Full Resolution Simulation for Evaluation of Critical Care Imaging Interpretation; Part 1: Fixed Effects Identify Influences of Exam, Specialty, Fatigue, and Training on Resident Performance

Chris L. Sistrom, MD, MPH, PhD, Roberta M. Slater, MD, Dhanashree A. Rajderkar, MD, Joseph R. Grajo, MD, John H. Rees, MD, Anthony A. Mancuso, MD

Rationale and Objectives: To describe our full-resolution simulation of critical care imaging coupled with posthoc grading of resident's interpretations and present results from the fixed effects terms in a comprehensive mixed regression model of the resulting scores.

Materials and Methods: The system delivered full resolution DICOM studies via clinical-grade viewing software integrated with a custom built web-based workflow and reporting system. The interpretations submitted by participating residents from 47 different programs were graded (scores of 0–10) on a case by case basis by a cadre of faculty members from our department. The data from 5 yearly (2014–2018) cycles consisting of 992 separate 65 case, 8 hour simulation sessions were collated from the transaction records. We used a mixed (hierarchical) statistical model with nine fixed and four random independent variables. In this paper, we present the results from the nine fixed effects.

Results: There were 19,916/63,839 (27.0%, CI 26.7%–27.4%) scores in the 0–2 range (i.e., clinically significant miss). Neurological cases were more difficult with adjusted scores 2.3 (CI 1.9–3.2) lower than body/musculoskeletal cases. There was a small (0.3, CI 0.20–0.38 points) but highly significant ($p<0.0001$) decrease in score for the final 13/65 cases (fifth quintile) as evidence of fatigue during the last hour of an 8 hour shift. By comparing adjusted scores from mid-R1 (quarter 3) to late-R3 (quarter 12) we estimate the training effect as an increase of 2.2 (CI 1.90–2.50) points.

Conclusion: Full resolution simulation based evaluation of critical care radiology interpretation is being conducted remotely and efficiently at large scale. Analysis of the resulting scores yields multiple insights into the interpretative process.

Key Words: Radiology education; Critical care imaging; Simulation; Competency milestones; Entrustable professional activities; Certification.

INTRODUCTION

The Accreditation Council for Graduate Medical Education (ACGME)—in its "envelope of expectations"— indicates that a trainee in Diagnostic Radiology should reach the level of "competent" late in the second to beginning of the third year of training (1). At a level of "competent," the resident is theoretically prepared to engage in the entrustable professional activity (EPA) of distance (out of institution) supervision of image interpretation with preliminary reporting and consultation (1–5). This generally occurs in the setting of after-hours coverage with faculty directly available for backup consultation from home. Such an independent experience is essential and required, as called for in section VI.A.2 of the ACGME Common Program Requirements document for radiology residency programs (1).

Anticipating the importance of this transition point in training, the ACGME has asked for objective documentation, within the Milestones Project (1), that residents are prepared for this EPA, which provides for advancement in graded authority and responsibility. Based on current milestone projections, such an objective assessment methodology of this preparedness was to be in place by 2018. In order to fulfill this mandate, our group developed a Critical Care Radiology Simulation (CCRS) which involved interpreting full sets of

Acad Radiol 2020; ■:1–10

From the Department of Radiology, University of Florida Health Center, P.O. Box 100374, JHMHC, 1600 SW Archer Road, Gainesville, FL 32610-0374. Received August 31, 2019; revised November 1, 2019; accepted November 1, 2019. **Address correspondence to:** C.L.S. e-mail: sistrc@radiology.ufl.edu

https://doi.org/10.1016/j.acra.2019.11.023

DICOM images in the domain of critical care radiology. Sixty-five cases across all radiology subspecialties, with a plausible range of difficulty including normal studies constitute this 8 hour simulation experience.

Simulation has a historic yet evolving role in graduate medical education. Introduced as a means for teaching manual dexterity and hand-eye coordination for interventional procedures, simulation has become a tool for delivering educational content, facilitating systems-based practice, and training for standardized examinations (6,7). Simulation-based training in radiology is in its infancy but offers immense potential for objective evaluation of interpretative competence (8,9). Others have begun to realize the potential of simulation for assessing milestones and executing the EPA of managing acute radiologic emergencies (10,11). In order to test our residents for readiness of graded responsibility and authority, we developed the CCRS to evaluate competency in managing critical care radiology cases in a high fidelity environment.

The CCRS has now been delivered 8 times over the last 8 years. The results of five most recent iterations (2014−2018) of the CCRS, after a testing and maturation period of 3 years, are the subject of this two-part report. This paper (Part 1) will lay out our analytic framework in a mixed model and quantify the fixed effects of 9 factors (e.g., exam type, clinical scenario, level of training, and etc.) on resident performance. Part 2 will report on four identifying variables (case, resident, program, and grader) estimated jointly as uncorrelated random effects in the same mixed model.

MATERIALS AND METHODS

Ethics

This research was approved by our Institutional Review Board and certified as being compliant with the Health Insurance Portability and Accountability Act). Additionally, the identities of residents being evaluated, their respective residency programs, and the radiologist graders were blinded (with an 'honest broker' mechanism) for analysis and will be presented only as unlabeled scatterplot points for Part 2.

Case Creation and Curation

The case material used for testing residents consisted of original full DICOM image sets obtained on patients being seen in our emergency department and/or during a hospital admission. Gender specific patient names were assigned from a random name generator. Unique random numbers were used to represent medical record number and accession number for each patient and exam. Patient age at the time of original presentation was recorded and used to calculate an offset from the simulation date to form a dynamic date of birth.

In addition to the full DICOM image sets and demographics, we constructed a clinical scenario from the electronic medical record using the originally recorded "Indications for Exam" free text, ICD10 codes associated in the order composer, and clinician notes. These elements were abstracted into separate "Exam Indications" and "Additional History" properties for display in the case presenter interface. For each case we assigned an "acuity" level by investigator consensus in keeping with the ACR Actionable Findings Initiative (13). Finally, a structured answer key was constructed as a guide for scoring a narrative report about the case on a 0−10 scale. This consisted of a list of observational and/or diagnostic findings, each having a full or partial credit point value. These values were weighted by relative clinical importance such that they summed to 10 in aggregate.

