The Use of Artificial Intelligence in Residency Application Evaluation—A Scoping Review

Maxwell D. Sumner, BS
T. Clark Howell, MD, MSHS
Alexandria L. Soto, BS
Samantha Kaplan, PhD
Elisabeth T. Tracy, MD
Aimee K. Zaas, MD
John Migaly, MD
Allan D. Kirk, MD, PhD
Kevin Shah, MD

ABSTRACT

Background Several residency programs have begun investigating artificial intelligence (AI) methods to facilitate application screening processes. However, no unifying guidelines for these methods exist.

Objective We sought to perform a scoping review of AI model development and use in residency/fellowship application review, including if bias was explored.

Methods A scoping review was performed according to PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines where a systematic search strategy identified relevant literature within the databases MEDLINE, Embase, and Scopus from inception to September 29, 2023. No limitations on language, article type, or geographic affiliation were placed on the search parameters. Data were extracted from relevant documents, and study findings were synthesized by the authors.

Results Twelve studies met inclusion criteria. Most used AI to predict interviews or rank lists (9 of 12, 75%), while the remaining 3 articles (25%) evaluated letters of recommendation with natural language processing. Six articles (50%) compared the model's output to a human-created rank list. Most of the reviewed articles (9 of 12, 75%) mention bias; however, few explicitly modeled biases by accounting for or examining the effect of demographic factors (3 of 12, 25%).

Conclusions Few studies have been published on incorporating AI into residency/fellowship selection, and bias remains largely unexplored. There is a need for standardization in bias and fairness reporting within this area of research.

Introduction

Residency and fellowship programs strive to use a holistic review to find the optimal pairing between applicants and programs.[1,2] However, holistic review's widespread implementation has been limited by its time-consuming nature, constraints on resources, human biases, and the absence of a universally accepted methodology.[3-5] These challenges have been exacerbated by the surge in applicant submissions over the last 5 years.[6,7] Therefore, residency programs have returned to metric-based cutoffs to narrow the number of applicants considered by each program.[8-12] Consequently, across specialties, between 30% to 65% of residency applications encounter rejection before undergoing holistic review.[13] This may inadvertently uphold inequitable practices in residency and fellowship recruitment.[14,15]

In parallel, there has been rapid adoption and innovation within artificial intelligence (AI)—namely natural language processing, generative pre-trained transformers, and machine learning models—due to its ability to quickly aggregate and summarize large amounts of data. With AI's mainstream ascendence, concerns have been raised about algorithmic bias and regulation.[16] These concerns motivated the passing of regulatory legislation at multiple levels of government.[17-19] As a result, over 31 states have adopted resolutions or enacted legislation regarding the use of AI.[20]

Within this broader context, several residency programs have investigated using generative AI methods to facilitate applicant screening with the primary aims of reducing the resources required to equitably screen all applications,[3,15,21] making application review more standardized,[22-25] identifying resident values in unstructured data,[26-28] and avoiding bias.[29,30] Institutions innovating with generative AI deserve recognition for their early efforts to utilize AI in practice, particularly because these methods will likely become a standard part of application review. However, there are growing concerns about the presence of bias, whether implicit or explicit, in the AI models incorporated into determining the future of physicians-in-training.[16] For this reason, we aimed to perform a scoping review of English

DOI: http://dx.doi.org/10.4300/JGME-D-24-00604.1

Editor's Note: The online supplementary data contains the search strategies used in the study and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews checklist.

language manuscripts documenting how AI, broadly defined, has been developed for uses within residency or fellowship application review and how those models account for bias.

Methods

The use of AI in trainee application evaluation is a broad and understudied topic, making a scoping review the most suitable methodology for identifying key trends and drawing overall conclusions within the field. The scoping review was performed according to published PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Scoping Reviews) guidelines.[31] A systematic search strategy identified relevant literature. Data was extracted from eligible documents, and findings were synthesized. Results have been reported in alignment with the PRISMA-ScR Checklist (online supplementary data).

Search Strategy

A medical librarian with expertise in systematic searching composed a sensitive search utilizing a mix of keywords and subject headings that represented the concepts of recruitment, residency, and artificial intelligence. A second medical librarian peer-reviewed the search strategy in accordance with the modified PRESS (Peer Review of Electronic Search Strategies) checklist.[32] No limitations on language, article type, or geographic affiliation were placed on the search parameters. The databases MEDLINE, Embase, and Scopus were searched from inception to September 29, 2023. All search results were compiled in EndNote (Clarivate) and imported into Covidence for deduplication and subsequent screening. The complete search strategies are shown in the online supplementary data.

Study Selection: Title and Abstract Screening

Articles were deemed eligible for this study if they met the following criteria: (1) reported on research; (2) mentioned or alluded to residency recruitment or selection; and (3) mentioned or alluded to AI. Articles that solely discussed AI in medical school admissions and did not comment on residency recruitment were excluded from this study. A single study member (M.D.S.) evaluated all titles and abstracts identified by the database screen for eligibility.

Study Selection: Full-Text Screening

The articles identified as eligible in the preliminary title and abstract screen underwent full-text screening by 2 study members (M.D.S., A.L.S.). The reviewers independently screened the full text of each article through Covidence. Any disagreements were resolved through discussion and the contribution of a third reviewer (T.C.H.). During this stage of review, an article on the use of ChatGPT (OpenAI) by residency applicants met the original inclusion criteria but was excluded from this study. This article was determined to be outside of the study scope because it did not describe how residency programs utilized AI for applicant evaluation. All other articles meeting eligibility criteria during full-text screen were marked for data extraction.

