HHS Public Access

Author manuscript

Acad Med. Author manuscript; available in PMC 2022 July 01.

Published in final edited form as:

Acad Med. 2021 July 01; 96(7): 1026–1035. doi:10.1097/ACM.0000000000004010.

Machine Scoring of Medical Students' Written Clinical Reasoning: Initial Validity Evidence

Anna T. Cianciolo, PhD [associate professor of medical education],
Southern Illinois University School of Medicine, Springfield, Illinois

Noelle LaVoie, PhD [president],
Parallel Consulting, Petaluma, California

James Parker, MA [senior research associate]
Parallel Consulting, Petaluma, California.

Abstract

Purpose—Developing medical students' clinical reasoning requires a structured longitudinal curriculum with frequent targeted assessment and feedback. Performance-based assessments, which have the strongest validity evidence, are currently not feasible for this purpose because they are time-intensive to score. This study explored the potential of using machine learning technologies to score one such assessment—the diagnostic justification essay.

Method—In May to September 2018, machine scoring algorithms were trained to score a sample of 700 diagnostic justification essays written by 414 third-year medical students from the Southern Illinois University School of Medicine classes of 2012–2017. The algorithms applied semantically based natural language processing metrics (e.g., coherence and readability) to assess essay quality on 4 criteria (differential diagnosis, recognition and use of findings, workup, and thought process); the scores for these criteria were summed to create overall scores. Three sources of validity evidence (response process, internal structure, and association with other variables) were examined.

Results—Machine scores correlated more strongly with faculty ratings than faculty ratings did with each other (machine: .28–.53, faculty: .13–.33) and were less case-specific. Machine scores and faculty ratings were similarly correlated with medical knowledge, clinical cognition, and prior diagnostic justification. Machine scores were more strongly associated with clinical communication than were faculty ratings (.43 vs .31).

Conclusions—Machine learning technologies may be useful for assessing medical students' long-form written clinical reasoning. Semantically based machine scoring may capture the communicative aspects of clinical reasoning better than faculty ratings, offering the potential for

Correspondence should be addressed to Anna T. Cianciolo, Department of Medical Education, Southern Illinois University School of Medicine, PO Box 19681, Springfield, IL 62794-9681; telephone: (217) 545-0123; acianciolo@siumed.edu; Twitter: @siusom.

Ethical approval: This study was deemed exempt from oversight by the Springfield Committee on Research Involving Human Subjects (Reference No. 016402, July 14, 2018).

Disclaimers: The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Supplemental digital content for this article is available at [LWW INSERT LINK].

automated assessment that generalizes to the workplace. These results underscore the potential of machine scoring to capture an aspect of clinical reasoning performance that is difficult to assess with traditional analytic scoring methods. Additional research should investigate machine scoring generalizability and examine its acceptability to trainees and educators.

Gathering information from patients, forming and testing diagnoses, and devising treatment plans lie at the heart of what medical educators expect trainees to learn to do. Accordingly, this clinical reasoning process[1] is included in national[2] and international[3] residency performance assessment frameworks and is a testing objective for senior undergraduate trainees seeking certification.[4] Moreover, supervisors use trainees' clinical reasoning performance to judge their competence in the workplace and entrust them with autonomous patient care.[5,6] Yet, both exam data[7] and supervisor reports[8] indicate alarming deficiencies in this core physician competency, which has become a target of urgent calls to improve patient safety.[9] Developing this key aspect of physician expertise requires a structured longitudinal curriculum with frequent targeted assessment and feedback.[8,10,11] However, the clinical reasoning assessment methods with the strongest validity evidence (i.e., performance-based assessments)[1] are currently not feasible to use for this purpose as they are time-intensive to score.

Among these assessment methods is the diagnostic justification essay, an exercise designed to elicit medical students' thinking as they reason through their findings and form a final diagnosis following a simulated patient encounter.[12] Research has shown that diagnostic justification essay scores reflect students' medical knowledge as well as their clinical pattern recognition and information gathering skills.[13] In addition, diagnostic justification essay scores can be used to identify curriculum strengths and weaknesses with respect to clinical reasoning development.[7,11] However, scoring the diagnostic justification essay is time-intensive for faculty, which has prevented the use of this assessment method for anything other than a high-stakes clinical performance examination at one institution.[7,12,13]

More efficient, analytic methods of written clinical reasoning assessment, which emphasize content and involve counting keywords or phrases (e.g., differential diagnoses and pertinent findings), have been adopted for large-scale and high-stakes examinations.[4,14,15] Yet, even these methods place sufficient burden on expert raters.[16–18] Moreover, these methods, by design, do not target the communicative aspects of clinical reasoning, such as organization, concision, and tailoring to audience and context,[19] which are important to demonstrating competent reasoning in the workplace.[5,6,20] A rigorous approach to scoring diagnostic justification essays that does not require faculty raters would allow for efficient reasoning assessment that not only captures trainees' medical knowledge and clinical skills but also their ability to convey their thought process to others.

Machine learning technologies have promise for achieving this aim. Machine learning involves computer algorithms that "learn" by identifying patterns in data sets and using these patterns to make inferences about new data (much as expert physicians do).[21] This form of artificial intelligence has already been applied to scoring essays in a variety of educational fields, with agreement between machines and humans that is comparable to agreement between human raters.[22–24] Machines have been used in lieu of humans to score essays for

high-stakes admissions exams, including the Scholastic Aptitude Test, the Graduate Record Examination, and the Graduate Management Admission Test,[24,25] as well as for lower-stakes educational purposes.[23,26] In medical education, machine learning has been used to evaluate aspects of clinical competence, including clinical note writing, exam interpretation, and medical knowledge.[16,27–29] Using machine learning to score diagnostic justification essays therefore represents a novel, yet plausible, approach to expanding clinical reasoning assessment options.

This study began an investigation into whether machine learning technologies could be used to score diagnostic justification essays in lieu of faculty raters. The goal was to produce initial validity evidence regarding human rating and machine score comparability by examining response process (i.e., how well machine scores and human ratings correspond), internal structure, and association with other academic performance data. Thus, the present study seeks to provide the foundation for a program of research into computer-assisted assessment of medical trainees' clinical reasoning.

