Full Resolution Simulation for Evaluation of Critical Care Imaging Interpretation; Part 2: Random Effects Reveal the Interplay Between Case Difficulty, Resident Competence, and the Training Environment

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 further characterize empirical data from a full-resolution simulation of critical care imaging coupled with post hoc grading of resident's interpretations by senior radiologists. To present results from estimating the random effects terms in a comprehensive mixed (hierarchical) regression model.

Materials and methods: After accounting for 9 fixed effects detailed in Part 1 of this paper, we estimated normally distributed random effects, expressed in terms of score offsets for each case, resident, program, and grader.

Results: The fixed effects alone explained 8.8% of score variation and adding the random effects increased explanatory power of the model to account for 36% of score variation. As quantified by intraclass correlation coefficient (ICC = 28.5%; CI: 25.1−31.6) the majority of score variation is directly attributable to the case at hand. This "case difficulty" measure has reliability of 95%. Individual residents accounted for much of the remaining score variation (ICC = 5.3%; CI: 4.6−5.9) after adjusting for all other effects including level of training. The reliability of this "resident competence" measure is 82%. Residency training program influence on scores was small (ICC = 1.1%; CI: 0.42−1.7). Although a few significantly high and low ones can be identified, reliability of 73% militates for caution. At the same time, low intraprogram variation is very encouraging. Variation attributable to differences between graders was minimal (ICC = 0.58%; CI: 0.0−1.2) which reassures us that the method of scoring is reliable, consistent, and likely extensible.

Conclusion: Full resolution simulation based evaluation of critical care radiology interpretation is being conducted remotely and efficiently at large scale. A comprehensive mixed model of the resulting scores reliably quantifies case difficulty and resident competence.

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

INTRODUCTION

In Part 1 of this paper, we introduced our Critical Care Radiology Simulation (CCRS) competency evaluation system and the resulting scores assigned by practicing radiologist graders to 63,839 interpretations rendered during 992 simulation sessions by 773 residents from 47 programs over 5 years. A mixed regression model with 9 fixed and 4 random effects was introduced. Part 1 describes statistics and inferences related to the fixed variables. This paper (Part 2) describes the model estimates for the 4 random effect variables and the error term.

MATERIALS AND METHODS

Data Source

Case selection, case curation, simulation testing paradigm, and grading of resident responses were fully described in Part 1 of this paper.

Statistical Analysis

Having already covered the fixed effects results in Part 1, we turned our attention to quantifying the random effects (case,

Acad Radiol 2020; ■:1–9

From the University of Florida, Gainesville, Florida (C.L.S.); Department of Radiology University of Florida Health Center, Gainesville, Florida (R.M.S., D.A.R., J.R.G., J.H.R., A.A.M.). 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.2019.11.025

TABLE 1. Random Effects Parameter Estimates

Random Effect	Case	Resident	Program	Grader
Distinct count	321	773	47	16
Minimum	−5.4	−2.4	−0.6	−0.4
Maximum	4.6	2.2	0.6	0.4
Variance estimate	3.78	0.53	0.10	0.06
Standard error (of variance estimate)	0.31	0.04	0.03	0.03
p value	<0.0001	<0.0001	0.0006	0.0339
Intraclass correlation coefficient (ICC)	28.5%	5.3%	1.1%	0.58%
ICC 95% confidence interval	25.0–31.6%	04.6–05.9%	00.4–01.7%	00.0–01.2%
Reliability (average)	94.8%	81.50%	72.6%	72.2%
Reliability 95% confidence interval	94.6–95.1%	81.3–81.8%	70.0–75.1%	66.0–78.4%

resident, program, and grader) in order to explore their utility as measurements and implications for training, certification, and practice. We obtained observation level ($N = 63{,}839$) score predictions and calculated residuals (observed minus predicted score) to confirm our model's assumption that errors were normally distributed despite the bi-modal distribution of the observed scores. We also used predictions and residuals from partial (e.g., leaving out fixed and/or random effects) and null (intercept only) models to look at explained score variance (i.e., R-Squared). We used the variance and/or co-variances reported by SAS (Version 9.4, Cary, North Carolina) PROC MIXED along with standard errors and residuals to calculate the intraclass correlation coefficients (ICC) with 95% confidence intervals for each of the random effects.

By requesting a solution for the random effects, we obtained unbiased joint estimates of the linear "intercept" for

