ORIGINAL
ARTICLES

Identifying Students at Risk of Failing the USMLE Step 2 Clinical Skills Examination

Susan Rosenthal, MD; Stefani Russo, MD; Katherine Berg, MD; Joseph Majdan, MD; Jennifer Wilson, MS; Charlotte Grinberg, MD; Jon Veloski, MS

BACKGROUND AND OBJECTIVES: New standards announced in 2017 could increase the failure rate for Step 2 Clinical Skills (CS). The purpose of this study was to identify student performance metrics associated with risk of failing.

METHODS: Data for 1,041 graduates of one medical school from 2014 through 2017 were analyzed, including 30 (2.9%) failures. Metrics included Medical College Admission Test, United States Medical Licensing Examination Step 1, and clerkship National Board of Medical Examiners (NBME) Subject Examination scores; faculty ratings in six clerkships; and scores on an objective structured clinical examination (OSCE). Bivariate statistics and regression were used to estimate risk of failing.

RESULTS: Those failing had lower Step 1 scores, NBME scores, faculty ratings, and OSCE scores ($P<.02$). Students with four or more low ratings were more likely to fail compared to those with fewer low ratings (relative risk [RR], 12.76, $P<.0001$). Logistic regression revealed other risks: low surgery NBME scores (RR 3.75, $P=.02$), low pediatrics NBME scores (RR 3.67, $P=.02$), low ratings in internal medicine (RR 3.42, $P=.004$), and low OSCE Communication/Interpersonal Skills (RR 2.55, $P=.02$).

CONCLUSIONS: Certain medical student performance metrics are associated with risk of failing Step 2 CS. It is important to clarify these and advise students accordingly.

(Fam Med. 2019;51(6):483-8.)
doi: 10.22454/FamMed.2019.429968

Multiple studies have analyzed predictors of performance on the United States Medical Licensing Examination (USMLE) Step 1 and Step 2 Clinical Knowledge (CK) examinations.[1-3] However, very limited information is available regarding the risk factors associated with failing the Step 2 Clinical Skills (CS) examination.

An early study of one medical school's objective structured clinical examination (OSCE) conducted in collaboration with investigators from the National Board of Medical Examiners (NBME) reported low correlations between OSCE subtest scores and subtest scores on the Step 2 CS.[4] Subsequently, Dong et al reported evidence of a stronger association between another school's OSCE and Step 2 CS subtest scores. However, their stepwise linear regression analysis revealed that preclinical grades and NBME Subject Examination scores in clerkships were more important predictors of Step 2 CS performance than were OSCE scores.[5] Neither study directly addressed the question of the student risk factors associated with failure.

Historically, the failure rate on Step 2 CS was low. Annual first-attempt passing rates for North American medical students ranged from 91% to 97% between 2004 and 2016, according to the USMLE.[6] However, in 2017 the USMLE Management Committee announced an increase in the passing standards after the routine review of examinee performance that it performs every 3 to 4 years. The USMLE announcement noted that if the new standards had been applied to the scores of earlier examinees evaluated under the previous standards, the national passing rate would have been three percentage points lower.[7]

From the Office of Student Affairs and Career Counseling (Dr Rosenthal), the Rector Clinical Skills Simulation Center (Drs Russo, Berg, and Majdan), the Center for Teaching and Learning (Ms Wilson), and the Center for Research in Medical Education and Healthcare (Mr. Veloski), Sidney Kimmel Medical College at Thomas Jefferson University, Philadelphia, PA; and Mount Auburn Hospital, Cambridge, MA (Dr Grinberg).

