Development and Validation of a Machine Learning-Based Decision Support Tool for Residency Applicant Screening and Review

Jesse Burk-Rafel, MD, MRes, Ilan Reinstein, MS, James Feng, MD, MS, Moosun Brad Kim, MS, Louis H. Miller, MD, Patrick M. Cocks, MD, Marina Marin, MSc, and Yindalon Aphinyanaphongs, MD, PhD

Abstract

Purpose
Residency programs face overwhelming numbers of residency applications, limiting holistic review. Artificial intelligence techniques have been proposed to address this challenge but have not been created. Here, a multidisciplinary team sought to develop and validate a machine learning (ML)-based decision support tool (DST) for residency applicant screening and review.

Method
Categorical applicant data from the 2018, 2019, and 2020 residency application cycles (n = 8,243 applicants) at one large internal medicine residency program were downloaded from the Electronic Residency Application Service and linked to the outcome measure: interview invitation by human reviewers (n = 1,235 invites). An ML model using gradient boosting was designed using training data (80% of applicants) with over 60 applicant features (e.g., demographics, experiences, academic metrics). Model performance was validated on held-out data (20% of applicants). Sensitivity analysis was conducted without United States Medical Licensing Examination (USMLE) scores. An interactive DST incorporating the ML model was designed and deployed that provided applicant- and cohort-level visualizations.

Results
The ML model areas under the receiver operating characteristic and precision recall curves were 0.95 and 0.76, respectively; these changed to 0.94 and 0.72, respectively, with removal of USMLE scores. Applicants’ medical school information was an important driver of predictions—which had face validity based on the local selection process—but numerous predictors contributed. Program directors used the DST in the 2021 application cycle to select 20 applicants for interview that had been initially screened out during human review.

Conclusions
The authors developed and validated an ML algorithm for predicting residency interview offers from numerous application elements with high performance—even when USMLE scores were removed. Model deployment in a DST highlighted its potential for screening candidates and helped quantify and mitigate biases existing in the selection process. Further work will incorporate unstructured textual data through natural language processing methods.

Holistic review uses multiple applicant data points—including experiences, attributes, and academic metrics—to gain a “whole” person perspective, with the aim of identifying applicants who align with an institution’s mission. In contrast to selection that relies on just a few academic performance metrics, holistic review provides a more balanced consideration that promotes diversity.

Although holistic review is the gold standard for selection of medical residents, U.S. residency programs face a mismatch between application numbers and program review resources that has led many programs to shift away from holistic review. For example, the average internal medicine program received 2,220 applications for 220 interview invitations and 14 positions (or about 160 applications per position). In place of holistic review, many programs use a variety of simplistic filters—such as United States Medical Licensing Examination (USMLE) scores—to obtain a more manageable pool of applicants that can then undergo more thorough review to determine who to interview.

Use of licensure exam scores as single measures for residency applicant screening has been widely criticized, contributing to the recent change of USMLE Step 1 to pass/fail and recommendations by the cross-organizational Coalition for Physician Accountability to redesign the overall process. Yet, absent immediate changes that dramatically reduce application numbers or increase the resources available to programs for screening, programs will likely continue to be burdened by large numbers of applications per position that threaten their ability to conduct holistic review on all applications.

Conceptually, an approach that could reduce the resources needed to assimilate many applicant data points and provide guidance on which applicants should be offered interviews would be beneficial. Machine learning (ML) is a subset of artificial intelligence (AI) whereby a computer “learns” patterns in large datasets. In “supervised” ML, a set of input variables (called “features” in ML) and an output variable (called a “label”) are provided to computer algorithms that employ regression.
ML and other AI techniques are increasingly being used in health care to analyze “big data.” Recently, there have been calls for AI training in health professions education and big data analytics to improve graduate medical education. With regard to selection, ML has been applied in undergraduate, graduate school, pharmacy school, and medical school admission processes. AI has been proposed as a potential tool in resident selection, but implementation of such a tool has not yet been described.

Here, we sought to address this gap by developing and validating an ML-based decision support tool (DST) for residency program directors to use for interview invitation decisions. We hypothesized that by incorporating as much data as possible from the Electronic Residency Application Service (ERAS) into the ML-based DST, we might partially automate our human review process where 13 program directors use multiple data points and principles of holistic review—but no standard rubric—to select applicants for interview.

