doi:10.22100/ijhs.v12i5.1446
Review Article

SJMS 2026;12(5):29-49
sjms.shmu.ac.ir
SJMS
Shahroud Journal of Medical Sciences

Predictive Models for Performance in High-Stakes Medical Assessments: A Systematic Review within Health Professions Education

Haniye Mastour[1, 2], Toktam Dehghani[3*], Mahdie Jajroudi[3, 4], Mohammadreza Farrokhnia[5], Mitra Zarei[6], Saeid Eslami[3, 4, 7]

[1] Department of Medical Education, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.
[2] School of Medical Education and Learning Technologies, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
[3] Department of Medical Informatics, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.
[4] Pharmaceutical Sciences Research Center, Institute of Pharmaceutical Technology, Mashhad University of Medical Sciences, Mashhad, Iran.
[5] Department of Learning, Data Analytics and Technology, Instructional Technology Section, University of Twente, Enschede, the Netherlands.
[6] Medical Librarian and Information Counselor, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.
[7] Department of Medical Informatics, UMC-Location AMC, University of Amsterdam, Amsterdam, Netherlands.

Received: 4 December 2025
Accepted: 4 April 2026

Abstract

Background: *High-stakes medical assessments play a critical role in certifying future healthcare professionals. Enhancing these assessments through predictive models offers a promising way to identify at-risk students and predict exam outcomes. The current study aims to address this need by providing a comprehensive overview of the application and effectiveness of predictive models, particularly in high-stakes medical assessments.*
Methods: *This systematic review searched four major databases (MEDLINE, PubMed, Scopus, and Web of Science) for studies using predictive models to evaluate performance in high-stakes medical assessments, following PRISMA guidelines and the Best Evidence in Medical Education framework.*
Results: *The review included 14 studies across various stages of professional training, including medical students, residents, chiropractors, and physician assistants. Sample sizes ranged from 147 to 5,825 participants, and most studies used local datasets. Regression models were the most commonly employed (78%), while random forests and artificial neural networks were frequently used for classification tasks. Although deep learning models were not observed, 14% of studies incorporated explainable AI (XAI) to improve interpretability. To provide a broader context, four systematic reviews of predictive modeling in nonmedical high-stakes assessments were also analyzed.*
Conclusion: *While predictive methods show significant promise for enhancing the predictive ability of high-stakes medical assessments, there is still ample room for improvement. Integrating comprehensive feature sets and longitudinal tracking is crucial for increasing accuracy. Expanding studies to more extensive, diverse populations can improve generalizability and pave the way for more precise evaluations and better educational outcomes.*

Keywords: Systematic review, Machine learning, Assessment, Medicine, Medical education.
[*]**Corresponding to:** T Dehghani, **Email:** DehghaniT@mums.ac.ir

Please cite this paper as: Mastour H, Dehghani T, Jajroudi M, Farrokhnia M, Zarei M, Eslami S. Predictive Models for Performance in High-Stakes Medical Assessments: A Systematic Review within Health Professions Education. Shahroud Journal of Medical Sciences 2026;12(5):29-49.

Introduction

High-stakes examinations are standardized tests with significant consequences for individuals and institutions. These assessments, often administered nationwide or worldwide, measure competency for professional licensure or academic progression [1]. These exams, which are conducted on large cohorts, often utilize various question formats, including multiple-choice and short-answer questions [2]. Such assessments act as critical gatekeepers, ensuring that prospective healthcare professionals possess the necessary competencies to deliver care that is not only effective but also safe [3]. In addition, these exams have a lasting impact on educational paths and opportunities for advancement [4]. Administered nationally and internationally, these assessments are pivotal in shaping educational policies that establish rigorous standards for healthcare training in the 21st century [5]. Building on this importance, many international medical training programs have adopted these assessments to ensure the development of a competent and qualified workforce [6].

However, concerns have been raised regarding the sole reliance on high-stakes assessments owing to issues surrounding their reliability and validity [7]. There is a need to enhance the effectiveness of these assessments and ensure that they accurately reflect candidates' competencies in delivering safe and effective patient care [7]. Moreover, these exams are time-consuming and costly for universities, placing immense pressure and stress on students [5]. Consequently, medical institutions worldwide seek more accurate ways to predict their students' performance on high-stakes examinations, underscoring the growing need to establish a robust standard of validity and reliability to minimize measurement errors and biases [8].

To address these challenges, predictive models have emerged as valuable tools in educational evaluation. Predictive models refer to computational approaches designed to anticipate outcomes based on input data, ranging from traditional statistical methods, such as linear and logistic regression, to advanced techniques, such as artificial intelligence (AI), data mining (DM), machine learning (ML), deep learning (DL), and large language models (LLMs). These advanced methods empower experts with predictive capabilities that often transcend their initial expectations, revealing patterns and insights that were previously overlooked [9]. In particular, integrating AI and ML models into educational data analysis has given rise to educational DM [10]. These models assess the likelihood of students failing high-stakes assessments by analyzing their progressive assessment scores [11]. Moreover, implementing these early warning systems can help identify at-risk students, provide targeted feedback, and reduce associated costs [12].

Several systematic reviews have aimed to provide an overview of empirical findings on the use of predictive models in high-stakes assessments [13-16]. Findings from these reviews reveal that predictive models are widely used in various educational contexts, including physics, chemistry, general science, computer science, biology, and earth science [13, 16]. Additionally, the reviews highlight that employing tailored classification and regression models for specific predictive tasks allows researchers to perform comprehensive assessments of student performance [13-16]. This approach plays a critical role in designing more effective educational strategies, enhancing predictive accuracy, and improving overall academic outcomes [14]. Moreover, the ongoing refinement and evolution of these models underscore their transformative potential in predictive assessments, paving the way for more targeted interventions and personalized learning opportunities [15]. On the other hand, the aforementioned systematic reviews also highlight several critical limitations in the field that require attention. A key challenge is the lack of action-oriented intervention studies that bridge the gap between theoretical insights and practical applications [15]. Additionally, there is a significant underutilization of advanced computational approaches, such as DL and reinforcement learning [13].

Previous systematic review studies have provided a comprehensive overview of various predictive models used for high-stakes assessments. However, they have often fallen short of offering a thorough overview within the context of medical education. Many reviews tend to include findings from a limited number of studies specific to the medical field—often only one [16] or two studies [13]—resulting in a limited ability to draw definitive conclusions. This limitation is evident in their findings, which frequently underscore the scarcity of robust empirical studies that specifically address the application of predictive models in high-stakes medical assessments [17]. One contributing factor to this limitation may be the exclusion of specialized databases, such as MEDLINE and PubMed, during the review processes of previous studies. These repositories are invaluable, as they hold significant potential to advance our understanding by offering critical insights into performance predictors at the intersection of medical and health education. Moreover, the limitations may also stem from the rigorous inclusion criteria used in these reviews, which heavily focus on specific types of regression models [13]. Such strict inclusion criteria would have decreased the chances of identifying papers across various contexts, including medical education.

Given these considerations, applying insights from predictive model-enhanced high-stakes assessments in other disciplines to medical education requires caution. While these models demonstrate strong technical potential, the unique complexities of healthcare training, which requires assessments to integrate theoretical knowledge, practical skills, and clinical reasoning, demand a more comprehensive and context-sensitive framework. These distinctive requirements, which safeguard patient safety and ensure competent clinical practice, set medical education apart from most other academic domains [18]. Accordingly, predictive analytics in this setting should not only aim for accuracy but also uphold educational validity, ethical responsibility, and interpretability. To address this critical requirement, the present study conducts a systematic review to synthesize existing evidence on predictive modeling in medical education, thereby establishing a theoretically grounded, field-specific foundation for future AI-based assessment and learning innovations.

The current study aims to address this need by providing a comprehensive overview of the application and effectiveness of predictive models, particularly in high-stakes medical assessments. To this end, the findings of robust empirical studies were synthesized to address the following research questions:

RQ1: What is the current state of the art regarding the use of predictive models in high-stakes medical assessments?

RQ2: What are the predominant predictors utilized by predictive models in high-stakes medical assessments?

RQ3: What preprocessing and predictive models are commonly employed in high-stakes medical assessments?

RQ4: What performance metrics and evaluation methods are used to assess predictive models in high-stakes medical assessments, and what are the key findings regarding their performance?

RQ5: What is the potential of predictive models to complement traditional high-stakes medical assessments?

Materials and Methods

This systematic review was conducted in accordance with the published protocol [19] and followed the methodological recommendations outlined therein. This review follows the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [19], which are explained in more detail in the following section.

Identification phase: The review team thoroughly investigated four electronic databases: MEDLINE/PubMed, SCOPUS, and Web of Science. The search strategy was meticulously crafted by an experienced academic medical sciences librarian. All the citations were subsequently imported into EndNote for systematic processing. This systematic review adheres to the PICO framework [20]. Studies were extracted from databases using terms that included all four relevant subject areas within their titles or abstracts:

- Medical Students: This includes terms such as trainees, interns, clerkships, residents, and fellows.
- High-Stakes Exams: Relevant terms include qualifications, licensing exams, certification exams, and assessments.
- Predictive models: The focus was limited to terms and Medical Subject Headings (MeSH) related to predictive models, ML, DL, and other pertinent methodologies.
- Prediction and Forecasting: This encompasses terms associated with predictive analytics.

A decade-long systematic review was conducted, covering 2014-2024. Forward and backward citation searches were implemented to identify relevant gray literature, with the references and citations of included articles and pertinent review papers examined to uncover additional relevant studies. The first inclusion criterion (Table 1) focused exclusively on peer-reviewed scientific journal articles. Furthermore, the search was restricted to English-language articles published from 2014 onward. The initial search yielded a total of 19,538 articles. As Figure 1 indicates, after applying filters for the specified time frame (2014–2024), peer-reviewed journals, and the English language, 11,524 articles remained. Among these, 1,082 articles were identified as duplicates.