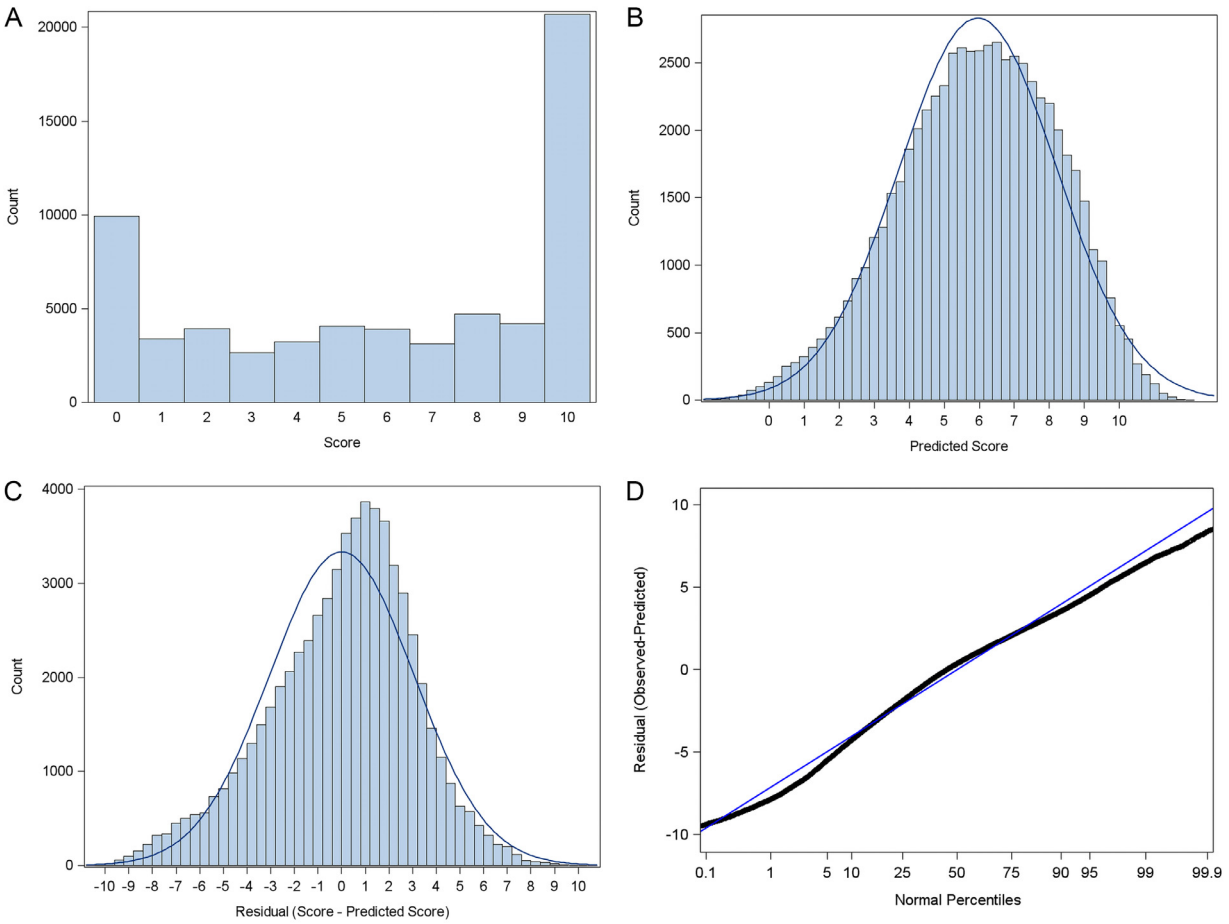

Figure 1. Histograms of observed scores (A), predicted scores from the model (B), Residuals (C). The residuals (observed-predicted) are plotted (D) against normal percentiles. The solid curves (B and C) and line (D) represent corresponding normal distributions for comparison. Aside from mild rightward skew (due to truncating predictions at 0), the residuals (error term) have near-normal distribution, which is a crucial assumption of regression models like ours.