Method

Context and overview

This study took place at Southern Illinois University School of Medicine (SIUSOM) from August 2017 to February 2019. SIUSOM is a community-based medical school that pioneered performance-based clinical skills examinations more than 30 years ago[30] and has been recognized internationally for excellence in student performance assessment.[31] SIUSOM's summative clinical competency exam (SCCX) has included the diagnostic justification essay since 2010.[12]

The investigation began in August 2017 with an evaluation of the suitability of archival SCCX diagnostic justification essay data for training machine scoring algorithms based on the quantity of essays available and the representativeness of the standardized patient cases on which the essays were based (Figure 1). Based on this review, 5 sets of essays each corresponding to a different patient case were selected for study. Faculty ratings of these essays were insufficiently reliable for training machine scoring algorithms, so trained research assistants were employed to re-rate the essays using a more rigorous process. Machine scoring training using these new ratings was validated by examining the association of computer-generated scores with the research assistants' ratings, the original faculty ratings, and archival academic performance data from the same students who wrote the essays selected for study.

This study was deemed exempt from oversight by the Springfield Committee for Research Involving Human Subjects (reference no. 016402).

Review of archival SCCX diagnostic justification data and patient case selection

The SCCX, administered to third-year medical students following core clerkship instruction, comprises 14 standardized patient cases in which students must gather clinical information and use their findings to achieve and justify a final diagnosis. All patient cases feature common diagnoses linked to curriculum objectives for classroom and clinical teaching.

Students must pass this committee-managed exam to graduate. For security purposes, patient cases on this high-stakes exam are reused for a maximum of 3 consecutive cohorts.

Following a simulated patient encounter (20 minutes maximum), the diagnostic justification essay exercise requires students to write a short (2,900-character or ~600-word limit) argumentative essay justifying their final diagnosis based on pertinent findings:

> Prompt: Explain your thought processes: i.e., how you used the data you collected from the patient to move from your initial list of possible diagnoses to your final diagnosis. Please include both positive and negative findings in this discussion. Please be thorough.

Students have up to 45 minutes to type their essay into locally developed exam software in addition to entering (for checklist evaluation purposes) their differential diagnosis, pertinent findings, requests for labs and imaging, final diagnosis, problem list, and initial management plans.

Two faculty (the dean for education and curriculum and the patient case author) independently rate each essay on 3 criteria (differential diagnosis, recognition and use of findings, and thought process) using 4 levels of quality (poor, borderline, competent, and excellent; see Supplemental Digital Appendix 1 at [LWW INSERT LINK]). To rate differential diagnosis, raters count the number of differentials in the essay that also appear on a list generated by the case author, assigning a higher rating for more complete coverage of the case author's list without penalty for extraneous differentials. Rating recognition and use of findings proceeds in an analogous, analytic fashion. By contrast, rating thought process is holistic, reflecting raters' general impression of argumentation quality. To create an overall essay rating, individual raters' criterion ratings are summed and then averaged across both raters. Faculty rater training takes place annually within a month before students complete the SCCX. It comprises an optional 1-hour session to explain the rubric and provide hands-on practice with 3 sample essays from a prior exam. Post-exam reports of the SCCX include rater agreement metrics without a minimum agreement standard, and the SCCX committee does not monitor, enforce, or remediate agreement in real-time.

Ten of the 14 SCCX patient cases include the diagnostic justification essay, which counts for 20% of the overall case score. The remaining 80% comprises checklist items entered into the exam software by students (differentials, findings, labs, final diagnosis, problem list, management plan) and by standardized patients (students' history and physical information gathering) and evaluated by exam staff. Although standardized patients also rate their satisfaction with several interpersonal aspects of the encounter (e.g., rapport, focus, closing), these ratings are not included in the overall case score but used to flag students in need of interpersonal skills remediation.

In September 2017, from a population of 33 unique SCCX diagnostic justification patient cases administered to the classes of 2012–2017, 5 cases (15%) were selected, creating a sample of 700 essays completed by 414 third-year medical students (Table 1). Selected cases had at least 2 cohorts of essay data, as a minimum sample of ~140 essays for each case was sought.[26] The cases featured chief complaints and diagnoses frequently seen in SIUSOM's

service area and case assessment objectives focused on clinical reasoning (rather than, e.g., cultural competence).

Research assistant re-ratings of selected cases

Faculty interrater reliabilities for the selected diagnostic justification cases, although higher than for unselected cases, were variable and often relatively low (Table 1). Because machine scoring reliability can only be as good as the human ratings on which it is based,[24] a refined rubric[32] was developed (see Supplemental Digital Appendix 2 at [LWW INSERT LINK]) and the essays were re-rated from February to April 2018. The refined rubric enabled holistic rating of 4 criteria (differential diagnosis, recognition and use of findings, workup, and thought process) using 4 levels of quality (poor, borderline, competent, and excellent). A holistic approach was chosen because it (1) maps better to semantically based natural language processing technologies than does analytic scoring, (2) better reflects rater expertise without sacrificing interrater agreement,[18] and (3) represents an informal gold standard among commercial machine scoring centers. The workup criterion was added to enhance the scoring system's potential generalizability to settings, including other medical schools, where workup is an assessment priority.[14,15] For each case, the rubric featured sample essays to illustrate the anchors for the rating levels for each criterion (see Supplemental Digital Appendix 3 at [LWW INSERT LINK]).

Ten paid research assistants—fourth-year medical students from the class of 2018 who achieved above-average scores on their own SCCX diagnostic justification essays—used the refined rubric to rescore all 700 essays in the sample. Pairs of students were assigned one case to rate after completing training with a medical education faculty member (A.T.C.), who has served on the SCCX committee for 9 years and coauthored multiple validation studies on the diagnostic justification essay.[7,13] Rater pair training began with a face-to-face meeting to collaboratively review the refined rubric and case blueprint and to score an initial set of 5 essays. Next, all 3 raters (both students and A.T.C.) scored a complete set of 10 essays individually then met to discuss their ratings, the rubric, and sources of disagreement. Sources of disagreement could include the rubric itself, and in one case, the rubric was revised. Once all 3 raters achieved a minimum level of agreement (correlation ≥ .75, exact agreement ≥ 65%, adjacent agreement = 100%)[33,34] on up to 2 consecutive sets of 10 essays, the students could proceed with independent rating.

