Correlation of Ophthalmology Residency Application Characteristics with Subsequent Performance in Residency

Brett M. Gudgel, MD1 Andrew T. Melson, MD1 Justin Dvorak, PhD1 Kai Ding, PhD1 R. Michael Siatkowski, MD1

1 University of Oklahoma Health Science Center, Oklahoma City, Oklahoma

Address for correspondence Andrew T. Melson, MD, Dean McGee Eye Institute, 608 Stanton L. Young Boulevard, Oklahoma City, OK 73104 (e-mail: andrew-melson@dmei.org).

Abstract

Purpose Only from reviewing applications, it is difficult to identify which applicants will be successful ophthalmology residents. The change of the USMLE Step 1 scoring to “Pass/Fail” removes another quantitative metric. We aimed to identify application attributes correlated with successful residency performance. This study also used artificial intelligence (AI) to evaluate letters of recommendation (LOR), the Dean’s letter (MSPE), and personal statement (PS).

Design Retrospective analysis of application characteristics versus residency performance was conducted.

Participants Residents who graduated from the Dean McGee Eye Institute/University of Oklahoma Ophthalmology residency from 2004 to 2019 were included in this study.

Methods Thirty-four attributes were recorded from each application. Residents were subjectively ranked into tertiles and top and bottom deciles based on residency performance by faculty present during their training. The Ophthalmic Knowledge Assessment Program (OKAP) examination scores were used as an objective performance metric. Analysis was performed to identify associations between application attributes and tertile/decile ranking. Additional analysis used AI and natural language processing to evaluate applicant LORs, MSPE, and PS.

Main Outcome Measures Characteristics from residency applications that correlate with resident performance were the primary outcome of this study.

Results Fifty-five residents and 21 faculty members were included. A grade of “A” or “Honors” in the obstetrics/gynecology (OB/GYN) clerkship and the presence of a home ophthalmology department were associated with ranking in the top tertile but not the top decile. Mean core clerkship grades, medical school ranking in the top 25 U.S. News and World Report (USNWR) primary care rankings, and postgraduate year (PGY)-2 and PGY-3 OKAP scores were predictive of being ranked in both the top tertile and the top decile. USMLE scores, alpha-omega-alpha (AOA) status, and number of publications did

DOI https://doi.org/10.1055/s-0041-1733932. ISSN 2475-4757. © 2021. The Author(s). This is an open access article published by Thieme under the terms of the Creative Commons Attribution-NonDerivative-NonCommercial-License, permitting copying and reproduction so long as the original work is given appropriate credit. Contents may not be used for commercial purposes, or adapted, remixed, transformed or built upon. (https://creativecommons.org/licenses/by-nc-nd/4.0/) Thieme Medical Publishers, Inc., 333 Seventh Avenue, 18th Floor, New York, NY 10001, USA
Correlation of Residency Application and Residency Performance

Gudgel et al.

Ophthalmology is one of the most competitive residencies within medicine. As such, the number and competitiveness of applicants continue to grow. In 2019, for the 790 medical students who initially registered for the ophthalmology match, 649 applicants submitted an application. Of these applicants, 25% did not match into an ophthalmology residency program. Residency programs are tasked with sorting through hundreds of applications to fill a mean of four residency slots per year. In an effort to better screen the large number of applicants, studies have attempted to identify which selection factors are most valuable in screening for top-tier candidates. Factors, such as interview performance, clinical course grades, recommendation letters, and USMLE scores have all been determined to be among the most important factors for applicant selection. Though these factors have been identified as important among interview selection committees, no studies have been done within ophthalmology to show if these attributes have any predictive value in applicant performance throughout residency. This uncertainty is highlighted by the fact that fewer than 50% of respondents felt that there was a strong correlation between the highest ranked applicants and the best performers during residency. To further complicate the situation, the United States Medical Licensing Examination (USMLE) program has elected to change its score reporting for the step-1 examination from a three-digit numeric score to a pass/fail outcome, eliminating one of the most commonly used application screening metrics. A clear need for long-term studies correlating applicant attributes with successful residency performance has been identified.