Data Extraction and Synthesis

The data extraction sheet was developed iteratively to ensure relevant and consistent information capture. Three articles were randomly selected for pilot testing of the data extraction sheet, and revisions were made as needed. Eligible articles were then randomly assigned to 2 reviewers, with the third reviewer resolving disagreements. Reviewers (M.D.S., A.L.S., T.C.H.) extracted data independently, each reviewing 8 of the following groupings: study ID, title, authorship, corresponding author appointment and contact details, country, if the study was multi-institutional, primary institution, study aims, study design, date published, study funding sources, specialty of interest, study inclusion and exclusion criteria, total participants/applications used, type of AI model, model performance factors, model endpoints, model deployment, if model is currently in use, mention of bias, bias explicitly modeled, narrative data model used, comparison to human ranking, and interpretable associations noted. Bias for this study was defined as either acknowledging or evaluating, directly or indirectly, systemic differences between protected or demographic groups (race, gender, etc). Results were summarized as data extraction fields and a selected cross-tabulation analysis was performed (TABLE 1).

Results

Search Results

We identified 1048 articles through the systematic database search. Following the removal of duplicates, 816 (78%) records remained and underwent screening. After this review, 799 (98%) records were excluded during the title and abstract screen, and the remaining 17 (2%) underwent a full-text screen for eligibility. Of these 17 articles, 5 (29%) were excluded for reasons highlighted in FIGURE 1, and therefore 12 (71%) were included in the scoping review.

TABLE 1
Characteristics of Examined Full-Text Articles

Study Characteristics	Full-Text Articles (N=12), n (%)
Years published	
<2000	1 (8)
2000-2010	1 (8)
2011-2020	0 (0)
2021-2023	10 (83)
Location	
United States	12 (100)
Institution type	
Academic medical centers	12 (100)
Multi-institutional	0 (0)
Study design	
Cohort	6 (50)
Cross-sectional	5 (42)
Nonrandomized experimental study	1 (8)
Primary outcome	
Creation of predicted rank list	9 (75)
LOR evaluation	3 (25)
Model development	
Reported model performance	6 (50)
Compared to human output	6 (50)
Listed attributable features	6 (50)
Bias	
Bias included in text	9 (75)
Only included as limitation	6 (42)
Bias explicitly modeled	3 (25)

Abbreviation: LOR, letters of recommendation.

Study Characteristics

Most articles (10 of 12, 83%) identified for this review were published in the last 3 years, with 6 (50%) of them having been published in 2023. All 12 articles were from the United States and were published by large academic health systems with residency or fellowship programs. However, none of the articles reviewed were multi-institutional. A range of specialties were represented in the literature, including internal medicine, general surgery, internal medicine/pediatrics, neurosurgery, emergency medicine, psychiatry, and pediatric otolaryngology. Within the reviewed literature, investigators used AI to complete one of 2 tasks. They either trained AI to analyze application material to predict outcomes (9 of 12, 75%) or used it to conduct textual/sentiment analysis on the letters of recommendation (3 of 12, 25%). An overview of study characteristics and findings can be found in TABLE 1.

Model Development

Authors used either free text or discrete data fields from submitted applications to develop their AI model. Some models (3 of 12, 25%) attempted to integrate both data types to represent a "holistic AI review." Models utilizing free text often worked to transform the text into discrete fields (eg, hobbies, text snippets). While half of the articles reviewed (6 of 12, 50%) did not include summary statistics of model performance, the articles that did report on model performance had only moderate average precision (TABLE 2). Only half of the models developed (6 of 12, 50%) compared model output to an analogous human generated output. Of the 12 articles identified, 6 reported feature importance, revealing the weights of which features are most important for the outcome variable of interest (TABLE 3).

Acknowledgment of Bias

Only a small proportion of reviewed articles explicitly address bias in their final models (3 of 12, 25%). In the remaining articles, 6 articles (50%) listed bias as a limitation without explicitly modeling bias, while the last 3 articles (25%) did not discuss bias.

Discussion

This scoping review aimed to assess how AI has been developed for residency or fellowship application evaluation and how model bias has been explored within the research literature. Our review identified a limited body of literature, with only 12 studies meeting inclusion criteria. These studies primarily focused on 2 AI use cases in academic residency and fellowship programs: (1) predicting rank lists from application features and (2) conducting textual analysis such as summarizing personal statements or letters of recommendation. Although these technologies hold promise for improving the residency selection process, few studies directly address bias.