Post-training, the principal investigators (A.T.C. and N.L.) evaluated agreement every 20 essays (i.e., on sets of 20 essays). If agreement fell below the minimum standard noted above, an investigator discussed sources of disagreement with the students, who then rescored the essays until satisfactory agreement was reached. Supplemental Digital Appendix 4 (at [LWW INSERT LINK]) summarizes the data used for machine scoring training generated by this rescoring effort. Student raters were more conservative than faculty, used a fuller range of the rating scale, and achieved very high levels of agreement. The correlation between average student ratings using the refined holistic rubric and average faculty ratings using the legacy rubric ranged from .47 to .69 (median = .53).

Machine scoring training

Overview.—The machine scoring system developed in this study incorporated several semantically based natural language processing metrics to assess the quality of students' SCCX diagnostic justification essays.[29] Using a background language model and the sample of essays reliably rated by humans described above (i.e., the research assistant ratings), the machine was trained to recognize the association between essays' linguistic and semantic properties (e.g., coherence, readability) and human ratings from May to September 2018. The resulting scoring algorithms combined these linguistic and semantic properties into linear regression models to analyze essays and generate scores.

Background language model.—The background language model comprised a very large set of training texts designed to give the scoring system a full context for evaluating essays.[35] Rather than a repository of diagnostic justification essays or lists of key medical terms, the background language model included nearly 200,000 paragraphs of text from (1) credible online biomedical and clinical knowledge sources to represent general medical knowledge and vocabulary and (2) archived and anonymized medical student service learning reflections to represent typical student writing. These source materials were chosen to include all language potentially encountered in the diagnostic justification essays, as this would enable the scoring system to measure similarity of meaning[36]—rather than similarity of words—among the essays.

Generating machine scores.—Machine scoring algorithms were developed independently for each assessment criterion for each patient case, totaling 4 algorithms per case (20 algorithms total). Machine scores ranged from 1 (= poor) to 4 (= excellent) for each assessment criterion (differential diagnosis, recognition and use of findings, workup, and thought process) and were summed to create an overall score. To make the best use of the limited data available and achieve a sufficient sample size for statistical analysis, the algorithms were trained using the entire set of 700 essays from the classes of 2012–2017 and then used to score the same essays. On the one hand, this approach sacrificed some degree of generalizability; scores produced this way will achieve higher, and potentially less reproducible, agreement with human ratings than would be expected if the algorithms were trained on one data set and applied to an independent test data set. On the other hand, this approach enabled the production of scores for pilot study; that is, using best case[18] machine scores to explore associations with human ratings and other academic performance data provides a reasonable test of whether this line of research warrants further pursuit.

Validation

The sources of validity evidence examined in this initial study were response process, internal structure, and association with other variables.[37] All analyses were performed from January to October 2019 and conducted using SPSS Statistics 26 (IBM Corporation, Armonk, New York). All correlations were Pearson correlations.

Response process.—To evaluate response process, the correlation of average research assistant (or student) and average faculty ratings with machine scores was examined. The first analysis evaluated the accuracy of machine scores relative to the student ratings on

which they were based. The second analysis explored whether machine scores were consistent with faculty ratings, based on the legacy rubric, of the same essays.

Internal structure.—To evaluate internal structure, Cronbach alpha was calculated for (1) assessment criterion scores within cases and (2) overall sums across all cases. The first analysis assessed the reliability of the scoring rubric, specifically the replicability of scores across applications of the same rubric (i.e., across the legacy or revised rubric).[38] The second analysis estimated the reliability of a hypothetical multicase diagnostic justification essay. Both internal consistency metrics were compared using machine scores and average faculty ratings.

Association with other variables.—A previous validation study of the diagnostic justification essay[13] offered a conceptual guide for examining the association of machine scores with archival academic performance data. Specifically, Cianciolo and colleagues[13] demonstrated that medical knowledge and clinical cognition (i.e., pattern recognition and information gathering skills) each predicted SCCX diagnostic justification essay scores. In the present study, composite scores for medical knowledge and clinical cognition as well as for prior diagnostic justification and clinical communication were entered into a stepwise regression model (probability of F to enter ≤ .05) to predict overall sums averaged across all cases. Two models were independently fit and compared: one for machine-scored overall sums and the other for faculty-rated overall sums.

The medical knowledge composite score comprised averaged first-take United States Medical Licensing Examination (USMLE) Step 1 scores and first-take USMLE Step 2 Clinical Knowledge scores obtained from archival student records. The clinical cognition composite score comprised the average checklist percent-correct scores for clinical data gathering and use (history and physical, findings, differential diagnosis, and patient satisfaction) from 3 sources: the 5 SCCX cases used in this study and 7 year 1 and year 2 summative standardized patient exam cases that included diagnostic justification essays (3 cases in year 1, 4 cases in year 2). Summative standardized patient exams in year 1 and year 2 were conducted similarly to the SCCX and administered by the same staff. The prior diagnostic justification composite score comprised average overall diagnostic justification essay scores from the same year 1 and year 2 summative standardized patient exam cases as those used for the clinical cognition composite. The clinical communication composite score comprised clerkship performance ratings pertinent to data gathering and use (history and physical, diagnosis and management, and interpersonal relationships with patients) averaged across 6 core rotations (family medicine, internal medicine, pediatrics, surgery, psychiatry, and obstetrics/gynecology). Preceptors assigned these ratings at the end of students' time on their service; thus, they represent supervising physicians' judgment of student competency based on oral case presentations, responses to directed questioning, and observed interactions with patients.[39] For all composite scores, data were obtained from the same students who wrote the essays selected for study (i.e., the classes of 2012–2017).