Table 1. Summary of inclusion and exclusion criteria

Phase	Inclusion criteria	Exclusion criteria
Identification phase	Peer-reviewed scientific venues English articles Publication date from 2014-2024	Non-peer-reviewed scientific venues, books, book chapters, proceedings, editorials, PhD dissertations, MSc theses, commentaries, and others non-english publication date before 2014
Screening phase	Empirical studies Higher education Medical sciences students in medical sciences universities, schools, faculties, and academic hospitals	Non-primary empirical research and conceptual/theoretical Primary, secondary, or high schools Non-medical sciences students, Continuing Medical Education (CME), and Continuing Professional Development (CPD)
Eligibility phase	High-stakes examination Predictive models Reporting quality of models and dominant predictors for students' performance prediction (score, status, etc.)	Low-stakes assessments

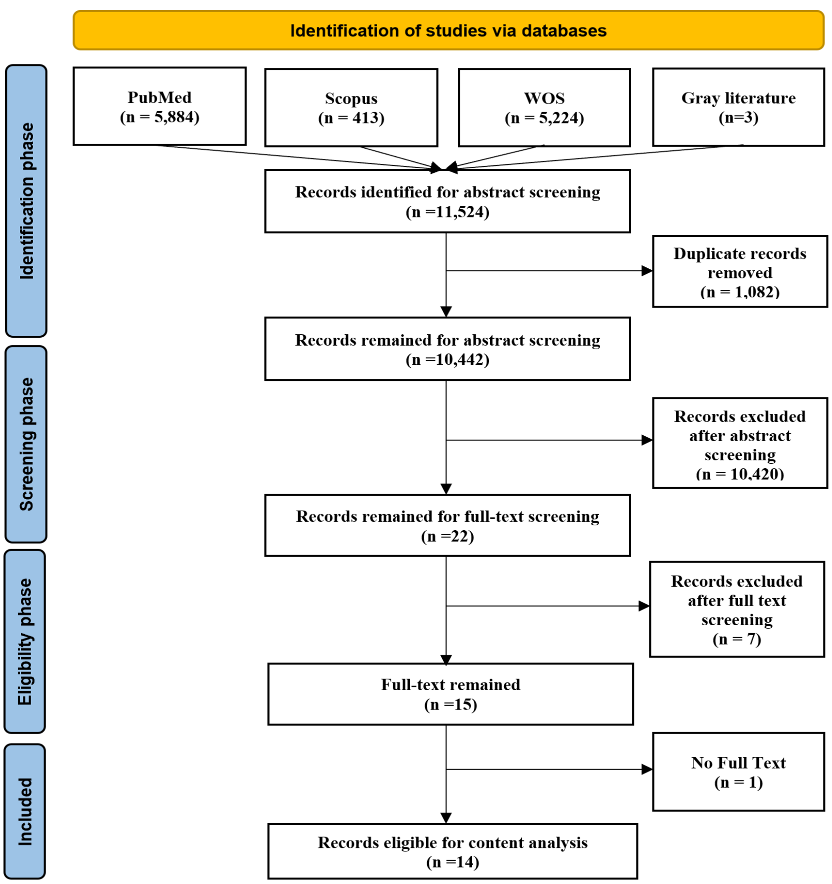

Figure 1. Adapted PRISMA flow diagram

Screening phase: The second inclusion criterion was employed for screening the abstracts and collecting relevant studies in the next step. During this step, the second (T.D.) and third authors (M.J.) reviewed the articles' titles, abstracts, and full texts when required and excluded several unrelated studies that did not satisfy the inclusion criteria (n=10,420). Both reviewers agreed on the relevance of the study and its inclusion. If both reviewers agreed to exclude a paper, it was rejected. If there was a disagreement, the reviewers were asked to discuss the issue and consult a reviewer (first author (H.M.)). The articles were excluded from further analysis for the following reasons: (1) they lacked empirical findings, such as meta-analyses, reviews, or conceptual studies (n=1,250); (2) they did not specifically address examinations related to medical science students (n=6,877); and (3) they failed to incorporate predictive methods (n=2,293).

Eligibility phase: In this phase, the first and second authors read the full texts of the remaining 22 articles to identify the most relevant ones. After the third set of inclusion criteria was applied, seven articles were deemed inappropriate for further analysis based on the following criteria: (1) a lack of focus on medical high-stakes assessments, (2) the absence of methodologies, and (3) insufficient reporting on model quality and the key predictors influencing students' performance outcomes, such as scores, academic status, and rankings. Additionally, one article was excluded because the full text was not accessible. The steps for selecting relevant articles are depicted in the adapted PRISMA flow diagram shown in Figure 1.

In the next step, the quality of the remaining articles was evaluated using a modified version of the Best Evidence in Medical Education (BEME) quality framework [21, 22]. The BEME framework offers a structured and transparent approach for evaluating empirical studies based on predefined criteria, helping to minimize subjective interpretation and potential bias during the appraisal process. Our adapted version includes 24 quality indicators, each assessed as "Met," "Unmet," or "Unclear." For scoring purposes, a value of 1 was assigned to "Met," 0.5 to "Unclear," and 0 to "Unmet." Therefore, the maximum possible score was 24, and studies scoring 12 or higher were classified as high-quality. Despite the comprehensive evaluation of the high-quality papers, we did not exclude any studies based on inadequate quality. The quality assessment was carried out independently by the second (T.D.) and third authors (M.J.), and any discrepancies were resolved through discussion and consensus with a third reviewer (the first author (H.M.)). The methodological quality assessment of the included studies revealed a high level of rigor. None of the studies failed to meet the quality criteria, as all scored 14 or more out of the total quality assessment points of 24. The scores ranged from 14 to 21. The studies demonstrated good quality in terms of most criteria; however, most studies still needed to fully meet the risk of bias, Triangulation, and prospective criteria.

The corresponding information was designated "not reported" if any required data were unavailable. Additionally, to establish coding reliability, the second (T.D.) and third authors (M.J.) randomly selected three articles, which constituted 50% of the sample, for blind coding. The interrater reliability of the coding quality was evaluated via Cohen's kappa statistic. According to Cohen, there was a high level of agreement between the coders (κ=0.80, P-value<0.001), further validating the reliability of the final coding framework [23, 24]. After resolving discrepancies through discussion and refining the coding scheme, the first author (H.M.) coded all identified articles to synthesize their findings.

Results

This systematic review aimed to provide a comprehensive overview of the current research landscape regarding applying predictive models to predict medical student performance in high-stakes assessments. The review included 14 studies summarized in Table 2.

Table 2. Characteristics of the included studies

Paper no.	P1	P2	P3	P4	P5	P6	P7	P8
Title	A generalizable approach to predicting performance on USMLE Step 2 c.k.	Application of DM techniques for predicting residents' performance on preboard examinations: a case study	Early prediction of medical students' performance in high-stakes examinations using machine learning approaches	Early prediction of the risk of scoring lower than 500 on the COMLEX 1	Examining the validity of chiropractic grade point averages for predicting National Board of Chiropractic Examiners Part I exam scores	The application of adaptive minimum match k-nearest neighbors to identify at-risk students in health professions education	Prediction of admission tests for medical students' academic performance	Predictors of medical school clerkship performance: a multispecialty longitudinal analysis of standardized examination scores and clinical assessments
First author	Jeffrey B Bird	Leila Amirhajlou	Haniye Mastour	Qing Zhong	Igor Himelfarb	Anshul Kumar	Abdulaziz Althewini	Petra M. Casey

Authors' departments	Science education, school of medicine	Medical education, school of medicine	Medical education, faculty of medicine	Pharmacology, physiology, statistics, and biomedical science	NBCE	Health professions education, school of healthcare leadership	Medical education	Gynecology, psychiatry and psychology, general surgery, emergency medicine
Publication year	2022	2019	2023	2021	2022	2023	2022	2016
Country	USA	Iran	Iran	USA/China	USA	USA	Kingdom of Saudi Arabia	USA
Journal name	Advances in Medical Education and Practice	Journal of Education and Health Promotion	Heliyon	BMC Medical Education	Journal of Chiropractic Education	Journal of Physician Assistant Education	Advances in Medical Education and Practice	BMC Medical Education
IF	2	1.4	4	3.6	0.8	0.6	2.2	2.7
Q	Q2	Q2	Q1	Q1	Q3	Q2	Q2	Q1
H-index	42	24	88	97	12	17	42	97
SJR 2023	0.43	0.48	0.62	0.94	0.25	0.27	0.49	0.93
Study design	Retrospective cohort	Retrospective cohort	Retrospective cohort	Retrospective cohort	Retrospective cohort	Retrospective observational	Retrospective cohort	Longitudinal retrospective cohort
Quality assessment	19	17	20	16	14	17	18	21

Paper no.	**P9**	**P10**	**P11**	**P12**	**P13**	**P14**
Title	In search of black swans: identifying students at risk of failing licensing examinations	Multivariable analysis of factors associated with USMLE scores across U.S. Medical schools	Development of statistical models to predict medical student performance on the USMLE Step 1 as a catalyst for deployment of student services	Predicting medical student success on licensure exams	Discovering unknown response patterns in progress test data to improve the estimation of student performance	Validity and reliability of prematriculation and institutional assessments in predicting USMLE Step 1 success: lessons from a traditional 2 x 2 curricular model
First author	Cassandra Barber	Arash Ghaffari-Rafi	Michael W. Lee	Charles A. Gullo	Miriam Sieg	Nitin Puri
Authors' department	Medicine and dentistry	-	Medical education	School of medicine	-	Internal medicine
Publication year	2018	2019	2017	2015	2023	2021
Country	Canada	USA	USA	USA	Germany	USA
Journal name	Academic Medicine	BMC Medical Education	Medical Science Educator	Medical Science Educator	BMC Medical Education	Frontiers in Medicine
IF	5.35	3.6	1.9	1.9	3.6	3.1
Q	Q1	Q1	Q2	Q2	Q1	Q1
H-index	188	97	22	22	97	86
SJR 2023	1.56	0.94	0.56	0.56	0.94	0.91
Study design	Retrospective cohort	Cross-sectional analytical	Retrospective cohort	Retrospective cohort	Longitudinal observational	Retrospective cohort
Quality assessment	20	18	19	18	21	18

RQ1. What is the current state of the art regarding the use of predictive models in high-stakes medical assessments?

Overview of the studies: The range of publication impact factors spans an average of 2.6±1.3. This variability indicates the diversity of journals focused primarily on medical education. These journals include "BMC Medical Education" [25-28], "Advances in Medical Education and Practice" [29, 30], "Medical Science Educator" [31, 32], "Journal of Education and Health Promotion" [33], "Journal of Chiropractic Education" [34], "The Journal of Physician Assistant Education" [35], "Academic Medicine" [36], "Heliyon" [17], and "Frontiers in Medicine" [37]. The representation across these various publications suggests widespread interest and discussion of the topic within the field of medical education.