While AI tools are still in the early stages of development, their use in evaluating residency and fellowship applications could introduce bias if not carefully designed.[16] Only a few studies in our review considered bias as a part of their methodology, a concerning omission given that AI models trained on skewed historical training data can perpetuate harmful patterns. Furthermore, there is no evidence that current AI models can create an output akin to that produced by a program's current methods. The lack of inter-reliability and comparison between AI and human methods prevents conclusions about AI's ability to replicate or improve decision-making. This is especially true when

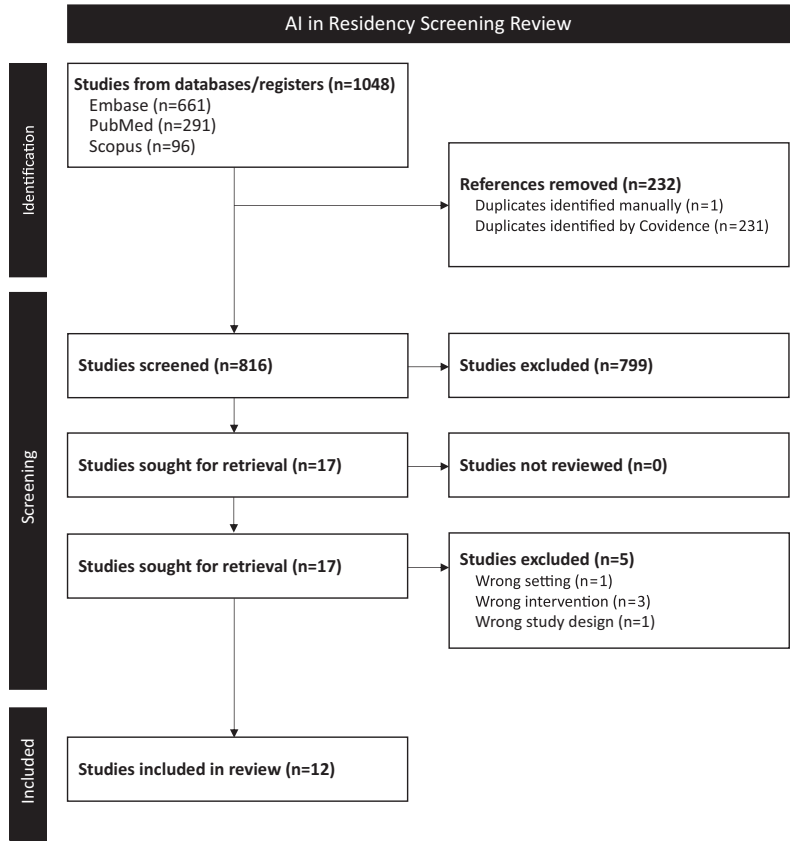

FIGURE 1
PRISMA Flow Diagram
Abbreviations: PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; AI, artificial intelligence.

explainability measures, like feature importance, remain absent from published analyses.

The variability in AI model development and validation also highlights the need for unified guidelines and reporting standards.[33] Half of the studies we reviewed included performance summary statistics. However, what was reported varied, meaning there is still a pressing need for universally accepted reporting guidelines. Other academic domains have published standards, and these guidelines can act as a guide for AI data reporting in medical education.[34-37] We advocate for the inclusion of accuracy, sensitivity, specificity, precision, feature importance, explicit bias and fairness assessments, program disclosure, and exploring potential implications of model implementation in future publications.[38] Without such standards, accepting black box models where internal processes are not observable poses risks to transparency and fairness, potentially hindering the positive benefits of diversity in residency programs.[39,40]

Explainable AI, which can clarify how specific aspects of applications are weighted in decision-making, holds promise for improving equity and transparency. One potential workflow of explainable AI is shown in FIGURE 2. However, transparency also presents the risk that applicants might tailor their materials to game the system.[41-43] Nevertheless, AI offers the potential for more consistent, transparent, and equitable application review. Further, feature importance can provide internal transparency to confirm emphasis on mission-driven values, such as research experience or clinical interest. It should be noted that bias against protected categories, such as race or gender, should be carefully mitigated. It is crucial to address algorithmic bias through explainability tools, such as SHAP (SHapley Additive exPlanations) values and LIME (Local Interpretable Model-agnostic Explanations), and to incorporate fairness audits into model development, such as IBM's AI Fairness 360 and the University of Chicago's Aequitas Bias and Fairness Audit Toolkit.[37,44-46]

Despite the limitations of our study, including that the referenced articles did not have a single standard method of model evaluation (limiting the comparisons of performance), the use of a single reviewer for our initial screening of titles and abstracts rather than 2, and the exclusion of gray literature, this scoping review offers a comprehensive assessment of AI use

