Can USMLE and COMLEX-USA Scores Predict At-Risk Emergency Medicine Residents' Performance on In-Training Examinations?

Michael C. Plewa [1], David J. Ledrick [1], Kenneth Jenkins [2], Aaron Orqvist [1], Michael McCrea [1]

1. Emergency Medicine, Mercy Health - St. Vincent Medical Center, Toledo, USA 2. Emergency Medicine, Ohio University Heritage College of Osteopathic Medicine, Athens, USA

Corresponding author: Michael C. Plewa, plewamc@gmail.com

Abstract

Purpose: The United States Medical Licensing Examination (USMLE) and Comprehensive Osteopathic Medical Licensing Examination (COMLEX) scores are standard methods used to determine residency candidates' medical knowledge. The authors were interested in using the USMLE and COMLEX part 2 scores in our emergency medicine (EM) residency program to identify at-risk residents who may have difficulty on the in-training exam (ITE) and to determine the cutoff values under which an intern could be given an individualized study plan to ensure medical knowledge competency.

Methods: The authors abstracted the USMLE and COMLEX part 2 scores and the American Board of Emergency Medicine (ABEM) ITE scores for a cohort of first-year EM residents graduating years 2010-2022, converting raw scores to percentiles, and compared part 2 and ABEM ITE scores with Pearson's correlation, a Bland-Altman analysis of bias and 95% limits of agreement, and ROC analysis to determine optimal the cut-off values for predicting ABEM ITE < 50th percentile and the estimated test characteristics.

Results: Scores were available for 152 residents, including 93 USMLE and 88 COMLEX exams. The correlations between part 2 scores and ABEM ITE were r = 0.36 (95%CI: 0.17, 0.52; p < 0.001) for USMLE and r = 0.50 (95%CI: 0.33, 0.64; p < 0.001) for COMLEX. Bias and limits of agreement for both part 2 scores were -14 ± 63% for USMLE and 13 ± 50% for COMLEX in predicting the ABEM ITE scores. USMLE < 37th percentile and COMLEX < 53rd percentile identified 42% (N = 39) and 27% (N = 24) of EM residents, respectively, as at risk, with a sensitivity of 61% and 49% and specificity of 71% and 92%, respectively.

Conclusion: USMLE and COMLEX part 2 scores have a very limited role in identifying those at risk of low ITE performance, suggesting that other factors should be considered to identify interns in need of medical knowledge remediation.

Introduction

Emergency medicine (EM) program directors and core faculty are interested in identifying residents at risk of poor academic performance early in their training, preferably early in the first post-graduate year (PGY1). The Council of Emergency Medicine Residency Directors (CORD) concluded that successful remediation is dependent on early identification of poor performance [1] and created both a remediation task force and a remediation consult service to address these issues [2]. A 2014 survey [3] found that 90% of EM programs had at least one resident undergoing remediation, with medical knowledge as the most common indication for remediation.

Performance on the American Board of Emergency Medicine (ABEM) in-training exam (ITE), which also predicts EM board certification success [4], is one method to assess academic progress and often the first indicator of identifying residents with knowledge deficits. Unfortunately, these results are not available until two months before the end of the PGY1 year. Aldeen et al. [5,6] demonstrated that EM faculty are only moderately able to predict resident performance on the ITE, with less accuracy for interns versus non-interns.

One of the primary factors in the residency selection process is a candidate's performance on one of the licensure exams, either the United States Medical Licensing Examination (USMLE) or the Comprehensive Osteopathic Medical Licensing Examination - USA (COMLEX) [7]. Program directors often embrace the adage that the "best predictor of future performance is past performance" as licensure exam scores have been shown to correlate with ABEM ITE exam scores [8,9] and predict passing the ABEM qualifying exam

following residency [10-13]. However, the relative importance of USMLE scores when EM program directors determine their applicant rank list has varied in published reports [14-16]. Multiple other specialties, including family medicine [17,18], general surgery [19], internal medicine [20,21], internal medicine/pediatrics [22], and obstetrics [23], have all examined the correlations between either licensure examinations or specialty ITE with specialty board certification examinations, with varying degrees of prediction [24].

Our primary goal was to determine the predictive value of the USMLE and COMLEX part 2 exam scores for the ABEM ITE scores for PGY1 residents in our EM program. Secondly, we sought to determine cutoff values from USMLE and COMLEX for predicting an ABEM ITE score below the 50th percentile to identify a PGY1 resident who would benefit from a medical knowledge remediation program before receiving the results of the ITE in the spring of the PGY1.