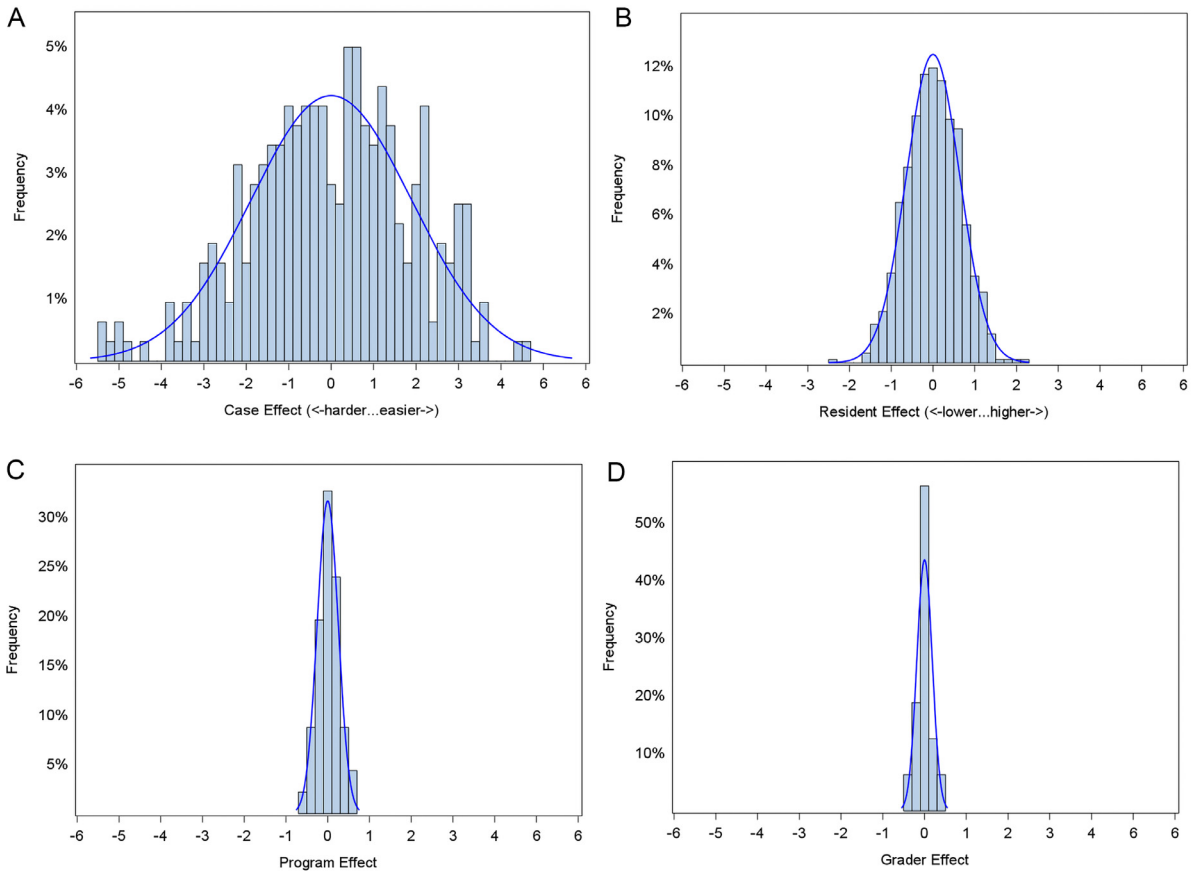

Figure 2. Histograms from estimates of influence on scores for the 321 cases (A), 773 residents (B), 47 programs (C), and 16 graders (D). The solid curves represent the underlying normal distributions, all centered on 0 with decreasing variances which are listed in Table 1 along with standard errors, intraclass correlation coefficients (ICC), reliability, and associated 95% confidence intervals.

each unique case ($N = 321$), resident ($N = 773$), radiology program ($N = 47$), and grader ($N = 16$). They were scaled as positive or negative score "offsets" centered on 0 with a standard error for each one. These "intercept/offset" values represent an individual case difficulty, resident competence, program effect, and grader bias with all else equal (including fixed and other random effects). Finally, we used the standard errors for each case, resident, program, and grader along with the associated overall random effect covariance to calculate individual reliability. The mean and 95% confidence intervals of these were reported as case, resident, program, and grader measurement reliabilities.

RESULTS

The R-Squared of the full model with 9 fixed effects (Part 1) and 4 random effects (Part 2) was 36.3%. The reduced model with only the 9 fixed effects had R-Squared of 8.8%. Table 1 shows count, range, covariance, standard error, p value, ICC, and average reliability for each of the four random effects, arranged in descending column order of variance (cases, residents, programs, and graders). This shows their relative and unique contributions to overall score variance after adjusting for all the fixed effects.

There were a total of 63,839 observations ranging from 0 to 10 having the bimodal distribution shown in Figure 1A. Histograms of the predicted scores and residuals (observed minus predicted score) are shown in Figures 1B and C respectively along with estimated normal probability distribution functions (solid curves). Figure 1D is a plot of the residuals versus percentiles and shows that they correspond well to the theoretical normal distribution (solid line). This reassures us about an important assumption (errors are normally distributed) underlying the overall statistical model. Figure 2 contains histograms of the individual solutions for each of the four random effect terms: cases (A), residents (B), programs (C), and graders (D). These help to visualize our model assumptions about and estimates of these parameters as listed in Table 1.

Figure 3 shows the case random effect estimates with 95% confidence intervals stratified by specialty: body (A), neurological (B), and musculoskeletal (C). Since the case difficulty was estimated jointly with all other fixed effects (including specialty) and the other random effects, they can be interpreted as adjusted difficulty within specialty. The 10 unique cases at each difficulty extreme for each specialty are indicated by dotted lines (difficult = red, easy = green). These are called out, supplemented with average scores, and rendered as