Results

Response process

Table 2 shows the correlation of machine scores with average student ratings and with average faculty ratings for each assessment criterion and the overall sum. Excepting the edema case, correlations between machine scores and student ratings were generally satisfactory (.56–.82, median = .69). Machine scores were less strongly correlated with faculty ratings (.25–.68, median = .53), but this association generally was equal to or higher than the correlation between faculty raters themselves (Table 1). Machine score correlations with both student and faculty ratings were higher for overall sums than for individual assessment criterion scores (overall: .49–.82, median = .65; assessment criteria: .25–.77, median = .58). Differences between machine-student and machine-faculty correlations were greatest for the differential diagnosis assessment criterion (differential diagnosis: .22–.40, median = .30; recognition and use of findings and thought process: .01–.22, median = .12).

Internal structure

Whereas the internal consistency reliability of faculty ratings was relatively consistent across the cases (.68–.85, median = .78, data not shown), machine score reliabilities varied as did the student ratings on which they were based. That is, internal consistency was high for the abdominal pain #1 (.81), chest pain (.81), and fever (.87) cases, but low for the edema and abdominal pain #2 cases (.46 and .56, respectively). The examination of assessment criterion correlations (data not shown) revealed a general pattern of higher correlations between recognition and use of findings and thought process than among the other assessment criteria, regardless of scoring method, suggesting that diagnostic justification scores were multidimensional.[38] This pattern of variable correlations among criteria was particularly pronounced for the edema and abdominal pain #2 cases.

Without exception, the correlation of overall sums among cases was higher for machine scores than for average faculty ratings (.28–.53 vs .13–.33, respectively), suggesting less case-specificity in machine scoring. Using data from the class of 2016 only (n = 69), the uncorrected internal consistency reliability of 3 of the machine-scored cases (abdominal pain #2, chest pain, and edema) was .65.

Association with other variables

Machine scores and faculty ratings were similarly correlated with medical knowledge, clinical cognition, and prior diagnostic justification (Table 3, bold values). By contrast, machine scores were more strongly correlated with clinical communication than were faculty ratings (.43 vs .31). Comparison of the final regression models (i.e., Model 3, Table 4) indicates that the composition of machine scores was different than that of faculty ratings; that is, medical knowledge, clinical communication, and prior diagnostic justification were significant predictors of machine scores (adjusted R^2 = .273), but medical knowledge, prior diagnostic justification, and clinical cognition were significant predictors of faculty ratings (adjusted R^2 = .262).

Discussion

This study demonstrates the potential of machine learning technologies to enable valid formative assessment of medical students' long-form written clinical reasoning. Even with a small sample of training data, machine scoring of diagnostic justification essays was (1) moderately correlated with and was as or more reliable than faculty ratings, showing evidence of response process validity, (2) generally consistent within and across cases, showing evidence of internal structure, and (3) associated with other academic performance data, showing evidence of relationships with other variables that are conceptually and empirically linked to clinical reasoning.[13] These findings are consistent with literature documenting the validity of machine scoring for essay-based standardized testing in higher education[24,25] and for assessment of physician clinical competence.[27]

Re-rating the diagnostic justification essays using a refined rubric was somewhat unusual for machine scoring projects and would not have been feasible outside of a funded research project. Ideally, faculty ratings would have been sufficiently reliable to offer a high ceiling for agreement with machine scores[24] and the legacy rubric would have been more amenable to training semantically based scoring algorithms. This would have allowed direct comparison of machine scores to faculty ratings, mirroring the validation efforts typically used to vet rater-based clinical performance assessments for high-stakes summative evaluations, such as examination of pass/fail decisions[17,18] and generalizability analysis.[40] However, poor rater agreement threatens valid clinical reasoning assessment, and the scoring system used here, which relies on holistic and highly reliable human ratings, improves the likelihood of generalization across assessment contexts.

The validity evidence reported here supports adoption of the machine scoring system described above—comprising case information (presentation, diagnosis, findings, differentials), the refined holistic scoring rubric, and the machine scoring algorithms—to assess diagnostic justification essays prospectively. This would enable efficient, if imperfect, formative assessment of long-form written clinical reasoning, likely for the first time at most institutions. However, a feasible machine scoring system must not require additional large-scale research and development efforts each time cases are modified or added to an assessment program. In the present study, the scoring algorithms were unaffected by minor case modifications across cohorts, and additional data collection would only improve their reliability and generalizability. Prospective use of this scoring system would allow for additional research, development, and validation as part of a longitudinal program of research. For instance, ongoing investigation is exploring whether the machine scores generated by this scoring system remain valid across case format (e.g., standardized patient encounter vs written patient summary) and learner level.

Although the initial results are promising, the data do not suggest that the machine scoring system described here is a perfect substitute for faculty raters. The scoring algorithms were as or more reliable than faculty raters, and machine scores were moderately correlated with faculty ratings, but they diverged from faculty ratings on the differential diagnosis assessment criterion. Machine scores also had slightly different patterns of association with academic performance data than did faculty ratings. However, these differences may be

conceptually meaningful and potentially worth preserving; the evidence is threefold that using a holistic scoring rubric to generate training data and analyzing these data with semantically based natural language processing metrics taught the machine scoring system to emphasize the communicative aspects of clinical reasoning, offering the potential for automated assessment that generalizes to the workplace.

First, machine scores diverged from faculty ratings most significantly on the differential diagnosis assessment criterion. Faculty rated differential diagnosis analytically; they compared students' answers to a list of items generated by the case author without acknowledging extraneous responses. The holistic rubric, by contrast, provided guidance that could be followed when students' responses went beyond the case blueprint; students earning a perfect score using the legacy rubric for being comprehensive would not achieve a perfect score using the holistic rubric if their response was not also concise and targeted.[18] In this way, the legacy rubric for differential diagnosis obviated communication skills that the holistic rubric held as a performance standard. Although faculty also rated recognition and use of findings analytically, machine scores may have diverged less on this criterion because evaluating what counted as a finding could be influenced by how the finding was used. For instance, faculty raters may have found vaguely stated patient information to be unconvincing evidence that a particular finding had been gathered or sufficiently understood to be used diagnostically. The stronger correlation between recognition and use of findings and thought process than between differential diagnosis and either of these criteria lends support to this notion. In sum, machine scores tended to correlate less strongly with faculty ratings that emphasized reasoning content over reasoning process.