Test Case Administration

The American College of Radiology (ACR) maintains a web-based, multi-purpose and customizable educational software stack called Radiology Content Management System (RCMS, a.k.a. Cortex). Full DICOM image sets are de-identified and uploaded by contributors with the ACR-TRIAD system (12). Using the basic case material as substrate, RCMS provides a flexible and dynamic authoring system for didactic instruction about imaging interpretation and reporting with built in self-testing as well as support for summative evaluation. In support of our project, ACR customized, according to our unique specifications, a function set of the RCMS enabling us to remotely administer a full resolution radiology simulation, collect interpretative responses, record scores assigned by graders, and download raw instance data. We used those results to make customized reports to individual residents and their program directors. Figure 1 depicts the simulation work queue, case detail, and response capture form. Each resident was given 8 hours to complete 65 cases assigned to them in the simulation work queue. The order of cases was randomized for each resident session. This allowed us to test the effect of cognitive fatigue on performance (score) in our statistical analysis.

For each of the 65 cases in a simulation session, once an interpretation had been entered and submitted, the case moved over to a completed queue and remained visible. Residents could bring any of the completed cases back, review the images, and submit revised or additional findings and/or acuity assertions. These were recorded, marked as "Addendum" and shown to graders along with the first submission.

Grading of Responses

Simulation case responses were collected as they were submitted and placed into a web-based grading queue for subsequent scoring. Graders score responses within their general subspecialty area (i.e., body, neuro, musculoskeletal, pediatric) by selecting from a worklist to score them. The list was organized so that graders work through all unscored submissions for each case. The scoring form showed responses entered by the trainee alongside structured key findings for the case including true/false assertions and weights for scoring. Figure 2 shows examples of the grading workflow.

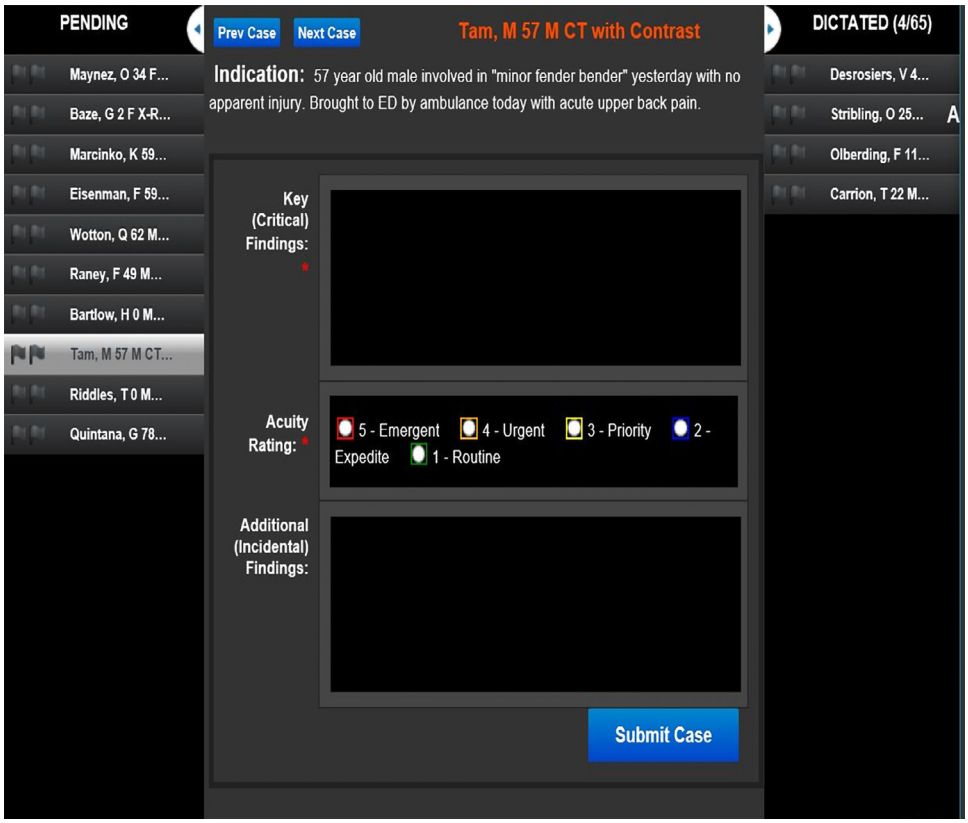

Figure 1. Screen shot of worklist, case presentation, and interpretation submission screen as seen by residents taking the simulation. Note that the patient names were randomly generated.

Measurement Paradigm

Our framework starts with a simulation case consisting of a specific clinical scenario and a diagnostic imaging examination, both involving the same real patient. We stipulate a "gold standard" set of findings accurately described and characterized in an "ideal" interpretative report agreed upon by consensus of expert, sub-specialty trained radiologists. The clinical scenario and findings are typical in that they did occur in actual clinical experience and represent cases that practicing radiologists will see and would be expected to correctly interpret.

An experimental unit is instantiated when a simulation subject is given the clinical scenario, asked to review the full set of images on a diagnostic workstation, and create an interpretation by dictating and/or typing. The outcome is created when a grader reviews the subject's interpretation, refers to the findings key, and records a 0−10 score. The score for a resident's report quantifies the grader's assessment of the clinical relevance and correctness of that response.

Graders were trained to consider their scores in the context of a clinical "read-out" of the simulation cases, conducted by themselves, immediately following an unsupervised "work shift" by the resident trainee. The scores are considered to be absolute such that the response is scored at face value regardless of the training or experience of the subject being evaluated. In real life situations, a score <3 would often prompt a call to the ordering service to remediate a clinically significant error; note that this error could be either a miss or an overcall. Score >7 would mean that the trainee's response (i.e., their preliminary reading) was essentially correct and subsequent report completion and finalizing workflow could proceed routinely. Scores between 3 and 7 represent a somewhat broader middle ground. Here, the interpretation in question requires editing to correct problematic assertions and/or rectify substantive omissions prior to being finalized and distributed. In our hypothetical situation, the supervising radiologist would seek to remediate the trainee in question.