**Method**

Raw data files for all applicants were exported in comma-separated values format from the ERAS Program Director Workstation for the NYU Langone Health internal medicine residency program (categorical track) for the 2017–2018, 2018–2019, and 2019–2020 application cycles (hereafter simply 2018, 2019, and 2020, respectively). The interview invite decision for each applicant as determined by the program leadership was also tabulated (Figure 1A). Medical students graduating from the NYU Grossman School of Medicine were excluded, as they underwent a separate review process.

In total, n = 8,243 applicants were included by combining the 2018, 2019, and 2020 cycles, of which n = 1,235 applicants (15.0%) received interview invitations, the outcome of interest.

**Figure 1** Approach for developing an artificial intelligence-based decision support tool for residency applicant screening. A. Raw applicant data are downloaded from the Electronic Residency Application Service (ERAS) Program Director Workstation in comma separated value format. The actual interview invite decision for each applicant by the program leadership is also tabulated. B. ERAS data are cleaned and reduced to 61 variables (or “features”). Where applicable, data are normalized, encoded numerically, and missing data imputed. C. The data are split into a training dataset (80% of applicants) and a validation dataset (20% of applicants). Four algorithms are evaluated on the training data using 10-fold cross-validation, monitoring evaluation metrics like area under the receiver operating characteristic and precision recall curves (AUROC, AUPRC). D. The performance of the optimized model is evaluated on the held-out validation data, including AUROC and AUPRC. E. ERAS data, machine learning algorithm predictions, and human invites (when available) are uploaded into an interactive decision support tool which can augment program directors’ selection decisions. Abbreviations: LR, Logistic Regression; RF, Random Forest; LightGBM, Light Gradient Boosting Machine; XGBoost, eXtreme Gradient Boosting.
dataset was deidentified and randomly split into 2 nonoverlapping sets of data: a training set (80% of applicants, n = 6,586, what the computer iteratively learns from) and a validation set (20% of applicants, n = 1,657, for testing the final ML model). The NYU Grossman School of Medicine IRB approved the study as exempt.

Data cleaning and feature extraction
The raw ERAS file includes over 640 data fields, which were reduced by removing largely blank fields, reducing dimensionality (e.g., converting multiple publication fields to a single count), and using domain knowledge to exclude duplicative or noncontributory fields (Figure 1B).

Ultimately 61 variables (or “features”) were identified. The selected features included demographics (age, sex, self-reported race), applicant type (couples match, visa status, track(s) applied to of several offered by our program), degree type, and the number of letters of recommendation (and were coded as having zero probability of interview invitation. Typically, such applicants undergo manual human screening. We kept these applicants in the analysis pipeline to enable a future use case where the DST could replace this manual screening step.

For the remainder of applicants, an ML model was built to predict the probability of interview invitation (Figure 1C). Four algorithms (or “classifiers” in ML) were tested: logistic regression and decision-tree-based methods (random forest, \( \text{LightGBM}, \text{XGBoost} \)). Tree-based algorithms use a series of decision rules to split a dataset into smaller, nonoverlapping sets where the response variable is predominantly one “class.” Modern ML-tree-based algorithms build groups of trees (or “ensembles”) and can employ “gradient boosting” for powerful fitting. Unlike logistic regression, tree-based classifiers are nonparametric, handling any data structure and incorporating abrupt variance in predictors without user input. For our binary classification task (i.e., prediction of the categorical outcome of interview invite vs not invite), the loss function that requires optimization corresponds to a logistic function and several evaluation metrics such as precision, recall, accuracy, and area under the receiver operating characteristic (AUROC) curve were used to monitor the performance during the model selection procedure.

The optimal model was selected using 10-fold cross-validation, where we monitored each of the evaluation metrics (i.e., accuracy, AUROC) by repeatedly randomly splitting the training data further into 2 sets (in a 4:1 ratio) and testing different combinations of hyperparameters (i.e., parameters of the algorithm) with each split. The algorithm and combination of hyperparameters that performed best was an XGBoost-based model (for \text{eXtreme Gradient Boosting}).