Such policy changes raise concerns among many faculty members and students. Passing the Step 2 CS examination is required for licensure. Soon after its introduction in 2004 many schools introduced changes to their curriculum to better prepare students for the examination.[8] Furthermore, just 5 years after its introduction, the Step 2 CS examination was identified by residency program directors as the sixth most important criterion in their process for selecting residents. They ranked it more important than medical school class rank, membership in A O A, and medical school research experience.[9] Students who fail and need to repeat the exam must pay a second fee, and many also face significant additional expenses for travel and lodging because of the limited number of testing sites.[10]

We designed this study to determine which student assessments routinely collected in the first 3 years of medical school can be used to estimate students' risk of failing the Step 2 CS examination.

Methods

The sample included all 1,041 members of the graduating classes of 2014 through 2017 at Sidney Kimmel Medical College at Thomas Jefferson University. We used data extracted from the Jefferson Longitudinal Study of Medical Education.[11] The University's institutional review board approved the data collection and its use in this type of study.

We evaluated the following: Medical College Admission Test (MCAT) scores, USMLE Step 1 scores, NBME Subject Examination scores in six major clerkships (family medicine, internal medicine, obstetrics/gynecology, pediatrics, psychiatry, surgery), subtest scores on an OSCE with standardized patients administered at the end of the third year,[4] and faculty ratings of clinical clerkship performance on a 4-point scale (Honors, Excellent, Good, Marginal).

The six clerkships used the same rating form, which included a mix of objective rating items and prompts for subjective comments. The clerkship directors decided which faculty members and residents would be assigned to rate students. The directors managed the process of integrating the individual ratings and comments from multiple raters into final global faculty ratings for the students in the clerkships. The specific process for final grading and associated standards varied across clerkships and hospitals. The ob/gyn and surgery clerkships also included a one-half day final OSCE, and the scores were included in the determination of the final grade.

We established statistical significance at 0.05. We used independent t-tests and χ^2 tests of independence to assess the bivariate descriptive statistics comparing the students who failed with the students who passed. We used logistic regression analysis to formulate a multivariate model that could be used to estimate a student's risk of failing. Calculations were performed using Stata SE 14.2 for Windows (Stata Corp, College Station, Texas).

Results

Table 1 shows the means for key performance metrics available prior to the exam for 30 (2.9%) students who failed the Step 2 CS examination in comparison to 1,011 (97.1%) who passed.

The Step 1 mean for students who failed was significantly lower ($P<.003$). The means for the six NBME clerkship examinations were also significantly lower for those who failed ($P<.01$). The mean scores for Data Gathering Skills, Patient Notes, and Communication and Interpersonal Skills (CIS) on the OSCE at the end of the third year were also significantly lower for those who failed ($P<.02$). It is noteworthy that the mean MCAT scores for Biological Sciences, Physical Sciences, and Verbal Reasoning were higher for the students who failed, which was inconsistent with the direction of the differences in the other metrics. However, these unusual differences in the MCAT scores were not statistically significant.

Not summarized in Table 1 are the different distributions of the clinical ratings in the 6 clerkships. The vast majority of students earned the high ratings of either Excellent or Honors in all clerkships, with the largest number earning these ratings in family medicine (95.0% of students), followed by pediatrics (91.2%), ob/gyn (90.6%), internal medicine (90.0%), psychiatry (89.2%), and finally surgery (89.0%). Correspondingly, the low ratings of Good or Marginal were rare, ranging from the fewest in family medicine (5.0% of students) up to surgery (11.0%).

In each clerkship the overall association between the four levels of clinical ratings and Step 2 CS outcomes was statistically significant ($P<0.05$) by χ^2 analysis. Inspection of the cross-tabulations for the four levels of ratings in each clerkship and Step 2 CS revealed little difference in outcomes between students who earned Honors compared to those who earned Excellent. However, the differences in outcomes between the high ratings of Honors or Excellent and the low ratings of Good or Marginal varied across clerkships. Table 2 shows the outcomes for students with low ratings of Good or Marginal in each clerkship. Students with low ratings in internal medicine had the greatest chance of failing (9.7%), whereas students with low ratings in family medicine had a 4.0% risk of failing that was only slightly higher than the 2.9% failure rate for all 1,041 students. Students with low clinical performance ratings in psychiatry had the lowest risk of failing Step 2 CS (1.8%).