Statistical Analysis

We consider the scores as discrete measurements taken from an underlying continuous distribution. Each score represents the interplay between three elements: case ($n = 321$), resident ($n = 773$), and grader ($n = 16$). At the time they participated, each subject was enrolled in one of 47 residency programs, representing a fourth. We estimated these traits as four independent random effects in a multi-level (hierarchical) model with fixed effects including factors such as exam modality, sub-specialty, care setting, clinical acuity, resident training level, order of the case within the simulation session, patient gender, and age group. The model is written as:

$$RS_{ijk} = \delta X_i + \theta Y_j + \lambda Z_k + B_{00} + B_{0c} + B_{0r} + B_{0p} + B_{0g} + \varepsilon_{ijk} \quad (1)$$

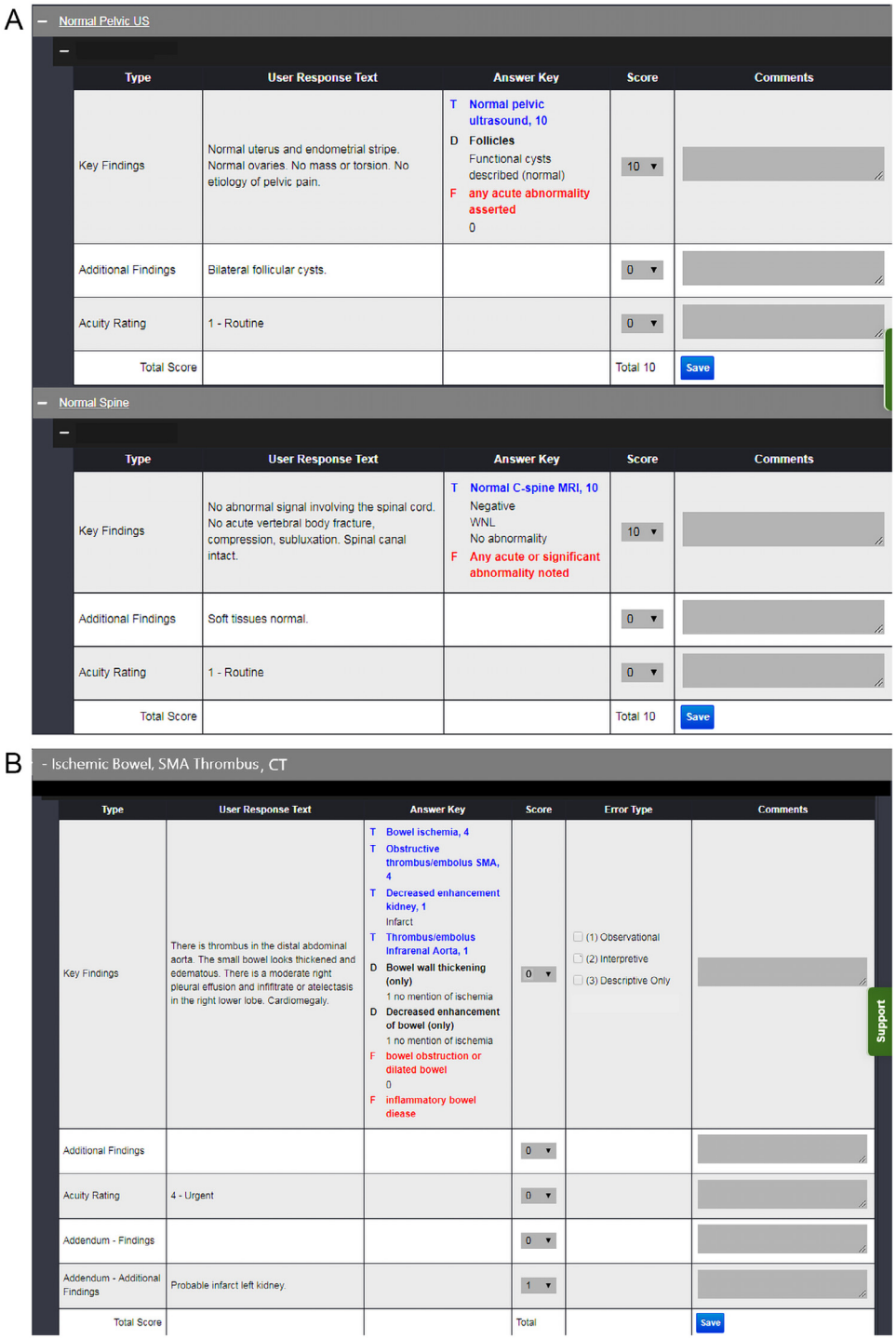

Figure 2. Screen shots of worklist, resident responses, answer key, and score entry form as seen by the graders. Examples of two relatively simple cases (A) and a rather more complex one (B) are shown.

Table 1 lists and defines all the symbols, including the index subscripts (i, j, k) fixed (X, Y, Z) and random (c, r, p, g) effects. We used PROC MIXED in SAS (Version 9.4, Cary, NC) to estimate the model and Tableau (Version 2019.1, Seattle, WA) for tabulation and graphs. It is important to note that including the four random effects (case, resident, program, and grader) in the model ensures that the calculation of coefficients, adjusted mean scores, and p-values for the fixed effects would be unbiased. That is because when random effects are included in a mixed model specification for PROC MIXED, the repeated measurement across them is fully accounted for which insures accurate (i.e., not inflated) calculation of all the standard errors.