Model performance, sensitivity analysis, and feature importance
For each applicant, the model produces a probability score for interview invite. If this probability is above a dimensionless threshold value—which can be changed to adjust the performance characteristics of the model—the candidate is labeled as predicted to receive an interview. Using the validation dataset, the predicted interview invites can be compared with actual interview invites. ROC curves were generated and AUROC calculated (Figure 1D). Additionally, given the tendency of ROC curves to mask poor performance in imbalanced datasets (i.e., many more applicants uninvited than invited), precision recall curves (PRCs) were also generated and the area under the PRC (AUPRC) calculated as a measure of average precision.

Sensitivity analysis was performed by rebuilding the models without USMLE Step 1 and 2 CK scores (but retaining pass/fail status) or without the 1,774 applicants who did not meet program requirements, and recalculating AUROC and AUPRC.

To understand the relative importance of each of the input features in the tree-based XGBoost model, SHapley Additive exPlanation (SHAP) values were calculated and displayed at the individual and cohort level (i.e., invited vs not invited applicants). SHAP values are analogous to coefficients in a logistic regression model, but vary for each individual prediction—features important for one applicant may be noncontributory for other applicants in nonparametric tree-based models.
Interactive decision support tool

A DST was developed by first loading raw ERAS application files, algorithm-based invite predictions, and actual program director invites (once available) into an Education Data Warehouse (EDUDW). Tableau analytics software (Seattle, Washington) was then used to connect directly to the EDUDW to visualize applicant data, predicted invites, and actual invites. The tool allows users interactive cohort-level exploration of the data by various applicant factors, modifying the invite cutoff threshold to adjust the performance of the predictions (i.e., tuning for high sensitivity or specificity), and drill-down features to the level of applicant with direct hyperlinks to both the original ERAS application and the algorithm explanation of feature importance. The DST was published on a centralized secure server and could be accessed by program leadership via a browser.

Model integration into applicant review

A growing concern with using machine-based models is their potential for unseen biases. To mitigate this concern, a 2-stage applicant review process was implemented in the 2020–2021 application cycle. In Stage 1, human reviewers evaluated applicants as they normally would. After all reviews were complete, in Stage 2, reviewers were then allowed to review the DST. In this pilot phase, applicants were only screened in, with the program reserving about 20 interview spots to fill with augmentations from the ML recommendations.

Results

The model ROC curves and PRCs are shown in Figure 2. For the validation data, the AUROC was 0.95 and the AUPRC was 0.76. Sensitivity analysis removing all USMLE scores resulted in a revised AUROC of 0.94 and AUPRC of 0.72 (Figure 2); excluding applicants who did not meet program requirements resulted in a revised AUROC of 0.93 and AUPRC of 0.76 (data not shown). Of note, a no-skill model would produce an AUROC of 0.5 and AUPRC of 0.15.

The model threshold can be changed to optimize for sensitivity (screen out task) or specificity (screen in task). For example, for a pre-screening (screen out) task, the sensitivity can be tuned to 91% with a specificity of 85%; for an applicant pool of 2,000 applicants and an invite rate of 15%, this results in 1,475 applicants screened out with a negative predictive value of 98%, with 525 applicants remaining for further review for 300 interview positions.

Feature importance analysis using SHAP values indicates that the model was substantially reliant on applicants’ medical school information such as location and research ranking, but that numerous predictors contributed to interview invite probability. As the importance of each feature changes for
every individual prediction in tree-based models, visualizations were created to show the feature importance at both the individual (Figure 3) and cohort levels (not shown).

An interactive DST was designed that incorporated the ML interview invite predictions and filters for key applicant information (Supplemental Digital Appendix 2, top, at http://links.lww.com/ACADMED/B155). In addition to displaying invite suggestions before human review, the interactive tool can also incorporate the initial round of human review invite decisions before applicants are notified. In this way, for the 2020–2021 cycle, the program directors examined gaps between the demographics of those individuals initially invited through human review and statistically expected demographics. The DST was then used to identify individuals who had not been invited in Stage 1 human review but were rated highly by the model in Stage 2 (Supplemental Digital Appendix 2, bottom, at http://links.lww.com/ACADMED/B155), and the relative importance of each feature for each of these individuals was explored (Figure 3). Overall, 20 applicants were selected for interview that had not been initially selected through human review.