TABLE 2
Summary of Reviewed Articles

Publication	Model Type	Input	Output	Model Performance						Brief Description
				Sensitivity[a]	Specificity	AUROC	PR Curve	Accuracy	Precision[b]	
St John et al (2022)[25]	Machine learning/ natural language processing	PS	Interview invite Rank list							ML and NLP extracted metrics from PS, and regression analysis correlated output with interview offer and final rank position. ML/NLP can identify meaningful variables that differentiate candidates. Proposed use: identify applicants who would be a good fit for training programs, which can result in a more holistic and accurate match. Bias not mentioned.
Stedman et al (2009)[24]	LIWC	LoR Gender	Positive and negative attribute frequency							LIWC scanned LoRs to produce a quantitative profile for 16 categories. Analysis demonstrated minimal variance between LoRs. Authors question the utility of LoRs as presently structured. Possible changes are discussed. Bias incorporated into model.
Sarraf et al (2021)[30]	IBM Watson Natural Language Understanding Tone Analyzer	LoR	Word frequency Word sentiment							AI and computer-based algorithms reviewed the LoRs of previously matched applicants and detected linguistic differences and gender bias. Results persisted following stratification by clerkship grades and when analyzed by decade. Bias incorporated into model.
Rees et al (2023)[3]	Random forest	Demographics School Details Applicant Achievement (test scores, awards, grades, publications) Hobbies/interests	Rank list Matriculation			Model 1: 0.93 Model 2: 0.60	Model 1: 0.65 Model 2: 0.11			A random Forest algorithm assessed 72 variables to predict ranked applicants. The produced rank list was compared to actual outcome and demonstrated impressive accuracy. Methods were repeated to predict ranked matriculants, which demonstrated modest but better-than-random accuracy. Bias incorporated into discussion and limitations but not modeled. Model 1: Model of ranked vs unranked candidates Model 2: Model of ranked matriculants vs nonmatriculants
Pilon et al (1997)[23]	Artificial neural network	PS SCG/USMLE LoR Interview score	Rank list					Model 1: CC: 0.74 R^2: 59.0% Model 2: CC: 0.77 R^2: 59.4%		Using linear regression and then a neural network, 2 rank lists were produced. Both outputs were compared to the actual rank order submitted to the NRMP. When compared to the NRMP rank order, the 2 forecasting models performed equally well. Bias not mentioned. Model 1: Linear regression Model 2: Neural network
Ortiz et al (2023)[28]	Python NLTK LASSO Demographics model	Demographics ERAS features (gender, IMG, AOA, GHHS, location, PhD, no. of publications, activities) LoR	Matched applicants			Model 1: 0.75 Model 2: 0.80 Model 3: 0.72				NLORs and demographic data similarly discriminate whether applicants will or will not match into their neurosurgical residency program. NLORs potentially provide further insight regarding applicant fit. As NLORs are predictive of both Match outcomes and SLOR rankings, continuing to include narrative evaluations may be invaluable to the Match process. Bias not mentioned. Model 1: NLOR model to predict Match Model 2: Demographics model (ERAS features) Model 3: NLOR model to predict SLOR ranking

TABLE 2
Summary of Reviewed Articles (continued)

Publication	Model Type	Input	Output	Model Performance						Brief Description
				Sensitivity[a]	Specificity	AUROC	PR Curve	Accuracy	Precision[b]	
Mahtani et al (2023)[15]	LASSO "Bag-of-words" with SpaCy v3.0 TF-IDF	Demographics ERAS features (see Burk-Rafel et al) Notable experience Work experience Volunteer experience Research experience	Interview invite			Model 1: 0.80 Model 2: 0.92 Model 3: 0.92	Model 1: 0.49 Model 2: 0.74 Model 3: 0.73			NLP-based artificial intelligence tool to promote holistic residency application review. Proposed use: identify applicants screened out using traditional metrics. Bias acknowledged as a limitation only. Model 1: NLP model of narrative experience Model 2: ERAS structured data only Model 3: Combined Model 1 + Model 2
Kibble et al (2023)[27]	MonkeyLearn Lexalytics MeaningCloud	MSPE	Rank list							The rubric for manual grading provided reliable interfaculty scoring and ranking of MSPEs. While the MLMs accurately detected positive sentiment in the MSPEs, they were unable to provide reliable rank ordering compared to human raters. Bias indirectly acknowledged in discussion.
Gray et al (2022)[29]	VADER	Demographics LoR	Polarity score							Bias in LORs, as reflected as differences in polarity, is likely a result of the intensity of the emotions being used and not the types of emotions being expressed. NLP shows promise in identification of subtle areas of bias that may influence an individual's likelihood of successful matching. Bias incorporated into model.
Drum et al (2023)[26]	RoBERTa	PS LoR MSPE Notable experience	Interview Invite	0.62	0.97				0.64	MLM created that can identify several values important for resident success in internal medicine–pediatrics programs with moderate sensitivity and high specificity using text snippets. Bias acknowledged as a limitation only.
Chillakuru et al (2022)[22]	BERT PCA XGBoost	PS No. of publications No. of presentations ENT residency Doximity rating USMLE scores Gender	Applicant fit Post-fellowship achievement	Model 1: 0.00 Model 2: 0.20 Model 3: 0.20		Model 1: 0.42 Model 2: 0.76 Model 3: 0.75		Model 1: 0.71 Model 2: 0.76 Model 3: 0.81	Model 1: 0.00 Model 2: 0.50 Model 3: 1.00	Demonstrated ability for document embeddings to capture meaningful information in personal statements from pediatric otolaryngology fellowship applicants. Bias acknowledged as a limitation only. Model 1: Applicant characteristics and PS cluster Model 2: raw BERT vector representation of PS Model 3: raw BERT vector and applicant characteristics
Burk-Rafel et al (2021)[21]	Random forest LightGBM XGBoost	Demographics ERAS features (USMLE result, awards, LoR, experiences [count/hours], medical school)	Interview invite	0.91	0.85	Model 1: 0.95 Model 2: 0.94 Model 3: 0.93	Model 1: 0.76 Model 2: 0.72 Model 3: 0.76			Developed MLM that assessed 61 ERAS features to predict probability of interview invite. Integrated MLM into interactive decision strategy tool for second pass review of applications previously excluded by human reviewers. Bias incorporated into discussion and limitations but not modeled. Model 1: All variables included Model 2: USMLE scores excluded Model 3: Incomplete applications excluded

[a] Recall and Sensitivity are synonymous in literature.