The studies were published from 2016 to 2023, and geographically, the majority of the included studies (9 out of 14) were performed in the United States. In contrast, the remaining studies were conducted in Iran (2 of 14), Canada, Germany, the Kingdom of Saudi Arabia, and China. This distribution suggests that using predictive models to forecast medical student examination outcomes has garnered significant attention and momentum within the North American context while highlighting the emerging interest in this research area among scholars in other regions. Interestingly, the authors of 66% of these studies were affiliated with medical/science education departments within medical schools. The remaining studies were conducted by researchers in other departments closely related to examinations, such as gynecology, psychiatry and psychology, general surgery, emergency medicine, internal medicine, dentistry, and physiology. Notably, one study (P6) was led by managers/directors of the National Board of Chiropractic Examiners (NBCE), further emphasizing the cross-disciplinary nature of this research area and its relevance to various stakeholders in the medical education landscape.

Regarding study design, the majority of the included studies employed retrospective cohort designs (10 of 14, 71%), using previously collected academic and examination records to develop predictive models of student performance in high-stakes assessments. Additionally, two studies (14%) adopted longitudinal observational designs, analyzing multi-year examination datasets to evaluate performance trajectories over time, while two studies (14%) used cross-sectional analytical approaches.

Aim and scope of the studies: The primary aim of the fourteen included studies was to predict medical students' or residents' performance on high-stakes assessments, such as the United States Medical Licensing Examination (USMLE Step 2 CK) (6 of 14 studies), preboard examinations, the NBCE, the Comprehensive Medical Basic Sciences Examination (CMBSE), the Physician Assistant National Certifying Examination (PANCE), the Standard Achievement Admission Test (SAAT), the General Aptitude Test (GAT), the National Board of Medical Examiners (NBME), the Medical Council of Canada Qualifying Examination Part 1 (MCCQE1), and the Progress Test Medizin (PTM). Twelve studies focused on students during basic medical science and residency programs, and two concentrated explicitly on physician and chiropractic assistant programs. The main objectives of these studies were to predict either the scores or the pass/fail status in these examinations (refer to Table 3).

Notably, some high-stakes exams, such as the USMLE Step 1, have transitioned from score-based to pass/fail reporting. Consequently, a subset of studies (5 of 14) concentrated exclusively on predicting the pass/fail status. In contrast, other studies (7 of 14) aimed to predict actual examination scores, which are typically reported as three-digit numbers. Two studies (2 of 14) predicted scores and pass/fail statuses and identified students at risk of underperforming in these high-stakes exams.

Table 3. Aim and scope of included studies

Paper no.	P1	P2	P3	P4	P5	P6	P7	P8
Aim	Predict student performance	Predict residents' performance	Predicting students' performance and identifying at-risk students.	Identify at-risk students	Evaluate the validity of GPA[f] for predicting exam scores	Predicting students' scores and identifying at-risk students.	Predict students' performance in basic and advanced science courses.	Predicted NBME Subject Exam Scores
Name of examination	USMLE Step2 CK[a]	PGY-4, called Preboard examinations	CMBSE[b]	COMLEX-USA 1[c]	NBCE[d] Part I	PANCE[e]	GAT[g] and SAAT[h]	NBME Subject Exams[i]
Scope of examination	Internal medicine residency	Medicine residency	Medicine basic sciences	Internal medicine residency	Chiropractic	Physician assistant	Medical students	Medical students

Paper no.	P9	P10	P11	P12	P13	P14
Aim	Predictive factors of student risk for failing an exam	Determining medical student success	Predictive models for USMLE Step 1 scores/Identify students needing support early	Determine students at risk for poor performance	Identify student response patterns in the test	Assess the validity and reliability of pre-matriculation and course-assessment data in predicting future performance
Name of examination	MCCQE1[j]	USMLE	USMLE Step 1	USMLE	PTM[k]	USMLE Step 1
Scope of examination	Medical student	Medical student	Medical student	Medical student	Medical student	Residents

a The United States Medical Licensing Examination (USMLE Step 2 C.K.)
b The Comprehensive Medical Basic Sciences Examination (CMBSE)
c The Comprehensive Osteopathic Medical Licensing Examination (COMLEX-USA)
d The National Board of Chiropractic Examiners (NBCE)
e The Physician Assistant National Certifying Examination (PANCE)
f Grade Point Average (GPA)
g The Standard Achievement Admission Test (SAAT)
h The General Aptitude Test (GAT)
i National Board of Medical Examiners (NBME) Medical Council of Canada Qualifying Examination Part 1 (MCCQE1)
k Progress Test Medizin (PTM)

RQ2: What are the predominant predictors utilized by predictive models in high-stakes medical assessments?

The sample sizes in the included studies varied considerably, ranging from 147 to 5,852 students. Some studies reported high exclusion rates due to missing values, such as course scores, ranging from 2% to 36%. The data collection periods differed among the studies: the most significant sample of 5,852 students was gathered within one year at a single exam center. In comparison, other studies spanned 2- 10 years and had sample sizes between 147 and 1,317 students. The data sources were primarily domestic cohorts, comprehensive education systems, or local databases, with the exception of P5, which utilized domestic and international chiropractic education institution databases, and P10, which used data from multiple U.S. medical schools (see Table 4).

The algorithms predict outcomes (target features) for each predictive model on the basis of input parameters (input features). The input parameters considered in the predictive models of the included studies were categorized into two primary groups: (1) Course/Exam scores, which include previous exam results and grade point averages (GPAs), and (2) demographic characteristics, which encompass factors such as age, ethnicity, sex, specialty, residency status, admission type, and school location.

Demographic characteristics were not consistently reported across the studies. Specifically, sex was included as a predictor in 6 of the 14 studies, whereas age was considered in only 3 studies. Among the included studies, five (P2, P3, P5, P9, and P11) reported the sex distribution of participants, which was approximately balanced. In addition, one study (P6) described the ethnic composition of its sample, reporting that 68.2% of participants identified as White, 8.8% as Asian/Pacific Islander, 4.5% as African American, 11.7% as Latino, 1% as Native American, and 6.1% as other ethnicities. Furthermore, two studies (P3 and P11) reported the pass–fail distribution of examination outcomes, indicating class imbalance, with 82% and 98% passing and 18% and 2% failing, respectively. Most studies focused on medical students or residents because the target examinations were primarily medical licensing examinations. However, studies P5 and P6 examined chiropractic and physician assistant education and therefore included a broader range of educational levels, spanning bachelor's, master's, and doctoral programs (Table 4).

A synthesis of predictor variables across the included studies revealed that prior academic performance indicators were the most consistently used predictors of high-stakes examination outcomes. GPA or cumulative academic performance was reported in 9 of the 14 studies (64%), making it the most frequently used predictor. MCAT scores were included in 8 studies (57%), primarily to predict outcomes in examinations such as the USMLE, Comprehensive Osteopathic Medical Licensing Examination (COMLEX), and MCCQE. Preclinical course grades and institutional academic scores were reported in 7 studies (50%) and were commonly used to predict performance in licensing examinations, including USMLE Step examinations, NBME assessments, and NBCE Part I. Additionally, standardized internal assessments such as Comprehensive Basic Science Examination (CBSE), Objective Structured Clinical Exam (OSCE), In-training examinations (ITEs), and Physician Assistant Clinical Knowledge Rating And Assessment Tool (PACKRAT) scores were incorporated in 5 studies (36%) to predict performance in board or licensing examinations. In contrast, demographic variables (including sex, age, and ethnicity) were used less frequently and were generally incorporated as supplementary predictors rather than primary determinants of examination outcomes (Figure 2).

In most studies, the importance and impact of input parameters on the outcome of models and their predictions are academic metrics such as course scores and GPAs, which are correlated with the target feature. In one study (P5), additional factors such as sex, ethnicity, and age were analyzed, revealing positive and negative correlations with the outcome (e.g., male and younger students were more successful on exams, whereas Asian/Pacific Islander and White students achieved the highest scores). Interestingly, while MCAT performance had no bearing on predicting USMLE 1 (according to study P1), it was significantly correlated with the prediction of COMLEX 1 (as reported in study P2).

The target features were either examination scores (typically reported as three-digit numbers or normalized scores) or exam statuses, categorized as either pass/fail or rated as excellent, good, medium, poor, and further classified into risk levels such as very likely to fail, moderate risk of failing, and low risk of failing.

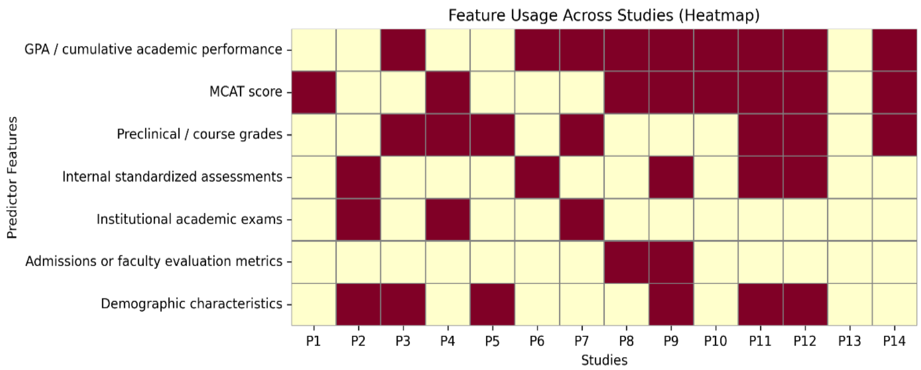

Figure 2. Distribution of predictor features across the included studies

Table 4. Details of the sample data in the included studies

Paper no.	P1	P2	P3	P4	P5	P6	P7	P8
Sample size	228 students in three cohorts	1317 after removing missing data: 841	1005 students	904 after removing missing data: 885	2278 students	224 students	650 students	435 students
Year of study	2018-2021	2004-2014	2017-2022	2011- 2017	2019	2015-2023	2018-2019	2000–2009
Data collection method	-	From the residency training database using a researcher-made checklist	From medical schools' CESa and manual data entry for the missing data	From seven cohorts	Computer-based testing administration of 18 domestic and four international institutions	From five cohorts	-	-
Location	ZSOMb	IUMSc	MUMSd	RVUCOMe	NBCEf	-	ksau-hsj	MMSk
Input features								
Scores	(1) MCATg (2) CASh exams	(1) Score of ITEs in PGY-1 and PGY-3 (2) Annual written examinations	(1) Grades in the primary medical science curriculum (2) GPA	(1) Course average scores (2) MCAT (3) Score in preclinical courses	(1) Score in general anatomy, spinal anatomy, physiology, chemistry, pathology, and microbiology	(1) iRAT quiz (2) PACKRATi scores (3) GPA	(1) English courses (2) GPA in basic science courses in biology (3) GPA in advanced science courses	(1) GPA (2) MCAT (3) USMLE 1 and 2 (4) Faculty assessment scores (FAS)
Demographic	-	(1) Gender (2) Type of specialty training	(1) Age (2) Gender (3) Residency status (4) Entrance semester (5) Type of admission	-	(1) Age (2) Ethnicity (3) Gender	-	-	(1) Gender (2) Age
Target	(1) Scores of Step 2 C.K. at five points	(1) Status of preboard exam	(1) Normalized exam score (2) Pass or fail	(1) COMLEX 1 status on the first attempt	(1) Status of NBCE Part I exam	(1) Score and status in PANCE	(1) GAT and (2) SAAT test scores	(1) Poor performance in NBME
Ratio of fail/pass	-	-	Pass: 82% Fail: 18%	-	-	-	-	-
Range/type of scores	Continues	Categorical	Continues Categorical	Continues Categorical	Continues Categorical	Continues Categorical	Continues	Continues
Males/females	-	-	M:52% F:48%	-	M: 41% F: 59%	-	-	M:53% F:47%