Discussion

In this proof-of-concept work, an ML model incorporating over 60 elements of the ERAS residency application was designed that predicted interview invites with high performance. To our knowledge, this is the first report of the development, validation, and deployment of an ML model for medical resident selection.

In contrast to the frameworks of Messick and Kane for validation in the absence of a reference standard, supervised ML relies on a gold standard label (e.g., the invite decision made by program directors) that is assumed to be the “ground truth.” Validation in the context of a clear reference standard is predominantly concerned with criterion validity (i.e., how well the model performs in predicting the ground truth). We generated several types of criterion validity evidence. First, we retrospectively validated the model performance on data not seen before by the ML model. Although not perfectly matching the human review process, the most important features of the ML model aligned with our program’s stated screening factors, indicating face validity.

Second, we prospectively validated the ML model by deploying an interactive DST in an ongoing resident selection cycle. Prospective validation is a critical step in determining if an ML model has real-world validity; in our case, the model showed utility in identifying applicants who might otherwise have been screened out. The 20 applicants who were screened in using the DST varied in their characteristics. In most cases, program leadership felt the human reviewer had either overlooked the applicant in the setting of the large application burden or overweighted a single factor (like a low USMLE score). Identifying such “diamonds in the rough” for further holistic review is a promising aspect of our approach and highlights how such ML models can be adjusted for different uses (e.g., widescale screening vs more directed selection decisions).
Our program uses applicant experiences, attributes, and academic metrics—consistent with holistic review—in selecting applicants for interview but does not employ a strict rubric. As the ML model learns based on prior program director decisions, creation of the model enabled quantification of the weights of different applicant factors in our selection process, at both the applicant and cohort levels. The high fidelity between the ML model and the human invites suggests the 13 program directors employ a shared selection framework.

Our methodology iteratively tested multiple types of algorithms; a decision tree-based algorithm called eXtreme Gradient Boosting (XGBoost) outperformed the more familiar logistic regression algorithm. XGBoost is one of the most powerful ML algorithms and consistently wins data competitions. Decision trees, and consequently tree-based estimators like XGBoost, make no assumptions about the relationships between features, innately incorporating interaction effects between features. Moreover, XGBoost includes other powerful features such as boosting (i.e., where a set of weak models are combined into a stronger ensemble), directly modeling bias-variance trade off during model selection (i.e., balancing over- and underfitting), and multiple regularization methods (i.e., penalizing complexity). Although tree-based methods are susceptible to overfitting, our study design employed an iterative cross-validation approach within the training set to mitigate the impact of overfitting and only validated performance on held out data (i.e., not seen before by the model).

Despite our program's use of USMLE Step 1 and 2 CK scores as an element of selection—which was substantiated in the feature importance evaluation showing these scores to be important features—the ML model performance did not substantially decline when USMLE scores were removed as predictors. This may be related to the collinearity of other predictors (e.g., demographics, medical school) with USMLE scores. This finding suggests USMLE score information may be captured by other elements of the application, which has important implications for programs as Step 1 goes pass/fail in January 2022.

While we generated robust internal validity evidence, the model has not yet been externally validated, which would add further generalizability evidence to our approach.

ML models, by definition, must learn from training data. Thus, our model would need to be "retrained" when applied to other residency programs' selection processes, which presumably weigh selection factors differently. Nonetheless, such retraining is feasible and the overall ML pipeline we developed (Figure 1)—which led to strong retrospective validity evidence in the held-out validation dataset—is generalizable to other programs using ERAS data.

Conceivably, a centralized approach could be implemented. Using programs' National Resident Matching Program rank order lists (as a proxy for their interview list) from prior years, ERAS could develop and validate ML algorithms that predict each applicant's probability of receiving an interview invite at a given residency program. Such a centralized prediction tool could augment decisions made by both applicants (for selecting where to apply) and programs (for selecting which applicants to interview). However, a potential concern is the "adversarial" nature of residency selection. Revealing too much of the "secret sauce" may cause applicants to game the system and models must be built in reaction. Adversarial applications are well studied in ML with spam detection being a quintessential example.