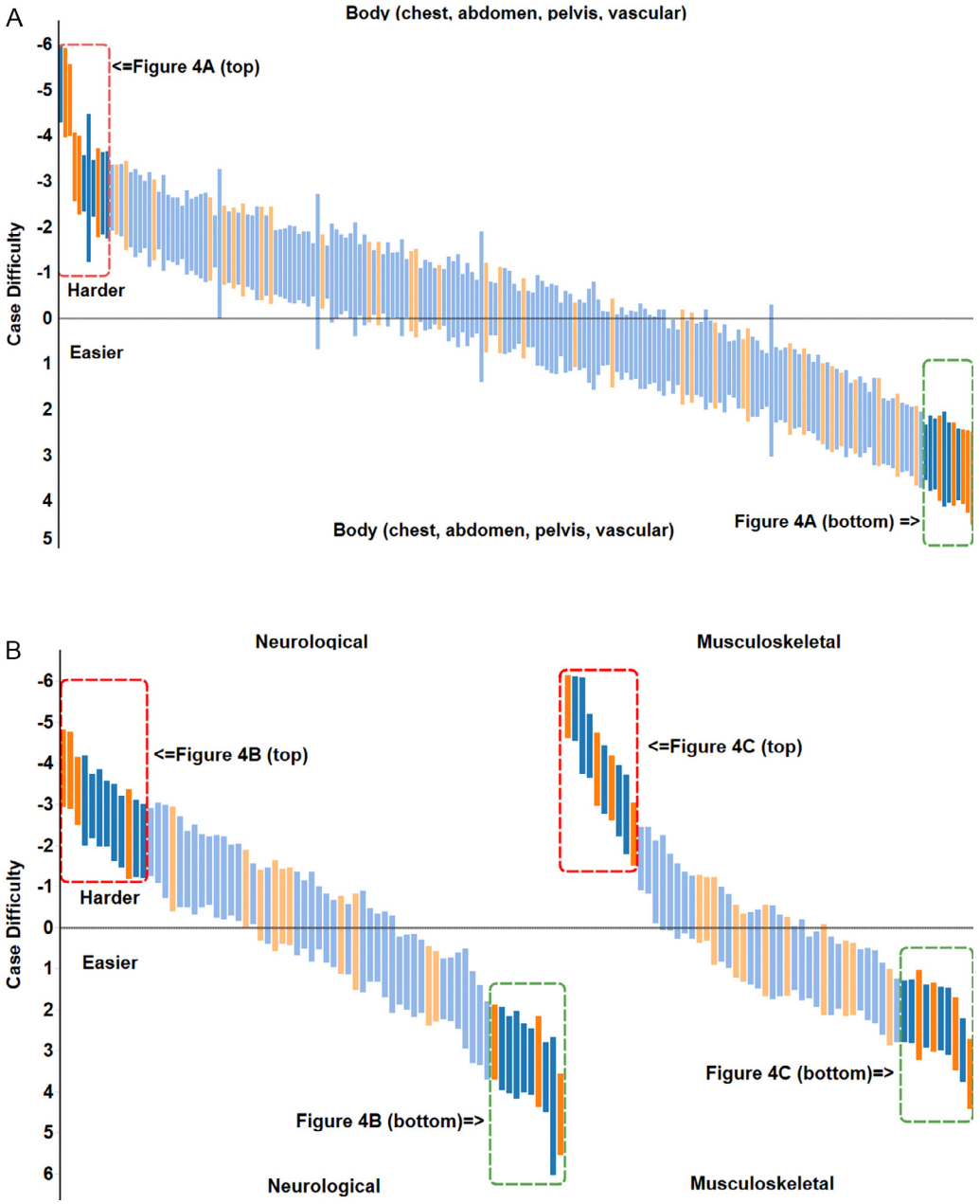

Figure 3. Estimated effect (difficulty) of body (A) neurological (B) and musculoskeletal (B) cases color coded by age group (adult = blue, pediatric = orange). The length of each bar represents the 95% confidence interval of the difficulty estimate. The boxes outlined with red dashes indicate the most difficult cases for body ($N = 11$), neurological ($N = 12$), and musculoskeletal ($N = 10$). The green boxes outline the easiest cases ($N = 10$) in the same specialties ($N = 10$ for all). Annotations next to each box connect it to further detail about these same cases found in Figure 4A, B, and C. (Color version of figure is available online.)

Forest plots in Figure 4 stratified by specialty: body (A), neurological (B), and musculoskeletal (C).

Note that there are "duplicate" case concept description labels in each specialty (body, neurological, and musculoskeletal) of Figure 4. These are indicated by asterisk and supplemented with parenthesized year each case was tested. In only one of these instances was the same exam from the same patient presented in more than one year. This was a CT scan of a child with pancreatic laceration tested in 2014 and 2017 (Fig 5). The resulting difficulty estimates (−4.8 points vs −3.3 points) and average scores (1.6 and 2.7) show the "repeatability" of a single case.

In the neuroradiology group, there were three distinct cases of suppurative adenitis shown in CT scans performed on different children tested in 2014 (Fig 6A), 2016 (Fig 6B), and 2017 (Fig 6C). With difficulty estimates of −3.8, −3.9, −3.2, and average scores of 0.56, 0.25, 1.02, they cluster at the upper end of the neurological case difficulty spectrum with rank orders of 2, 1, and 3, respectively (Fig 4B). By way of contrast, there were two cases of Legg Calve Perthes

Disease (LCPD) on hip radiographs that had difficulty estimates at opposite ends of the musculoskeletal difficulty spectrum (Fig 4C). A study with typical advanced destructive changes of the right femoral head (Fig 7A) given in 2018 was the third easiest musculoskeletal case with difficulty estimate of 2.6 and average score of 8.91 points. A much more challenging case of LCPD was given in 2016 and had difficulty estimate and average scores of −2.3 and 3.9, respectively. The radiographic findings (Fig 7B) in the Left Femoral Head, representing early stage disease, were subtle though quite specific.

Figure 8A depicts the random effect (vertical axis) from each of the 773 participating residents plotted against their average score (horizontal axis). In Figure 8B, the average score percentile has replaced the horizontal axis. Note that the overall average resident score (5.88) is nearly identical to