Table 3 summarizes the number of low clinical performance ratings per student across the six clerkships. The majority of students, 689 of 1,041 (66%), never had a low rating of Good or Marginal. Their failure rate on the Step 2 CS examination was 2.5%. The overall failure rate for students with three or fewer low ratings was 2.35%. However, for those with four or more low ratings

Table 1: Mean Scores on MCAT, USMLE Step 1, NBME Subject Exams, and OSCE by Pass/Fail on the USMLE Step 2 CS

	Failed (n=30)	Passed (n=1,011)	*P**
MCAT Biological Sciences	11.3	10.9	.12
MCAT Physical Sciences	11.0	10.5	.13
MCAT Verbal Reasoning	10.5	10.1	.07
USMLE Step 1	220.2	230.4	.003
NBME Family Medicine	81.4	84.7	.009
NBME Internal Medicine	82.1	87.0	.0002
NBME Ob/Gyn	81.4	84.3	.01
NBME Pediatrics	81.2	85.5	.0014
NBME Psychiatry	83.6	88.7	.0001
NBME Surgery	81.7	85.7	.0027
OSCE–Data Gathering**	80.4	83.4	.002
OSCE–CIS**	81.5	85.6	.002
OSCE–Patient Notes**	80.0	82.3	.02

* *P* values based on independent *t*-tests.

** Objective structured clinical examination (OSCE) scores are based on a comprehensive clinical skills assessment with standardized patients administered at the end of the third-year required clerkships.

Table 2: Pass/Fail Outcomes on the USMLE Step 2 CS for Students With Low Clinical Ratings Performance in Six Clerkships

Clerkship	Results for Step 2 CS			*P**
	Failed No. (%)	Passed No. (%)	Total	
Internal medicine	10(9.7)	93 (90.3)	103 (100)	.001
Pediatrics	7 (7.8)	83 (92.2)	90 (100)	.004
Ob/gyn	7 (7.1)	91 (92.9)	98 (100)	.008
Surgery	7 (6.1)	107 (93.9)	114 (100)	.03
Family medicine	2 (4.0)	48 (96.0)	50 (100)	.63
Psychiatry	2 (1.8)	110 (98.2)	112 (100)	.46

Note: A rating of Good or Marginal was considered low clinical performance.

* *P* values are based on *z*-tests comparing the failure rate of students with low ratings in that clerkship to the failure rate for the total sample.

Table 3: Pass/Fail Outcomes on the USMLE Step 2 CS by Total Number of Low Clinical Ratings in Six Clerkships

Number of Low Ratings in Six Clerkships	Results for Step 2 CS		
	Failed No. (%)	Passed No. (%)	Total (%)
None	17 (2.5)	672 (97.5)	689 (100)
1	5 (2.4)	206 (97.6)	211 (100)
2	1 (1.2)	82 (98.8)	83 (100)
3	1 (2.6)	37 (97.4)	38 (100)
4 or more	6 (30.0)	14 (70.0)	20 (100)
Total	30 (2.9)	1,011 (97.1)	1,041 (100)

Note: Faculty rated students' performance in clerkships on a 4-point scale (Honors, Excellent, Good, Marginal). The majority (70% to 80%) of students earned Excellent or Honors in each. A rating of Good or Marginal was considered low clinical performance.

the failure rate rose dramatically to 30%. A comparison of this 30% failure rate with the 2.35% failure rate revealed that students with four or more low ratings in clerkships have a 12.76 relative risk (RR) of failing Step 2 CS.

When we performed logistic regression analysis with failing (1) or passing (0) the Step 2 CS examination as the dependent variable, we found a significant (P<.0001) value of McFadden's pseudo r^2 estimate of 0.19 (Table 4).