TABLE 1. Key to Terms in Hierarchical Statistical Model (Equation 1)

Symbol	Meaning	Details
RS_{ijk}	response score	Assigned by grader to response for each of 65 cases submitted by a resident during a simulation session.
I	the ith item	$n = 63,839$
δX_i	Item level fixed effect coefficients & variables	case order during simulation session (quintiles: 1–5) had addendum (yes, no)
J	the jth case	$n = 321$
θY_j	Case level fixed effect coefficients & variables	specialty (neuro, mks, body) modality (XR, CR, MR, NUC, US) patient age group (adult, pediatric) patient gender (male, female) patient setting (emergency, inpatient) acuity (routine/negative, priority, urgent, emergent)
K	the kth simulation session	$n = 992$
λZ_k	Session level fixed coefficients and variables	resident training (quarters: 2–16)
ε_{ijk}	error term (residual)	centered on 0, normal distribution
B_{00}	Intercept	grand mean of all scores
B_{0C}	Case random intercept	score offsets ($n = 321$)
B_{0R}	Resident random intercept	score offsets ($n = 773$)
B_{0P}	Program random intercept	score offsets ($n = 47$)
B_{0G}	Grader random intercept	score offsets ($n = 16$)

RESULTS

The program (and our analytic sample) grew steadily over the five years of study (2014–2018). Table 2 lists the unadjusted mean scores (with 95% confidence interval) for each fixed effect variable and cycle year. Figure 3A is a stacked histogram, colored by year, with score frequency and percentage. The scores have a bimodal distribution and in Figure 3B, we have binned them into three groups. These reflect the clinical logic we asked graders to use as described previously. Overall, 46.4% (CI 46.0–46.8%) of the resident interpretations were correct (score=8–10) while 27.0% (CI 26.7–27.4%) were in the significant miss range (score = 0–2).

Table 3 lists the jointly estimated omnibus (Type 3) tests of significance for the nine fixed effect variables in the mixed model. Six of nine variables had significant influence on the score. These included case order within the simulation session ($p<0.0001$), whether or not the resident made an addendum to their interpretation after initial submission ($p=0.0018$), the specialty (neurologic, body, musculoskeletal) of the case ($p<0.0001$), our consensus case acuity ($p<0.0001$), and

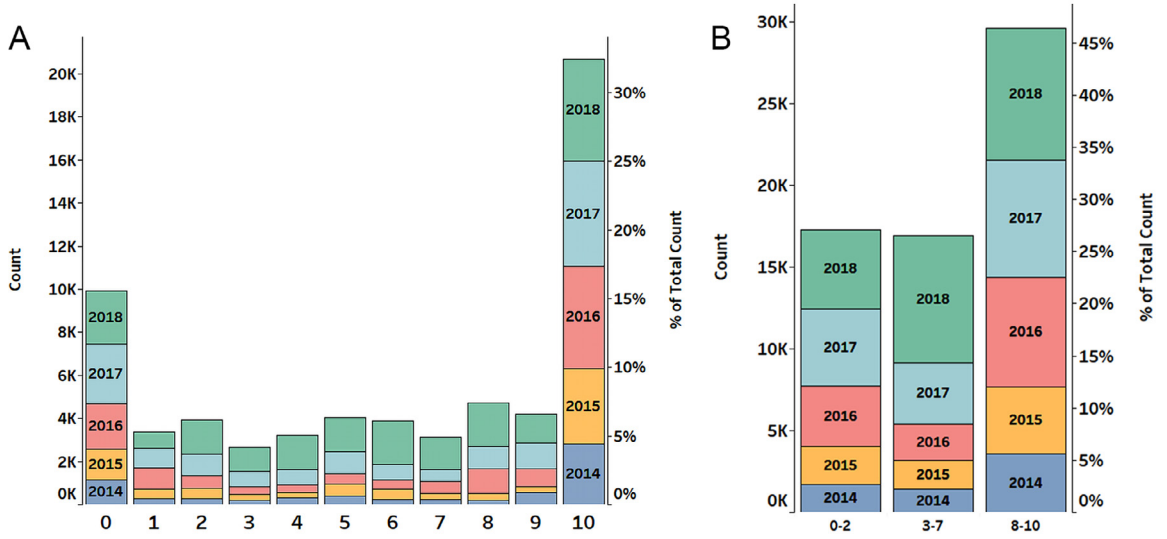

Figure 3. Histogram of individual (A) and grouped (B) scores. Cumulative counts (left vertical axis) and percentages (right vertical axis) over five simulation cycles from 2014 through 2018.

TABLE 2. Unadjusted Score Statistics and Counts of Sessions, Cases, Residents, Programs, and Graders, Stratified by Nine Fixed Effects Variables and Year of Study

		Scores		Number Of Unique				
Variable	Level	Mean (95% CI)	*n*	Sessions	Cases	Residents	Programs	Graders
Case Order (quintile)	1	6.13 (6.07–6.20)	12,818	992	291	773	47	16
	2	6.09 (6.02–6.15)	12,872	992	304	773	47	16
	3	6.19 (6.13–6.26)	12,869	992	311	773	47	16
	4	5.83 (5.76–5.90)	12,837	991	310	773	47	16
	5	5.52 (5.45–5.59)	12,443	985	311	771	47	16
Had Addendum	no	5.95 (5.92–5.98)	61,754	992	321	773	47	16
	yes	6.04 (5.89–6.20)	2,085	681	314	573	45	15
Case Specialty	neuro	4.56 (4.50–4.63)	13,165	992	69	773	47	10
	body	6.29 (6.25–6.33)	39,847	992	196	773	47	15
	msk	6.41 (6.34–6.49)	10,827	992	56	773	47	10
Exam Modality	xr	6.11 (6.04–6.17)	17,803	992	92	773	47	14
	ct	5.87 (5.83–5.91)	30,310	992	154	773	47	16
	mr	5.37 (5.27–5.46)	6,081	988	31	770	47	10
	nu	7.47 (7.26–7.69)	1,301	986	6	770	47	9
	us	6.15 (6.06–6.24)	8,344	992	38	773	47	14
Patient Age Group	adult	6.07 (6.04–6.11)	48,292	992	241	773	47	16
	ped	5.59 (5.53–5.66)	15,547	992	80	773	47	16
Gender	Male	6.05 (6.01–6.09)	31,891	992	166	773	47	16
	Female	5.86 (5.82–5.90)	31,948	992	155	773	47	16
Patient Setting	Emergency	5.98 (5.95–6.01)	58,075	992	291	773	47	16
	Inpatient	5.74 (5.63–5.84)	5,764	992	30	773	47	14
Case Acuity	1-routine	7.33 (7.23–7.42)	8,148	992	43	773	47	14
	2-priority	6.12 (6.06–6.19)	14,153	670	83	538	34	15
	3-urgent	5.70 (5.66–5.74)	32,214	992	150	773	47	16
	4-emergent	5.39 (5.32–5.46)	9,324	992	45	773	47	16
Resident Training (quarters)	3	5.49 (5.36–5.61)	3,929	61	321	61	3	10
	4	5.66 (5.62–5.70)	40,472	631	321	587	45	16
	5	6.43 (6.04–6.82)	388	6	255	6	4	3
	7	6.91 (6.77–7.05)	2,466	38	258	38	2	8
	8	6.55 (6.49–6.62)	11,590	179	321	178	18	16
	9	6.38 (6.16–6.60)	1,224	19	258	19	4	4
	11	6.58 (6.27–6.90)	390	6	65	6	1	4
	12	6.97 (6.80–7.14)	1,820	28	193	28	3	11
	15	6.72 (6.33–7.10)	260	4	65	4	1	4
	16	7.15 (6.96–7.34)	1,300	20	258	20	2	11
Year	2014	6.39 (6.29–6.48)	6,674	103	65	103	9	4
	2015	6.17 (6.09–6.26)	8,202	127	65	127	16	2
	2016	6.15 (6.08–6.22)	12,671	197	65	197	25	11
	2017	5.73 (5.67–5.80)	15,623	243	65	243	29	10
	2018	5.78 (5.73–5.83)	20,669	322	65	322	41	5
Grand	Total	**5.955 (5.95–5.96)**	**63,839**	**992**	**321**	**773**	**47**	**16**