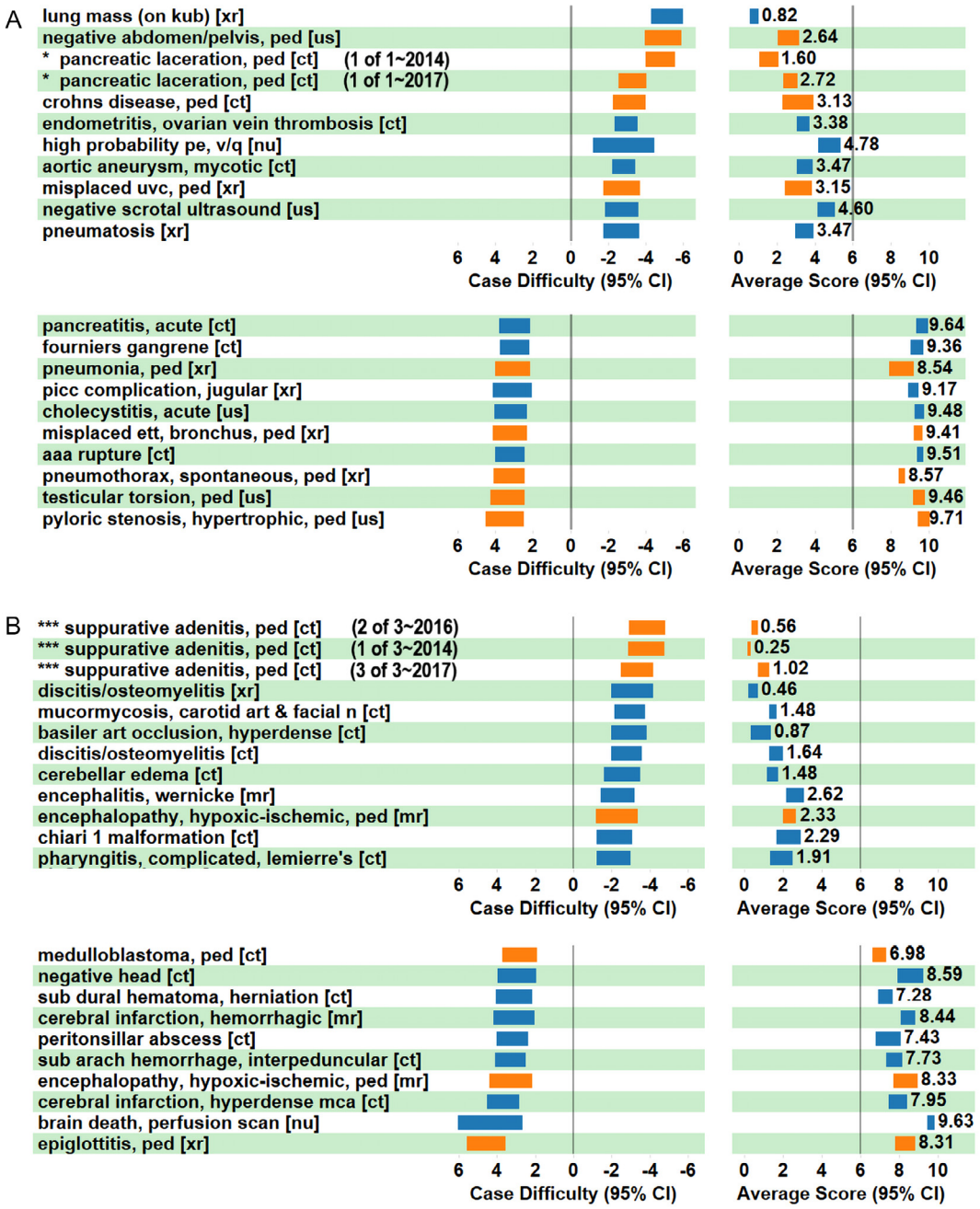

Figure 4. Hardest (10) and easiest (10) cases for body (A), neurological (B), and musculoskeletal (C) imaging. Horizontal bars on the left indicate case difficulty (width = 95% confidence interval). The horizontal bars to right represent the mean score for each case (width = 95% confidence interval). As with Figure 3, adult cases are colored blue and pediatric are colored orange. Note that the rows ($N = 7$) marked with asterisk representing cases that had similar exams and findings but were given to different groups of residents during separate years during the study. These will be called out in results with further explication in the discussion. (Color version of figure is available online.)

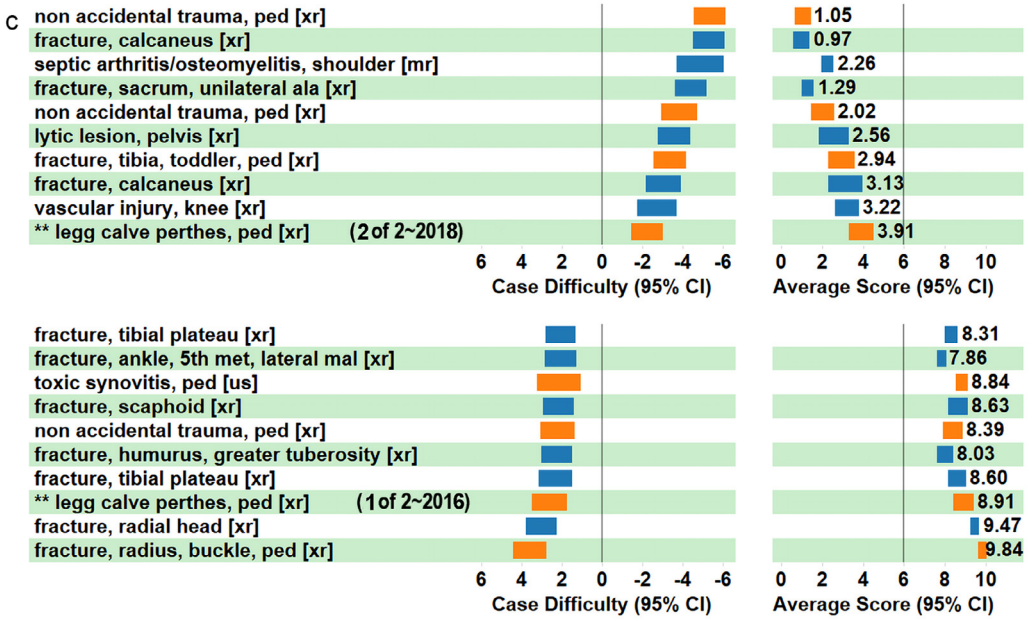

Figure 4 Continued.