Second, the correlations among cases were higher when machine scoring was used, indicating less case-specificity. This finding suggests that machine scores picked up on aspects of clinical reasoning that were somewhat distinct from case content, possibly communicative aspects such as organization and concision,[19,41] because the holistic rubric used to produce training data emphasized these features via descriptive anchors and illustrative sample essays. Yet, machine scores correlated as well with medical knowledge as did faculty ratings, suggesting that assessing clinical reasoning's communicative aspects did not come at the cost of assessing relevant knowledge.

Third, the regression models indicated that after medical knowledge, clinical communication (i.e., preceptors' ratings of students' clinical reasoning based on interactions) had the strongest association with machine scores. By contrast, clinical communication was not significantly associated with faculty ratings. This suggests that machine scores may better reflect the kind of clinical reasoning performance that preceptors see and that fosters workplace entrustment.[5,6]

Limitations

Additional data would improve the machine scoring system described here and provide further evidence that machine-scored diagnostic justification essays are valid measures of both content and communicative aspects of clinical reasoning. In this study, a small sample of training data (~140 essays per case), relative to what is typically needed to build

automated essay scoring systems (300 essays per case),[33] necessitated research design choices that may limit the generalizability of the findings. The strength and pattern of associations found in this study, therefore, reflect a best case[18] scenario that may not play out at other institutions or at the same institution at later times.

A traditional hold-out evaluation, frequently used in machine learning to evaluate the generalizability of automated scoring, was applied to probe this limitation. For each patient case, 60% of the essays (~84 essays per case) were randomly selected to train the scoring algorithms, and new scores were generated for the remaining 40% (~56 essays per case). Then, the correlation between machine scores and faculty ratings for this subset of essays was calculated. This process was conducted twice for each case, and the correlation between machine scores and faculty ratings was averaged. Across all cases, this analysis reduced the machine-human correlation by just 13%, demonstrating that even when using a more conservative approach, machine scores were more strongly correlated with faculty ratings than faculty ratings were correlated with each other. Nevertheless, extending the present investigation would require scoring more essays from more cases using the refined holistic rubric. This larger set of human scores would allow for the creation of larger independent data sets for training and testing the machine scoring algorithms. Ideally, these data would also feature greater spread, as the data used here had relatively few samples of high and low values; a more normal distribution would be more useful for training automated essay scoring systems.[33]

The results of this study also are limited to one form of clinical reasoning assessment—long-form (~600 words) written diagnostic justification essays.[1] Because the diagnostic justification essay is situated within a performance-based exam and requires students to articulate their thinking, it ranks among the stronger methods for assessing clinical reasoning.[1] It also has direct validity evidence warranting its use for high-stakes assessments and curriculum evaluation.[7,12,13] However, the scoring algorithms used in this study, which were trained on these long-form essays to apply a semantically based scoring approach, could not be applied effectively to evaluate the brief patient notes required by the USMLE Step 2 Clinical Skills exam. It remains an open question as to whether machine scoring of written clinical reasoning assessments akin to those on the USMLE Step 2 Clinical Skills exam[14] could achieve similar efficiency and validity.[16]

Next steps

Further research should seek additional sources of validity evidence to inform the larger question of whether or how to incorporate machine learning technologies into clinical reasoning curricula. Machine learning technologies are gaining increased attention in medicine[21] and clinical reasoning assessment,[16,29] but the acceptability and utility of machine scoring to trainees and educators remain important questions. The algorithms developed in the present study provide a solid foundation, but to meet efficiency expectations and merit widespread adoption, the machine scoring system for SCCX diagnostic justification essays should be robust across diverse applications, fitting into clinical reasoning curricula with different case content, case formats, and learner levels. Multi-institutional study not only would provide the opportunity to apply machine learning

technologies where approaches to teaching clinical reasoning and openness to advanced technologies differ; it also would allow the development of very large data sets for training and testing scoring algorithms that would generalize more broadly.

Conclusions

Using machine learning technologies for efficient yet rigorous assessment of medical trainee competence is of growing interest.[16,21,29] This single-institution pilot study offers an important proof-of-concept that machine scoring techniques already used for essay-based standardized testing in higher education[24,25] could be a valid method for assessing medical students' long-form written clinical reasoning. The results also suggest that using a holistic rubric and semantically based natural language processing metrics to score the diagnostic justification essay may capture the communicative aspects of clinical reasoning better than analytic faculty ratings do, thus, underscoring the potential of machine scoring to capture aspects of clinical reasoning performance that are difficult to assess with traditional analytic scoring methods. This assessment capability can support the deliberate practice[10] of a physician competency critical to patient safety[9] and workplace entrustment.[5,6] Additional multi-institutional validation research is needed to determine the generalizability of these findings and to refine the proof-of-concept into a robust machine scoring system that is acceptable to trainees and educators.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments:

The authors gratefully thank the Southern Illinois University School of Medicine Clinical Competency Exam Committee for sharing data and discussing results, especially Dr. Erica Nelson and Ms. Mary Aiello. They also would like to thank Dr. Debra L. Klamen for her consultation on this project. This article is dedicated to Dr. Reed G. Williams, whose vision for clinical performance assessment created the kind of community of inquiry and living laboratory where a study like this one could take place.

Funding/Support: The research reported in this article was supported by the National Institutes of Health, National Institute of General Medical Sciences Award No. 2R42GM108104–02A1.

Other disclosures: The grant that funded this study includes small business development and commercialization as a project goal. To prevent the intrusion of commercial bias, the study conduct and results were presented at multiple stages to curriculum committees and at academic conferences.

References

1. Daniel M, Rencic J, Durning S, et al. Clinical reasoning assessment methods: A scoping review. Acad Med. 2019;94(6):902–912. [PubMed: 30720527]
2. Accreditation Council for Graduate Medical Education. ACGME Common Program Requirements. http://www.acgme.org/Portals/0/PFAssets/ProgramRequirements/CPRs_2017-07-01.pdf. Accessed January 15, 2021.
3. Royal College of Physicians and Surgeons of Canada. CanMEDS: Better Standards, Better Physicians, Better Care. http://www.royalcollege.ca/rcsite/canmeds/canmeds-framework-e. Accessed January 15, 2021.
4. United States Medical Licensing Examination. Step 2 CS. https://www.usmle.org/step-2-cs/. Accessed January 15, 2021.

