www.aamc.org/admissions/admissionslifecycle/409100/situationaljudgmenttest.html.

14. Society for Medical Decision Making. Definition of medical decision making. Available at: http://smdm.org/hub/page/definition-of-medical-decision-making/about.

15. Mishra A, Catchpole K, McCulloch P. The Oxford NOTECHS System: reliability and validity of a tool for measuring teamwork behavior in the operating theatre. Qual Saf Health Care. 2009; 18(2):104–8.

16. Steinemann S, Berg B, DiTullio A, et al. Assessing teamwork in the trauma bay: introduction of a modified ''NOTECHS'' scale for trauma. Am J Surg. 2012;203(1):69–75.

17. Wynd CA, Schmidt B, Schaefer MA. Two quantitative approaches for estimating content validity. West J Nurs Res. 2003;25:508–18.

18. American Trauma Society. Trauma center levels explained. Available at: http://www.amtrauma.org/?page=TraumaLevels

19. Mossman D. Three-way ROCs. Med Decis Making. 1999; 19(1):78–89.

20. Wayne DB, Didwania A, Feinglass J, Fudala M, Barsuk J, McGaghie W. Simulation based education improves quality of care during cardiac arrest team responses at an academic teaching hospital: a case-control study. Chest. 2008;133:56–61.

21. McGaghie WC, Draycott TJ, Dunn WF, Lopez CM, Stefanidis D. Evaluating the impact of simulation on translational patient outcomes. Simul Healthc. 2011;6(suppl):S42–7.

22. Barsuk JH, McGaghie WC, Cohen ER, O'Leary KJ, Wayne DB. Simulation based mastery learning reduces complications during central venous catheter insertion in a medical intensive care unit. Crit Care Med. 2009;37:2697–701.

23. Shay LA, Lafata JE. Where is the evidence? A systematic review of shared decision making and patient outcomes. Med Decis Making. 2015;35(1):114–31. doi:10.1177/0272989X14551638.