Everyone Is Awesome: Analyzing Letters of Reference in a General Surgery Residency Selection Process

Chelsea Towaij, MD
Isabelle Raîche, MD, MAEd, FRCSC
Julia Younan, MD
Nada Gawad, MD, MAEd

ABSTRACT

Background The resident selection process involves the analysis of multiple data points, including letters of reference (LORs), which are inherently subjective in nature.

Objective We assessed the frequency with which LORs use quantitative terms to describe applicants and to assess whether the use of these terms reflects the ranking of trainees in the final selection process.

Methods A descriptive study analyzing LORs submitted by Canadian medical graduate applicants to the University of Ottawa General Surgery Program in 2019 was completed. We collected demographic information about applicants and referees and recorded the use of preidentified quantitative descriptors (eg, best, above average). A 10% audit of the data was performed. Descriptive statistics were used to analyze the demographics of our letters as well as the frequency of use of the quantitative descriptors.

Results Three hundred forty-three LORs for 114 applicants were analyzed. Eighty-five percent (291 of 343) of LORs used quantitative descriptors. Eighty-four percent (95 of 113) of applicants were described as above average, and 45% (51 of 113) were described as the "best" by at least 1 letter. The candidates described as the "best" ranked anywhere from second to 108th in our ranking system.

Conclusions Most LORs use quantitative descriptors. These terms are generally positive, and while the use does discriminate between different applicants, it was not helpful in the context of ranking applicants in our file review process.

Introduction

The process of selecting medical students for residency positions across all specialties is a complex and subjective exercise involving the analysis of multiple different data points. In Canada this system is organized and overseen by the Canadian Resident Matching Service (CaRMS). Through this application portal, programs look at applicants' personal statements, CV, medical school records, and letters of reference (LORs), and try to draw meaningful comparisons from these documents in order to select the best-suited candidates for their programs. Currently, recommendations for the content of LORs are provided by CaRMS, but the content of LORs remains variable. While there are some institutional differences, most programs accept 3 LORs per applicant.[1]

In urology and plastic surgery studies, LORs from known sources were often considered the most important factor in selecting residents for interviewing and ultimately matching to a residency position.[2,3] Studies on the value of narrative LORs found that LORs from unknown writers are generally found to hold less weight.[2,4] With increasing subspecialization,[5] programs will inevitably have to interpret LORs from faculty who are unknown to them. Some programs, including emergency medicine, otolaryngology, and dermatology, have implemented standardized letters of reference (SLORs) in order to mitigate high interreader variability and ambiguity of terminology and provide easier comparison between candidates.[6,7] These SLORs are also thought to decrease gender bias that has been described in the literature about Otolaryngology residency selection.[8] Many programs, however, continue to use narrative LORs that are inherently subjective in nature.

Previous studies investigating the content and value of narrative LORs have described liberal use of glowing single word summary statements, such as "outstanding" to describe candidates,[9] and many editorials have criticized both the level of inflation[10] and the poor quality of the letter writing.[11] De Zee et al[12] surveyed 110 institutional members of the

DOI: http://dx.doi.org/10.4300/JGME-D-20-00034.1

Editor's Note: The online version of this article contains information about the relationship between the use of different quantitative descriptors in the study.

Clerkship Directors in Internal Medicine and found that numeric comparisons of applicants to other students (eg, top one-third of students) was the second most important factor when rating LORs (the first was perceived depth of understanding of the candidate). No studies to date have examined how frequently these quantitative comparisons are used, and it is unclear if the use of these terms is descriptive of the applicants. Therefore, the true value of these quantitative descriptors remains uncertain.

The first objective of our study was to assess the frequency with which LORs use quantitative descriptors, such as "above average" or "in the top third" to describe applicants. The second objective was to assess whether the use of these terms reflects the ranking of trainees in the selection process.

What was known and gap
Selecting medical students for residency positions involves analyzing multiple subjective data points, including letters of reference (LORs).

What is new
A descriptive study analyzing LORs submitted by Canadian medical graduate applicants to the University of Ottawa General Surgery Program in 2019.

Limitations
Single center, single specialty study limits generalizability.

Bottom line
Most LORs frequently used quantitative descriptors to compare applicants, and their usage demonstrates inflation that makes it difficult to discriminate between applicants in a resident selection process.

Methods

Study Population

This retrospective cohort study included all Canadian medical graduates (CMGs) in the 2019 CaRMS cycle applying to the University of Ottawa General Surgery Program—an urban, university-based program with 32 residents. There were 6 available residency spots, 5 of which were for CMGs and 1 of which was for an internal medical graduate (IMG). IMGs were excluded from our study since their applications are reviewed by different criteria and within a separate stream to account for differences in applicant profiles. For example, IMGs are less likely to have completed multiple Canadian clinical experiences, and these are often observerships with no direct patient contact. Their LORs are also more heterogeneous and frequently written by referees outside of the specialty. Therefore, it would be difficult to compare IMG LORs to CMG LORs in a meaningful way, and they do not compete for the same residency spots.