msk, musculoskeletal, ped, pediatric.

resident length of training ($p<0.0001$) as measured by how many quarters had elapsed between starting residency and when they took the simulation. Variables with no significant effect on score were exam modality and care setting as well as patient age group and gender. Figure 4 is a Forest plot of adjusted mean scores (with 95% confidence intervals) for each level of the fixed effects variables. Figure 5 shows the model coefficients (with 95% confidence intervals) from the same fixed effect variables and levels as in Figure 2. These coefficients have the same scale as the analyzed scores. Thus, they can be interpreted as "score offsets" for a given variable level compared to reference offset of zero (uppermost levels with no bar on Figure 5).

This paper is the first of two, and, in Part 2, we will fully report results for each of the random effect variables, indicating how they contribute to score variation. The mixed model estimates a score offset for each case, resident, program, and grader. This allows us to reliably quantify case difficulty, resident competence, program effectiveness, and grader influence.

TABLE 3. Fixed Effects Joint Significance (Type 3)

Variable name	Levels	F Value	p Value
Case order (quintile)	5	15.34	<0.0001**
Had addenda Yes, no	2	9.72	0.0018**
Specialty Neuro, msk, body	3	22.11	<0.0001**
Exam modality xr, ct, mr, nuc, us	5	2.33	0.0532
Patient age group Adult, pediatric	2	0.06	0.8069
Patient gender Male, female	2	0.04	0.844
Patient setting Emergency, inpatient	2	0.57	0.4497
Case acuity (consensus definition)	4	7.25	<0.0001**
Resident training 2–17 quarters	10	76.04	<0.0001**

** Statistically significant.

DISCUSSION

Case Order and Fatigue

There was a decrease in reader performance throughout an eight hour shift reaching significance ($p<0.037$) around 6 hours (4th quintile). By the last hour (5th quintile) scores were 0.3 (95% CI 0.20−0.38) points lower ($p<0.0001$) than at start of shift (1st quintile). We attribute this to cognitive fatigue and believe that there is a temporal dose effect, reaching significance between 6 and 7 hours. This has been studied by others who suggest a tipping point for degraded interpretive accuracy around 10 hours (13−15). Our results are consistent with theirs perhaps with more power to detect small incremental declines earlier.

Report Addendums

When residents did go back to review a case and made an addendum during the same simulated shift ($n = 2085/63{,}839$, 3.3%), they scored significantly higher ($p = 0.002$). The effect size was rather small: 0.22 (95% CI 0.08−0.36) points. The use of addenda was widespread across sessions, cases, and residents. For example, at least one addendum was made during 68% (681/992) of sessions and for 98% (314/321) of cases. Likewise, 74% (573/773) of residents made at least one addendum.

Case Characteristics

Neuroradiology cases were the most difficult to interpret correctly ($p<0.0001$) with effect size of 2.3 (95% CI 1.9−3.2) points lower than body and musculoskeletal oriented cases,