Multiple studies have been conducted within other specialties to attempt to identify the attributes that best correlate with resident performance. Many of these studies fail to agree on common predictive resident attributes, and often contradict each other. There is significant variability in predictive factors from specialty to specialty which further highlights the need for the specialty of ophthalmology to attempt to identify ophthalmology-specific predictors of success. Better determining such predictors could help programs identify those applicants who are most likely to become excellent residents. Additional difficulty exists with the interpretation and utilization of letters of recommendation (LOR) and the Dean’s letter (MSPE) in the resident applicant evaluation process. Program directors and resident selection committees have indicated that LOR are among the most important factors in resident selection. Despite this, there remains inconsistency regarding interreviewer reliability, and there is some doubt that LOR are actually useful in predicting future applicant success as a resident. The same holds true with respect to the personal statement (PS), with no identifiable studies correlating its contents to success in residency.

The goals of this study are to (1) identify resident attributes that are predictive of success in an ophthalmology residency, (2) validate or invalidate the selection metrics that are currently most highly valued, and (3) attempt to utilize artificial intelligence (AI) and machine learning to develop a more meaningful and consistent method of LOR, MSPE, and PS interpretation and utilization. In doing so, we hope to improve the resident selection process within ophthalmology by giving programs additional evidence to guide application evaluation.

Materials and Methods

Application Characteristics Collected

This study was granted exempt status by the Institutional Review Board of the University of Oklahoma Health Sciences Center. The residency applications of residents graduating from the Dean McGee Eye Institute/University of Oklahoma Ophthalmology Department from 2004 through 2019 (55 residents) were obtained through the San Francisco Match database and Dean McGee Eye Institute records. The applications were reviewed, and the applicant characteristics that were recorded included: gender, number of preresidency publications, USMLE step-1 score, USMLE step-2 score (if listed), alpha-omega-alpha (AOA) status, number of “away” or “audition” rotations, advanced degrees (e.g., PhD, MS, and others), preclinical medical school grade point average (GPA), core clerkship grades, ophthalmology rotation grade, military service, previous nonmedical employment, participation in a varsity sport in college, significant experience in art or music, undergraduate GPA, ranking of medical school (based on 2019 U.S. News and World Report [USNWR] rankings), ranking of undergraduate school (2019 USNWR ranking), majoring in a science; presence of a home ophthalmology department, ranking of ophthalmology department (based on 2019 Doximity rankings), attendance of a medical school in a nearby state, and attendance of an undergraduate school in a surrounding state. The clerkships of medicine, surgery, pediatrics, obstetrics/gynecology (OB/GYN), and psychiatry were considered to be “core.” The highest possible

not correlate with subjective resident performance. AI analysis of LORs, MSPE, and PS did not identify any text features that correlated with resident performance.

Conclusions

Many metrics traditionally felt to be predictive of residency success (USMLE scores, AOA status, and research) did not predict resident success in our study. We did confirm the importance of core clerkship grades and medical school ranking. Objective measures of success such as PGY-2 and PGY-3 OKAP scores were associated with high subjective ranking.
grade (A, Honors, etc.) was considered to be Honors for the rotation. Both the USNWR primary care and research rankings were used for medical school ranking. Significant experience in art or music was defined as consistent involvement for at least a decade, a history of multiple public performances/exhibitions/competitions, or at least one prize, such as a state, regional, or national level.

Resident Performance Determination
Resident performance was composed of both a subjective and an objective component. For the subjective component, a survey was built on the RedCap platform and was sent to every current faculty member at the Dean McGee Eye Institute/University who was also a member of the faculty during the review period. The survey asked, “In which tertile would you rank the above resident as compared to your expectation for a graduating resident at completion of training?” If the top tertile was selected, the survey asked, “Is the resident in the top 10%?” and if the bottom tertile was selected, the survey asked, “Is the resident in the bottom 10%?” The survey then assessed each resident for deficiencies in each of the Accreditation Council for Graduate Medical Education (ACGME) core competencies both during residency and at the completion of residency. Space was provided for additional comments. The faculty members were blinded to the resident applications. At our institution, the faculty members are also blinded to the resident Ophthalmic Knowledge Assessment Program (OKAP) scores. Resident scores on postgraduate year (PGY)-2, PGY-3, and PGY-4 OKAP examinations were used as an objective measurement of residency performance, and had no relationship to the subjective faculty ranking.

Statistical Analysis
Data were descriptively summarized (e.g., mean, SD, range, count, and percentage). The effect of each predictor on the resident's ranking (top vs. bottom tertiles and top 10% vs. bottom 90%) was assessed with a logit model with the compound-symmetry covariance structure, accounting for multiple ratings per resident.

Interrater reliability was assessed using Fleiss' kappa statistic for multiple categories and raters, which was computed according to Chen et al.32

The effect of predictors on numerical OKAP scores was investigated using a series of linear regression models, as each resident had only a single OKAP score for each clinical year.