Ethnicity of population	-	-	-	-	White, Asian/Pacific, Islander, African/American, Latino, Others	-	-	-
Level of education	Resident	Resident	Medical student	Resident	Bachelor: 84.6% Master: 4.0% Doctoral: 1.9%	Master of Physician Assistant	Medical student	Medical student MD: 83% Joint-degree programs: 17%
Specialty	Medicine	Medicine	Medicine	Medicine	Chiropractic	Physician	Medicine	Medicine

Paper no.	P9	P10	P11	P12	P13	P14
Sample size	788 students (five cohorts)+A test set of 157 students	100 U.S. medical schools	377	344 +198 students	5,852 students	76 students (Class of 2020)+71 students (Class of 2021)
Year of study	2011–2016	2014	2013-2017	2008-2012	2020	2020-2022
Data collection method	-	U.S. News and World Report	UCFCOM's Office	AMCAS	Eight university	Cohort
Location	Schulich School of Medicine and Dentistry, Western University, Canada	-	UCFCOM	Joan C. Edwards School of Medicine	-	Joan C. Edwards School of Medicine
Input features						
Scores	(1) MCAT (2) Cumulative averages (3) OSCE score (4) Admissions interview scores	(1) GPA (2) MCAT total score (3) NIH funding (4) Residency assessments (5) Faculty-to-student ratio (6) Class size	(1) MCAT scores (2) GPA (3) Internal course grades (4) CBSE scores	(1) MCAT scores (2) GPA (3) Internal exams (Family Medicine Clinical Sciences, Internal Medicine Clinical Science, Obstetrics and Gynecology Clinical, Pediatrics, Psychiatry, Surgery) (4) CBSSA	200 questions of PTM	(1) MCAT (2) GPA (3) Exam scores (anatomy, physiology, pathology, pharmacology)
Demographic	(1) Gender (2) School location (urban/rural)	-	Gender	Age	-	-
Target	MCCQE Part 1 score (cutoff score 450 to determine pass/fail).	Score	USMLE Step 1 score	USMLE Step 1 and Step 2 performance	PTM questions that distinguish the groups	(1) NBME (2) STEP1 scores
Ratio of fail/pass	-	-	10/367	-	-	-
Range/type of scores	Continues	Continues	Categorical	Continues	Categorical	Continues
Males/females	M:49% F:51%	-	M:48% F:52%	-	-	-
Ethnicity of population	-	-	-	-	-	-
Level of education	Medical school: primarily pre-clerkship and clerkship	Medical student	Medical student	Medical student	Medical student	Medical student and resident
Specialty	Medicine	Medicine	Medicine	Medicine	Medicine	Medicine

a Comprehensive Education System (CES)
b Donald and Barbara Zucker School Of Medicine At Hofstra (ZSOM)
c Iran University Of Medical Sciences (IUMS)
d Mashhad University Of Medical Sciences (MUMS)
e Rocky Vista University College Of Osteopathic Medicine (RVUCOM)
f National Board of Chiropractic Examiners (NBCE)
g Medical College Admission Test (MCAT)
h Customized Assessment Service (CAS)
i Physician Assistant Clinical Knowledge Rating And Assessment Tool (PACKRAT)
j King Saud Bin Abdulaziz University For Health Sciences (KSAU-HS)
k Mayo Medical School (MMS)
l Comprehensive Basic Science Examination (CBSE) Scores

RQ3: What preprocessing and predictive models are commonly employed in high-stakes medical assessments?

Preprocessing in the predictive model is a critical step that involves transforming raw data into a format suitable for modeling. The input parameters in the included studies varied, with an average of 15±9 features. Notably, some studies (6 of 14) did not specify the exact number of features utilized (Table 5). The outcomes of the predictive models were classified as either continuous (exam scores) or categorical (pass/fail status, percentile, as well as ratings such as excellent, good, medium, and poor, and risk levels such as very likely to fail, moderate risk of failing, and low risk of failing).

In the preprocessing stage, five studies (P3, P4, P5, P7, P8, and P9) opted to transform the continuous target variable into a binary pass/fail status by defining a specific cutoff value (Table 5). The included studies involved various data-related challenges, such as missing values, high correlations between features, and imbalanced datasets, necessitating preprocessing techniques. In these studies, researchers employed various preprocessing methods to prepare the data for predictive modeling:

1. Descriptive analysis: Several studies (P3, P4, P6, P7, P8, and P11) conducted statistical tests, such as the chi-square test, Pearson correlation, and ANOVA, to analyze significant differences in feature values between groups. The chi-square test was used to determine the independence of categorical variables, the Pearson correlation was used to assess the linear relationship between continuous variables, and ANOVA was used to examine the differences in means across multiple groups.

2. Data Scaling: Five studies (P1, P3, P6, P7, and P11) performed data scaling via techniques such as z-score transformation and standard scaling (MinMax).

3. Missing Value Imputation: The studies (P2, P6, and P10)addressed missing values manually or by replacing them with mean values.

4. Feature Selection: Two studies (P3 and P13) utilized the Shapley Additive exPlanations (SHAP) method to determine the significance of features and select the most relevant features for the final models. In contrast, another study (P10) adopted a more traditional approach, employing univariate analysis to assess feature importance. An alternative clustering-based approach was used in study P13, which utilized the k-means algorithm to segment the dataset and identify groups of students with similar characteristics.

5. Imbalanced Data Handling: One study (P3) recognized the imbalanced distribution of classes (pass/fail) in high-stakes exam datasets and employed resampling techniques such as the Synthetic Minority Oversampling Technique and SMOTE Tomek to address this issue.

Most studies (78%) utilized regression models to predict scores on medical high-stakes exams, with linear regression, multivariable linear regression, and logistic regression used to describe the relationship between a dependent variable (exam score) and independent variables (student features). In addition to regression, the studies employed classic and ensemble ML classification models to predict student outcomes, such as pass/fail status (Table 6). The investigations included in this systematic review demonstrate a preference for support vector machines (SVMs), extreme gradient boosting (XGB), k-nearest neighbors (KNNs), artificial neural networks (ANNs), and ensemble models such as random forest (RF), which are the most popular ML models. SVM, XGB, KNN, ANN, and RF were utilized in 21%, 14%, 14%, 14%, and 14% of the studies, respectively (Table 6). One study (P6) also presented a novel model called AMM-KNN, an adaptive minimum match version of the KNN algorithm. Additionally, one study (P2) used a frequent pattern growth algorithm to extract frequent patterns in the dataset and generate rules.

Notably, the platforms used for modeling and evaluation are diverse. SPSS (4 studies), SAS (4 studies), and Python (3 studies) were the most popular, while R packages (2 studies), RapidMiner (1 study), and MATLAB (1 study) were also utilized. Only two studies, P3 and P6, provided code availability.

Table 5. Details of the preprocessing methods of the included studies

Paper No.	P1	P2	P3	P4	P5	P6	P7	P8
Preprocessing method	(1) Transform (z-score) (2) Cumulative average of the score (at five different time points)	(1) Replacing missing values with the mean value (2) ANOVAa	(1) Correlation using a Pearson and a Chi-square test (2) Data Scaling and Normalization (3) Feature Importance (SHAPb) (4) Resampling Imbalanced Data (SMOTEb)	(1) Correlation using Pearson correlation	-	(1) Manual preprocessing (2) Correlation using Pearson correlation (3) Data Normalization	(1) Normality analysis (2) Log-transformed (3) Assessing Multicollinearity	(1) Correlation of student characteristics
Features' types	Continues	Continues Categorical Binary	Continues Categorical Binary	Continues	Categorical	Continues Categorical	Categorical: Grades (A, B, C, and D)	Continues Categorical
No. features	3	8	18	-	15	-	-	-
Target feature	Continues (3-digit number)	Categorical (excellent, good, medium, and poor)	Continues Categorical	Continues with cutoff value (binary)	Continues with a cutoff value (binary)	Continuous Categorical (very likely, moderate, low risk of failing)	Continues	Continues with a cutoff value (at or below the 10th percentile)

Paper No.	P9	P10	P11	P12	P13	P14
Preprocessing method	-	(1) Excluded missing values (2) Univariate analysis	(1) Standardized (2) Analyzed statistically	-	(1) SHAP (2) k-means	-
Features' types	Continues Categorical	Continues Categorical	Continues	Continues	Categorical	Continues
No. features	14	9	22	37 preadmission and medical school variables	-	-
Target feature	Continues Categorical	Continues Categorical	Continues	Continues Categorical	Categorical	Continues

a One-way analysis of variance (ANOVA)
b SHapley Additive exPlanations (SHAP)
c Synthetic minority oversampling technique (SMOTE)

Table 6. Details of predictive models in included studies

Paper no.	P1	P2	P3	P4	P5	P6	P7	P8
Approach of ML	Regression	Regression, classification, and frequent pattern	Regression and classification	Regression	Regression	Classification	Regression	Regression and classification
Algorithms	-	-	-	-	-	-	-	-
Classification	Linear regression	MLP-ANNa and SVMb	SVM, KNNc, BGd, RFe, ADAf, XGBg, and stacking	-	-	AMM-KNN, RF, SVM, and KNN	-	-
Regression	Linear regression	Linear regression	Logistic regression	Multiple logistic regression, backward logistic regression, and logistic regression	Six regression models for each domain	-	Multivariate regression	Multivariable linear regression

Frequent pattern	-	Frequent pattern-growth algorithm	-	-	-	-	-	-

Paper no.	P9	P10	P11	P12	P13	P14
Approach of ML	Classification	Regression	Regression	Regression	Classification	Regression
Algorithms	-	-	-	-	-	-
Classification	HGLMsh	-	-	-	XGB	-
Regression	-	Multivariable linear regression	Multiple linear regression	Linear regression (stepwise)	-	linear regression
Frequent pattern	-	-	-	-	-	-

a Artificial Neural Network (ANN)
b Support Vector Machines (SVM)
c K-Nearest Neighbors (KNN)
d Bagging (Bootstrap Aggregation)
e Random Forest (RF)
f Adaptive Boosting (ADA)
g Extreme Gradient Boosting (XGB)
h Hierarchical Generalized Linear Models (HGLMs)

RQ4: What metrics and evaluation methods are used to assess predictive models in high-stakes medical assessments, and what are the key findings regarding their performance?