5. Kennedy TJT, Regehr G, Baker GR, Lingard LA. Point-of-care assessment of medical trainee competence for independent clinical work. Acad Med. 2008;83(10 Suppl):S89–S92. [PubMed: 18820510]
6. Landreville JM, Cheung WJ, Hamelin A, Frank JR. Entrustment checkpoint: Clinical supervisors' perceptions of the emergency department oral case presentation. Teach Learn Med. 2019;31(3):250–257. [PubMed: 30706726]
7. Williams RG, Klamen DL, Markwell SJ, Cianciolo AT, Colliver JA, Verhulst SJ. Variations in senior medical student diagnostic justification ability. Acad Med. 2014;89(5):790–798. [PubMed: 24667511]
8. Rencic J, Trowbridge RL, Fagan M, Sautzer K, Durning S. Clinical reasoning education at US medical schools: Results from a national survey of internal medicine clerkship directors. J Gen Intern Med. 2017;32:1242–1246. [PubMed: 28840454]
9. Connor DM, Durning SJ, Rencic JJ. Clinical reasoning as a core competency. Acad Med. 2020;95:1166–1171. [PubMed: 31577583]
10. Ericsson A. Acquisition and maintenance of medical expertise: A perspective from the expert-performance approach and deliberate practice. Acad Med. 2015;90(11):1471–1486. [PubMed: 26375267]
11. Klamen DL. Getting real: Embracing the conditions of the third-year clerkship and reimagining the curriculum to enable deliberate practice. Acad Med. 2015;90(10):1314–1317. [PubMed: 25901873]
12. Williams RG, Klamen DL. Examining the diagnostic justification abilities of fourth-year medical students. Acad Med. 2012;87(8):1008–1014. [PubMed: 22722355]
13. Cianciolo AT, Williams RG, Klamen DL, Roberts NK. Biomedical knowledge, clinical cognition and diagnostic justification: A structural equation model. Med Educ. 2013;47(3):309–316. [PubMed: 23398017]
14. Park YS, Lineberry M, Hyderi A, Bordage G, Riddle J, Yudkowsky R. Validity evidence for a patient note scoring rubric based on the new patient note format of the United States Medical Licensing Examination. Acad Med. 2013;88:1552–1557. [PubMed: 23969362]
15. Park YS, Hyderi A, Heine N, et al. Validity evidence and scoring guidelines for standardized patient encounters and patient notes from a multisite study of clinical performance examinations in seven medical schools. Acad Med. 2017;92(11 Suppl):S12–S20. [PubMed: 29065018]
16. Salt J, Harik P, Barone MA. Leveraging natural language processing: Toward computer-assisted scoring of patient notes in the USMLE Step 2 Clinical Skills exam. Acad Med. 2019;94(3):314–316. [PubMed: 30540567]
17. Yudkowsky R, Hyderi A, Holden J, et al. Can nonclinician raters be trained to assess clinical reasoning in postencounter patient notes? Acad Med. 2019;94(11 Suppl):S21–S27. [PubMed: 31663941]
18. Slater SC, Boulet JR. Predicting holistic ratings of written performance assessments from analytic scoring. Adv Health Sci Educ. 2001;6:103–119.
19. Chan MY. The oral case presentation: Toward a performance-based rhetorical model for teaching and learning. Med Educ Online. 2015;20:28565. doi:10.3402/meo.v20.28565 [PubMed: 26194482]
20. Lancaster I, Basson MD. What skills do clinical evaluators value most in oral case presentations? Teach Learn Med. 2019;31(2):129–135. [PubMed: 30551724]
21. Rowe M An introduction to machine learning for clinicians. Acad Med. 2019;94(10):1433–1436. [PubMed: 31094727]
22. Landauer TK, Laham RD, Foltz PW. Automated scoring and annotation of essays with the Intelligent Essay Assessor. Assess Educ. 2003;10(3):295–308.
23. Shermis MD, Burstein J, Higgins D, Zechner K. Automated essay scoring: Writing assessment and instruction. In: Peterson PL, Baker EL, McGaw B, eds. International Encyclopedia of Education, 3rd ed. Oxford, England: Elsevier; 2010:75–80.
24. Shermis MD. State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. Assess Writing. 2014;20:53–76.

25. Zhang M Contrasting automated and human scoring of essays. R &D Connections. 2013;21:1–11. https://www.ets.org/Media/Research/pdf/RD_Connections_21.pdf. Accessed January 15, 2021.
26. Streeter L, Bernstein J, Foltz P, DeLand D. Pearson's automated scoring of writing, speaking, and mathematics. San Antonio, TX: Pearson; 2011. https://images.pearsonassessments.com/images/tmrs/PearsonsAutomatedScoringofWritingSpeakingandMathematics.pdf. Accessed January 15, 2021.
27. Dias RD, Gupta A, Yule SJ. Using machine learning to assess physician competence: A systematic review. Acad Med. 2019;94(3):427–439. [PubMed: 30113364]
28. Chen Y, Wrenn J, Xu H, et al. Automated assessment of medical students' clinical exposures according to AAMC geriatric competencies. AMIA Annu Symp Proc. 2014;2014:375–384. [PubMed: 25954341]
29. Swygert K, Margolis M, King A, et al. Evaluation of an automated procedure for scoring patient notes as part of a clinical skills examination. Acad Med. 2003;78(10 Suppl):S75–S77. [PubMed: 14557102]
30. Williams RG, Barrows HS, Vu NV, et al. Direct, standardized assessment of clinical competence. Med Educ. 1987;21(6):482–489. [PubMed: 3696021]
31. Cianciolo AT, Klamen DL, Beason AM, Neumeister EL. ASPIRE-ing to excellence at SIUSOM. MedEdPublish. 2017. doi.org/10.15694/mep.2017.000082
32. LaVoie N, Parker J, Cianciolo AT. Why rater cognition matters: A process to improve interrater agreement. Paper presented at: 126th Annual American Psychological Association Convention; August 9–12, 2018; San Francisco, CA.
33. Hellman S, Rosenstein M, Gorman A, et al. Scaling up writing in the curriculum: Batch mode active learning for automated essay scoring. Paper presented at: 2019 Learning @ Scale conference; June 24–25, 2019; Chicago, IL.
34. Williamson DM, Xi X, Breyer FJ. A framework for evaluation and use of automated scoring. Educ Meas Issues Prac. 2012;31(1):2–13.
35. LaVoie N, Parker J, Legree PJ, Ardison S, Kilcullen RN. Using latent semantic analysis to score short answer constructed responses: Automated scoring of the Consequences Test. Educ Psychol Meas. 2019;80(2):399–414. [PubMed: 32158028]
36. Martin DI, Berry MW. Latent semantic indexing. In: Bates MJ, Maack MN, eds. Encyclopedia of Library and Information Sciences, 3rd ed. New York, NY: Taylor & Francis; 2010:3195–3204.
37. Downing SM. Validity: On the meaningful interpretation of assessment data. Med Educ. 2003;37:830–837. [PubMed: 14506816]
38. Schmitt N. Uses and abuses of coefficient alpha. Psychol Assess. 1996;8(4):350–353.
39. Cianciolo AT, Hingle S, Hudali T, Beason AM. Evaluating clerkship competency without exams. Clin Teach. 2019;16:1–5.
40. Yudkowsky R, Park YS, Hyderi A, Bordage G. Characteristics and implications of diagnostic justification scores based on the new patient note format of the USMLE Step 2 CS exam. Acad Med. 2015;90(11 Suppl):S56–S62. [PubMed: 26505103]
41. Bordage G, Lemieux M. Semantic structures and diagnostic thinking of experts and novices. Acad Med. 1991;66(9 Suppl):S70–S72. [PubMed: 1930535]