Data Extraction

Letters of reference for each applicant in the study population were identified through the CaRMS database. A predefined, pre-piloted data extraction form was used to gather data points related to applicant gender and home school, referee gender and home school, the title of the referee (program director, division chief, staff surgeon), the type of exposure the referee had to the student (clinical or research), and finally the quantitative descriptors used in the letters. The LORs were available online during the CaRMS application window and were not preserved or downloaded to ensure the confidentiality of the applicants.

Quantitative descriptors were identified a priori based on an initial review of 10 sample letters. Quantitative descriptors were defined as any term meant to compare candidates in an objective way, and included references to the "best" applicants, those who were average or above average, those who functioned at the level of a resident, or those described with a global percentage (ie, as being in the top "x" percent of applicants; TABLE 1).

Data extraction with the form was then completed by one author (C.T.). To ensure accuracy, an independent, duplicate 10% audit was completed by a second author (N.G.). There was greater than 90% agreement on all data, and all identified discrepancies were minor, consisting of typographical errors that were then corrected. No new quantitative descriptors were identified in the remaining data extraction.

Program requirements were 3 LORs. One person submitted 4 LORs, which were included in all analyses except those looking at consistencies across 2 or more letters.

The file review process in our institution includes review of each applicant's file by 3 reviewers consisting of faculty and senior residents, each applicant's personal statement, CV, elective experience, and LORs, which are scored to generate a final ranking that determines which applicants are offered an interview.

Statistical Analysis

Descriptive statistics were used to analyze the frequency of quantitative descriptor use. Categorical variables were described as proportions. A 1-way analysis of variance was used to compare the mean file review rankings between groups of applicants described by different quantitative descriptors. Chi-square analysis was used to evaluate if there was any statistical relationship between the use of different quantitative descriptors. P values $< .05$ were considered statistically significant. IBM SPSS Statistics for

TABLE 1
Quantitative Descriptors of Applicants

Quantitative Descriptor	Synonyms Included
"Best"	The finest, the top of their class, the strongest
Above average	Above their peers, excellent for their level
Average	At the level of their peers, good for their level
At a resident level	N/A
In the top "x" percent	In the top (fraction)

Abbreviation: N/A, not applicable.

Windows, Version 25.0 (IBM Corp, Armonk, NY) was used for all analyses.

Ethics approval was waived by the Ottawa Health Science Network Research Ethics Board.

Results

The study cohort included 343 letters for 114 applicants. The majority of letters were clinical (87%, n = 300), written by men (70%, n = 241) who self-identified as staff surgeons (82%, n = 282), and were from the applicant's home school (58%, n = 200) as described in TABLE 2.

The majority of LORs used quantitative descriptors (85%, n = 291). TABLE 3 describes the frequency of use of different quantitative descriptors to describe applicants. Most applicants were described as above average (84%, n = 95) and working at the level of a resident (73%, n = 82) by at least 1 LOR. Just under half (45%, n = 51) of applicants were described as the "best," or a synonym thereof (TABLE 1), by at least 1 letter. Half of applicants were described as being above average (48%, n = 54), one-third were described as functioning at a resident level (35%, n = 40), and only 8% (n = 9) were described as being the "best" by at least 2 LORs.

TABLE 2
Letter of Reference Characteristics (N = 343)

Characteristics	n (%)
Referee gender	
Male	241 (70)
Female	102 (30)
Title of referee	
Program director	32 (9)
Division chief	25 (7)
Staff surgeon	282 (82)
Fellow	3 (1)
Type of letter	
Clinical	300 (87)
Research	6 (2)
Both	37 (12)
School of referee	
Same as applicant	200 (58)
Other	143 (42)

TABLE 3
Frequency of Use of Quantitative Descriptors to Describe Applicants (N = 113)

Quantitative Descriptor	n (%)	
	At Least 1 Letter	2 or 3 Letters
Average	52 (46)	6 (5)
Above average	95 (84)	54 (48)
"Best"	51 (45)	9 (8)
"At resident level"	82 (73)	40 (35)

Over half of applicants (58%, 64 of 113) were described using a global percentage, which is to say that they were described as in the top "x" percent of their peers. When used, global percentages ranged from the top 1% to the top 33%, with a mean (± SD) of 8.9% (± 6.8%).