The reviewed studies have utilized various model evaluation methods. K-fold cross-validation, a technique in which the dataset is divided into k equal-sized subsets and the model is trained and tested k times via a different subset for testing each time, was used in 3 out of the 14 studies (21%). Leave-one-out cross-validation (LOOCV), a particular case of k-fold where each data point is used as the test set while the remaining data points are used for training, was used in 1 out of the 14 studies (7%). Some studies have used internal validation for training and test data partitioning, where the same dataset is split into training and testing portions. In contrast, others employed external validation by using data from different cohorts. Specifically, P3, P4, and P13 used 67%, 82%, and 75% of the records for training, respectively, and 33%, 18%, and 25% for testing, respectively. In contrast, P1, P4, and P9 used data from different years or cohorts for training and testing, with the most recent data used for the testing phase (Table 7).

The performance metrics employed were quite heterogeneous, particularly for the regression models, which poses a significant challenge in directly comparing the effectiveness of the different approaches. The metrics reported for classification tasks vary and include precision, sensitivity, specificity, accuracy, the F-measure, the MCC, the area under the receiver operating characteristic curve (AUC-ROC), the area under the precision-recall curve (AUC-PRC), the calibration plot, and the Brier score (BS). Additionally, confusion matrix metrics such as true positive (TP), false positive (FP), true negative (TN), and false negative (FN) are also reported. However, only three studies reported their evaluation metrics for classification tasks, with the most commonly used metrics being sensitivity, specificity, and AUC-ROC. For regression models, the reported evaluation metrics include the mean squared error (MSE), root mean squared deviation (RMSD), mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2). However, the three studies that reported evaluation metrics for regression tasks disagreed on a standard set of metrics, with each study reporting different evaluation measures. Figure 3 provides an overview of the evaluation metrics used in the literature and highlights the predominance of AUC, sensitivity, and specificity in classification-based studies.

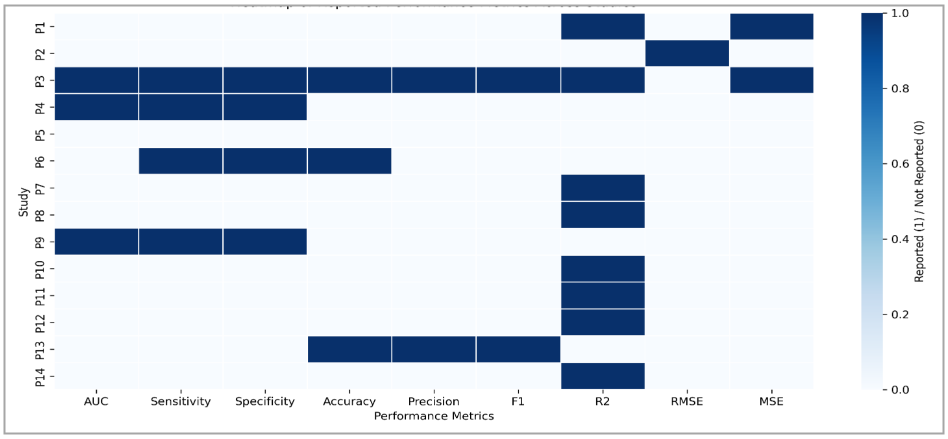

Figure 3. Distribution of performance metrics across the included studies

Table 7. Details of the evaluation methods of the included studies

Paper no.	P1	P2	P3	P4	P5	P6	P7	P8
Train	Classes of 2018, 2019, and 2020	-	67%	82%	-	181 students in four cohorts	-	-
Test	Class of 2021	-	33%	18%	-	43 students	-	-
Classification	-	-	Precision[e], Sensitivity[f], Specificity[g], Accuracy[h], F-measure[i], MCC[j], AUC-ROC[k], AUC-PRC[l], Calibration Plot, and BS[m]	Sensitivity, Specificity, TP[a], FP[b], TN[c], FN[d], and AUC-ROC	-	Confusion matrix, Sensitivity, Accuracy, Specificity, and F.P.	-	Sensitivity, Specificity
Regression	R^2 and MSE	RMSE[r] and MAE[q]	MSE[n], RMSD[o], and R^2[p]	-	-	-	R^2	R^2
Evaluation methods	10-fold cross-validation	K-fold cross	-	-	LOOCV, External validation	-	-	-
Platform	SAS	SPSS, Rapid Miner	Python	SPSS, SAS, Python	-	R	R	SAS
Code availability	No	No	Yes	No	No	Yes	No	No

Paper no.	P9	P10	P11	P12	P13	P14
Train	Five cohorts	-	-	-	75%	-
Test	2016 cohort	-	-	-	25%	-
Classification	AUC, sensitivity, specificity, and odds ratios.	-	-	-	Accuracy, Recall, Precision, F1	-
Regression	-	R^2	R^2	R^2	-	R^2
Evaluation methods	-	-	-	-	-	-
Platform	SPSS	SAS	SPSS	MATLAB	Python (Sklearn)	-
Code availability	No	No	No	No	No	No

a False positive (F.P.)
b True positive (T.P.)
c False negative (F.N.)
d True negative (T.N.)
e Precision (PPV=TP/(TP+FP))
f Sensitivity (Recall=TP/(TP+FN))
g Specificity (Spe=TN/(TN+FP))
h Accuracy (Acc= (TP+TN)/(TP+FP+FN+TN))
i F-measure (F1=2*(Precision*Recall)/(Precision+Recall))
j Matthew's correlation coefficient (MCC=(TP*TN – FP*FN)/√(TP+FP)(TP+FN)(TN+FP)(TN+FN))
k Area under the curve of the receiver operating characteristic (AUC-ROC)
l The area under the curve of precision-recall (AUC-PRC)
m Brier score (BS=mean squared error)

n Mean Squared Error (MSE)
o Root mean square deviation (RMSD)
p Coefficient of determination (R²)
q Mean absolute error (MAE)
r Root mean square error (RMSE)

The studies included in this review report findings across two aspects: the best-performing algorithms (methods and settings) and the evaluation results (Table 8). The top-performing models for regression and classification tasks varied across the studies. These models included logistic regression, multivariable linear regression, XGB, RF, stacking, ANN, and KNN. Moreover, the evaluation results indicated that in the classification task, study P3 achieved an AUC of 0.81, a precision–recall curve (PRC) of 0.93, a precision of 0.87, and an MCC of 0.52 when the RF was used. The stacking model in the same study attained a sensitivity of 0.83, an accuracy of 0.83, and an F1 score of 0.84. In study P4, the three predictive models had sensitivities ranging from 0.65 to 0.71 and specificities between 0.83 and 0.88 in predicting a score lower than 500 on COMLEX 1. Furthermore, in study P6, the AMM-KNN model achieved an impressive accuracy of 0.93 when LOOCV was used.

With respect to the regression models, study P1 predicted the target variable at the end of years 1 and 3, with MSEs of 8.3 points (SD=6.8) and 5.4 points (SD=4.1), respectively. Study P2 employed a multilayer perceptron (MLP-ANN) model, achieving an RMSE of 0.325 and an MAE of 0.212. Additionally, study P3 utilized a linear regression model, which attained an MAE of 0.09 and an RMSD of 0.13. Across all the studies, the researchers concluded that ML models enable student performance prediction, and even early identification of at-risk students is possible. As a result, instructors can offer constructive advice promptly, and these models can serve as suitable alternatives or additional tools alongside existing exams.

Table 8. Overview of findings from included studies

Paper no.	P1	P2	P3	P4	P5	P6	P7	P8
Importance and impact of input parameters	(1) NBME CAS exam scores are a significant factor in predicting performance. (2) MCAT performance had no bearing on the model. (3) Predict student performance without the USMLE Step 1 score.	A significant difference was observed among the six specialties regarding the mean scores of PGY-1 and preboard examinations.	The Anatomical Sciences, Biochemistry, Parasitology, and Entomology grades demonstrated the strongest positive correlation with the outcomes.	(1) MCAT scores and courses' final grades were statistically significantly correlated with the target. (2) Third-semester courses were most important in prediction.	(1) Age revealed a negative correlation with all the domains of the Part I exam. (2) Asian/Pacific Islander and White statuses were associated with the exam scores. (3) Educational level did not correlate significantly.	-	(1) Basic science courses (biology, chemistry, and physics) (2) Advanced science courses (medical terminology, behavioral science, biochemistry, biological statistics, and medical computer science)	NBME examination scores were correlated with MCAT, USMLE 1 and 2 scores, most clerkship scores (Medicine, Neurology, and Psychiatry), and FAS.
Best-performing algorithms	LR	The overall best result was MLP-ANN	RF and stacking demonstrate the best performance	The backward logistic regression model was the best.	LR	AMM-KNN, RF, SVM, and KNN	LR	LR

Evaluation results	Step 2 C.K. performance is predicted at the end of years 1 and 3 with an MSE of 8.3 points (SD=6.8) and 5.4 points (SD=4.1).	The overall best result was MLP-ANN with RMSE=0.325 and MAE=0.212	(1) Best AUC=0.81, PRC=0.93, precision=0.87, MCC=0.52 by RF and sensitivity=0.83, accuracy=0.83, and F1=0.84 by Stacking (2) LR model: MAE=0.09 and RMSD=0.13	The three predictive models had sensitivities of 65.8 - 71% and specificities of 83.2 - 88.2% in predicting scores.	Outstanding students tended to be underpredicted, while others were overpredicted.	(1) AMM-KNN achieved an accuracy of 93% in LOOCV. (2) RF, SVM, and KNN had high false-positive predictions.	(1) R^2= 27% for the basic science course (2) R^2= 22% for the advanced science course.	(1) 77% to 83% sensitivity and specificity (2) R^2=61.6%

Paper no.	P9	P10	P11	P12	P13	P14
Importance and impact of input parameters	Year 1 and 2 performance is a better predictor than admissions	MCAT and NIH funding impact scores	Early performance in the curriculum	(1) Microbiology exam (2) Pathology exam (3) CBSSA	-	Course assessments have a strong correlation with STEP 1 score.
Best-performing algorithms	Pre-MCCQE1 model	Multivariable linear regression.	Multiple linear regression	Linear regression (stepwise)	XGB	linear regression
Evaluation results	AUC=0.66–0.93	R^2=0.584	(1) GPA (r=0.13, R^2=0.02, P-value=0.01) (2) Undergraduate science GPA (r=0.11, R^2=0.01, P-value=0.03) (3) MCAT total score (r=0.26, R^2=0.07, P-value<0.001) 4) MCAT physical sciences score (r=0.15, R^2=0.02, P-value=0.01) (5) MCAT biological sciences score (r=0.23, R^2=0.05, P-value<0.001) (6) MCAT verbal reasoning score (r=0.09, R^2=0.02, P-value=0.09).	R^2=0.77	(1) Accuracies ranged from 0.855 to 0.899 (2) Weighted precision and weighted recall 0.856 to 0.899 (3) Weighted F1 score 0.899	R^2=0.66

RQ5: What is the potential of predictive models to complement traditional high-stakes medical assessments?

