Original Investigation

Attending Radiologist Variability and Its Effect on Radiology Resident Discrepancy Rates

Joseph C. Wildenberg, MD, PhD, Po-Hao Chen, MD, MBA, Mary H. Scanlon, MD, Tessa S. Cook, MD, PhD

Rationale and Objectives: Discrepancy rates for interpretations produced in a call situation are one metric to evaluate residents during training. Current benchmarks, reported in previous studies, do not consider the effects of practice pattern variability among attending radiologists. This study aims to investigate the impact of attending variability on resident discrepancy rates to determine if the current benchmarks are an accurate measure of resident performance and, if necessary, update discrepancy benchmarks to accurately identify residents performing below expectations.

Materials and Methods: All chest radiographs, musculoskeletal (MSK) radiographs, chest computed tomographies (CTs), abdomen and pelvis CTs, and head CTs interpreted by postgraduate year-3 residents in a call situation over 5 years were reviewed for the presence of a significant discrepancy and composite results compared to prior findings. Simulations of the expected discrepancy distribution for an "average resident" were then performed using Gibbs sampling, and this distribution was compared to the actual resident distribution.

Results: A strong inverse correlation between resident volume and discrepancy rates was found. There was wide variability among attendings in both overread volume and propensity to issue a discrepancy, although there was no significant correlation. Simulations show that previous benchmarks match well for chest radiographs, abdomen and pelvis CTs, and head CTs but not for MSK radiographs and chest CTs. The simulations also demonstrate a large effect of attending practice patterns on resident discrepancy rates.

Conclusions: The large variability in attending practice patterns suggests direct comparison of residents using discrepancy rates is unlikely to reflect true performance. Current benchmarks for chest radiographs, abdomen and pelvis CTs, and head CTs are appropriate and correctly flag residents whose performance may benefit from additional attention, whereas those for MSK radiographs and chest CTs are likely too strict.

Key Words: Resident Education; Discrepancies; Call; Practice Variability.

INTRODUCTION

There is increasing use of metrics to demonstrate competency and measure performance of trainees across all specialties of medicine (1–3). Within radiology, discrepancy rates for study interpretations made in call situations, without immediate attending input or oversight, are commonly incorporated into resident feedback (4,5). This feedback may be occasional, in the form of missed case conferences or a semiannual meeting with the program director, or in near real-time implemented in a dashboard that residents can access via a web interface (5–7).

Many radiology residency programs calculate on-call discrepancy rates of trainees to understand trainee performance (8,9). At our institution, a previous study investigated the distribution of individual discrepancy rates across all residents (10). Although many factors contribute to trainee discrepancies, rates greater than anticipated, falling above 1.5% for radiographs and 4.0% for computed tomographies (CTs) (approximately two standard deviations above the mean), are considered outliers and indicate the potential need for intervention to ensure the resident performs at a level adequate for training (11–13). These values are in line with discrepancy rates reported elsewhere (14–17). To ensure adequate sampling of an individual resident, the cutoffs were applied only after residents had interpreted greater than 200 radiographs and 50 CTs in total.

However, wide variation in attending radiologist practice patterns may also contribute to the distribution of major discrepancies, the extent of which is currently unknown. There is a possibility that this metric is more reflective of which particular attending radiologist is responsible for overreading studies on a given shift than the resident themselves. This is particularly concerning for study types that occur infrequently and for which annual interpreted volume is typically low for individual residents.

Understanding the contribution of dynamic variables such as attending practice behavior and scheduling requires probabilistic modeling. Monte Carlo simulations are a common

Acad Radiol 2017; 24:694–699

From the Department of Radiology, Hospital of the University of Pennsylvania, 3400 Spruce St., Philadelphia, PA 19104. Received April 5, 2016; revised November 23, 2016; accepted December 1, 2016. **Address correspondence to:** J.C.W. e-mail: joseph.wildenberg@uphs.upenn.edu

http://dx.doi.org/10.1016/j.acra.2016.12.004

method to explore the magnitude and distribution of probabilistic effects (18). This procedure, which involves simulating a system many thousands of times, builds an expected probability distribution of the simulated system for comparison to actual observations. Understanding the impact of attending behavior is necessary to determine if discrepancy rates are an accurate metric of resident performance.