Materials And Methods

This was an IRB-approved (Mercy Health North Institutional Review Board 2020-5) retrospective analysis of all EM residents who graduated from 2010 to 2022 with complete data. Our program has 14 residents per year in a community urban teaching center. Although our residency historically participated only in the National Resident Matching Program, our hospital is also a core clinical training site for multiple osteopathic medical schools. As such, we have always interviewed and matched both allopathic and osteopathic medical students. Given that many osteopathic residents take both the COMLEX and USMLE exams, our program considers both exams.

Deidentified data from residency files were abstracted by a research assistant and reviewed by two of the authors to ensure a complete dataset that was corrected. We converted all part 2 USMLE, part 2 COMLEX, and the ABEM ITE exam scores into a percentile ranking appropriate for the year the test was taken. All data were entered into an Excel spreadsheet for analysis by a biostatistician.

Data were analyzed for association and agreement using several different approaches. For association, a Pearson correlation coefficient between the COMLEX or USMLE scores and the ABEM ITE scores was calculated with MedCalc® Statistical software (version 19.7.2; MedCalc Software Ltd., Ostend, Belgium). For agreement, we calculated bias and limits of agreement according to the methods of Bland et al. [25]. Bias is defined as the mean of the difference between scores, and the limits of agreement were defined as bias ± two standard deviations [25]. Additionally, a receiver operating curve (ROC) analysis was used to identify the optimal cut points for USMLE step 2 and COMLEX level 2 in predicting an ITE score below the 50th percentile for the year of training. Using these cutoff points, we calculated the percentage of EM1 residents at risk, the test characteristics, and the associated positive and negative likelihood ratios (LR) and odds ratios.

Results

Complete data were available for 152 of 165 EM residents, including 93 USMLE and 88 COMLEX exams. Seven (3.8%) were excluded for missing data, including two ABEM ITE, one USMLE, and four COMLEX exams. The correlations between part two scores and ABEM ITE were r = 0.36 (95%CI: 0.17, 0.52; p < 0.001) for USMLE part 2 and r = 0.50 (95%CI: 0.33, 0.64; p < 0.001) for COMLEX part 2.

The mean of the differences (bias) between USMLE or COMLEX and ABEM ITE was -14.4 percentile points and 13.0 percentile points, respectively. The limits of agreement were ± 63 percentile points for USMLE part 2 and ± 50 percentile points for COMLEX part 2, in predicting the ABEM ITE.

ROC analysis identified cut-points of USMLE < 37th percentile and COMLEX < 53rd percentile to optimize the prediction of ABEM ITE < 50th percentile (Figure 1). These cut-points identified 42% (N = 39) and 27% (N = 24) of EM residents, respectively, as being at risk, with positive predictive values of 95.8% and 66.7% and negative predictive values of 59.4% and 70.3%, respectively (Table 1).
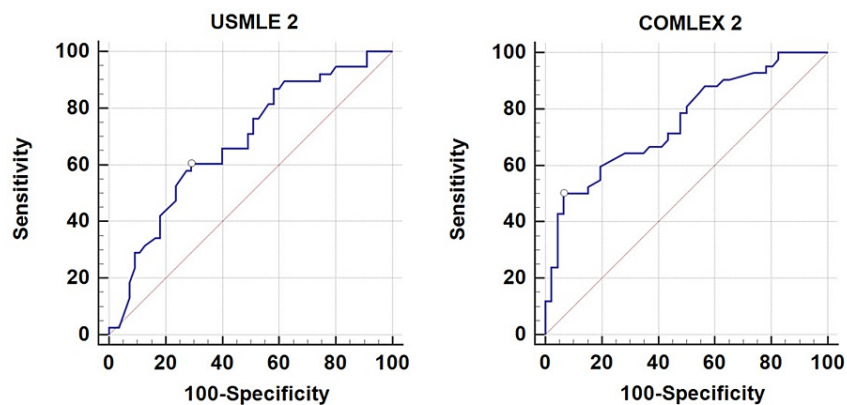
FIGURE 1: ROC curves for USMLE and COMLEX part 2 predicting ABEM ITE scores < 50%

ABEM ITE: American Board of Emergency Medicine In-Training Exam; COMLEX-USA: Comprehensive Osteopathic Medical Licensing Examination - USA; ROC: receiver operating curve; USMLE: United States Medical Licensing Examination