McFadden's r^2 is an index of model fit, similar to the r^2 calculated in ordinary least-squares regression that can range from 0 to 1. The model identified four statistically significant risks: (1) below-average performance on the NBME surgery examination (RR 3.75, P=.02), (2) below-average performance on the NBME pediatrics examination (RR 3.67, P=.02), (3) low rating in the internal medicine clerkship (RR 3.42, P=.004), and (4) below-average score on the OSCE CIS (RR 2.55, P=.02).

Discussion

Since the Step 2 CS examination was first administered in 2004, it has come to play an increasingly important role in licensure, medical schools' academic decisions, and candidate selection in residency programs. Already, Step 2 CS has significantly influenced the curricula of many medical schools. Despite early concerns about its justification and cost-effectiveness,[12,13] there has been growing recognition of the substantial positive influence of Step 2 CS on the clinical education of young physicians.[14,15] The markedly higher passing standards introduced in 2017 further underscore its value, but also increase concerns for students who fail the exam. Failure can delay graduation and limit professional career prospects. Consequently, it is becoming even more important to be able to identify any performance metrics that suggest that a student might be at a higher risk of failing.

This study showed significant associations between failing Step 2 CS and low Step 1 scores, low NBME clerkship scores, low faculty ratings of clerkship performance, and low scores on a third-year OSCE. These associations are consistent with the published reports of correlations between clinical assessments in medical schools and subtest scores on Step 2 CS.[4,5]

However, our study went beyond previous work in an attempt to identify more specific performance metrics suggesting that a student might be at a higher risk of failing. Students with four or more very low faculty ratings of clinical performance in required clerkships were nearly 13 times more likely to fail than students with three or fewer low ratings. Low faculty ratings in clerkships were so rare that a history of multiple low ratings should serve as a clear and logical red flag for poor performance on the Step 2 CS.

Students with low faculty ratings in the internal medicine clerkship were more likely to fail than students with low ratings in other clerkships. The importance of solid clinical performance in the internal medicine clerkship as a prerequisite for success on the Step 2 CS examination is hardly a surprise considering that CS measures key clinical skills. A task force convened by the Alliance for Academic Internal Medicine identified a subset of key entrustable professional activities considered the principal responsibility of the internal medicine clerkship, including obtaining focused histories and clinically-relevant physical examinations, generating complete differential diagnoses, and providing well-organized clinical documentation.[16]

While clinical skills in internal medicine appear to be important for success on Step 2 CS, it is not clear why above-average knowledge

Table 4: Logistic Regression Model for Failing the Step 2 CS Examination for 1,041 Students in the Classes of 2014 Through 2017

Predictor	Relative Risk	P*
USMLE Step 1	0.72	.49
NBME Family Medicine	0.84	.69
NBME Internal Medicine	0.67	.39
NBME Ob/Gyn	0.76	.57
NBME Pediatrics	3.67	.02
NBME Psychiatry	2.04	.12
NBME Surgery	3.75	.02
Family medicine ratings**	0.51	.43
Internal medicine ratings	3.42	.004
Obstetrics/gynecology ratings	1.62	.33
Pediatrics ratings	1.72	.31
Psychiatry ratings	0.25	.08
Surgery ratings	1.14	.81
OSCE data gathering	0.83	.69
OSCE communications/interpersonal	2.55	.02
OSCE patient notes	1.37	.49

Note: Dependent variable is 1 for those who failed Step 2 CS (n=30), and 0 for those who passed (n=1,011). McFadden's r^2 is 0.19 (P<.0001).

* P values based on two-tailed z-test that odds ratio (OR) in logistic regression model is equal to 0. Relative risk ratio (RR) is based on OR with a population prevalence of a failure rate of 4%.

** Ratings in six clerkships are faculty ratings of students' clinical performance.

of surgery and pediatrics, specifically, were also identified in this analysis. This finding may be related to the specific clinical content of these clerkships at this medical college or the content of the Step 2 CS examinations during the 4-year time period of this study and warrants further investigation. Nevertheless, the findings imply that outcomes on Step 2 CS are associated with test scores that measure specific clinical knowledge as opposed to the diverse clinical skills assessed in clerkships and OSCEs.