[b] Precision and PPV are synonymous in literature.

Abbreviations: AUROC, area under the receiver operating characteristic; PR, precision-recall; PS, personal statement; MLM/ML, machine learning model; NLP, natural language processing; LIWC, linguistic inquiry word count version 2.063; LoR, letters of recommendation; AI, artificial intelligence; SCG, scores and clerkship grades; USMLE, United States Medical Licensing Examination; CC, correlation coefficient; NRMP, National Resident Matching Program; NLTK, Natural Language Toolkit; ERAS, Electronic Residency Application Service; IMG, international medical graduate; AOA, Alpha Omega Alpha; GHHS, Gold Humanism Honor Society; NLOR, narrative letter of recommendation; SLOR, standardized letter of recommendation; TF-IDF, Term Frequency-Inverse Document Frequency; MSPE, Medical Student Performance Evaluation; ENT, ear, nose, and throat.

TABLE 3
Model Feature Importance for Articles Creating Rank List (N=9)

Author	Year	Title	Country	Institution	Clinical Area	Outcome	Features Explicitly Mentioned
St John, et al[25]	2022	Match Maker: Assessing Applicant Personal Statements With Artificial Intelligence	United States	University of Maryland Medical Center	General surgery	Interview invite Rank list	
Rees, et al[3]	2023	Machine Learning for the Prediction of Ranked Applicants and Matriculants to an Internal Medicine Residency Program	United States	Geisel School of Medicine at Dartmouth	Internal medicine	Rank list Matriculation	Ranked applicants: medical school type, medical school state, medical school country, medical degree type, USMLE Step 2 CK Ranked matriculants: medical school country, contact address state, permanent address state, medical school type, hobbies (creative writing)
Pilon, et al[23]	1997	Neural Network and Linear Regression Models in Residency Selection	United States	University of New Mexico School of Medicine	Emergency medicine	Rank list	Medical school grades, written autobiography, interviews, letters of recommendation, and part one of the National Board scores
Ortiz, et al[28]	2023	Words Matter: Using Natural Language Processing to Predict Neurosurgical Residency Match Outcomes	United States	Vanderbilt University Medical Center	Neurosurgery	Matched applicants	AOA, USMLE Step 1 score, no. of publications, PhD, no. of volunteer activities, no. of research activities, gender, Gold Humanism Honor Society member, no. of work activities, current resident status, international medical graduate
Mahtani, et al[15]	2023	A New Tool for Holistic Residency Application Review: Using Natural Language Processing of Applicant Experiences to Predict Interview Invitation	United States	New York University Grossman School of Medicine	Internal medicine	Interview invite	Phrases indicating active leadership, research, or work in social justice and health disparities were associated with interview invitation.

TABLE 3
Model Feature Importance for Articles Creating Rank List (N=9) (continued)

Author	Year	Title	Country	Institution	Clinical Area	Outcome	Features Explicitly Mentioned
Kibble, et al[27]	2023	Comparing Machine Learning Models and Human Raters When Ranking Medical Student Performance Evaluations	United States	University of Central Florida College of Medicine (UCF)	Medical students at UCF	Rank list	
Drum, et al[26]	2023	Using Natural Language Processing and Machine Learning to Identify Internal Medicine–Pediatrics Residency Values in Applications	United States	University of Utah School of Medicine	Internal medicine–pediatrics	Interview invite	Academic strength, communication, compassion, DEI, leadership, self-awareness, teamwork, work ethic
Chillakuru, et al[22]	2022	Deep Learning for Predictive Analysis of Pediatric Otolaryngology Personal Statements: A Pilot Study	United States	Children's National Health System	Pediatric otolaryngology fellowship	Applicant fit Post-fellowship achievement	Doximity residency ranking was the most significant multivariate regression variable for post-fellowship research output. Did not discuss feature importance other than personal statements.
Burk-Rafel, et al[21]	2021	Development and Validation of a Machine Learning–Based Decision Support Tool for Residency Applicant Screening and Review	United States	New York University Grossman School of Medicine	Internal medicine	Interview invite	Invite: school research rank, non-White/non-Asian race, Northeast area school, Gold Humanism Honor Society member, USMLE Step 1 Do not invite: school research rank, USMLE Step 1, USMLE Step 2 CK, internal medicine recommendation letter count, Northeast area school

Abbreviations: USMLE, United States Medical Licensing Examination; CK, clinical knowledge; AOA, Alpha Omega Alpha; DEI, diversity, equity, and inclusion; MSPE, Medical Student Performance Evaluation.

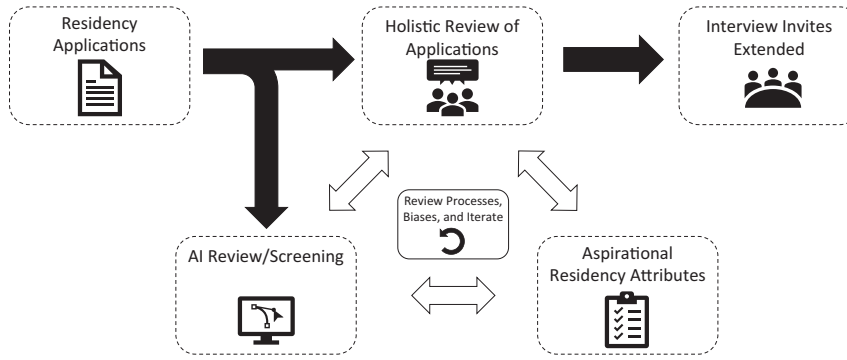

FIGURE 2
Potential Use of Artificial Intelligence in Residency Application Review

Abbreviation: AI, artificial intelligence.