Level 2 Exam	USMLE (N = 93)	COMLEX (N = 88)
Correlation with ABEM ITE	r = 0.36 (95%CI: 0.17, 0.52; p < 0.001)	r = 0.50 (95%CI: 0.33, 0.64; p < 0.001)
Mean of differences (bias)	- 14.4 percentile points	13.0 percentile points
Limits of agreement	± 63 percentile points	± 50 percentile points
Cut-point to predict residents at risk of ABEM ITE < 50th percentile†	< 37th percentile	< 53rd percentile
Area under ROC curve	0.68 (95% CI: 0.57, 0.77)	0.73 (95% CI: 0.63, 0.82)
No. residents identified as at-risk	39 (42%)	24 (27%)
Sensitivity	60.5%	48.8%
Specificity	70.9%	91.5%
Positive predictive value	95.8%	66.7%
Negative predictive value	59.4%	70.3%
Odds ratio for low ABEM using cut-point score	3.74 (95% CI: 1.56, 8.95)	14.33 (95% CI: 3.84, 53.52)

TABLE 1: Predictive ability of USMLE and COMLEX scores to identify PGY1 residents at risk of scoring low on the ABEM ITE

† A cut-point score was determined by the receiver operating characteristic methodology to maximize the sum of sensitivity and specificity to predict ABEM ITE < 50th percentile. The data have been represented as N, signifying the number per group, and percentiles. Values of p < 0.05 are considered significant.

ABEM ITE: American Board of Emergency Medicine In-Training Exam; COMLEX-USA: Comprehensive Osteopathic Medical Licensing Examination - USA; PGY1: First post-graduate year; USMLE: United States Medical Licensing Examination

Discussion

Our study was unique in several aspects. First, we used percentile scores instead of raw scores for each of the USMLE, COMLEX, and ABEM ITE exams, allowing comparison between exams and different examination years. Second, we assessed the relationship of the COMLEX with the ABEM ITE in 57.8% of our cohort, in addition to USMLE, which has been previously reported [9,10]. Last, we assessed associations with correlation coefficients, and agreement with the Bland-Altman analysis [25], between part two scores and ABEM ITE. Our findings confirm an association between part two exams and ABEM ITE scores but indicate

that these correlations were moderate (r = 0.36 for USMLE and r = 0.50 for COMLEX), where correlation values > 0.50 are considered strong, 0.30-0.49 are considered moderate, and < 0.30 is considered low [26]. Correlations were lower in our cohort than those described by Hiller et al. (r = 0.58) [10] and Thundiyil et al. (r = 0.76) [13]. The Bland-Altman analysis identified bias, with USMLE on average 14.4 percentile points lower and COMLEX on average 13 percentile points higher than the ABEM ITE scores. The limits of agreement were wide for both USMLE and COMLEX exams for ABEM ITE. Within two SDs, the USMLE part 2 would underestimate or overestimate ABEM ITE by as much as 77.3 and 48.5 percentile points, respectively. COMLEX part 2 would underestimate or overestimate ABEM ITE by as much as 37 and 63 percentile points, respectively.

Significant ORs for both USMLE part 2 and COMLEX part 2 again validate the association with the ABEM ITE scores but do not assess their predictive nature. Instead, the LR were not very useful, with only a slight to moderate effect on the probability of poor ABEM ITE performance. The LR positive for both USMLE and COMLEX were low (2.08 and 5.73, respectively), whereas the LR negative for both exams was large (0.56, for both).

Using our USMLE and COMLEX cut-offs would have resulted in large percentages (27-42%) of our PGY1 class undergoing medical knowledge remediation. By comparison, the 2014 survey of 160 allopathic EM programs found only 4.4% of EM residents in remediation [3]. Having approximately one-third of PGY1 residents in an individualized remediation program might be considered a significant burden since a tailored curriculum specific to the learner might include assigning a faculty mentor, assigned readings and exams, and an altered schedule allowing for more study early in the program.

Other methods should be considered in identifying interns with medical knowledge deficits. Although Aldeen et al. found an overall 60% accuracy, they admit that only 23% of faculty could correctly identify those residents scoring significantly below their class mean [5]. Goyal et al. discussed a variety of medical knowledge assessment tools, including USMLE, ABEM ITE, question bank scores, performance on virtual or simulation scenarios, oral exams, direct observation of skills, and objective structured clinical exams, among others [27].