To address these issues, we first analyze resident discrepancy rates and compare these to prior reports while measuring the variability in attending overread behavior. We then simulate the performance of an average resident to understand how the magnitude of attending behavioral variability impacts resident discrepancy rates. Finally, we propose updated benchmarks incorporating these data to ensure the benchmarks accurately identify residents performing below expectations.

MATERIALS AND METHODS

Institutional review board exemption was granted for the data collected in this study.

Resident Call Structure

Trainees' interpretations of evening and weekend examinations on call were evaluated at an academic program covering two large hospitals with busy emergency departments, one of which is a Level-I Trauma Center. Subspecialty attending radiologists provide immediate feedback and final review of trainee reports during regular business hours (weekdays from 7:30 AM to 5:00 PM). During evenings and weekends on-call trainees produce preliminary reports, whereas subspecialty attending radiologists are available for consultation from home. All trainees, beginning with the postgraduate year (PGY)-3 year, generate full-length preliminary reports for all emergency department and inpatient modalities other than magnetic resonance imaging and neurologic CT angiograms, including both stat and routine imaging. Preliminary interpretations for those studies are produced by radiology fellows.

Residents at our institution begin to take independent call at the start of the PGY-3 year, with two to three trainees on call at any given time. The majority of call shifts, divided into overnight and weekend duties, are covered by PGY-3 residents. Two limited afternoon or evening shifts stationed at the Level-I Trauma Center are covered by PGY-4 and PGY-5 residents. Throughout training all residents cover an approximately equal number of each shift.

Preliminary Report Overreading

All resident reports are reviewed by subspecialty attending radiologists the following morning, including on weekends. Attestation macros issued by attending radiologists include one of five grades of discrepancy ratings when reviewing a call study. "Great Call," "Agree," and "Addition" indicate general agreement with the resident's interpretation, possibly with small clarifications. A "minor" discrepancy is used when a non-critical finding, unlikely to affect patient care, was missed or a slight modification of a pertinent finding's interpretation is made in the final report. Examples would include a missed definitively benign lesion such as a liver hemangioma, the addition of a possible diagnosis to the differential, or a dictation error claiming that a surgically absent organ appears normal.

A "major" discrepancy is used when an interpretation error may cause significant clinical impact or if the error demonstrates performance clearly below expectation for the level of training. Although some studies have described consensus interpretations between radiologists well below 100% (19,20), for the purposes of this study the interpretation of the attending is considered correct. Note that no reference list of findings which definitively constitute a major discrepancy exists, and the variation in usage between attending radiologists is a focus of this paper.

Discrepancy Data Analysis

All chest radiographs, musculoskeletal (MSK) radiographs, chest CTs, abdomen and pelvis CTs, and head CTs interpreted by PGY-3 residents in a call situation between July 1, 2010 and June 30, 2015 were reviewed for the presence of a major discrepancy. The standardized nature of the attestation macros allowed this process to be automated with simple report parsing. A web-based case log, updated daily, was already in place, and the volume and discrepancy data were extracted from the database back-end (6). The bulk of the call burden at our institution falls on PGY-3 residents, and only studies interpreted by these residents were analyzed to eliminate variability caused by level of training. Angiographic CTs, other than those performed for the evaluation of pulmonary embolism, were not included secondary to the relatively low per-resident volume. Ultrasound studies were excluded because the performing technologist may influence the trainee's interpretation.

All calculations were performed using the R statistical package (21). Summary statistics were compiled by study type for every attending to determine their individual propensity for issuing a major discrepancy as well as the proportional volume of each study type they overread. Summary statistics were also calculated for residents to determine the average individual major discrepancy rates and the overall average volume interpreted for each study type. Residents were excluded from analysis for a particular study type if they had not interpreted sufficient examinations (200 for radiographs or 50 for CTs). Correlation between measured resident major discrepancy rates and individual volume interpreted for each study type was calculated using Pearson product moment correlation.