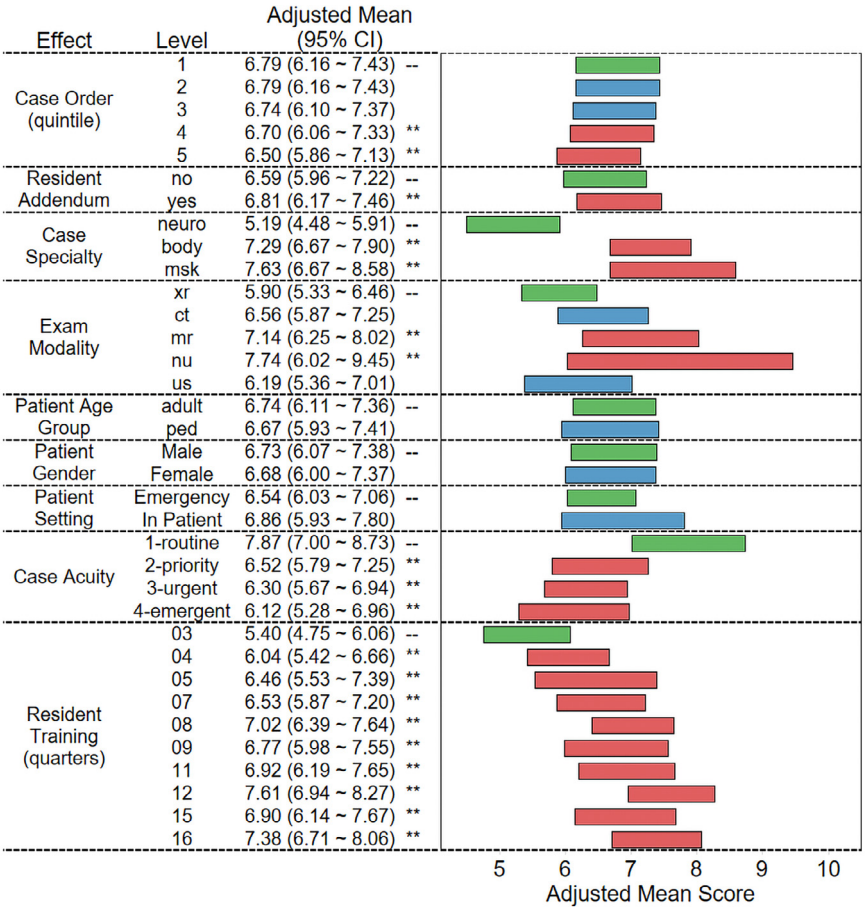

Figure 4. Adjusted mean scores (95% CI) for fixed effects variables. Dashes (–) represent the reference levels and asterisks (**) indicate significance. Bars also represent 95% confidence intervals. Green = reference level, Red = significantly different from reference level, Blue = not significantly different from reference level. "(Color version of figure is available online.)"

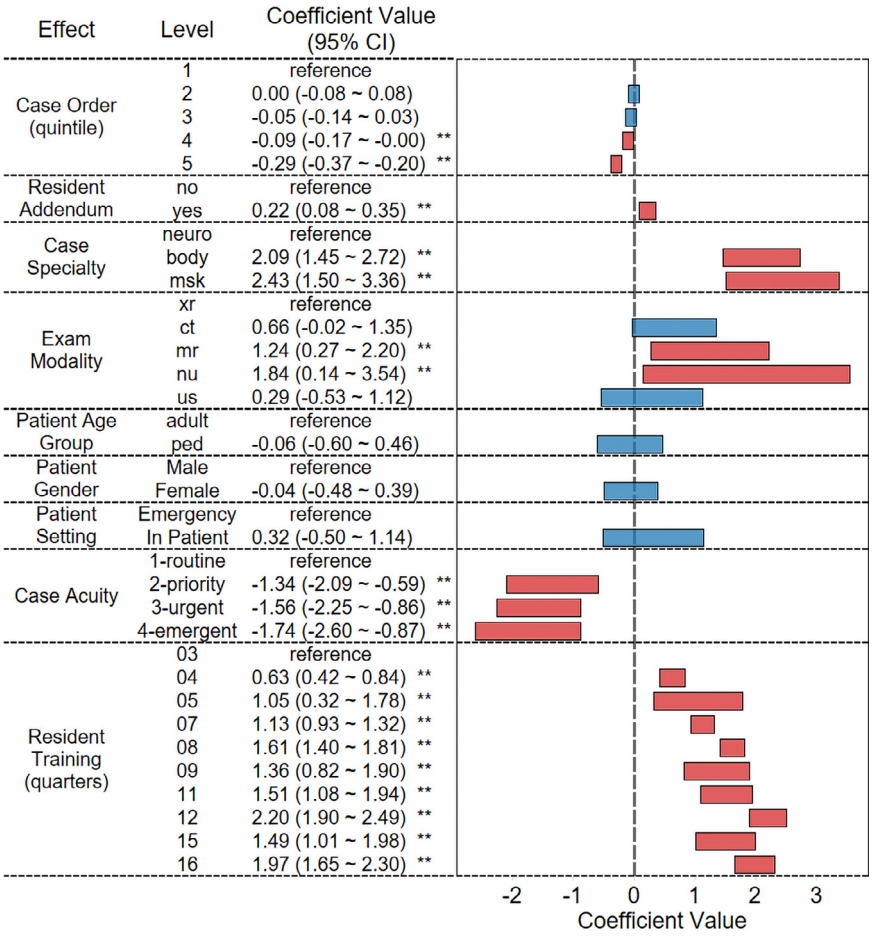

Figure 5. Coefficients for fixed effects variables (95% CI) for fixed effects variables. Dashs (–) represent the reference levels and asterisks (**) indicate significance. Bars also represent 95% confidence intervals. Red = significantly different from reference level, Blue = not significantly different from reference level. Reference level coefficient is 0 (vertical dotted line) by definition.

which were not different from each other (p = 0.36). Multiple analyses by other investigators of resident-attending discrepancy reflect this trend with neuroradiology overnight readings having significantly higher rates compared to other sub-specialties (16–21).

Case acuity had a significant effect (p<0.0001) on residence performance. Negative cases (*n* = 43/321, 13.4%) were always labeled with clinical acuity of "1-Routine" by our consensus and these had the highest adjusted average score of 7.9 points with 5698/8148 (70%) having unequivocally negative/normal interpretations (i.e., 10 points). The reciprocal (slightly less than 1 in 3) represents an upper bounds estimate of overcall rates in critical care radiology. Outside of the mammography literature, there are very few published estimates of the false positive rate in our specialty. The previously cited resident-attending discrepancy literature rarely addresses overcalls by residents (21).

As our consensus acuity cases increased to "2-Priority" in 83/321 (26%), "3-Urgent" in 150/321 (47%), and, finally, "4-Emergent" in 45/321 (14%); adjusted average scores decreased from 7.9 for negative (Routine) cases to 6.5, 6.3, and 6.1 points, respectively (p<0.0001). This correlates well with similar reports that errors of omission increase as severity and number of abnormalities increases which may partly be due to satisfaction of search (22–25).

The modality of the examination had borderline effect on score (p=0.053). We set radiography (XR on tables/figures) as the reference with the lowest adjusted mean score of 5.90 (CI 5.33–6.47) points. Adjusted mean scores were 1.24 (CI 0.27–2.21) points higher for MRI (p=0.012) and 1.84 (CI 0.14–3.55) higher for Nuclear Imaging (p=0.034). We did not include the body area(s) studied by the index exam as separate factors in the model because it is strongly correlated with–and a proxy for–specialty. That said, our observation that radiography is more "difficult" for 21st century radiology residents is relevant to concerns about inadequate training in and decreased utilization of these less costly modalities. The remaining case attributes; including patient gender (p = 0.844), age group (p = 0.807), and setting of care (p = 0.45) had no effect on interpretative performance. The lack of significant effect from these factors is consistent with other literature about errors in resident preliminary reporting during overnight shifts (13).