The potential of predictive models to complement traditional high-stakes medical assessments offers an intriguing prospective evolution in medical education. However, several common limitations were identified across the reviewed studies. Table 9 provides an overview of the reviewed articles' identified limitations.

Need for expanding the testing population: One major obstacle is the challenge of generalizing the proposed methods to other institutions and universities, primarily due to variations in curricula and grading systems. These differences were highlighted in several studies, including P1, P3, P4, P10, and P13. Another significant limitation is the use of relatively small datasets. Although adequate for preliminary exploration, the authors frequently emphasized the need for larger, more comprehensive datasets—ideally spanning multiple years and centers—to fully evaluate the models' efficacy.

Need for a broader array of features: Furthermore, many studies have highlighted the necessity of incorporating additional features to improve model accuracy and reliability. Nonacademic factors such as demographic information, mental health status, and previous educational achievements, such as scores on university entrance exams and cumulative GPAs, are crucial considerations for a holistic assessment approach. These factors can contribute significantly to predicting outcomes, emphasizing the need for a broadened scope that includes diverse data points.

Need for adaptable frameworks: The evolution of exams and testing protocols poses another layer of complexity. For example, Study P1 reported the shift of USMLE Step 1 from a scored exam to a pass/fail system, a change that requires recalibration of existing models to adapt to the new framework. Additionally, the COVID-19 pandemic has impacted data collection and assessment conditions, necessitating updates to the models to remain relevant amidst these evolving contexts.

Table 9. Overview of limitations, conclusions, and future work from included studies

Paper no.	P1	P2	P3	P4	P5	P6	P7	P8
Limitation	(1) USMLE Step 1 has changed, and the model needs to be updated. (2) Generalizability may be limited (not all institutions use NBME CAS exams)	(1) The present study did not include predicting board certification results based on all component scores of ITEs. (2) Medical education databases are suggested	(1) Model generalization due to the varying curricula across settings and countries is difficult. (2) No access to the students' performance in the university entrance exam.	(1) Used scores from the old MCAT. (2) Other medical schools may have different curricula and different scoring.	(1) The sample size used in this study is relatively small. (2) The assumption of normal distribution of the outcome variable is upheld after diagnostic analysis.	(1) Not considering nonacademic factors (personal, medical, and mental health). (2) May be affected by the COVID-19 pandemic.	-	(1) Single institution (2) Curriculum change (3) Not all scores were available for all students
Conclusion	Enable early student performance prediction on Step 2 C.K. without Step 1 quantitative data.	Identify at-risk residents, allowing instructors to offer timely constructive advice.	(1) Comparing classic and ensemble ML models revealed that ensemble models are superior. (2) A suitable alternative for medical licensing examinations.	Predictive models have similar accuracy in predicting poor performance on COMLEX 1.	The prediction patterns are systematically different for students with higher pre-chiropractic GPAs than others.	The developed ML model could be used as an additional tool along with already-existing tools.	GAT and SAAT may significantly predict students' GPA during the first two years of medical school study.	(1) Both standardized testing and FAS were highly predictive of NBME subject examination scores (2) Identification of students at risk for poor clinical Performance.
Future work	Considering demographic data	Including all residency programs, not only six	Testing on larger populations	Testing on a new version of the MCAT and grading	Testing on other exams as a model outcome	Generalize to other health professional programs	Explore the best possible preadmission variables that positively affect the performance of medical students	Generalize to other U.S. medical students

Paper no.	P9	P10	P11	P12	P13	P14
Limitation	(1) Variability among cohorts affects model accuracy (2) Requires periodic validation	(1) Single-year data limits generalizability (2) Outliers suggest curriculum impact	(1) Single-institution data. (2) Mixed models are needed for better accuracy.	(1) Pre-admission scores alone are weak predictors (2) Results specific to a single institution	(1) Limited demographic information (2) Generalizability due to the study's focus on a specific progress test and student group.	(1) Small sample size of students (2) GPA is not included.

Conclusion	Predictive modeling aids in early risk identification and intervention	Institutional variables like faculty ratio impact Step 1 scores	Early identification and intervention are recommended	Early medical school exam data improves predictions	ML clustering revealed distinct response patterns, helping improve the accuracy of student performance assessments in medical education	(1) Early academic assessments, like MCAT scores and GPA, are unreliable indicators of USMLE Step 1 success. (2) Integrating clinical and basic sciences assessments has better predictive value
Future work	Extend models to include MCCQE2 data.	Research on curriculum features of outperforming institutions.	Explore noncognitive factors (e.g., self-regulation, motivation) for improved predictive accuracy.	Use predictive model data to develop remediation programs for at-risk students for licensure exam failure.	(1) Expanding the approach to diverse medical programs and testing. (2) Additional predictive factors could enhance the model's reliability and broaden its applicability.	(1) Expanded Predictive Models. (2) Additional variables, such as demographics, study habits, and nonacademic skills. (3) Curriculum Variations. (4) Longitudinal Tracking.

Discussion

Predictive modeling represents a transformative approach in health professions education, offering new opportunities to enhance learning processes and support more responsive, learner-centered systems. Interpreting these models as practical tools rather than purely technical innovations highlights their direct implications for how future healthcare professionals are taught, supported, and evaluated. To synthesize the growing body of evidence in this area, the present systematic review examined the use of predictive models for forecasting medical students' performance on high-stakes assessments, drawing on 14 studies conducted across various healthcare education programs.

Predictive analytics offers a strategic mechanism for strengthening competency-based medical education (CBME). By identifying patterns of underperformance and potential skill gaps, these models can inform adaptive curriculum design and support mastery-oriented learning environments. Integrating predictive insights into student support systems may also promote individualized learning and well-being by enabling the early identification of students who may be at risk of academic difficulty or burnout, thereby facilitating timely mentoring and targeted interventions. Furthermore, predictive dashboards and feedback systems may empower students to engage more effectively in self-regulated learning by providing timely and data-informed feedback about their academic trajectories. At the institutional and policy levels, such tools have the potential to improve quality assurance and promote educational equity. However, their responsible implementation requires continuous attention to data governance, transparency, and algorithmic fairness to maintain institutional trust and protect student interests.

Despite the generally acceptable methodological quality of the included studies (each scoring above 14 out of 24 on the quality assessment), several important methodological limitations warrant consideration. Many studies did not fully meet key criteria related to risk of bias assessment, triangulation, and prospective study design. In particular, most of the included studies relied on retrospective datasets, which may introduce selection bias and limit the robustness of the reported predictive performance. In addition, the limited application of triangulation and the incomplete reporting of bias mitigation strategies may affect the reliability and interpretability of the findings. These methodological shortcomings should therefore be considered when interpreting the reported model performance, as retrospective designs and insufficient bias control may lead to overestimation of predictive accuracy and may restrict the generalizability of the results to broader educational and clinical contexts.

Beyond these methodological considerations, the findings of this review revealed distinct patterns and methodological complexities in the application of predictive models within medical education across four key domains: (1) the predominant predictors used in modeling student performance, (2) the types of predictive models employed, (3) the evaluation metrics and reported performance of these models, and (4) the potential of predictive analytics to complement or enhance existing high-stakes assessment systems. Examining these domains also enabled comparison of our findings with those reported in previous systematic reviews that investigated ML approaches for predicting academic performance in non-medical educational settings [13-16].

Predominant predictors utilized by predictive models: The results of the current study indicate that medical education tends to concentrate narrowly on demographic variables—such as age, gender, ethnicity, and residency status—and academic metrics such as GPA and standardized exam scores. These models primarily aim to identify at-risk students through examination performance and binary pass/fail outcomes. Notably, data in medical studies often derive from localized, private residency training databases, reflecting the field's rigorous privacy mandates and the sensitive nature of medical information. This reliance on private data sources, with datasets ranging from 147 [37] to 5,852 [38] records, highlights significant limitations, emphasizing the need for integrating additional features, including psychosocial factors and mental health indicators, to increase predictive accuracy.

Conversely, nonmedical educational domains include a broader array of variables. Systematic reviews in these fields reported academic and demographic indicators, socioeconomic factors, engagement metrics from learning management systems, family background, and psychosocial elements [13, 14]. This comprehensive approach seeks to predict academic outcomes and flag at-risk students, focusing on final grades and performance in traditional classrooms and massive open online courses (MOOCs) [15]. However, they often suffer from inadequate transparency regarding geographical data sources and methodological constraints in feature analysis, despite examining up to 150 distinct variables. These reviews primarily draw from a wide range of datasets, from private repositories with approximately 10,000 records [14] to large public databases with up to 155,000 [15] student entries. Addressing these challenges, recent studies have examined the performance of models in nonmedical high-stakes assessments via an extensive dataset of 1.3 million student records [39].