Simulations

Simulation of an "average resident" was performed for each study type using a Markov chain Monte Carlo algorithm known as Gibbs sampling (22). Specifically, simulation of a single interpretation consisted of randomly assigning the study to an attending weighted by relative attending overread volumes. After that assignment, a major discrepancy was randomly issued

based on that attending's major discrepancy issuance rate. This simulation algorithm was repeated for the average interpreted volume of that study type to generate a single instance of an average resident. For example, a single instance of an average resident interpreting chest radiographs includes 1157 simulated studies.

This procedure was repeated 100,000 times per study type to calculate a distribution of the average resident's expected discrepancy rate that only reflects attending variability. A binomial curve was fit to each simulated distribution.

Updated Benchmarks

The resulting distributions from the simulations were compared to the measured resident discrepancy rates for each study type. By integrating the area underneath the binomial curve, an updated benchmark was generated for each study type at the 95th percentile.

RESULTS

Measured Discrepancy Data

From a total of 42 residents spanning the five academic years, 18,599 imaging reports were included. The average resident interpretation volume, major discrepancy rate, and correlation of discrepancy rate and interpretation volume, separated by study type, are listed in Table 1. Most study types had nearly all residents meeting the minimum interpreted volume for analysis. Chest CT was the exception, with 33 out of the 42 meeting the minimum volume.

There was a statistically significant inverse relationship between interpretation volume and major discrepancy rate for all study types except chest CTs. In other words, the more cases of a given study type residents interpreted, the lower their major discrepancy rate.

Large variability was seen in all study types for both attending radiologist proportional overread volumes and discrepancy issuance rates (Fig 1). Note that only the five attendings with the largest proportional overread volumes are displayed in Figure 1, although all contributing attendings were included in the simulation. As an example, for chest CTs, the attending with the highest proportional overread volume of 18.3% had a major discrepancy issuance rate of 4.7%, whereas the attending with the next highest volume of 13.6% had a major discrepancy issuance rate of 0.4%.

Simulation Data

Simulations show that the mean simulated discrepancy rates were nearly equal to the resident discrepancy rates, as would be expected given that the discrepancy distributions used in the simulation were derived from the resident data. That the mean simulated rates are slightly lower is a reflection of the correlation between discrepancy rate and resident volume, which is not modeled in the simulation.

Plots of the actual resident and simulated distributions show large overlap of the bulk of the distributions, with several residents falling above current benchmarks as well as the 95th percentile of the simulated results (Fig 2). The current benchmarks of 1.5% for radiographs and 4.0% for CTs matched well with the 95th percentile for chest radiographs, and are slightly lenient for abdomen and pelvis CTs and head CTs. For MSK radiographs and chest CTs, however, the current benchmarks fall near the middle of the distributions.

DISCUSSION

Average resident major discrepancy rates were in line with expectations based on the prior study from our institution used to generate the original benchmarks as well as other published reports, although other recent work has suggested underreporting of significant discrepancies (10,14–16,23). All study types except chest CT showed an inverse relationship between resident volume and discrepancy rate. This finding could reflect more comfort with a particular study type with increased practice, a correlation between resident interpretation speed and accuracy, or a combination of the two.

There was wide variability among the practice patterns of attending radiologists, both in the proportional volume of studies they overread and the rate at which they issued major discrepancies. This variability holds true for the five study types investigated.

The Monte Carlo simulations reveal the extent to which attending practice variations impact the discrepancy rate variability seen among residents. Figure 2 demonstrates that there

TABLE 1. Volume and Discrepancy Rate Information for PGY-3 Residents. The Correlation of the Resident Volume and Resident Discrepancy Rate Was Performed Using Pearson Product Moment Correlation

Study Type	Residents Meeting Minimum Volume	Average Volume	Resident Discrepancy Rate	Volume/Rate Correlation	Volume/Rate P Value
Chest radiograph	41	1157 ± 523	1.11 ± 0.82%	–0.651	$P < .0001$
MSK radiograph	38	474 ± 130	1.73 ± 1.32%	–0.613	$P < .0001$
Chest CT	33	103 ± 31	3.01 ± 2.31%	–0.297	$P = .093$
Abdominal and pelvis CT	42	223 ± 78	2.13 ± 1.26%	–0.381	$P = .013$
Head CT	42	169 ± 58	1.86 ± 1.33%	–0.354	$P = .021$

CT, computed tomography; PGY, postgraduate years.