This study had several limitations. Our sample was from a single program, and the results may not apply to programs with different selection or advancement criteria. Our sample size was modest but similar in size to that of Thundiyil et al. [13] and larger than that of Hiller et al. [10]. We also did not have complete data on 13 of 265 residents, though there was no clear pattern to the missing numbers. We did not include USMLE step 1 or COMLEX level 1 exam scores since these are now pass/fail as of 2022 and will do little to aid programs in assessing performance. We did not include part three exams since these are usually taken during the PGY1 and therefore irrelevant to our purpose.

Conclusions

The USMLE and COMLEX exams are among the most standardized and universally applied measurements of medical knowledge competency. Our results concur with prior studies in that the USMLE and COMLEX part 2 exam scores do correlate with performance on ABEM ITE scores. However, we find this association is imperfect and demonstrates substantial bias and inadequate agreement. Because these scores have poor predictive value, and likely overestimate the number of interns in need of early academic intervention, programs should consider using alternative evaluation methods to be more specific in identifying those requiring additional resources.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Acquisition, analysis, or interpretation of data: Michael C. Plewa, David J. Ledrick, Kenneth Jenkins, Aaron Orqvist, Michael McCrea

Drafting of the manuscript: Michael C. Plewa, David J. Ledrick, Michael McCrea

Critical review of the manuscript for important intellectual content: Michael C. Plewa, David J. Ledrick, Kenneth Jenkins, Aaron Orqvist, Michael McCrea

Concept and design: David J. Ledrick, Michael McCrea

Supervision: David J. Ledrick

Disclosures

References

1. Katz ED, Dahms R, Sadosty AT, Stahmer SA, Goyal D: Guiding principles for resident remediation: recommendations of the CORD remediation task force. Acad Emerg Med. 2010, 17 Suppl 2:S95-103. 10.1111/j.1553-2712.2010.00881.x
2. Egan DJ, Gentges J, Regan L, Smith JL, Williamson K, Murano T: An emergency medicine remediation consult service: access to expert remediation advice and resources. AEM Educ Train. 2019, 3:193-6. 10.1002/aet2.10330
3. Silverberg M, Weizberg M, Murano T, Smith JL, Burkhardt JC, Santen SA: What is the prevalence and success of remediation of emergency medicine residents?. West J Emerg Med. 2015, 16:839-44. 10.5811/westjem.2015.9.27357
4. Frederick RC, Hafner JW, Schaefer TJ, Aldag JC: Outcome measures for emergency medicine residency graduates: do measures of academic and clinical performance during residency training correlate with American Board of Emergency Medicine test performance?. Acad Emerg Med. 2011, 18 Suppl 2:S59-64. 10.1111/j.1553-2712.2011.01116.x
5. Aldeen AZ, Quattromani EN, Williamson K, Hartman ND, Wheaton NB, Branzetti JB: Faculty prediction of in-training examination scores of emergency medicine residents: a multicenter study. J Emerg Med. 2015, 49:64-9. 10.1016/j.jemermed.2015.01.015
6. Aldeen AZ, Salzman DH, Gisondi MA, Courtney DM: Faculty prediction of in-training examination scores of emergency medicine residents. J Emerg Med. 2014, 46:390-5. 10.1016/j.jemermed.2013.08.047
7. Sandella JM, Gimpel JR, Smith LL, Boulet JR: The use of COMLEX-USA and USMLE for residency applicant selection. J Grad Med Educ. 2016, 8:358-63. 10.4300/JGME-D-15-00246.1
8. Caffery T, Fredette J, Musso MW, Jones GN: Predicting American Board of Emergency Medicine qualifying examination passage using United States Medical Licensing Examination step scores. Ochsner J. 2018, 18:204-8. 10.31486/toj.17.0101
9. Harmouche E, Goyal N, Pinawin A, Nagarwala J, Bhat R: USMLE scores predict success in ABEM initial certification: a multicenter study. West J Emerg Med. 2017, 18:544-9. 10.5811/westjem.2016.12.32478
10. Hiller K, Franzen D, Heitz C, Emery M, Poznanski S: Correlation of the National Board of Medical Examiners emergency medicine advanced clinical examination given in July to intern American Board of Emergency Medicine in-training examination scores: a predictor of performance?. West J Emerg Med. 2015, 16:957-60. 10.5811/westjem.2015.9.27303
11. Levy D, Dvorkin R, Schwartz A, Zimmerman S, Li F: Correlation of the emergency medicine resident in-service examination with the American Osteopathic Board of Emergency Medicine Part I. West J Emerg Med. 2014, 15:45-50. 10.5811/westjem.2013.7.17904
12. Li F, Gimpel JR, Arenson E, Song H, Bates BP, Ludwin F: Relationship between COMLEX-USA scores and performance on the American Osteopathic Board of Emergency Medicine Part I certifying examination. J Am Osteopath Assoc. 2014, 114:260-6. 10.7556/jaoa.2014.051
13. Thundiyil JG, Modica RF, Silvestri S, Papa L: Do United States Medical Licensing Examination (USMLE) scores predict in-training test performance for emergency medicine residents?. J Emerg Med. 2010, 38:65-9. 10.1016/j.jemermed.2008.04.010
14. Breyer MJ, Sadosty A, Biros M: Factors affecting candidate placement on an emergency medicine residency program's rank order list. West J Emerg Med. 2012, 13:458-62. 10.5811/westjem.2011.1.6619
15. Kane KE, Yenser D, Weaver KR, et al.: Correlation between United States Medical Licensing Examination and Comprehensive Osteopathic Medical Licensing Examination scores for applicants to a dually approved emergency medicine residency. J Emerg Med. 2017, 52:216-22. 10.1016/j.jemermed.2016.06.060
16. Katzung KG, Ankel F, Clark M, et al.: What do program directors look for in an applicant? . J Emerg Med. 2019, 56:e95-e101. 10.1016/j.jemermed.2019.01.010
17. Hudson KM, Feinberg G, Hempstead L, Zipp C, Gimpel JR, Wang Y: Association between performance on COMLEX-USA and the American College of Osteopathic Family Physicians in-service examination. J Grad Med Educ. 2018, 10:543-7. 10.4300/JGME-D-17-00997.1
18. O'Neill TR, Peabody MR, Song H: The predictive validity of the National Board of Osteopathic Medical Examiners' COMLEX-USA Examinations with regard to outcomes on American Board of Family Medicine Examinations. Acad Med. 2016, 91:1568-75. 10.1097/ACM.0000000000001254
19. Sutton E, Richardson JD, Ziegler C, Bond J, Burke-Poole M, McMasters KM: Is USMLE Step 1 score a valid predictor of success in surgical residency?. Am J Surg. 2014, 208:1029-34. 10.1016/j.amjsurg.2014.06.032
20. Kay C, Jackson JL, Frank M: The relationship between internal medicine residency graduate performance on the ABIM certifying examination, yearly in-service training examinations, and the USMLE Step 1 examination. Acad Med. 2015, 90:100-4. 10.1097/ACM.0000000000000500