Another challenge to wider use of ML for resident screening is that ML models benefit from large datasets. Our program's large applicant pool—despite only a 15% interview rate—permits robust analysis even with over 60 features (equating to 135 observations per predictor). Smaller programs may need to use more years of data, fewer variables, or incorporate data augmentation methods such as up-sampling, down-sampling, or creating "synthetic" applicants. The models could adapt to changes to a program's selection process through retraining on new selection data.

One critique of AI techniques like ML is that they propagate and automate bias. Indeed, as previously mentioned, ML models will reflect the same biases in the data and decisions that they are trained on. This challenge is particularly relevant to high-stakes admissions decisions, which are known to have racial, gender, and other inequities. Yet, conversely, ML models can make underlying biases in selection processes more explicit. Our ML model allowed our program directors to quantify their underlying decision processes and, through use of the DST, to target desirable candidates that may have been overlooked owing to human or metrics biases. An emerging area of research is building detection and mitigation of latent biases into ML models, which we hope to pursue in future iterations.

A core limitation of our model is that it does not yet include unstructured application data (much in PDF format), such as the Medical Student Performance Evaluation (e.g., overall adjective, grades, rotation comments), letters of recommendation (e.g., strength of recommendation, concerns stated), personal statement (e.g., writing clarity), volunteer experiences (e.g., evidence of service or leadership), and research experiences (e.g., publication impact, author position). Ongoing work is using natural language processing approaches—another domain of AI—to conduct this analysis and incorporate it into the prediction algorithm.

Conclusions

We developed and validated a high-performing ML algorithm that incorporated over 60 elements of the ERAS application to predict residency applicants' probability of receiving an interview invitation at one large internal medicine program. The algorithm was integrated into a DST, which program leaders deployed in a staged approach to screen in applicants for interview that may have been overlooked on initial human review. The approach showed promise in augmenting existing human review. The model performed well when USMLE scores were removed, highlighting that such "holistic" screening approaches may obviate the need for USMLE scores in screening—an important outcome, as Step 1 will go pass/fail in January 2022. Further model refinement is needed, including incorporation of unstructured textual information.
data through natural language processing methods.

Acknowledgments: The authors wish to thank Jonathan Chun, MD, for early input on this work and Akshay Ravi, MS, for help with dashboard development.

Funding/Support: None reported.

Other disclosures: None reported.

Ethical approval: This study was approved as exempt by the institutional review board of NYU Grossman School of Medicine.

Previous presentations: Preliminary work was presented as a virtual research abstract at the Association of American Medical Colleges Learn Serve Lead 2020 conference and as a virtual poster at Alliance for Academic Internal Medicine Academic Internal Medicine Week 2021.

J. Burk-Rafel is assistant professor of medicine and assistant director of UME-GME innovation, Institute for Innovations in Medical Education, NYU Grossman School of Medicine, New York, New York. At the time this work was completed, he was an internal medicine resident at NYU Langone Health, New York, New York; ORCID: https://orcid.org/0000-0003-3785-2154.

I. Reinstein is a research scientist, Institute for Innovations in Medical Education, NYU Grossman School of Medicine, New York, New York.

J. Feng is an orthopedic surgery resident, Beaumont Health, Royal Oak, Michigan. At the time this work was completed, he was a student's master in biomedical informatics, NYU Grossman School of Medicine Vileck Institute of Graduate Biomedical Sciences, New York, New York.

M.B. Kim is a biostatistician at Aprogen, Seongnam, Republic of Korea. At the time this work was completed, he was an internal medicine resident at NYU Langone Health, New York, New York.

L.H. Miller is assistant professor of cardiology and assistant dean for career advisement, Zucker School of Medicine at Hofstra/Northwell, New York, New York.

P.M. Cocks is the Abraham Sunshine Assistant Professor of Medicine, and program director of the internal medicine residency program, NYU Langone Health, New York, New York.

M. Marin is director of the division of educational analytics, Institute for Innovations in Medical Education, NYU Grossman School of Medicine, New York, New York.

Y. Aphinyanaphongs is director of operational data science and machine learning, NYU Langone Health, New York, New York.

References


8. Triola MM. Director, Institute for Innovations in Medical Education, NYU Grossman School of Medicine. Personal communication with J. Burk-Rafel, April 26, 2021.