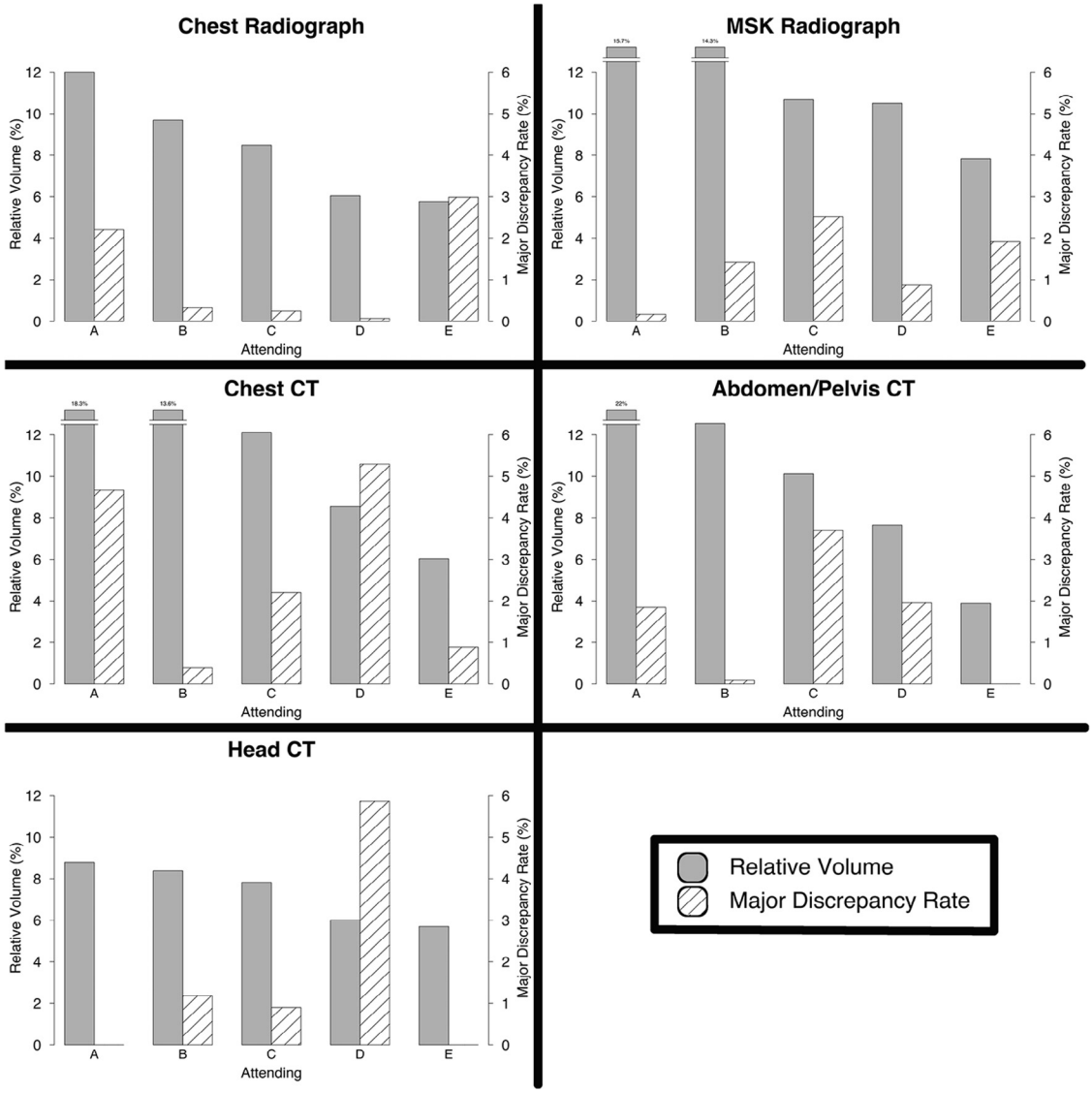

Figure 1. Attending radiologists' relative overread volume and propensity to issue a major discrepancy, separated by study type. Attending radiologists are sorted by relative volume within each study. High percentages that map off the scale of the graph are represented by a break in the bar with the value printed at the top. Note that only the top five radiologists by volume are shown for each study, and the letter designations are separate for each study type.

is extensive overlap between the actual resident distributions and the simulated distributions for all study types. These findings imply that individual discrepancy rates exhibit substantial uncertainty because of attending variability. Importantly, if discrepancy rates are used to compare the performance of two residents, this uncertainty will often be larger than the difference in rates. Major discrepancy rates, as currently implemented, are not an appropriate method to rank the performance of individual residents.