Resident Training

There was a highly significant (p<0.0001) increase in performance during the 3.5 years of resident training we were able to observe. The apparent magnitude of the effect spans just over 2 out of 10 points. Specifically, comparing the average of quarter

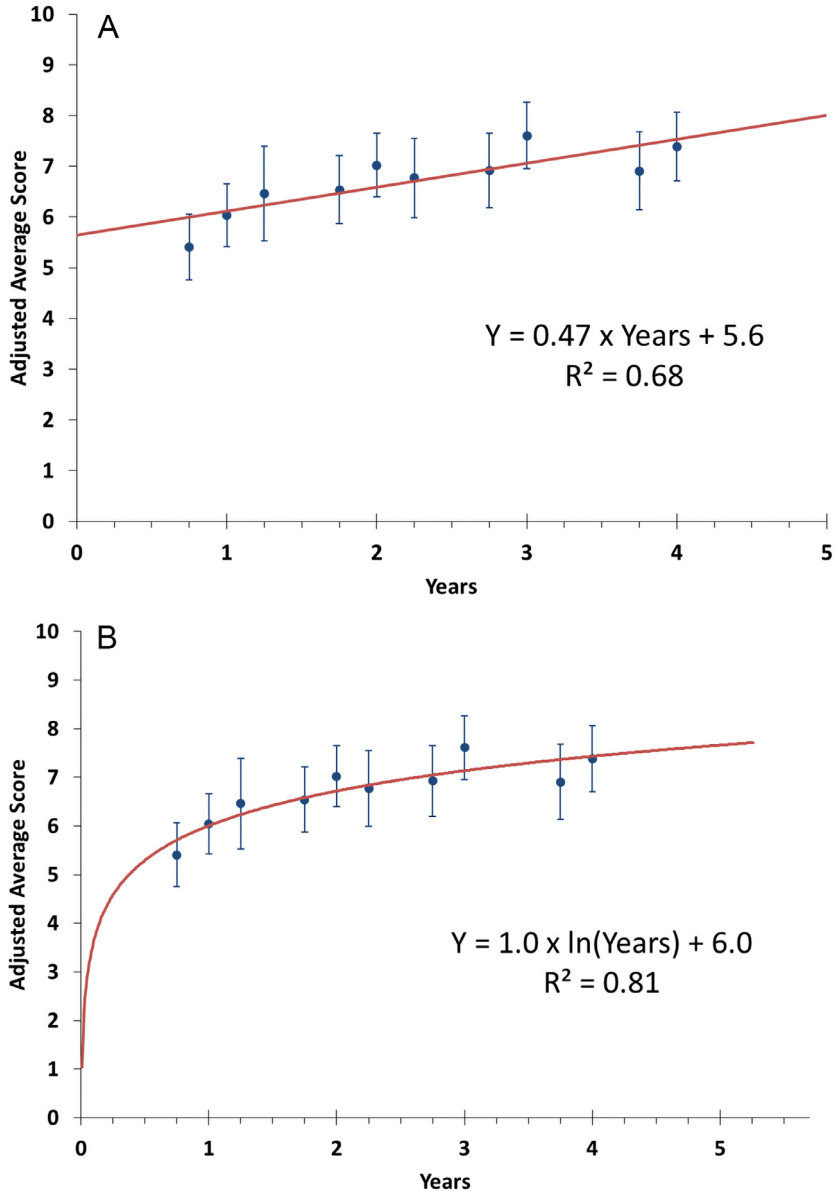

Figure 6. Adjusted mean scores and 95% Confidence Intervals (Blue circles and bars) by resident training (3-16 quarters). Linear trend (A) solid red line from residency start (0 years/quarters) to end (4 years/16 quarters). Learning effect is just under 0.5 points per year. Log growth model (B) fits the data better (R-Squared 0.81 compared to 0.68). The asymptote is 10, by definition. "(Color version of figure is available online.)"

12 (late R3) with quarter 3 (mid R1), the modeled score difference is 2.2 (CI 1.90−2.50) points. The average of quarters 12 and 16 (late R3/R4) is 1.74 (CI 1.41−2.47) points higher than quarter 3 (mid R1). Visually (Figs 4−6) we might be tempted to infer a "plateau" effect in the R4 year. Quantitatively, there is a tiny, nonsignificant ($p = 0.73$), decrease from quarters 12 to 16 of 0.06 (CI -0.38−0.27) points.

Figure 6 depicts the adjusted mean scores for each quarter with 95% confidence intervals. The horizontal axis is set to years for clarity. In Figure 6A, a simple linear trend with zero intercept is shown. Note that the R-Squared (0.68) is low. This is not surprising because human skill sets are described as having a "learning curve". On the other hand, Figure 6B has been annotated with a basic log growth trend. Not only does the associated R-Squared (0.81) indicate a better fit, but the curve is more consistent with our conceptual understanding and educational theory.

Our results complement and extend those of Ganguli et. al., who conducted a 20 case simulation based assessment of R1-R4 residents (n∼56) from a single program for 5 years. Their resident's interpretative responses were graded by senior radiologists on a 3 level ordinal scale and reported as percent correct in the papers. The percent correct showed average increase of 71% to 86% from R1−R4. Statistically significant ($p<0.05$) temporal learning effect was demonstrable between the R1 cohort and the upper levels (26−28).

General Considerations

This full resolution Critical Care Radiology Simulation (CCRS) evaluation encompasses three essential component EPAs: observational skills, interpretive skills (once proper observations are made), and professional behavior. It departs

from other competency evaluation rubrics in four critical elements of its methodology. First, it provides the full set of DICOM images generated in "real life" for the critical care radiology scenario, i.e., an imaging study and defined clinical situation at hand. Second, the participants must assert, in their own words, whether the study is normal or abnormal by delivering their consultation in a short written response format. Third, the test set includes normal studies. These first three key elements eliminate the shortcomings of "pretest bias" of leading questions and artificial thought processes imposed by multiple-choice questions and selected images as well as the "given", in other testing methods, that the images in question are known to be abnormal. Fourth and finally, the CCRS embodies the professional behavioral objectives of the consultative process by requiring an understanding of clinical context, production of a duly expected work product and properly communicating the acuity of the clinical scenario at hand, inclusive of proper advice going forward, if required by scenario and study findings. All of this is required to get a full credit "10" score on each case.