It is noteworthy and quite surprising that low clinical ratings in family medicine were not one of the more important risk factors identified in the regression model. This finding is likely related to the fact that family medicine assigned such a very high percentage of Honors/Excellent ratings, and correspondingly the lowest number of Good and Marginal ratings. A similar phenomenon was seen in the ratings for the psychiatry clerkship. Although all clerkships use the same rating form, it is possible that faculty used different standards in the ambulatory settings of these clerkships. It is also possible that faculty raters in these specialties that demand strong interpersonal skills were reluctant to document students' lower clinical performance. Whatever the explanation, their policy appears to be consistent with the findings of a national study of 119 medical schools in which the authors reported that the widest variation and greatest percentages of very high clinical ratings were awarded in the family medicine and psychiatry clerkships.[17] Similarly, a recent proposal for changes in student performance assessment reported that "there is no commonly accepted standard for how to assign clerkship grades" in medical schools.[18]

Low CIS scores in the third-year OSCE also indicated a higher risk of failure, but low scores in data gathering skills or the quality of patient notes in the OSCE did not. It is noteworthy that throughout the 4-year period of this study, one of the authors (J.M.) was conducting a remedial program for about 20 students per year with very low OSCE scores in data gathering, patient notes, or CIS. Each student completed several weeks or more of a required individual remediation program before attempting the Step 2 CS examination. The remedial program for each student was developed to address each student's particular weaknesses. It is possible that this diverse intervention may have confounded some of the associations with the OSCE scores in this study. None of the 30 students who failed Step 2 CS had participated in the remedial program. Also, it is noteworthy that during the period of the study, the students with low clerkship ratings or low NBME scores identified were not being systematically screened and identified for remediation before attempting the Step 2 CS examination. The students were selected for the remedial program based only on their low OSCE scores. The small number of failures and nonrandom assignment to the remedial program precluded any statistical analysis to investigate and try to adjust for its effect. Furthermore, the intervention varied for subgroups of students and there was variation in student adherence to faculty recommendations.

There are other limitations in this study. Although it was conducted using data from a single institution, it does include four classes and more than 1,000 students. While the number of students who failed was small, this sample is representative of the national rates during that time period. This limitation is reflected in the low value of the r^2 estimate of fit for the logistic regression model. The amount of information on failures in the sample is limited. Although the modest r^2 value is also partially due to measurement error in variables—especially the clinical ratings and OSCE scores—it also suggests that there are other independent variables that were not included in the model. For example, failure may be associated with noncognitive variables (eg, professionalism, work ethic, professional responsibility) that were not explicitly measured in the present study. Chang and colleagues at the University of California San Francisco reported that such measures were associated with the patient-physician interaction portion of their Clinical Performance Examination.[19] Finally, the primary language of the vast majority of students in this study was English. No one in the sample failed the Spoken English Proficiency component of the Step 2 CS examination.

Conclusions

Overall, the findings imply that students who encounter difficulty in multiple clerkships and assessments of clinical skills are at the greatest risk. The relative risk of the individual clerkship metrics is significant, but less certain. The findings for students who failed and the most important risk factors reported here are logically related to the proficiencies measured by the Step 2 CS examination. These findings are consistent with previous correlational studies. As the importance of Step 2 CS grows, meaningful information is needed about the risks of failure to students. Future studies involving larger samples would be helpful to examine the risk factors reported here and to investigate the predictive value of other student metrics.

ACKNOWLEDGMENT: The authors sincerely thank the three peer reviewers whose suggestions helped improve and clarify this manuscript.