Figure 5. Abdominal CT scan with pancreatic laceration that was shown in two separate testing years (see Figure 4A).

the median resident score (P50) which is 5.87. This reflects the model assumption of resident performance being normally distributed. Also shown on Figure 8B are the 5th (P5) and 95th (P95) percentiles with average resident scores of 4.25 and 7.60 respectively.

In both Figure 8A and B, the marks for residents who significantly underperformed (all else equal) are downward oriented red triangles (118 of 772 = 15.3%). Likewise, residents who significantly outperformed peers are marked with upward oriented blue triangles (110 of 772 = 14.2%). Finally, green circles represent residents with competence (performance) estimates that were not significantly different from zero (544 of 772 = 70.5%). As with the case random effects, the significance (of being different from 0) for each resident's competence estimate was determined by the 95% confidence intervals associated with each one. For all marks, the size is proportional to the number of completed and scored cases each resident took during their simulation session(s).

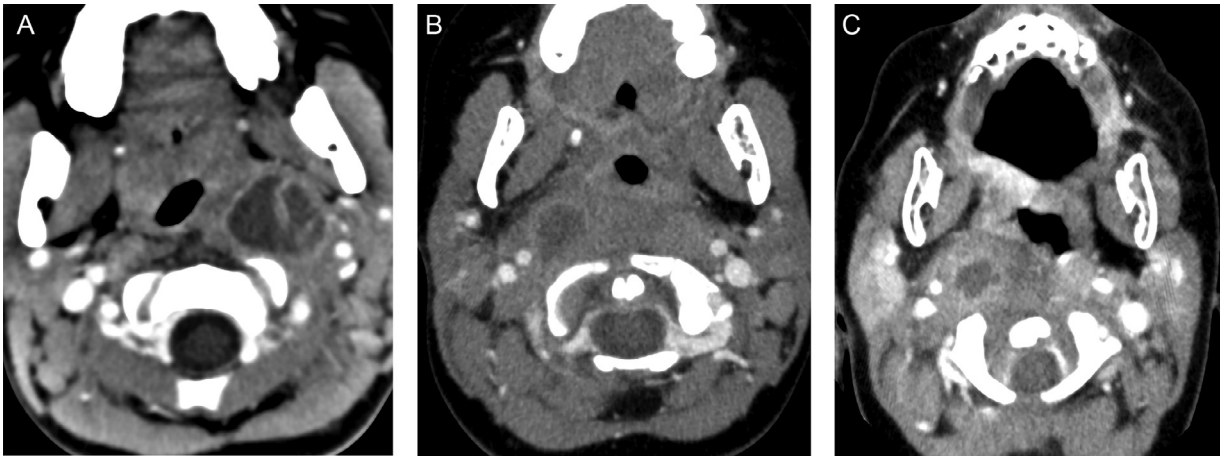

Figure 6. Neck CT scans from three separate patients with suppurative adenitis. These were shown in 2014 (A), 2016 (B), and 2017 (C) respectively and clustered at the upper end of the case difficulty spectrum (Fig 4B).

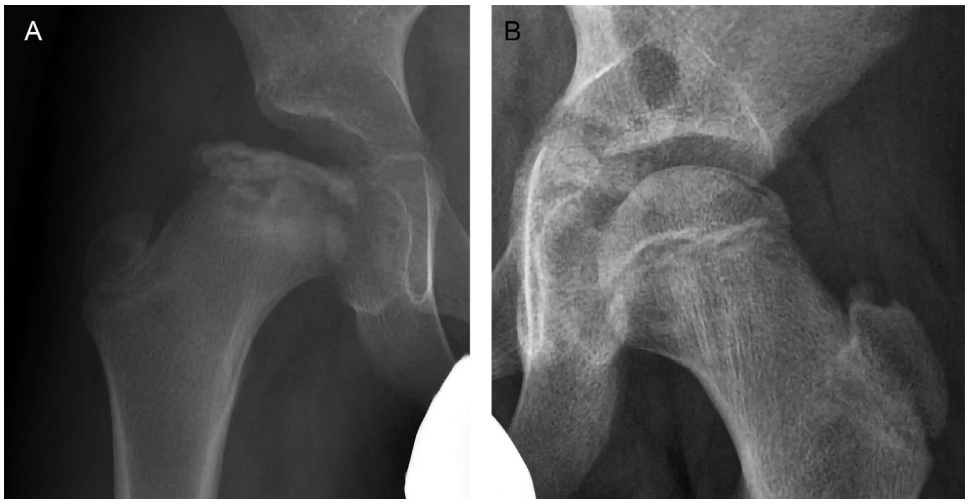

Figure 7. Hip radiographs from two separate cases of Legg Calve Perthes Disease. The condition was present in different stages of severity when tested; advanced in 2016 (A) and early in 2018 (B). The advanced disease radiograph was the third easiest of all musculoskeletal ones and the radiograph with early findings was tenth hardest (see Fig 4C).