CONCLUSION

We have shown that it is possible to simulate and evaluate competence in the interpretative tasks faced by radiology residents working in critical care (overnight call) settings. Accounting for and measuring fixed effects (e.g., case mix, level of training, fatigue, and etc.) is necessary for unbiased measurement of resident preparedness and provide many insights about study characteristics, work flow and process. In the second part of this paper, we will report on estimates of case difficulty and resident competence as random effects.

ACKNOWLEDGMENTS

Since 2012, the American College of Radiology (ACR) has provided a customized institutional "sandbox" for our program on their Radiology Content Management System (RCMS, a.k.a. "Cortex"). Further details are in Materials and Methods. Visage Imaging Inc. (San Diego, CA) donated use of their enterprise imaging platform server and web based client software (Version 7) for fully functional diagnostic quality study visualization during the simulations.

REFERENCES

1. ACGME Common Program Requirements [Internet]. 2017[cited 2018 Nov 20].
2. Vydareny KH, Amis ES, Becker GJ, et al. Diagnostic Radiology Milestones. J Grad Med Educ 2013; 5(1s1):74–78.
3. ten Cate TJO, Snell L, Carraccio C. Medical competence: the interplay between individual ability and the health care environment. Med Teach 2010; 32(8):669–675.
4. ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? Acad Med 2007; 82(6):542–547.
5. ten Cate O. Entrustability of professional activities and competency-based training. Med Educ 2005; 39(12):1176–1177.
6. Willis RE, Van Sickle KR. Current Status of Simulation-Based Training in Graduate Medical Education. Surg Clin North Am 2015; 95(4):767–779.
7. Orledge J, Phillips WJ, Murray WB, et al. The use of simulation in healthcare: from systems issues, to team building, to task training, to education and high stakes examinations. Curr Opin Crit Care 2012; 18(4):326–332.
8. Chetlen AL, Mendiratta-Lala M, Probyn L, Auffermann WF, et al. Conventional medical education and the history of simulation in radiology. Acad Radiol 2015; 22(10):1252–1267.
9. Klein KA, Neal CH. Simulation in radiology education: thinking outside the phantom. Acad Radiol 2016; 23(7):908–910.
10. Sabir SH, Aran S, Abujudeh H. Simulation-based training in radiology. J Am Coll Radiol 2014; 11(5):512–517.
11. Sarwani N, Tappouni R, Flemming D. Use of a simulation laboratory to train radiology residents in the management of acute radiologic emergencies. Ame J Roentgenol 2012; 199(2):244–251.
12. Giaddui T, Yu J, Manfredi D, et al. Structures' validation profiles in Transmission of Imaging and Data (TRIAD) for automated National Clinical Trials Network (NCTN) clinical trial digital data quality assurance. Pract Rad Oncol 2016; 6(5):331–333.
13. Ruutiainen AT, Durand DJ, Scanlon MH, et al. Increased error rates in preliminary reports issued by radiology residents working more than 10 consecutive hours overnight. Acad Radiol 2013; 20(3):305–311.
14. Hanna TN, Lamoureux C, Krupinski EA, et al. Effect of shift, schedule, and volume on interpretive accuracy: a retrospective analysis of 2.9 million radiologic examinations. Radiology 2018; 287(1):205–212.
15. Krupinski EA, Schartz KM, Van Tassell MS, et al. Effect of fatigue on reading computed tomography examination of the multiply injured patient. J Med Imaging 2017; 4(03):1.
16. Miyakoshi A, Nguyen QT, Cohen WA, et al. Accuracy of preliminary interpretation of neurologic CT examinations by on-call radiology residents and assessment of patient outcomes at a level I trauma center. J Am Coll Rad 2009; 6(12):864–870.
17. Filippi CG, Schneider B, Burbank HN, et al. Discrepancy rates of radiology resident interpretations of on-call neuroradiology MR imaging studies. Radiology 2008; 249(3):972–979.
18. Sistrom CL, Deitte L. Factors affecting attending agreement with resident early readings of computed tomography and magnetic resonance imaging of the head, neck, and spine. Acad Radiol 2008; 15(7):934–941.
19. Ruchman RB, Jaeger J, Wiggins EF, et al. Preliminary radiology resident interpretations versus final attending radiologist interpretations and the impact on patient care in a community hospital. AJR. 2007; 189(3):523–526.
20. Ruutiainen AT, Scanlon MH, Itri JN. Identifying benchmarks for discrepancy rates in preliminary interpretations provided by radiology trainees at an academic institution. J Am Coll Radiol 2011; 8(9):644–648.
21. Provenzale JM, Kranz PG. Understanding errors in diagnostic radiology: proposal of a classification scheme and application to emergency radiology. Emerg Radiol 2011; 18(5):403–408.
22. Velmahos GC, Fili C, Vassiliu P, et al. Around-the-clock attending radiology coverage is essential to avoid mistakes in the care of trauma patients. Am Surg 2001; 67(12):1175–1177.
23. Schartz KM, Madsen MT, Kim J, et al. Satisfaction of search revised. 2017;13(8):973–8.
24. Banaste N, Caurier B, Bratan F, et al. Whole-body CT in patients with multiple traumas: factors leading to missed injury. Radiology 2018:180492.
25. Novelline RA. CT in the patient with multiple trauma: risk factors for missed findings. Radiology 2018:181534. 7 nn.
26. Ganguli S, Pedrosa I, Yam C-S, et al. Part I: preparing first-year radiology residents and assessing their readiness for on-call responsibilities. Acad Radiol 2006; 13(6):764–769.
27. Ganguli S, Camacho M, Yam CS, et al. Preparing first-year radiology residents and assessing their readiness for on-call responsibilities: results over 5 years. Am J Roentgenol 2009; 192(2):539–544.
28. Yam C-S, Kruskal J, Pedrosa I, et al. Part II: preparing and assessing first-year radiology resident on-call readiness technical implementation. Acad Radiol 2006; 13(6):770–773.