There was no relationship between the use of the terms "best," "above average," ($P = .33$, compared with "best"), and functioning at a "resident level" ($P = .67$, compared with "best"; $P = .23$, compared with "above average") even when stratified by applicants who had been described by these terms in at least 2 letters. In other words, an applicant who was described as the "best" by 2 referees was not statistically more likely to be described as working at a resident level or being above average by another referee (provided as online supplemental material).

Candidates described as being the "best" in at least 1 LOR did score higher on average during the residency program's initial file review (20.4 vs 16.7, $P < .05$; TABLE 4); however, they ranked anywhere from 2 to 108 of 114 applicants and thus this did not help to discriminate between candidates (provided as online supplemental material).

Discussion

This study demonstrated that most LORs use numeric or other quantitative descriptors, and the majority of these are positive. It further suggests that the use of quantitative descriptors may be inflated given that

TABLE 4
File Review Scores of "Best" Applicants

Candidates	N	File Review Score[a] (mean ± SD)	*P* Value
"Best"	52	20.4 ± 4.5	< .001
Others	71	16.7 ± 5.8	

[a] Score based on file review that includes the candidate's personal statement, letters of reference, CV, and clinical electives.

most applicants were described as above average and nearly half of applicants were described as the "best" by at least 1 letter. The use of these quantitative descriptors also did not correlate with the final ranking of candidates.

While previous studies analyzing the content of LORs have suggested a level of inflation given the frequent use of positive one-word adjectives,[9] our study demonstrates that this degree of inflation may limit the interpretation of LORs. This study demonstrated that a more plausible percentage of applicants were described as above average (48% vs 84%) and the "best" (8% vs 45%) when considering 2 or more LORs. This suggests that when analyzing LORs, it may be prudent to focus on a consensus across LORs as opposed to focusing on the individual content of each letter.

Emergency medicine initially piloted the SLOR, recently renamed the standardized letter of evaluation (SLOE), as a substitute for traditional LORs. These rely on direct observation in predetermined competencies and have been shown to have improved interrater reliability.[13] There remains concern that these SLOEs are subject to inflation similarly to our findings with narrative LORs.[14] While most authors of SLOEs do not believe they grade inflate, surveys have revealed that they may use their own interpretation of adjectives in the SLOEs, and many authors have not read the instructions to authors.[15] The purpose of and barriers to training authors are therefore important considerations. In dermatology, an analysis of 141 SLORS demonstrated significant grade inflation where an "exceptional" grade (meant for the top 5% of students) was given 25% of the time.[16] Furthermore, at least 1 program has opined that the SLOR has been of limited utility given that most candidates remain clustered at the top of the scale.[17]

The utility of LORs and SLORs in residency may be more critical this year with changes in Canada, the United States, and elsewhere to the interview process. Medical student electives have been severely limited by the COVID-19 pandemic, and interviews will take place virtually for 2021.[18] Residency programs may need to rely more heavily on elements of the application file review, and understand the limitations of LORs.

This study is limited by the potential lack of generalizability to other residency programs given that the characteristics of both our applicants and referees may differ from other specialties, in that general surgery traditionally puts a lot of value on being the "best."[19] This may be reflected by the frequent use of superlatives. Given the deidentification of our retrospectively collected data we were not able to analyze the impact of the LOR score on the candidates' file review scores. However, given that this difference in scores failed to result in meaningful differences in ranking, this may be less relevant. We also did not assess whether the identity of the letter writer factored into interpretation of the quantitative descriptors in the LORs.

Analysis of the LORs of unmatched applicants may help clarify whether the absence of certain quantitative descriptors may in fact be interpreted as a cause for concern, but this is challenging given the importance of maintaining student confidentiality and anonymity. Analysis of the content of LORs from other specialties that are perhaps traditionally viewed to place a greater value on communication and relationships would also help generalize our findings.

Conclusions

Narrative LORs frequently use quantitative descriptors to compare applicants, and their usage demonstrates inflation that makes it difficult to discriminate between applicants in a resident selection process. Use of these quantitative descriptors was not found to correlate with candidate rankings.

References

1. CaRMS. Reference guidelines. https://www.carms.ca/match/r-1-main-residency-match/referee/reference-guidelines-r1-referee/. Accessed July 31, 2020.
2. Liang F, Rudnicki PA, Prince NH, Lipsitz S, May JW Jr, Guo L. An evaluation of plastic surgery resident selection factors. *J Surg Educ*. 2015;72(1):8–15. doi:10.1016/j.jsurg.2014.07.013.
3. Weissbart SJ, Stock JA, Wein AJ. Program directors' criteria for selection into urology residency. *Urology*. 2015;85(4):731–736. doi:10.1016/j.urology.2014.12.041.
4. Greenburg AG, Doyle J, McClure DK. Letters of recommendation for surgical residencies: what they say and what they mean. *J Surg Res*. 1994;56(2):192–198. doi:10.1006/jsre.1994.1031.
5. Webber EM, Ronson AR, Gorman LJ, Taber SA, Harris KA. The future of general surgery: evolving to meet a changing practice. *J Surg Educ*. 2016;73(3):496–503. doi:10.1016/j.jsurg.2015.12.002.
6. Kimple AJ, McClug SW, Del Signore AG, Tomoum MO, Lin FC, Senior BA. Standardized letters of recommendation and successful match into otolaryngology. *Laryngoscope*. 2016;126(5):1071–1076. doi:10.1002/lary.25637.
7. Keim SM, Rein JA, Chisholm C, Hendey GW, Jouriles NJ, King RW, et al. A standardized letter of