In many other paradigms of testing radiology knowledge or competence, unadjusted averages of the candidate scores are transformed into percentiles and used to categorize performance (e.g., below 5th, above 95th). Figure 8B allows us to compare and contrast that method with our model. Consider a cutoff of fifth percentile (resident average score = 4.25) as representing actionable underperformance. Perhaps a resident might be held back for remediation or a board candidate must try again to obtain their certificate. Using our results, 40 residents had average scores below the fifth percentile (4.25). However, four of those residents had performance estimates with 95% confidence intervals crossing zero and would have not been classified as underperformers using our case mix adjusted model. On the other hand, 35 residents had average scores above the 95th percentile and 10 of those would be classified as being same as peers by our model based competence measure. Between the 5th and 75th unadjusted average score percentiles, there were 82, 530, and 85 residents classified by the model as below, same, and above peers respectively.

DISCUSSION

Our model seeks to describe, explain, and measure several important concepts related to interpretation of radiology studies during emergency and critical care situations. Among these are that there is an objective and articulable set of findings (or pertinent negative observations) that should be included in a complete interpretative report of any imaging study. Also, we believe that experienced practicing radiologists who have reviewed and know the salient features of a study can read a report about it and render a 0−10 score which reliably and accurately quantifies its "correctness." Further, a given case (a single patient's imaging study) possesses an inherent difficulty to any clinician who interprets it. Radiologists themselves manifest different levels of competence around different classes (specialties) of cases, based on their own training and experience. Although these fundamental assumptions may seem obvious, they necessarily underpin all of diagnostic radiology education and evaluation.

After accounting for the intrinsic difficulty of the case itself and the competence of the interpreting resident; training program and grader have much less influence on scores. The relative contribution of these four factors to score variation is quantified by the ICC (Table 1). We interpret the case ICC to mean that knowing the identity (and associated relative difficulty) of a case, explains 28.5% of the score. Similarly, the ICC for residents means that if we know who read the study, our predictive accuracy about the score increases by another 5.3% to just over one-third (33.8%). On the other hand, the aggregate of program and grader ICC is less than 2%, indicating that they collectively contribute very little to score variation.

Interpreting the program effects calls for caution and circumspection. One way to describe this is as program "quality" or "effectiveness." However, it is very important to understand that the program effect in our model is adjusted for the mixture of competence among the residents they are able to recruit and retain. What we can say is that relatively low program ICC is validating from an overall measurement quality perspective. Had the program ICC been 5% or higher, before assuming that much difference between programs, we would be obliged to question the generalizability and reproducibility of the simulation and evaluation methodology itself. We are also encouraged that the factor which gives the least information about the score is the identity of the grader. Unlike the other 3 random effects (case, resident, and program), we expect and require that graders have little if any influence on scores. The grader ICC of 0.58% with 95% confidence interval that includes zero is quite reassuring in this regard.