Moreover, the current systematic review revealed marked heterogeneity across sample features and analytical approaches. Sample sizes varied considerably (from 147 [37] to 5,852 [38]), with significant attrition rates (2**%** to 36%) due to data incompleteness. A particularly troubling finding was the inconsistent reporting of demographic variables—only 42.9% of studies reported gender distributions, whereas ethnic composition was detailed in just 7.1% of studies [34]. This inconsistency in demographic documentation spans datasets ranging from single-year cohorts to longitudinal studies exceeding a decade across diverse educational systems. These methodological disparities emphasize significant challenges in generalizing findings across various educational contexts. Our findings underscore the urgent need for standardized demographic documentation and harmonized data collection protocols across educational institutions to increase the reliability and applicability of predictive modeling efforts in education.

An interesting divergence emerged regarding the predictive value of MCAT scores across the included studies. While one study (P1) reported that MCAT scores were not significant predictors of USMLE Step 1 performance, another study (P2) found a significant correlation between MCAT scores and COMLEX Level 1 outcomes. Several factors may explain this discrepancy. First, the two licensing examinations differ in structure and underlying educational philosophies: the USMLE primarily reflects the allopathic medical curriculum, whereas the COMLEX examination integrates osteopathic principles and practice. These curricular and conceptual differences may influence how pre-admission metrics such as MCAT scores relate to later examination performance. Second, differences in the characteristics of the study populations may also contribute to variations in predictive relationships, as admission profiles, academic preparation, and institutional contexts can differ across these programs. Finally, methodological differences between the studies, including sample size, model construction, and the set of covariates included in the predictive models, may also affect the observed strength of the MCAT–performance relationship. These findings suggest that the predictive value of pre-admission metrics may be context-dependent and highlight the importance of considering institutional and examination-specific factors when developing predictive models in medical education.

Preprocessing and predictive models: In the realm of medical high-stakes assessments, distinct preprocessing challenges persist, primarily revolving around missing value imputation, dataset balancing, and standardization requirements. A notable preference for regression-based approaches emerges, with approximately 78% of studies utilizing these methods. Advanced algorithms, including XGB, SVM, KNN, and RF, often supplement such foundational techniques.

Both medical and nonmedical educational contexts notably lack the application of DL and XAI, presenting fertile ground for future research initiatives. Traditional statistical packages such as SPSS and SAS play a vital role, especially for conventional statistical analyses. These tools are effectively integrated with contemporary computational platforms, including R, Python, WEKA, RapidMiner, and MATLAB, illustrating medical education's unique positioning at the intersection of traditional methodologies and modern computational techniques.

Performance metrics and evaluation methods: In the evolving field of medical education analytics, the adoption of advanced predictive modeling techniques continues to set a high standard of performance. Leveraging ensemble methods such as the RF and XGB models alongside classical approaches such as logistic regression and MLP-ANN [26], the current methodologies produce impressive analytical metrics. These results are demonstrated by an AUC of 0.81 [17], sensitivity of 0.83 [26], accuracy of 0.93 [35], and specificity of 0.88 [25], reflecting a robust ability to predict and manage educational outcomes effectively.

Conversely, earlier structured analyses of nonmedical educational data have established a solid methodological foundation through the rigorous application of test-train-validation strategies and cross-validation techniques. These studies report commendable classification metrics, with accuracy rates ranging from 77.69% to 85.19%, F1 scores between 0.767 and 0.847, and AUC values between 70.4% and 86.3% [13]. Notably, logistic regression has demonstrated exceptional efficacy in dropout prediction, achieving an accuracy of 97.13%, precision of 98.95%, recall of 96.73%, and specificity of 97.14% [13].

Within the context of this systematic review, compelling trends emerge that connect sample size with model performance across both classification and regression frameworks. Notably, within classification analyses, it appears that larger sample sizes do not inherently guarantee superior performance. For example, one study involving 1,005 students and 18 variables [17]—including demographic information, GPA, and course grades—yielded commendable results: an AUC of 0.81, a precision of 0.87, and an F1 score of 0.84. In contrast, a different study with just 224 students [35] achieved a remarkable accuracy of 93%, effectively challenging the prevalent assumption that an increase in dataset size necessarily enhances predictive outcomes. Interestingly, another study comprising

885 students [25] reported more moderate results, with sensitivities between 65.8% and 71% and specificities ranging from 83.2% to 88.2%.

Regression analyses further highlight a diverse spectrum of outcomes across varying sample sizes. One study, with a cohort of 650 students, reported a moderate R-squared value of 27%, whereas another study involving 435 participants achieved a more significant predictive capacity, with an R-squared of 61.6%. Perhaps most strikingly, an investigation that integrated two datasets, totaling 542 students [26], realized an impressive R-squared value of 0.77. Moreover, a study with only 147 students [37] demonstrated robust efficacy, with an R-squared of 0.66.

These observations compellingly argue against the simplistic view that larger sample sizes inherently correlate with more accurate model performances. This multifaceted landscape suggests that refined methodological approaches, rather than increasing data abundance, are more indicative of optimal model performance.

Transition of examinations from numerical scoring to a pass/fail reporting system: Most predictive models identified in this review were developed when USMLE Step 1 scores were reported numerically. Accordingly, many studies implemented either regression models to predict continuous scores (MSE, RMSE, MAE, or R^2) or classification models to predict score thresholds and risk of low performance. The recent transition of Step 1 to a Pass/Fail reporting system may limit the direct applicability of models designed specifically for numerical score prediction. However, the underlying methodological frameworks remain relevant. Several studies already approached prediction as a classification task aimed at identifying students at risk of poor examination outcomes rather than estimating exact scores. For example, study P3 demonstrated strong classification performance (AUC=0.81, PRC=0.93), while study P4 predicted the probability of scoring below a threshold on COMLEX Level 1 with sensitivities ranging from 0.65 to 0.71 and specificities from 0.83 to 0.88. Such approaches remain applicable in a Pass/Fail context, where outcomes can be reframed as the probability of passing or failing. In addition, some studies showed that predictive models can estimate downstream outcomes without relying on a numerical Step 1 score. For instance, study P1 successfully predicted USMLE Step 2 CK performance using earlier academic indicators, including NBME CAS scores and prior academic performance. These findings suggest that predictive modeling retains its value even when Step 1 provides only a binary outcome. More broadly, the transition to Pass/Fail reporting may shift the focus of predictive modeling in medical education from predicting exact examination scores toward early identification of students at risk of failing, prediction of subsequent licensing examination outcomes (Step 2 CK, COMLEX, MCCQE), and the integration of broader academic and curriculum-based indicators.

Potential of predictive models to complement traditional high-stakes assessments: In exploring predictive models as potential enhancers for high-stakes assessments, our systematic review of medical education research reveals both shared and unique challenges in these approaches. A key issue is the restricted generalizability across various institutions due to differences in curricula and grading systems, accompanied by datasets limited in scope by sample size and temporal coverage, which curtails comprehensive evaluation. This highlights the critical need for extensive, multiyear, and multicenter data collection to increase model robustness and applicability.

Conversely, nonmedical education research faces constraints such as inadequate integration of socioeconomic and mental health considerations and the underutilization of advanced computational models such as DL [13]. Additionally, a significant gap persists due to the scarcity of action-based intervention studies, demonstrating a disconnect between theoretical and practical applications [15]. Compounding these challenges is the prevalence of opaque "black box" models, which pose challenges for stakeholders in deriving actionable insights for decision-making processes within educational environments.

Although the reviewed studies demonstrate the considerable potential of predictive models to identify students at risk of underperforming on high-stakes assessments, a critical gap remains in translating these predictions into effective educational interventions. As summarized in Table 9, only a limited number of studies extended their analyses beyond risk identification to examine how predictive insights could be operationalized within remediation or support frameworks. Without clearly defined intervention pathways, the practical value of early warning systems remains constrained. Future research should therefore prioritize the design, implementation, and rigorous evaluation of targeted remediation strategies (such as academic coaching, adaptive learning modules, or structured mentoring) triggered by model-based risk predictions. Importantly, such intervention studies should evaluate not only short-term academic outcomes but also broader impacts on learner well-being, equity, and long-term professional development, thereby closing the loop between prediction, intervention, and meaningful educational impact.

Conclusion: Assessing medical students' proficiency through high-stakes assessments remains a significant challenge for educational systems worldwide. While these assessments are crucial in determining students' futures, institutions struggle with cost management and lengthy evaluation processes. This systematic review demonstrates the emerging potential of predictive models as valuable complementary tools or alternatives to traditional high-stakes assessments in medical education, particularly in their capacity to identify at-risk students and accurately predict performance outcomes.

Our systematic analysis revealed several crucial insights into predictive applications in medical education. Academic metrics, specifically course grades and GPAs, consistently emerged as the strongest predictors of examination outcomes across studies. While the prevalent use of private residency training databases reflects healthcare's necessary privacy mandates, this reliance on restricted data sources presents significant limitations. Future models would benefit from incorporating a broader range of variables, including psychosocial factors and mental health indicators, to enhance predictive accuracy. Furthermore, the diverse approaches to demographic documentation—from single-year cohorts to decade-long longitudinal studies across different educational systems—underscore the methodological challenges in drawing universally applicable conclusions.

The reviewed predictive models showed remarkable versatility, effectively predicting both numerical examination scores and binary pass/fail outcomes. This capability is especially valuable for identifying at-risk students early, enabling timely interventional support. Traditional regression-based approaches, particularly linear and logistic regression, dominate the analytical landscape and are chosen primarily for their interpretability and implementation simplicity. However, the successful application of more advanced algorithms (SVM, KNN, XGB, and RF) demonstrates the field's evolution toward more sophisticated predictive modeling approaches.

A comparative analysis with systematic reviews from nonmedical domains reveals striking similarities in limitations and challenges. Both medical and nonmedical fields require enhancement in three critical areas: the integration of comprehensive predictive variables (demographic, psychological, and pre-matriculation factors), the implementation of robust longitudinal tracking, and the development of more extensive multi-institutional studies. These parallel challenges emphasize the need for more sophisticated, transparent research methodologies (DL, reinforcement models, and XAI) that can effectively capture the complex determinants of student success.

In conclusion, this systematic review provides valuable insights for educators and policymakers to initiate meaningful reforms in medical education. The strategic implementation of methodologies offers a transformative opportunity to enhance assessment practices, enabling more personalized and effective educational experiences for future healthcare professionals. By addressing the identified limitations and leveraging these technological capabilities, institutions can optimize learning processes, refine curricula, and significantly improve student outcomes in high-stakes assessments.

Ethical Considerations

The study was approved by the Iran National Committee for Ethics in Biomedical Research (Approval ID: IR.NASRME.REC.1400.029) on May 23, 2021.