Although the simulations argue against direct comparison for near-average residents using discrepancy rates, they also provide an upper bound for the magnitude of this uncertainty and suggest these rates can be used to flag performance below expectations. The benchmarks, set at the 95th percentile for each study type, match well with the criteria reported in the prior study for chest radiographs, abdomen and pelvis CTs, and head CTs.

For MSK radiographs and chest CTs, however, the current benchmarks reside in the middle of the simulated distribution, and for MSK radiographs below the mean. As shown in Figure 2 and listed in Table 2, a more appropriate benchmark for these specific study types would be 2.5% for MSK radiographs and 5.8% for chest CTs. The original benchmarks were calculated by modality without sub-analysis by specialty so the findings here do not necessarily contradict what was reported previously (10).

The large effect of attending practice variability on resident discrepancy rates suggests that the benchmark rates reported here likely need to be updated when applied to other institutions, and even over time within each institution because

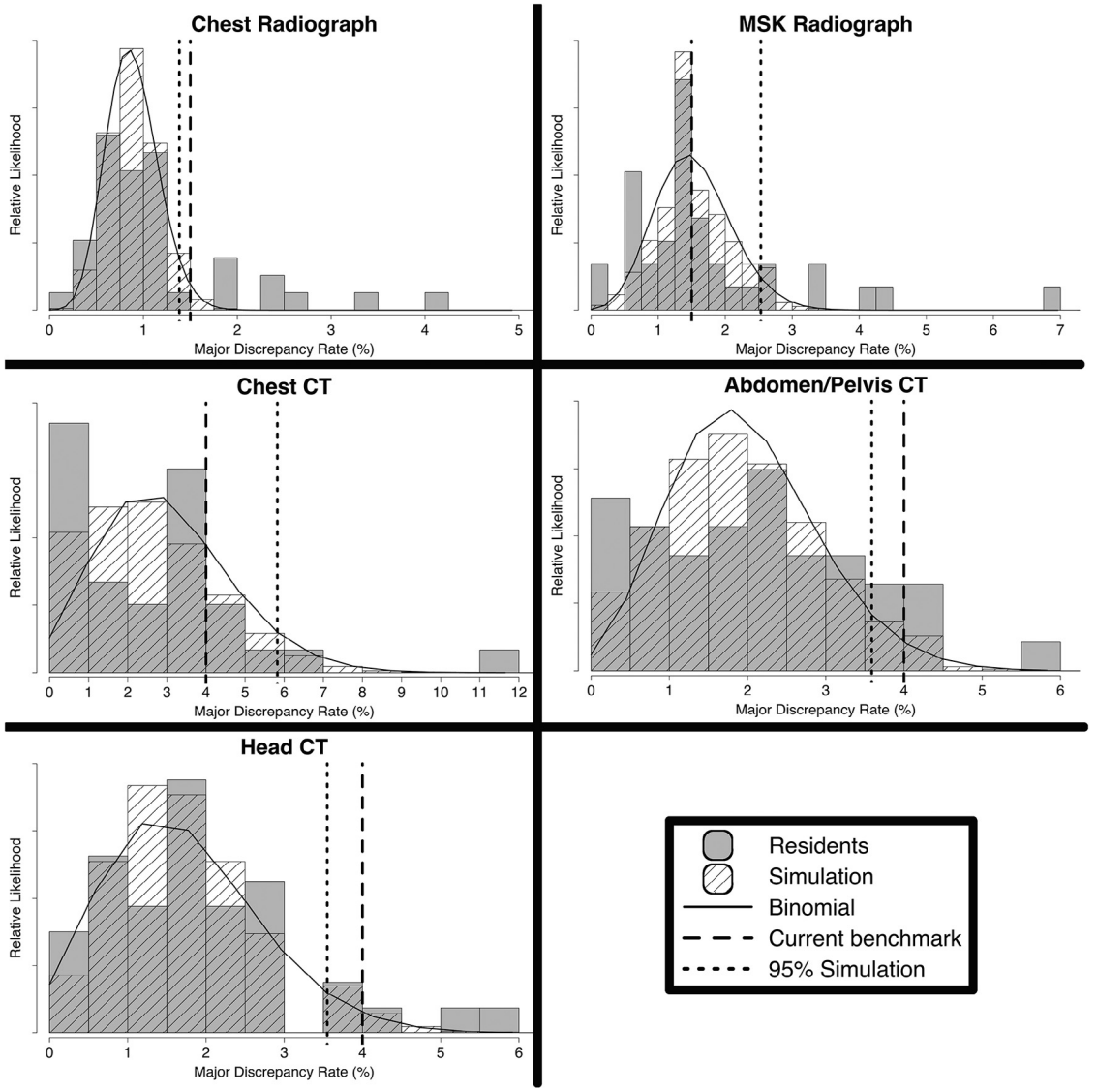

Figure 2. Histograms of the actual resident discrepancy rates (gray) and the simulated distributions (diagonal lines) for each study type. The black line represents a fitted binomial curve for the simulated data. The dashed line shows the current benchmark, whereas the dotted line shows the suggested benchmark at two standard deviations above the mean of the simulated data. The relative likelihood is a scaled measure such that the summed height of the bars in each group total 100%.