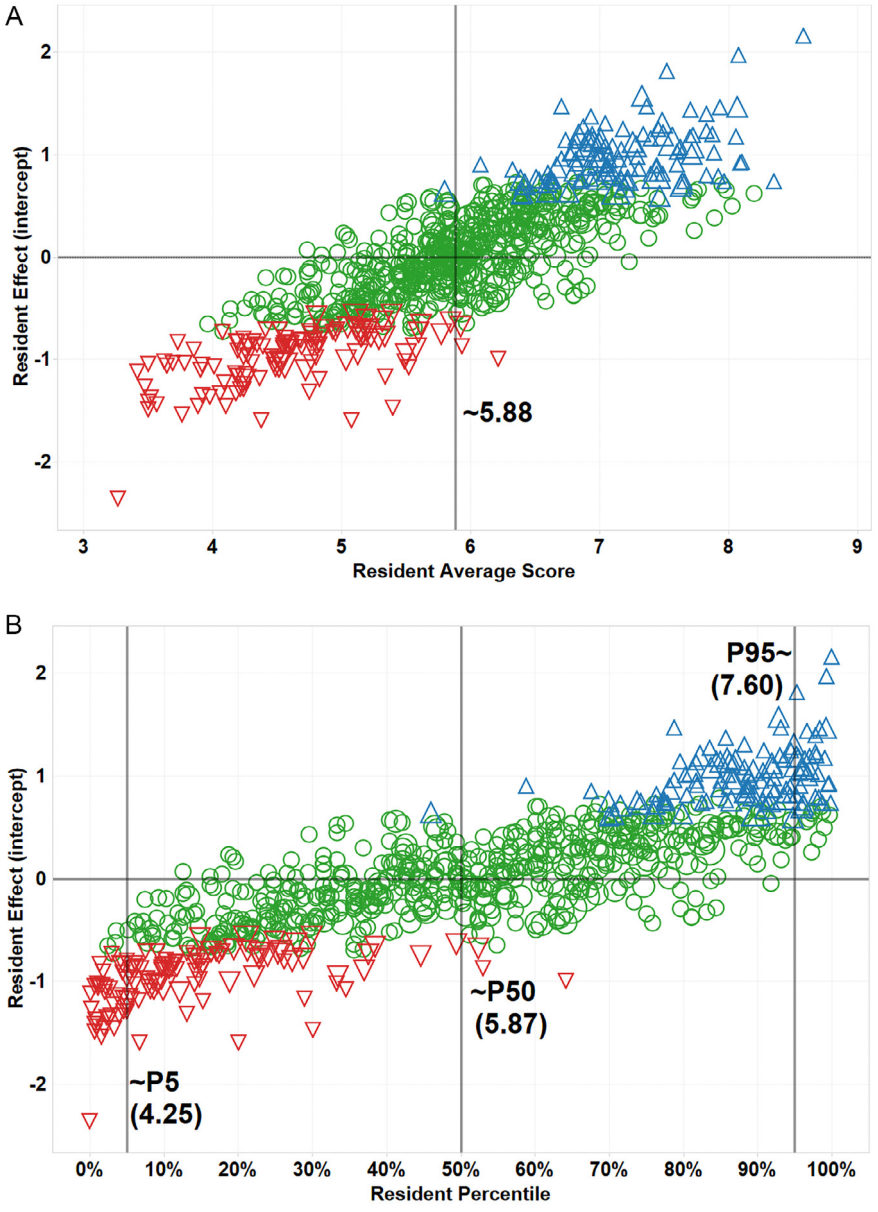

Figure 8. Scatterplot of the random effect (performance) estimate and average score (A) as well as percentile score (B) from each of 773 unique participating residents. Statistical significance of the resident's random effect is indicated by shape and color of marks. Red triangle (with apex down) represents that the resident was significantly less competent than "average" (green circles). Blue triangles (with apex up) depict residents that were significantly more competent. The overall average resident score (5.88) is shown (A) as are the 5th, 50th, and 95th percentiles (B). (Color version of figure is available online.)

The reliability of just under 95% associated with case difficulty is estimated with minimal error (94.8% CI: 94.6%–95.1%). This has several implications for radiology training, evaluation, and practice. For example, in Figures 4B and 6 the three hardest neuroradiology cases share the same concept of "suppurative adenitis" occurring in 3 different children all diagnosed from CT scan. Given the high reliability of case difficulty, we infer that pediatric suppurative adenitis on CT scan is difficult to interpret correctly. It is also reasonable (and testable) to believe that targeted, specific, and direct instruction of radiology residents about CT in children with acute pharyngitis/tonsillitis would be highly beneficial. On the other hand, musculoskeletal curricula concerning LCPD might benefit from greater emphasis on early stage radiographic findings.

Computer based curriculum delivery and competency evaluation is particularly well suited to education in diagnostic radiology because of the nature of the basic cognitive tasks. Full resolution simulation at a distance is increasingly practicable given widespread use of softcopy reading with server side rendering, self-edited report authoring, and improved workflows for results communication. Simulation enhanced methods have enjoyed some early successes and other investigators agree with us that there is considerable promise for further refinement and expansion (1–7). In our future research, we plan to develop a conceptual, taxonomic, and analytic framework that will serve as the basis for a general purpose rubric for curriculum development, competency assessment and targeted remediation. The same framework should inform innovative pedagogic approaches (e.g., flipped

classroom, direct instruction, etc.) to complement existing lecture-centric methods for teaching radiology interpretation at many levels of sophistication and expertise.

CONCLUSION

The CCRS program was initially started to satisfy two local purposes. First, to assess mid-level radiology resident's preparedness for remotely supervised overnight in-house duty shifts in our full-service academic teaching hospital. The second was to formally document this assessment in order to fulfill institutional quality initiatives and national educational mandates. The fact that 46 additional radiology residency programs in 23 states have chosen to use CCRS is a testament to the utility of the simulation and evaluation methodology and affords us a unique opportunity to analyze and share the results from our collective experience. We hope that the conceptual and statistical model we developed and presented will be useful to others working to understand and improve radiology residency education and evaluation.

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. Cook TS, Hernandez J, Scanlon M, et al. Why isn't there more high-fidelity simulation training in diagnostic radiology? Results of a survey of academic radiologists. Acad Radiol 2016; 23(7):870–876.
2. Sabir SH, Aran S, Abujudeh H. Simulation-based training in radiology. J Am Coll Radiol 2014; 11(5):512–517.
3. Nayahangan LJ, Nielsen KR, Albrecht-Beste E, et al. Determining procedures for simulation-based training in radiology: a nationwide needs assessment. Eur Radiol 2018; 28(6):2319–2327.
4. Chetlen AL, Mendiratta-Lala M, Probyn L, et al. Conventional medical education and the history of simulation in radiology. Acad. Radiol. 2015; 22 (10):1252–1267.
5. Klein KA, Neal CH. Simulation in radiology education: thinking outside the phantom. Acad Radiol 2016; 23(7):908–910.
6. Klein KA, Neal CH. Simulation in radiology education. Acad Radiol 2016; 23(7):908–910.
7. van der Gijp A, Ravesloot CJ, Tipker CA, et al. Increasing authenticity of simulation-based assessment in diagnostic radiology. Simul Healthc 2017; 12(6):377–384.