CORRESPONDING AUTHOR: Address correspondence to Jon Veloski, Director, Medical Education Research, Center for Research in Medical Education and Healthcare, Sidney Kimmel Medical College at Thomas Jefferson University, 1015 Walnut Street, Suite 319, Philadelphia, PA 19107. 215-955-7901. Jon.Veloski@jefferson.edu.

References

1. Coumarbatch J, Robinson L, Thomas R, Bridge PD. Strategies for identifying students at risk for USMLE step 1 failure. Fam Med. 2010;42(2):105-110.

2. Case SM, Ripkey DR, Swanson DB. The relationship between clinical science performance in 20 medical schools and performance on Step 2 of the USMLE licensing examination. 1994-95 Validity Study Group for USMLE Step 1 and 2 Pass/Fail Standards. Acad Med. 1996;71(1)(suppl):S31-S33.
3. Ripkey DR, Case SM, Swanson DB. Identifying students at risk for poor performance on the USMLE Step 2. Acad Med. 1999;74(10)(suppl):S45-S48.
4. Berg K, Winward M, Clauser BE, et al. The relationship between performance on a medical school's clinical skills assessment and USMLE Step 2 CS. Acad Med. 2008;83(10)(suppl):S37-S40.
5. Dong T, Swygert KA, Durning SJ, et al. Validity evidence for medical school OSCEs: associations with USMLE® step assessments. Teach Learn Med. 2014;26(4):379-386.
6. United States Medical Licensing Examination. Performance Data. www.usmle.org/performance-data. Accessed September 5, 2018.
7. United States Medical Licensing Examination. Change in Performance Standards for Step 2 CS. Announcements. www.usmle.org/announcements/?ContentId=210. Published August 4, 2017. Accessed September 5, 2018.
8. Gilliland WR, La Rochelle J, Hawkins R, et al. Changes in clinical skills education resulting from the introduction of the USMLE step 2 clinical skills (CS) examination. Med Teach. 2008;30(3):325-327.
9. Green M, Jones P, Thomas JX Jr. Selection criteria for residency: results of a national program directors survey. Acad Med. 2009;84(3):362-367.
10. National Board of Medical Examiners. USMLE Examination Fees. www.nbme.org/students/examfees.html. Accessed September 5, 2018.
11. Gonnella JS, Hojat M, Veloski J. AM last page. The Jefferson Longitudinal Study of medical education. Acad Med. 2011;86(3):404.
12. Lehman EP IV, Guercio JR. The Step 2 Clinical Skills exam—a poor value proposition. N Engl J Med. 2013;368(10):889-891.
13. First LR, Chaudhry HJ, Melnick DE. Quality, cost, and value of clinical skills assessment. N Engl J Med. 2013;368(10):963-964.
14. Saheb Kashaf M. Clinical skills in the age of Google: A call for reform and expansion of the USMLE step 2 CS. Acad Med. 2017;92(6):734.
15. Ecker DJ, Milan FB, Cassese T, et al. Step up—not on—the step 2 clinical skills exam: directors of Clinical Skills Courses (DOCS) oppose ending step 2 CS. Acad Med. 2018;93(5):693-698.
16. Fazio SB, Ledford CH, Aronowitz PB, et al. Competency-Based Medical Education in the Internal Medicine Clerkship: A Report From the Alliance for Academic Internal Medicine Undergraduate Medical Education Task Force. Acad Med. 2018;93(3):421-427.
17. Alexander EK, Osman NY, Walling JL, Mitchell VG. Variation and imprecision of clerkship grading in U.S. medical schools. Acad Med. 2012;87(8):1070-1076.
18. Hauer KE, Lucey CR. Core clerkship grading: the illusion of objectivity. [Published online ahead of print August 14, 2018]. Acad Med. 2018.
19. Chang A, Boscardin C, Chou CL, Loeser H, Hauer KE. Predicting failing performance on a standardized patient clinical performance examination: the importance of communication and professionalism skills deficits. Acad Med. 2009;84(10)(suppl):S101-S104.