TABLE 2. Discrepancy Rate Information for Residents and the Simulated Distribution. The 95th Percentile of the Binomial Distribution Was Calculated from a Fit of the Simulated Data

Study Type	Resident Discrepancy Rate %)	Simulated Discrepancy Rate (%)	Current Benchmark (%)	Binomial 95th Percentile (%)
Chest radiograph	1.11 ± 0.82	0.88 ± 0.28	1.5	1.38
MSK radiograph	1.73 ± 1.32	1.53 ± 0.56	1.5	2.53
Chest CT	3.01 ± 2.31	2.96 ± 1.67	4.0	5.83
Abdominal and pelvis CT	2.13 ± 1.26	1.97 ± 0.93	4.0	3.59
Head CT	1.86 ± 1.33	1.71 ± 1.00	4.0	3.55

CT, computed tomography; MSK, musculoskeletal.

of changes in staffing and scheduling. Future directions could include individual expectations calculated separately for each resident based on their study volume and adjusted for the attending radiologists who overread those studies.

CONCLUSIONS

Resident perception of wide variability in attending practice patterns is accurate, and simulations suggest that major discrepancy rates should not be used to directly compare resident performance. However, this metric can be used to flag outliers whose performance is below expectations.

Globally, measured resident discrepancy rates are in line with prior studies, although sub-analysis by study type has allowed calculation of study-specific statistics not previously reported. Benchmarks calculated from the simulations provide modality and subspecialty-specific cutoffs to identify residents who may benefit from additional attention or remediation.