To construct a single resident-quality metric for the purposes of natural language processing (NLP), the individual raters' assignments of tertiles and top versus bottom 10% were transformed into numerical percentile ratings, and then aggregated within residents. The transformations were designed such that the resulting numerical value represented the midpoint of the “bin” in which a given rater would place a resident. Note that top and bottom tertile assignments are further refined by top 10% and bottom 10% status, respectively; for example, a resident who was placed in the top tertile by a given rater, but not the top 10%, would receive a score of 78.3% from that rater because this is the midpoint between 66.7% (lower end of top tertile) and 90% (upper end of top tertile after excluding the top 10%).

For the AI component, we performed a token analysis and sentiment analysis to determine whether any text features were predictive of resident performance.

Analyses were performed on three document classes: MSPE, PS, and LOR. Token analysis was conducted by stemming all words in each document class and applying term-frequency and inverse-document frequency weighting. Sentiment analysis was conducted by mapping each word or token in a given document to one of four NLP corpora to generate a set of scores for each.33–36 Sentiment scores in each of the corpora were transformed to a numeric range, and harmonized for modeling purposes. Following this process, each sentiment score for the document was used in two different linear regression models as a predictor for each of the associated resident's outcomes (percentiles and OKAP scores). All statistical analyses, including NLP modeling, were conducted in R v3.5.1 and SAS v9.4, with an α level of 0.05 (two-sided) defining statistical significance.

Results
A total of 55 residents from 2004 to 2019 were included in the study. A summary of resident characteristics can be found in Table 1. The majority of residents in the study period were male (80%). The mean USMLE step-1 score was 243.42 (range: 203–270), and the mean USMLE step-2 score was 246.50 (range: 228 - 273, though 43 applications did not report a step-2 score). The majority of residents were elected to AOA (61%, with 11 applications not reporting any kind of AOA status). The mean number of prereidency publications

| Table 1 Baseline characteristics and attributes of residents evaluated |
|---------------------------------|----------------|
| Resident characteristics | |
| Total number of residents | 55 |
| Male | 44 (80.00%) |
| Female | 11 (20.00%) |
| Mean USMLE step 1 | 243.42 ± 13.31 |
| Mean USMLE step 2 | 246.50 ± 14.61 |
| AOA membership* | 34 (61.82%) |
| Mean pre-residency publications | 1.04 (range: 0–8) |
| Advanced degree | 3 (5.45%) |
| Military service | 2 (3.64%) |
| College varsity athlete | 3 (5.45%) |
| Significant art/music experience | 9 (16.36%) |
| Home ophthalmology program | 49 (89.09%) |
| Home ophthalmology program ranked within top 25 | 20 (36.36%) |

Abbreviations: AOA, alpha-omega-alpha Honor Medical Society; USMLE, United States Medical Licensing Examination.

*11 applications did not report any AOA status.
was 1, ranging from 0 to 8. Only three residents had advanced degrees, two residents served in the military, and three residents played a varsity sport in college. The majority of applicants had a home ophthalmology department (89%), with 36% of home ophthalmology departments ranking in the top 25 based on the 2019 Doximity rankings.

Twenty-one faculty members participated in the survey. The mean number of reviewers per resident was 14.6 (range: 9–21). Interreviewer reliability for tertile ranking, measured by Fleiss’ kappa, was 0.148 (p < 0.001), which suggests only a slight agreement among the reviewers. Predictors of subjective success based on the faculty survey can be found in Table 2. The analysis was done comparing residents in the top tertile to the bottom tertiles, as well as the top 10% to the bottom 90%. Applicant characteristics that were predictive of being ranked in the top tertile were a grade of A or Honors in OB/GYN clerkship (p = 0.040), presence of a home ophthalmology department (p = 0.036), medical school ranking in the top 25 USNWR primary care rankings (p = 0.035), and mean core clerkship GPA (p = 0.025). Characteristics that were predictive of being in the top 10% were attending a medical school in a surrounding state (p = 0.047), medical school ranked in the top 25 USNWR primary care rankings (p = 0.022), and mean core clerkship GPA (p = 0.010). Attending an undergraduate school ranking in the USNWR top 50 was inversely correlated with being ranked in both the top tertile and top 10% (p = 0.037 and 0.027, respectively). Performance on the PGY-2 (p = 0.021 and 0.006, respectively) and PGY-3 (p = 0.013 and p = 0.004, respectively) OKAP examinations was predictive of being subjectively ranked in both the top tertile and top 10%.