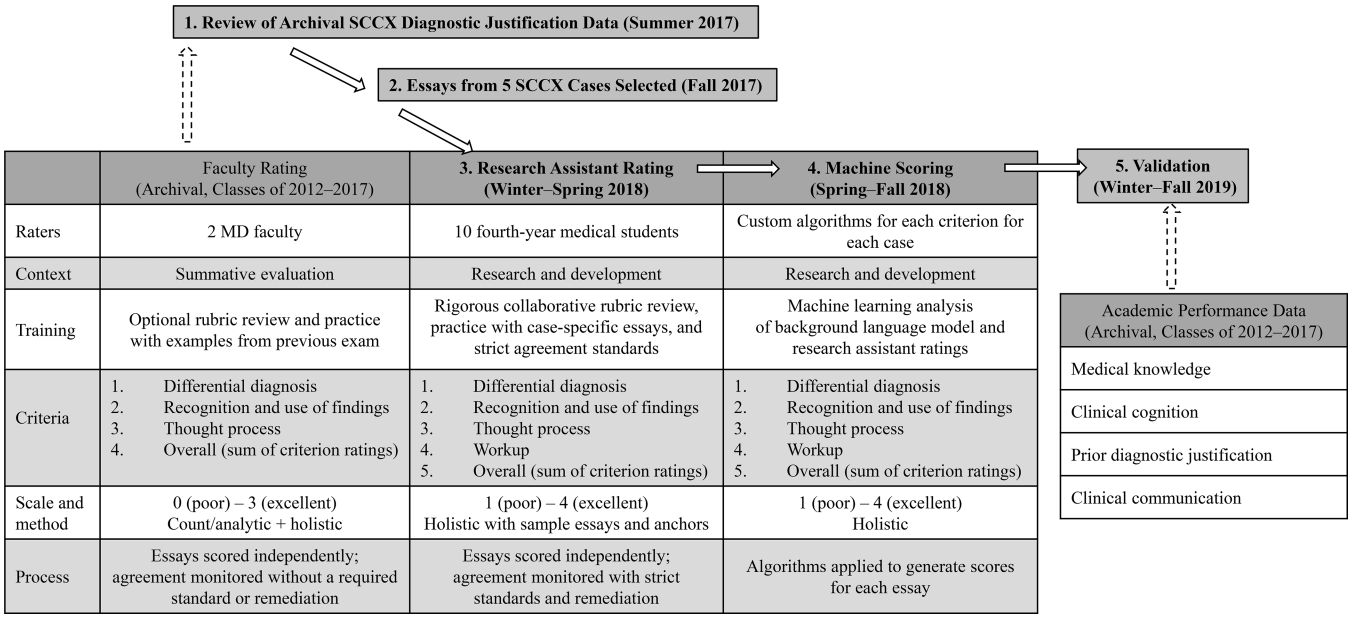

Figure 1.

Overview of the study timeline and comparison of faculty and research assistant ratings and machine scoring, from a study seeking to produce initial validity evidence regarding human rating and machine score comparability by examining response process, internal structure, and association with other academic performance data, Southern Illinois University School of Medicine, August 2017 to February 2019. Dashed arrows represent the contribution of archival data to research conduct. Solid arrows represent the flow from one study phase to the next. Abbreviation: SCCX, summative clinical competency exam.

Table 1

Summary of Content (Chief Complaint, Initial Presentation, and Diagnosis) and Administration (No. of Essays, Cohorts, and Faculty Interrater Reliability) Data of Selected Summative Clinical Competency Exam Diagnostic Justification Patient Cases[a]

Chief complaint	Initial presentation (diagnosis)[b]	No. of essays	Class of 2012 (n = 69)	Class of 2013 (n = 79)	Class of 2014 (n = 65)	Class of 2015 (n = 69)	Class of 2016 (n = 69)	Class of 2017 (n = 63)	Faculty interrater reliability (first cohort/second cohort)[c]
Abdominal pain #1	Amber Barlow, age 38, presents in the emergency department with abdominal pain (Acute cholecystitis)	144		X	X				.31/.36
Abdominal pain #2	Mike Mendenhall, age 55, presents in the emergency department complaining of abdominal pain (Acute pancreatitis)	138				X	X		.37/.61
Chest pain	Paul Knight, age 45, presents to the emergency department with chest pain (Acute coronary syndrome)	138				X	X		.44/.65
Edema	Hannah Bergman, age 28, presents to the clinic with left leg swelling (Deep vein thrombosis)	132					X	X	.82/.39
Fever	Ashley Carter, age 4 weeks, was brought to the emergency department by ambulance, after having a seizure at her doctor's office (HSV encephalitis)	148	X	X					.69/.67

Abbreviation: HSV, herpes simplex virus.