Note: Once applications are received, holistic review can establish applications that the program would like to interview. Within holistic review, AI review can play a part in evaluating and monitoring biases that occur during the holistic review process compared to predetermined aspirational residency attributes. Note, in this proposed framework, AI does not create a de novo rank list without human oversight.

in residency and fellowship recruitment and the explicit assessment of model bias. It is also important to note that while AI has been widely used in other fields for screening applications[18,47,48] and has shown both benefits and risk,[49-57] the residency selection process remains unique, and conclusions from other sectors may not be fully transferable.

Conclusions

AI is being developed for residency and fellowship application evaluation, with current applications including applicant outcome prediction and textual analysis. Our review reveals a limited body of literature on these methods, with insufficient exploration of bias and fairness in AI models. Review of the literature identified gaps in standardized reporting of model performance, including metrics such as feature importance and explainability.

References

1. LaFemina J, Rosenkranz KM, Aarons CB, et al. Outcomes of the 2021-2022 APDS general surgery recruitment process recommendations. *J Surg Educ*. 2023;80(6):767-775. doi:10.1016/j.jsurg.2023.02.019
2. Association of American Medical Colleges. Holistic review. Accessed April 14, 2025. https://www.aamc.org/services/member-capacity-building/holistic-review
3. Rees CA, Ryder HF. Machine learning for the prediction of ranked applicants and matriculants to an internal medicine residency program. *Teach Learn Med*. 2023;35(3):277-286. doi:10.1080/10401334.2022.2059664
4. Berk GA, Ho TD, Stack-Pyle TJ, et al. The next step: replacing Step 1 as a metric for residency application. *Laryngoscope Investig Otolaryngol*. 2022;7(6):1756-1761. doi:10.1002/lio2.947
5. Tidwell J, Yudien M, Rutledge H, Terhune KP, LaFemina J, Aarons CB. Reshaping residency recruitment: achieving alignment between applicants and programs in surgery. *J Surg Educ*. 2022;79(3):643-654. doi:10.1016/j.jsurg.2022.01.004
6. Singh NP, Boyd CJ. Rapidly increasing number and cost of residency applications in surgery. *Am Surg*. 2023;89(12):5729-5736. doi:10.1177/00031348231173947
7. Baimas-George M, Schiffern L, Yang H, et al. Deconstructing the roadmap to surgical residency: a national survey of residents illuminates factors associated with recruitment success as well as applicants' needs and beliefs. *Global Surg Educ*. 2022;1(1):66. doi:10.1007/s44186-022-00070-9
8. 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. doi:10.1097/ACM.0b013e3181970c6b
9. Makdisi G, Takeuchi T, Rodriguez J, Rucinski J, Wise L. How we select our residents-a survey of selection criteria in general surgery residents. *J Surg Educ*. 2011;68(1):67-72. doi:10.1016/j.jsurg.2010.10.003
10. Cohn MR, Bigach SD, Bernstein DN, et al. Resident selection in the wake of United States Medical Licensing Examination Step 1 transition to pass/fail scoring. *J Am Acad Orthop Surg*. 2020;28(21):865-873. doi:10.5435/JAAOS-D-20-00359
11. Dort JM, Trickey AW, Kallies KJ, Joshi AR, Sidwell RA, Jarman BT. Applicant characteristics associated with selection for ranking at independent surgery residency programs. *J Surg Educ*. 2015;72(6):e123-e129. doi:10.1016/j.jsurg.2015.04.021
12. Garber AM, Kwan B, Williams CM, et al. Use of filters for residency application review: results from the internal medicine in-training examination program director survey. *J Grad Med Educ*. 2019;11(6):704-707. doi:10.4300/JGME-D-19-00345.1