With respect to objective success as measured by OKAP examinations, USMLE step-1 score was predictive of performance in the PGY-2 (0.9 point increase in OKAP score per 1 point increase in USMLE step-1 score, p = 0.001), PGY-3 (0.79 point increase in OKAP score per 1 point increase in USMLE step-1 score, p < 0.0001), and PGY-4 years (0.51 point increase in OKAP score per 1 point increase in USMLE step-1 score, p = 0.014). Undergraduate GPA (p = 0.020) and attending a medical school ranked in USNWR top 25 primary care ranking (p = 0.006) were predictive of performance on PGY-2 OKAP examination. Male gender (p = 0.007), preclinical medical school GPA (p = 0.049), and Honors in surgery clerkship (p = 0.016) were predictive of performance on PGY-3 OKAP examination. Significant experience in art or music was inversely predictive of performance on both the PGY-2 (p = 0.028) and PGY-3 (p = 0.015) OKAP examinations. Characteristics that were notably not predictive of resident success based on the faculty survey and OKAP results included membership in AOA, number of publications, average number of nonmedical employments, and grade of Honors or A in the medicine clerkship. USMLE step-1 and -2 scores were not predictive of subjective resident success based on faculty survey but did predict OKAP performance in PGY-2/-3 years.

In the NLP analyses, we were unable to identify any words or phrases in any of the documents that correlated nonrandomly with either resident performance or OKAP scores. In the token analysis, only the MSPE had reasonable out-of-bag accuracy (random forest) and leave-one-out cross-validation accuracy; however, this high accuracy was primarily an artifact of the ranking distribution and resultant confusion matrix, rather than indicative of predictive keywords. In the sentiment analysis, only a limited number of sentiments were associated with the overall resident performance score, and the consistency of this association did not hold across document classes. No sentiment scores were consistently associated with OKAP scores across residency years.

### Discussion

To our knowledge, this represents the first published, large-scale, multifaculty review study within ophthalmology that correlates preresidency characteristics with residency performance. Additionally, this study is the first to evaluate the use of AI as a potential predictor of which applicants will be successful in residency. Prior studies have utilized machine learning to evaluate LORs for gender bias, racial bias, and even success rates for matching into residency, but none have evaluated for success during residency. Our study

### Table 2 Resident attributes predictive of tertile rating on faculty survey

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Top tertile (n = 406)</th>
<th>Bottom tertiles (n = 395)</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honors in OB/GYN clerkship</td>
<td>307 (75.99%)</td>
<td>233 (60.05%)</td>
<td>0.0400</td>
</tr>
<tr>
<td>Presence of home ophthalmology department</td>
<td>382 (94.09%)</td>
<td>337 (85.32%)</td>
<td>0.0356</td>
</tr>
<tr>
<td>Medical school ranked in USNWR top 25 (primary care)</td>
<td>67 (16.50%)</td>
<td>31 (7.85%)</td>
<td>0.0350</td>
</tr>
<tr>
<td>Mean core clerkship GPA</td>
<td>3.91 ± 0.17</td>
<td>3.79 ± 0.28</td>
<td>0.0248</td>
</tr>
<tr>
<td>Mean PGY-2 OKAP score (percentile)</td>
<td>67.19 ± 24.02</td>
<td>56.94 ± 28.22</td>
<td>0.0214</td>
</tr>
<tr>
<td>Mean PGY-3 OKAP score (percentile)</td>
<td>75.26 ± 18.68</td>
<td>68.32 ± 20.59</td>
<td>0.0133</td>
</tr>
<tr>
<td>Undergraduate school ranked in USNWR top 50</td>
<td>69 (17.47%)</td>
<td>117 (30.87%)</td>
<td>0.0371</td>
</tr>
</tbody>
</table>

Abbreviations: GPA, grade point average; OB/GYN, obstetrics and gynecology; OKAP, ophthalmic knowledge assessment program; PGY, postgraduate year; USNWR, United States News and World Report.

*Also predictive of ranking in top decile.

*Inversely correlated with ranking in top tertile and decile.
identified multiple factors that were predictive of subjective and objective resident success, some of which differed from studies in other subspecialties.