REFERENCES

1. Schmitt JE, Scanlon MH, Servaes S, et al. Milestones on a shoestring: a cost-effective, semi-automated implementation of the new ACGME requirements for radiology. Acad Radiol 2015; 22:1287–1293. doi:10.1016/j.acra.2015.02.013.
2. Carraccio C, Englander R, Holmboe ES, et al. Driving care quality: aligning trainee assessment and supervision through practical application of entrustable professional activities, competencies, and milestones. Acad Med 2015; 91:199–203. doi:10.1097/ACM.0000000000000985.
3. Cooney CM, Cooney DS, Bello RJ, et al. Comprehensive observations of resident evolution: a novel method for assessing procedure-based residency training. Plast Reconstr Surg 2016; 137:673–678. doi:10.1097/01.prs.0000475797.69478.0e.
4. Issa G, Taslakian B, Itani M, et al. The discrepancy rate between preliminary and official reports of emergency radiology studies: a performance indicator and quality improvement method. Acta Radiol 2015; 56:598–604. doi:10.1177/0284185114532922.
5. Itri JN, Kang HC, Krishnan S, et al. Using focused missed-case conferences to reduce discrepancies in musculoskeletal studies interpreted by residents on call. Am. J. Roentgenol. 2011; 197:W696–W705. doi:10.2214/AJR.11.6962.
6. Chen PH, Chen YJ, Cook TS. Capricorn—a web-based automatic case log and volume analytics for diagnostic radiology residents. Acad Radiol 2015; 22:1242–1251. doi:10.1016/j.acra.2015.06.011.
7. Kalaria AD, Filice RW. Comparison-Bot: an automated preliminary-final report comparison system. J Digit Imaging 2016; 29:325–330. doi:10.1007/s10278-015-9840-2.
8. Platon A, Becker M, Perneger T, et al. Emergency computed tomography: what is missed at first reading? J Comput Assist Tomogr 2016; 40:177–182. doi:10.1097/RCT.0000000000000317.
9. Mellnick V, Raptis C, McWilliams S, et al. On-call radiology resident discrepancies: categorization by patient location and severity. J Am Coll Radiol 2016; 13:1233–1238. doi:10.1016/j.jacr.2016.04.020.
10. 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; 8:644–648. doi:10.1016/j.jacr.2011.04.003.
11. Sistrom C, 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; 15:934–941. doi:10.1016/j.acra.2008.02.013.
12. 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; 20:305–311. doi:10.1016/j.acra.2012.09.028.
13. Bruni SG, Bartlett E, Yu E. Factors involved in discrepant preliminary radiology resident interpretations of neuroradiological imaging studies: a retrospective analysis. AJR Am J Roentgenol 2012; 198:1367–1374. doi:10.2214/AJR.11.7525.
14. Walls J, Hunter N, Brasher PMA, et al. The DePICTORS Study: discrepancies in preliminary interpretation of CT scans between on-call residents and staff. Emerg Radiol 2009; 16:303–308. doi:10.1007/s10140-009-0795-9.
15. Cooper VF, Goodhartz LA, Nemcek AA, et al. Radiology resident interpretations of on-call imaging studies: the incidence of major discrepancies. Acad Radiol 2008; 15:1198–1204. doi:10.1016/j.acra.2008.02.011.
16. Huntley JH, Carone M, Yousem DM, et al. Opportunities for targeted education: critical neuroradiologic findings missed or misinterpreted by residents and fellows. Am. J. Roentgenol. 2015; 205:1155–1159. doi:10.2214/AJR.15.14905.
17. Ruma J, Klein KA, Chong S, et al. Cross-sectional examination interpretation discrepancies between on-call diagnostic radiology residents and subspecialty faculty radiologists: analysis by imaging modality and subspecialty. J Am Coll Radiol 2011; 8:409–414. doi:10.1016/j.jacr.2011.01.012.
18. Harrison RL. Introduction to Monte Carlo simulation. AIP Conf Proc 2010; 1204:17–21. doi:10.1063/1.3295638.
19. Pow RE, Mello-Thoms C, Brennan P. Evaluation of the effect of double reporting on test accuracy in screening and diagnostic imaging studies: a review of the evidence. J Med Imaging Radiat Oncol 2016; 60:306–314. doi:10.1111/1754-9485.12450.
20. Wu MZ, McInnes MDF, Macdonald DB, et al. CT in adults: systematic review and meta-analysis of interpretation discrepancy rates. Radiology 2014; 270:717–735. doi:10.1148/radiol.13131114.
21. R Core Team. R: A language and environment for statistical computing. Vienna, Austria.: R Foundation for Statistical Computing, 2016. Available at: https://www.R-project.org/. Accessed January 18, 2017.
22. Hines KE. A primer on Bayesian inference for biophysical systems. Biophys J 2015; 108:2103–2113. doi:10.1016/j.bpj.2015.03.042.
23. Brown JM, Dickerson EC, Rabinowitz LC, et al. "Concordance" revisited: a multispecialty appraisal of "concordant" preliminary abdominopelvic CT reports. J Am Coll Radiol 2016; 13:1111–1117. doi:10.1016/j.jacr.2016.04.019.