[a]From a study seeking to produce initial validity evidence regarding human rating and machine score comparability by examining response process, internal structure, and association with other academic performance data, Southern Illinois University School of Medicine, August 2017 to February 2019.

[b]The information presented here is from standardized patient cases authored by faculty and is therefore fictitious.

[c]Faculty interrater reliabilities were calculated using Pearson correlation. Reliability among faculty ratings was calculated separately for each cohort because the case author differed across cohorts.

Table 2

Pearson Correlation of Machine Scores With Human Ratings (by Medical Students and Faculty)[a]

	Abdominal pain #1		Abdominal pain #2		Chest pain		Edema		Fever	
Criterion	**Student**	**Faculty**	**Student**	**Faculty**	**Student**	**Faculty**	**Student**	**Faculty**	**Student**	**Faculty**
Differential diagnosis	.65	.25	.72	.38	.64	.34	.48	.26	.76	.50
Recognition and use of findings	.61	.55	.77	.55	.71	.57	.49	.39	.63	.59
Workup[b]	.63	NA	.63	NA	.56	NA	.51	NA	.70	NA
Thought process	.65	.50	.67	.61	.63	.50	.46	.47	.72	.58
Overall sum	.77	.62	.73	.55	.77	.56	.55	.49	.82	.68

Abbreviation: NA, not available.

[a]From a study seeking to produce initial validity evidence regarding human rating and machine score comparability by examining response process, internal structure, and association with other academic performance data, Southern Illinois University School of Medicine, August 2017 to February 2019.

[b]Prior to this study, faculty assigned ratings using a legacy analytic rubric (see Supplemental Digital Appendix 1 at [LWW INSERT LINK]). In 2018, trained student raters used a refined holistic rubric designed for the purpose of developing machine scoring algorithms (see Supplemental Digital Appendix 2 at [LWW INSERT LINK]) to re-rate the essays. The legacy rubric did not include workup as an assessment criterion.

Table 3

Pearson Intercorrelation of Composite Scores (Medical Knowledge, Clinical Cognition, Prior Diagnostic Justification, and Clinical Communication), Machine Overall Sum Scores, and Average Faculty Overall Sum Ratings[a,b]

Composite score	Medical knowledge[c]	Clinical cognition[d]	Prior diagnostic justification[e]	Clinical communication[f]	Overall sum (machine score)	Overall sum (average faculty rating)
Medical knowledge[c]	1.00					
Clinical cognition[d]	.36	1.00				
Prior diagnostic justification[e]	.34	.39	1.00			
Clinical communication[f]	.60	.49	.38	1.00		
Overall sum (machine score)	**.47**	**.27**	**.33**	.43	1.00	
Overall sum (average faculty ratings)	**.47**	**.28**	**.35**	.31	.63	1.00

Abbreviations: USMLE, United States Medical Licensing Examination; SCCX, summative clinical competency exam.

[a] From a study seeking to produce initial validity evidence regarding human rating and machine score comparability by examining response process, internal structure, and association with other academic performance data, Southern Illinois University School of Medicine, August 2017 to February 2019.

[b] All correlations are significant at the $P < .001$ level. Bold values indicate that machine scores and faculty ratings were similarly correlated with academic performance data.

[c] Medical knowledge = averaged first-take USMLE Step 1 scores and first-take USMLE Step 2 Clinical Knowledge scores.

[d] Clinical cognition = average checklist percent-correct scores for clinical data gathering and use from the 5 SCCX cases used in this study and 7 year 1 and year 2 summative standardized patient exam cases that included diagnostic justification essays.

[e] Prior diagnostic justification = average overall diagnostic justification essay scores from the same year 1 and year 2 summative standardized patient exam cases as used for clinical cognition.

[f] Clinical communication = clerkship performance ratings pertinent to data gathering and use averaged across 6 core rotations.

Table 4

Stepwise Regression of Overall Sums (Machine Scores and Average Faculty Ratings) on Medical Knowledge, Clinical Cognition, Prior Diagnostic Justification, and Clinical Communication[a]

Regression model	Standardized beta	t	*P* value	Adjusted R^2
Overall sum (machine scores)				
Model 1				
Medical knowledge[b]	.474	10.876	.000	.222
Model 2				
Medical knowledge[b]	.327	6.191	.000	.260
Clinical communication[c]	.246	4.656	.000	
Model 3				
Medical knowledge[b]	.304	5.741	.000	.273
Clinical communication[c]	.203	3.737	.000	
Prior diagnostic justification[d]	.135	2.876	.004	
Overall sum (average faculty ratings)				
Model 1				
Medical knowledge[b]	.465	10.631	.000	.215
Model 2				
Medical knowledge[b]	.388	8.495	.000	.253
Prior diagnostic justification[d]	.214	4.690	.000	
Model 3				
Medical knowledge[b]	.360	7.666	.000	.262
Prior diagnostic justification[d]	.175	3.645	.000	
Clinical cognition[e]	.119	2.466	.014	

Abbreviations: USMLE, United States Medical Licensing Examination; SCCX, summative clinical competency exam.

[a]From a study seeking to produce initial validity evidence regarding human rating and machine score comparability by examining response process, internal structure, and association with other academic performance data, Southern Illinois University School of Medicine, August 2017 to February 2019.

[b]Medical knowledge = averaged first-take USMLE Step 1 scores and first-take USMLE Step 2 Clinical Knowledge scores.

[c]Clinical communication = clerkship performance ratings pertinent to data gathering and use averaged across 6 core rotations.

[d]Prior diagnostic justification = average overall diagnostic justification essay scores from the same year 1 and year 2 summative standardized patient exam cases as used for clinical cognition.

[e]Clinical cognition = average checklist percent-correct scores for clinical data gathering and use from the 5 SCCX cases used in this study and 7 year 1 and year 2 summative standardized patient exam cases that included diagnostic justification essays.