2024 Plewa et al. Cureus 16(4): e58684. DOI 10.7759/cureus.58684

5 of 6

21. Rayamajhi S, Dhakal P, Wang L, Rai MP, Shrotriya S: Do USMLE Steps, and ITE score predict the American Board of Internal Medicine certifying exam results?. BMC Med Educ. 2020, 20:79. 10.1186/s12909-020-1974-3

22. Ost SR, Wells D, Goedecke PJ, Tolley EA, Kleinman M, Thompson NS: Relationship between standardized test scores and board certification exams in a combined internal medicine/pediatrics residency program. Cureus. 2021, 13:e13567. 10.7759/cureus.13567

23. Lingenfelter BM, Jiang X, Schnatz PF, O'Sullivan DM, Minassian SS, Forstein DA: CREOG in-training examination results: contemporary use to predict ABOG written examination outcomes. J Grad Med Educ. 2016, 8:353-7. 10.4300/JGME-D-15-00408.1

24. McCrary HC, Colbert-Getz JM, Poss WB, Smith BK: A systematic review of the relationship between in-training examination scores and specialty board examination scores. J Grad Med Educ. 2021, 13:43-57. 10.4300/JGME-D-20-00111.1

25. Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986, 1:307-10.

26. Cohen J: A power primer. Psychol Bull. 1992, 112:155-9. 10.1037//0033-2909.112.1.155

27. Goyal N, Aldeen A, Leone K, Ilgen JS, Branzetti J, Kessler C: Assessing medical knowledge of emergency medicine residents. Acad Emerg Med. 2012, 19:1360-5. 10.1111/acem.12033

2024 Plewa et al. Cureus 16(4): e58684. DOI 10.7759/cureus.58684

6 of 6