13. National Resident Matching Program. Data Release and Research Committee: results of the 2022 NRMP program director survey. Accessed April 14, 2025. https://www.nrmp.org/wp-content/uploads/2022/09/PD-Survey-Report-2022_FINALrev.pdf
14. Li NY, Gruppuso PA, Kalagara S, Eltorai AEM, DePasse JM, Daniels AH. Critical assessment of the contemporary orthopaedic surgery residency application process. *JBJS*. 2019;101(21):e114. doi:10.2106/jbjs.18.00587
15. Mahtani AU, Reinstein I, Marin M, Burk-Rafel J. A new tool for holistic residency application review: using natural language processing of applicant experiences to predict interview invitation. *Acad Med*. 2023;98(9):1018-1021. doi:10.1097/ACM.0000000000005210
16. Ferryman K, Mackintosh M, Ghassemi M. Considering biased data as informative artifacts in AI-assisted health care. *N Engl J Med*. 2023;389(9):833-838. doi:10.1056/NEJMra2214964
17. The New York City Council. Automated Employment Decision Tools (AEDT), 2021. Accessed April 14, 2025. https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=ID%7CText%7C&Search=
18. Mearian L. NYC law governing AI-based hiring tools goes live. *Computerworld*. Published July 6, 2023. Accessed April 14, 2025. https://www.computerworld.com/article/1630802/nyc-law-governing-ai-based-hiring-tools-goes-live.html
19. The White House. Blueprint for an AI bill of rights (2022). Accessed April 14, 2025. https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/
20. National Conference of State Legislatures. Artificial Intelligence 2024 Legislation. Updated September 9, 2024. Accessed April 14, 2025. https://www.ncsl.org/technology-and-communication/artificial-intelligence-2024-legislation
21. Burk-Rafel J, Reinstein I, Feng J, et al. Development and validation of a machine learning-based decision support tool for residency applicant screening and review. *Acad Med*. 2021;96(suppl 11):54-61. doi:10.1097/ACM.0000000000004317
22. Chillakuru YR, Preciado DA, Cha J, Mann H, Behzadpour HK, Espinel AG. Deep learning for predictive analysis of pediatric otolaryngology personal statements: a pilot study. *Otolaryngol Head Neck Surg*. 2022;167(5):877-884. doi:10.1177/01945998221082535
23. Pilon S, Tandberg D. Neural network and linear regression models in residency selection. *Am J Emerg Med*. 1997;15(4):361-364. doi:10.1016/s0735-6757(97)90125-x
24. Stedman JM, Hatch JP, Schoenfeld LS. Letters of recommendation for the predoctoral internship in medical schools and other settings: do they enhance decision making in the selection process? *J Clin Psychol Med Settings*. 2009;16(4):339-345. doi:10.1007/s10880-009-9170-y
25. St John AJ, Abdou H, Kavic SM. Match maker: assessing applicant personal statements with artificial intelligence. *J Am Coll Surgeons*. 2022;235(suppl 5):224. doi:10.1097/01.XCS.0000894760.78127.f5
26. Drum B, Shi J, Peterson B, Lamb S, Hurdle JF, Gradick C. Using natural language processing and machine learning to identify internal medicine-pediatrics residency values in applications. *Acad Med*. 2023;98(11):1278-1282. doi:10.1097/ACM.0000000000005352
27. Kibble J, Plochocki J. Comparing machine learning models and human raters when ranking medical student performance evaluations. *J Grad Med Educ*. 2023;15(4):488-493. doi:10.4300/JGME-D-22-00678.1
28. Ortiz AV, Feldman MJ, Yengo-Kahn AM, et al. Words matter: using natural language processing to predict neurosurgical residency Match outcomes. *J Neurosurg*. 2023;138(2):559-566. doi:10.3171/2022.5.JNS22558
29. Gray GM, Williams SA, Bludevich B, et al. Examining implicit bias differences in pediatric surgical fellowship letters of recommendation using natural language processing. *J Surg Educ*. 2023;80(4):547-555. doi:10.1016/j.jsurg.2022.12.002
30. Sarraf D, Vasiliu V, Imberman B, Lindeman B. Use of artificial intelligence for gender bias analysis in letters of recommendation for general surgery residency candidates. *Am J Surg*. 2021;222(6):1051-1059. doi:10.1016/j.amjsurg.2021.09.034
31. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med*. 2018;169(7):467-473. doi:10.7326/m18-0850
32. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. *J Clin Epidemiol*. 2016;75:40-46. doi:10.1016/j.jclinepi.2016.01.021
33. Hashimoto DA, Varas J, Schwartz TA. Practical guide to machine learning and artificial intelligence in surgical education research. *JAMA Surg*. 2024;159(4):455-456. doi:10.1001/jamasurg.2023.6687
34. Sounderajah V, Ashrafian H, Golub RM, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open*. 2021;11(6):e047709. doi:10.1136/bmjopen-2020-047709
35. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):e048008. doi:10.1136/bmjopen-2020-048008

36. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020; 26(9):1351-1363. doi:10.1038/s41591-020-1037-7
37. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x
38. Foulds JR, Islam R, Keya KN, Pan S. An intersectional definition of fairness. *arXiv*. Published September 10, 2019. Accessed April 14, 2025. doi:10.48550/arXiv.1807.08362 https://arxiv.org/abs/1807.08362
39. Angus SV, Williams CM, Stewart EA, Sweet M, Kisielewski M, Willett LL. Internal medicine residency program directors' screening practices and perceptions about recruitment challenges. *Acad Med*. 2020;95(4): 582-589. doi:10.1097/acm.0000000000003086
40. Whitla DK, Orfield G, Silen W, Teperow C, Howard C, Reede J. Educational benefits of diversity in medical school: a survey of students. *Acad Med*. 2003;78(5): 460-466. doi:10.1097/00001888-200305000-00007
41. Abril D. Job applicants are battling AI résumé filters with a hack. *Washington Post*. Published July 24. Accessed December 1, 2024. https://www.washingtonpost.com/technology/2023/07/24/white-font-resume-tip-keywords/
42. Zhang JS, Yoon C, Williams DKA, Pinkas A. Exploring the usage of ChatGPT among medical students in the United States. *J Med Educ Curric Dev*. 2024;11: 23821205241264695. doi:10.1177/23821205241264695
43. Zumsteg JM, Junn C. Will ChatGPT match to your program? *Am J Phys Med Rehabil*. 2023;102(6): 545-547. doi:10.1097/phm.0000000000002238
44. Speicher T, Heidari H, Grgic-Hlaca N, et al. A unified approach to quantifying algorithmic unfairness: measuring individual & group unfairness via inequality indices. Presented at: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018; London, United Kingdom.
45. Saleiro P, Kuester B, Hinkson L, et al. Aequitas: a bias and fairness audit toolkit. *arXiv*. Published April 29, 2019. Accessed April 14, 2025. doi:10.48550/arXiv.1811.05577 https://arxiv.org/abs/1811.05577
46. Bellamy RKE, Dey K, Hind M, et al. AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. *Ibm J Res Dev*. 2019;99. doi:10.1147/JRD.2019.2942287
47. Abou Hamdan L. *The Role of Artificial Intelligence in the Recruitment and Selection Processes: A Systematic Review*. American University of Beirut; 2019.
48. Hunkenschroer AL, Luetge C. Ethics of AI-enabled recruiting and selection: a review and research agenda. *J Business Ethic*. 2022;178(4):977-1007. doi:10.1007/s10551-022-05049-6
49. Persson A. Implicit bias in predictive data profiling within recruitments. In: Lehmann A, Whitehouse D, Fischer-Hübner S, Fritsch L, Raab C, eds. *Privacy and Identity Management: Facing Up to Next Steps*. Springer International Publishing; 2016:212-230.
50. Polli F. Using AI to eliminate bias from hiring. *Harvard Business Review*. Published October 29, 2019. Accessed April 14, 2025. https://hbr.org/2019/10/using-ai-to-eliminate-bias-from-hiring
51. Hofeditz L, Clausen S, Riess A, Mirbabaie M, Stieglitz S. Applying XAI to an AI-based system for candidate management to mitigate bias and discrimination in hiring. *Electron Mark*. 2022;32(4):2207-2233. doi:10.1007/s12525-022-00600-9
52. Chen ZS. Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Hum Soc Sci Commun*. 2023;10(1). doi:10.1057/s41599-023-02079-x
53. Malik A. AI bias in recruitment: ethical implications and transparency. *Forbes*. Published September 25, 2023. Accessed April 14, 2025. https://www.forbes.com/councils/forbestechcouncil/2023/09/25/ai-bias-in-recruitment-ethical-implications-and-transparency/
54. Varsha PS. How can we manage biases in artificial intelligence systems—a systematic literature review. *Int J Inform Manag Data Insights*. 2023;3(1):100165. doi:10.1016/j.jjimei.2023.100165
55. Bogen M. All the ways hiring algorithms can introduce bias. *Harvard Business Review*. Published May 6, 2019. Accessed April 14, 2025. https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias
56. Cappelli P. Data science can't fix hiring (yet). *Harvard Business Review*. Published May-June 2019. Accessed April 14, 2025. https://hbr.org/2019/05/data-science-cant-fix-hiring-yet
57. Tambe P, Cappelli P, Yakubovich V. Artificial intelligence in human resources management: challenges and a path forward. *California Manag Rev*. 2019;61(4): 15-42. doi:10.1177/0008125619867910

Maxwell D. Sumner, BS, is an MD/MBA Candidate, Duke University School of Medicine, Durham, North Carolina, USA; **T. Clark Howell, MD, MSHS,** is a PGY-5 General Surgery Resident, Department of Surgery, Duke University School of Medicine, Durham, North Carolina, USA; **Alexandria L. Soto, BS,** is an MD Candidate, Duke University School of Medicine, Durham, North Carolina, USA; **Samantha Kaplan, PhD,** is a Research and Education Liaison Librarian, Duke Medical Center Library, Duke University School of Medicine, Durham, North Carolina, USA; **Elisabeth T. Tracy, MD,** is an Assistant Professor of Surgery and Outgoing Program Director, General Surgery Residency Program, Department of Surgery, Duke University School of Medicine, Durham, North Carolina, USA; **Aimee K. Zaas, MD,** is a Professor of Medicine and Program Director, Internal Medicine Residency Program, Department of Internal Medicine, Duke University School of Medicine, Durham, North Carolina, USA; **John Migaly, MD,** is an Associate Professor of Surgery and Vice Chair of Education,

Department of Surgery, Duke University School of Medicine, Durham, North Carolina, USA; **Allan D. Kirk, MD, PhD,** is the David C. Sabiston, Jr. Distinguished Professor of Surgery and Chair, Department of Surgery, Duke University School of Medicine, Durham, North Carolina, USA; and **Kevin Shah, MD,** is an Assistant Professor of Surgery and Program Director, Department of Surgery, Duke University School of Medicine, Durham, North Carolina, USA.

Funding: The authors report no external funding source for this study.

Conflict of interest: The authors declare they have no competing interests.

This work was previously presented at the Duke University AHEAD Education Conference, March 22, 2024, Durham, North Carolina, USA.

The authors would like to thank the members of the Laboratory of Transformative Administration for their support and critical feedback. Additionally, the authors wish to thank the Departments of Surgery and Internal Medicine for their insight and feedback.

Corresponding author: Maxwell D. Sumner, BS, Duke University School of Medicine, Durham, North Carolina, USA, max.sumner@duke.edu

Received July 26, 2024; revision received September 24, 2024, and March 9, 2025; accepted March 25, 2025.