recommendation for residency application. *Acad Emerg Med*. 1999;6(11):1141–1146. doi:10.1111/j.1553-2712.1999.tb00117.x.
8. Friedman R, Fang CH, Hasbun J, Han H, Mady LJ, Eloy JA, et al. Use of standardized letters of recommendation for otolaryngology head and neck surgery residency and the impact of gender: gender and letter of recommendations. *Laryngoscope*. 2017;127(12):2738–2745. doi:10.1002/lary.26619.
9. Fortune JB. The content and value of letters of recommendation in the resident candidate evaluative process. *Curr Surg*. 2002;59(1):79–83. doi:10.1016/S0149-7944(01)00538-4.
10. Friedman RB. Sounding board. Fantasy land. *N Engl J Med*. 1983;308(11):651–653. doi:10.1056/NEJM198303173081110.
11. Prayson RA, O'Toole EE. On the subject of writing letters of recommendation. *Ann Diagn Pathol*. 2019;40:180–181. doi:10.1016/j.anndiagpath.2018.08.006.
12. DeZee KJ, Thomas MR, Mintz M, Durning SJ. Letters of recommendation: rating, writing, and reading by clerkship directors of internal medicine. *Teach Learn Med*. 2009;21(2):153–158. doi:10.1080/10401330902791347.
13. Love JN, Doty CI, Smith JL, Deiorio NM, Jordan J, Van Meter MW, et al. The emergency medicine group standardized letter of evaluation as a workplace-based assessment: the validity is in the detail. *West J Emerg Med*. 2020;21(3):600–609. doi:10.5811/westjem.2020.3.45077.
14. Love J, Ronan S, Deiorio NM, Howell J, Doty C, Lane D, et al. Characterization of the CORD standardized letter of recommendation in 2011 to 2012. *Ann Emerg Med*. 2013;62(5):168–169. doi:10.1016/j.annemergmed.2013.06.030.
15. Love JN, Smith J, Weizberg M, Doty CI, Garra G, Avegno J, et al. Council of Emergency Medicine Residency Directors' standardized letter of recommendation: the program director's perspective. *Acad Emerg Med*. 2014;21(6):680–687. doi:10.1111/acem.12384.
16. Wang RF, Zhang M, Alloo A, Stasko T, Miller JE, Kaffenberger JA. Characterization of the 2016–2017 dermatology standardized letter of recommendation. *J Clin Aesthet Dermatol*. 2018;11(3):26–29.
17. Puscas L. Viewpoint from a program director they can't all walk on water. *J Grad Med Educ*. 2016;8(3):314–316. doi:10.4300/JGME-D-16-00237.1.
18. The Association of Faculties of Medicine of Canada. AFMC decision regarding the 2021 R-1 match. AFMC. https://afmc.ca/en/node/357. Published 2020. Accessed July 31, 2020.
19. de Santibañes E, Cano Busnelli V, Pellegrini C. Excellence in surgery: becoming the "best" you can be. *Bulletin of the American College of Surgeons*. https://bulletin.facs.org/2018/04/excellence-in-surgery-becoming-the-best-you-can-be/. Accessed July 31, 2020.

All authors are with University of Ottawa, Ottawa, Ontario, Canada. **Chelsea Towaij, MD,** is a Resident, Department of Surgery; **Isabelle Raîche, MD, MAEd, FRCSC,** is Assistant Professor of Surgery, Department of Surgery; **Julia Younan, MD,** is a Resident, Department of Surgery; and **Nada Gawad, MD, MAEd,** is a Resident, Department of Surgery.

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

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

This abstract was previously presented at the Canadian Surgery Forum, Montreal, Quebec, Canada, September 5–7, 2019, and the International Conference on Residency Education, Ottawa, Ontario, Canada, September 26–28, 2019.

Corresponding author: Chelsea Towaij, MD, University of Ottawa, Attn: General Surgery Program Admin, 725 Parkdale Avenue, Ottawa, ON K1Y 4E9, Canada, 613.889.9274, ctowaij@toh.ca

Received January 9, 2020; revision received June 13, 2020; accepted June 24, 2020.