Acknowledgment

This project was funded by the National Agency for Strategic Research in Medical Education (Tehran, Iran) under Grant No. 994079. The authors would like to express their sincere gratitude to the National Agency for Strategic Research in Medical Education and Mashhad University of Medical Sciences for their financial support and resources, which made this research possible and facilitated the development of this article. The authors used AI-assisted language tools (e.g., Grammarly and QuillBot) were strictly limited to language polishing and not used for data interpretation or synthesis. All aspects of research design, analysis, and interpretation remain the intellectual responsibility of the authors, in accordance with international authorship standards.

Conflict of Interest

The authors declare no competing financial or non-financial interests.

Funding

This project was funded by the National Agency for Strategic Research in Medical Education. Tehran. Iran. Grant Number 994079. We thank the National Agency for Strategic Research in Medical Education (NASR) for funding the project (Grant Number 994079), which provided the context for the writing of this article (Webpage of the grant number: http://rms.nasrme.ac.ir).

References

1. Bhanji F, Naik V, Skoll A, Pittini R, Daniels VJ, Bacchus CM, et al. Competence by Design: The Role of High-Stakes Examinations in a Competence Based Medical Education System. Perspectives on Medical Education. 2024;13(1):68-74. doi: 10.5334/pme.965

2. Mee J, Pandian R, Wolczynski J, et al. An experimental comparison of multiple-choice and short-answer questions on a high-stakes test for medical students. Advances in Health Sciences Education. 2024;29(3):783-801. doi: 10.1007/s10459-023-10266-3

3. Templeton K, Vanston P, Luciw-Dubas UA, et al. Navigating a High-Stakes Assessment in Medical School: Students' Lived Experiences During a Stressful Period of Preparation. Journal of Medical Regulation. 2022;108(2):7-18. doi: 10.30770/2572-1852-108.2.7

4. Gosnell PW. Bridging the Gap Between High-Stakes Testing and 21st Century Skills. Clemson University; 2023.

5. Thoma B, Monteiro S, Pardhan A, Waters H, Chan T. Replacing high-stakes summative examinations with graduated medical licensure in Canada. Canadian Medical Association journal. 2022;194(5):E168-E170. doi: 10.1503/cmaj.211816

6. Lucey CR, Hauer KE, Boatright D, Fernandez A. Medical education's wicked problem: Achieving equity in assessment for medical learners. Academic Medicine. 2020;95(12S):S98-S108. doi: 10.1097/ACM.0000000000003717

7. French S, Dickerson A, Mulder RA. A review of the benefits and drawbacks of high-stakes final examinations in higher education. Higher Education. 88, 893-918 (2024). doi: 10.1007/s10734-023-01148-z

8. Franco D'Souza R, Mathew M, Mishra V, Surapaneni KM. Twelve tips for addressing ethical concerns in the implementation of artificial intelligence in medical education. Medical Education Online. 2024;29(1):2330250. doi: 10.1080/10872981.2024.2330250

9. Lykourentzou I, Giannoukos I, Nikolopoulos V, Mpardis G, Loumos V. Dropout prediction in e-learning courses through the combination of machine learning techniques. Computers and Education. 2009;53(3):950-965. doi: 10.1016/j.compedu.2009.05.010

10. Asif R, Merceron A, Ali SA, Haider NG. Analyzing undergraduate students' performance using educational DM. Computers and Education. 2017;113:177-194. doi: 10.1016/j.compedu.2017.05.007

11. Ojajuni O, Ayeni F, Akodu O, et al. Predicting Student Academic Performance Using Machine Learning. Springer International Publishing; 2021:481-491. doi: 10.1007/978-3-030-87013-3_36

12. Sobiesuo EM, Edmond A, Issaka CA, Appiah SCY. Knowledge of AI use and prediction of AI adoption among selected residents in the greater Kumasi area of Ghana. Discover Artificial Intelligence. 2025;5(1):1-15. doi: 10.1007/s44163-025-00558-5

13. Alalawi K, Athauda R, Chiong R. Contextualizing the current state of research on the use of machine learning for student performance prediction: A systematic literature review. Engineering Reports. 2023;5(12):e12699. doi: 10.1002/eng2.12699

14. Fahd K, Venkatraman S, Miah SJ, Ahmed K. Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. Education and Information Technologies. 2022:1-33.

15. Sekeroglu B, Abiyev R, Ilhan A, Arslan M, Idoko JB. Systematic literature review on machine learning and student performance prediction: Critical gaps and possible remedies. Applied Sciences. 2021;11(22):10907. doi: 10.3390/app112210907

16. Zhai X, Yin Y, Pellegrino JW, Haudek KC, Shi L. Applying machine learning in science assessment: a systematic review. Studies in Science Education. 2020;56(1):111-151. doi: 10.1080/03057267.2020.1735757

17. Mastour H, Dehghani T, Moradi E, Eslami S. Early prediction of medical students' performance in high-stakes examinations using machine learning approaches. Heliyon. 2023;9(7):e18248. doi: 10.1016/j.heliyon.2023.e18248

18. Mastour H, Dehghani T, Moradi E, Eslami S. Explainable artificial intelligence for predicting medical students' performance in comprehensive assessments. Scientific Reports. 2025;15(1):23752. doi: 10.1038/s41598-025-07460-1

19. Mastour H, Dehghani T, Jajroudi M, Moradi E, Zarei M, Eslami S. Prediction of medical sciences students' performance on high-stakes examinations using machine learning models: a protocol for a systematic review. BMJ Open. 2023;13. doi: 10.1136/bmjopen-2022-064956

20. Amir-Behghadami M, Janati A. Population, Intervention, Comparison, Outcomes and Study (PICOS) design as a framework to formulate eligibility criteria in systematic reviews. Emergency Medicine Journal. 2020;37(6):387. doi: 10.1136/emermed-2020-209567

21. Buckley S, Coleman J, Davison I, Khan KS, Zamora J, Malick S, et al. The educational effects of portfolios on undergraduate student learning: a Best Evidence Medical Education (BEME) systematic review. BEME Guide No. 11. Medical teacher. 2009;31(4):282-298. doi: 10.1080/01421590902889897

22. Mortaz Hejri S, Jalili M, Shirazi M, Masoomi R, Nedjat S, Norcini J. The utility of mini-Clinical Evaluation Exercise (mini-CEX) in undergraduate and postgraduate medical education: protocol for a systematic review. Systematic Reviews. 2017;6(1):1-8. doi: 10.1186/s13643-017-0539-y

23. Li M, Gao Q, Yu T. Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters. BMC Cancer. 2023;23(1):799. doi: 10.1186/s12885-023-11325-z

24. Nichols TR, Wisner PM, Cripe G, Gulabchand L. Putting the kappa statistic to use. The Quality Assurance Journal. 2010;13(3-4):57-61. doi: 10.1002/qaj.481

25. Zhong Q, Wang H, Christensen P, McNeil K, Linton M, Payton M. Early prediction of the risk of scoring lower than 500 on the COMLEX 1. BMC Medical Education. 2021;21(1):70. doi: 10.1186/s12909-021-02501-5

26. Casey PM, Palmer BA, Thompson GB, et al. Predictors of medical school clerkship performance: a multispecialty longitudinal analysis of standardized examination scores and clinical assessments. BMC Medical Education. 2016;16:128. doi: 10.1186/s12909-016-0652-y

27. Ghaffari-Rafi A, Lee RE, Fang R, Miles JD. Multivariable analysis of factors associated with USMLE scores across US medical schools. BMC Medical Education. 2019;19(1):154. doi: 10.1186/s12909-019-1605-z

28. Sieg M, Roselló Atanet I, Tomova MT, et al. Discovering unknown response patterns in progress test data to improve the estimation of student performance. BMC Medical Education. 2023;23(1):193. doi: 10.1186/s12909-023-04172-w

29. Bird JB, Olvet DM, Willey JM, Brenner JM. A Generalizable Approach to Predicting Performance on USMLE Step 2 CK. Advances in Medical Education and Practice. 2022;13:939-944. doi: 10.2147/AMEP.S373300

30. Althewini A, Al Baz N. Prediction of Admission Tests for Medical Students' Academic Performance. Advances in Medical Education and Practice. 2022;13:1287-1292. doi: 10.2147/AMEP.S355474

31. Lee MW, Johnson TR, Kibble J. Development of statistical models to predict medical student performance on the USMLE Step 1 as a catalyst for deployment of student services. Medical Science Educator. 2017;27:663-671. doi: 10.1007/s40670-017-0452-y

32. Gullo CA, McCarthy MJ, Shapiro JI, Miller BL. Predicting medical student success on licensure exams. Medical Science Educator. 2015;25:447-453. doi: 10.1007/s40670-015-0179-6

33. Amirhajlou L, Sohrabi Z, Alebouyeh MR, et al. Application of DM techniques for predicting residents' performance on pre-board examinations: A case study. Journal of Education and Health Promotion. 2019;8:108. doi: 10.4103/jehp.jehp_394_18

34. Himelfarb I, Shotts BL, Gow AR. Examining the validity of chiropractic grade point averages for predicting National Board of Chiropractic Examiners Part I exam scores. Journal of Chiropractic Education. 2022;36(1):1-12. doi: 10.7899/JCE-20-5

35. Kumar A, DiJohnson T, Edwards RA, Walker L. The Application of Adaptive Minimum Match k-Nearest Neighbors to Identify At-Risk Students in Health Professions Education. Journal of Physician Assistant Education. 2023;34(3):171-177. doi: 10.1097/JPA.0000000000000513

36. Barber C, Hammond R, Gula L, Tithecott G, Chahine S. In search of black swans: identifying students at risk of failing licensing examinations. Academic Medicine. 2018;93(3):478-485. doi: 10.1097/ACM.0000000000001938

37. Puri N, McCarthy M, Miller B. Validity and Reliability of Pre-matriculation and Institutional Assessments in Predicting USMLE STEP 1 Success: Lessons From a Traditional 2 x 2 Curricular Model. Frontiers in Medicine. 2022;8:798876. doi: 10.3389/fmed.2021.798876

38. Haupt F, Kanzow P. The Relation Between Students' Theoretical Knowledge and Practical Skills in Endodontics: Retrospective Analysis. Interactive Journal of Medical Research. 2023;12:e46305. doi: 10.2196/46305

39. Rayhan M, Alam MGR, Dewan MAA, Ahmed MHU. Appraisal of high-stake examinations during SARS-CoV-2 emergency with responsible and transparent AI: Evidence of fair and detrimental assessment. Computers and Education: Artificial Intelligence. 2022;3:100077. doi: 10.1016/j.caeai.2022.100077