Our study used both a subjective and an objective component to define success in residency. The following attributes were predictive of subjective success as measured by our faculty survey: mean core clerkship GPA (top tertile and top decile), attending a medical school ranked in the USNWR top 25 primary care ranking (top tertile and top decile), attending a medical school in a surrounding state (top decile only), presence of a home ophthalmology program (top tertile only), and grade of Honors or “A” in the OB/GYN clerkship (top decile only). Performance on PGY-2 and PGY-3 OKAP examinations was also predictive of top tertile and top decile subjective rating, but attributes prior to residency training are not known. For the objective component of resident success, our study used performance on the OKAP examination as a metric. The only attribute attributed that was predictive of OKAP performance on all 3 years was the USMLE step-1 score.

There has been some concern over the usefulness of clerkship grades as an evaluation metric which is highlighted by the notable interapplicant variability in methods of clerkship grade reporting. Despite these concerns, performance on clinical clerkships appears to be one of the most consistently reported predictors of both resident clinical and academic performance. This held true in our study as well, with core clerkship performance being one of the most significant predictors of subjective resident success, even though the mean differences were rather small. Logically, the unique linguistic skills. Similarly, the MSPE style and format is very different at different levels of their career, and with different innate linguistic skills. Similarly, the MSPE style and format is very different at different institutions. Finally, there is likely an inherent attempt to put forth the best impression for each applicant, and relative reticence to note negatives unless they are egregious.

**Limitations and Strengths**

Our study has several limitations. One of the biggest difficulties common to studies of this kind is defining what actually constitutes a “successful” resident. Like many studies, we divided this definition into a subjective and objective component. While we would consider the number of faculty members captured in our subjective component survey as a strength, not every resident was evaluated by the same number of faculty members. Though our survey showed statistical significance and interrater reliability, suggesting...
only slight agreement among reviewers. Some faculty members may be harder graders than others, creating a nonstandard evaluation for the residents.

Ophthalmology is a relatively small specialty with relatively small residency classes. Because of this, we are unable to include as many residents in our study as larger specialties such as internal medicine or general surgery. This limits our ability to draw conclusions about less common applicant characteristics such as varsity sports participation ($n=3$). Additionally, ophthalmology is a very competitive specialty, and our residency typically receives 100 applications per slot available. Thus, the applicants who are interviewed have already been prescreened based on several application metrics. In light of this, our results are likely not generalizable to the entire, nonscreened ophthalmology applicant pool. Additionally, the subjective USNWR and Doximity rankings change over time, and the use of only the most recent year’s rankings does not take into account these changes. Furthermore, our study population was 80% male during the time period studied. Research in other specialties have demonstrated gender-based differences in LORs. Through gender was only found to be a significant predictor for PGY-3 OKAP scores in our study, gender differences between our study population and the general applicant pool serves as a potential source of bias.

Finally, this study is limited by the inability to understand the variation in applicant characteristics present among an entire applicant cohort. We obviously could not examine the documents of all applicants to our program, including applicants who were screened out early in the process without an interview, applicants who were interviewed but not offered a position, and applicants who were offered a position but matched elsewhere. It is certainly possible that a study of this nature, although virtually impossible to perform, may identify many more significant differences among these groups of candidates.

**Conclusion**

In summary, to our knowledge this represents the first large-scale, multifaculty review study within ophthalmology that correlates residency application characteristics with actual residency performance. To our knowledge, this is also the first study within any specialty to use machine learning and AI to analyze the quality of LOR, MSPE, and PSs with regard to their predictive nature for success in that specialty. While core clerkship performance (especially in OB/GYN) and medical school ranking correlated with future success as a resident in our program, many commonly valued application characteristics, such as USMLE scores, AOA status, and number of preresidency publications, were not predictive of subjective success in residency. USMLE step-1 scores were decreasingly predictive over time of OKAP performance throughout residency. Our inability to identify predictive language in the current LOR formatting suggests that this aspect of the application process may be less discerning. While a standardized LOR process has been proposed, it has not been widely adopted and faces barriers to widespread implementation. Finally, our study lends credence to the use of the OKAP examination as a surrogate of more global resident performance.

**Meeting Presentation**

This study was previously presented in part as a poster at the annual Association of University Professors of Ophthalmology (AUPO) meeting, 2019.

**Conflict of Interest**

No conflicting relationship exists for any author.

**Acknowledgments**

This research is supported in part by an unrestricted grant from Research to Prevent Blindness, Inc., New York City, NY. The sponsor or funding organization had no role in the design or conduct of this research.

**References**

Correlation of Residency Application and Residency Performance

Gudgel et al. e157