Journal of Clinical Neuroscience 101 (2022) 63–66

Contents lists available at ScienceDirect

Journal of Clinical Neuroscience

journal homepage: www.journals.elsevier.com/journal-of-clinical-neuroscience

Clinical study

Impact of virtual vs. in-person interviews among neurosurgery residency applicants

Safwan Alomari [1], Daniel Lubelski [1], James Feghali, Henry Brem, Timothy Witham, Judy Huang [*]

Department of Neurosurgery, Johns Hopkins University School of Medicine, Baltimore, MD, USA

ARTICLE INFO

Keywords:
Neurosurgery
Residency application
Interview
Virtual
Reliability

ABSTRACT

Background: The interview is considered a key factor in selecting residents in various medical and surgical specialties. However, the reliability of the interview process in selecting neurosurgery training program applicants remains largely under-investigated.

Objective: To investigate the reliability of the interview process for neurosurgery residency applicants and to evaluate the impact of virtual interviews on this process.

Methods: We analyzed the records of neurosurgery residency applicant interviews at our institution between 2016 and 2021. An average of 20 neurosurgery faculty members (clinical and research) interviewed each applicant and graded them 1 (best) to 4 (worst). Intraclass correlation coefficient (ICC) and Levene's test were used to assess the inter-rater and intra-rater reliability, respectively.

Results: 214 neurosurgery residency applicants were interviewed at a single institution between 2016 and 2021. The mean applicant rating each year ranged from 1.77 to 1.92. Inter-rater agreement was relatively poor in each year, (ICC < 0.5, $P < 0.05$). Among 60% of the raters, variability of scores significantly changed from year to year, ($p < 0.05$). When comparing the scores submitted during the virtual interview process (2021) with the scores submitted in the previous years (2016–2020), 2 interviewers (10%) had less variability using the virtual process.

Conclusion: Our analysis found that the current interview process for neurosurgery residency applicants' selection suffers from poor inter- and intra-rater reliability. Virtual interviews may be part of a cost-effective strategy to improve the reliability of the interview process. Further validation is needed, as well as identification of novel strategies to maximize the reliability of the selection process.

1. Introduction

Faculty and applicants from multiple medical and surgical specialties have identified the interview as an important factor in determining resident selection [1-4]. However, the interview process can be stressful and a significant financial burden for both applicants and programs. Applying for a neurosurgical residency has an average cost of $10,255 per applicant, making up 69% of total expenses [5].

Despite the substantial resources allocated to the interview process, studies have demonstrated that they can be highly subjective and have poor reliability [4,6]. Some have criticized the interview process for its inability to adequately assess or predict applicants' future performance [7,8].

In the present study, we investigated the inter-and intra-rater reliability of neurosurgery residency applicant interviews at a single neurosurgery residency program over a six-year period. We sought to evaluate (a) the variability of scores among different faculty for a given applicant in a given year, (b) variability of the average scores recorded for an individual faculty member within a given year and between years, and (c) did the virtual interviews conducted in the setting of the COVID-19 pandemic significantly affect the reliability and variability within the interview process.

Abbreviations: ICC, intraclass correlation coefficient; SD, standard deviation; I, increase; D, decrease; USMLE, United Stated Medical Licensing Exam.

* Corresponding author at: Department of Neurosurgery, Johns Hopkins Hospital, 1800 Orleans Street, Sheikh Zayed Tower 6115F, Baltimore, MD 21287, USA.
E-mail address: jhuang24@jhmi.edu (J. Huang).

[1] Co-first authors.

https://doi.org/10.1016/j.jocn.2022.05.005
Received 22 December 2021; Accepted 7 May 2022
Available online 10 May 2022

2. Material and methods

2.1. Data source

This study was conducted using the records of neurosurgery residency applicant interviews at our institution between the years 2016 and 2021. Applicants meet with an average of two faculty members for each interview, and have approximately 10 interviews throughout the day (20 total interviewers). The interviewers then submit their scores for each applicant. The scores submitted by the interviewers are meant to reflect the overall assessment, including both the application and the interview itself. There is an attempt to have most faculty interview each applicant, but sometimes certain individuals are unable to. The interview process was in-person during the years 2016–2020 and virtual in 2021. Institutional review board approval was not required.

2.2. Statistical analysis

Statistical analysis was conducted using R statistical software (R Foundation for Statistical Computing, Vienna, Austria). The intraclass correlation coefficient (ICC) was utilized to measure the inter-rater reliability for interviewers; this reflects the degree of agreement of different raters on the same applicant(s) within each year [9]. ICC values<0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability [9]. A p value ≤ 0.05 for ICC indicates that the resulting values of the ICC test were unlikely to be due to chance. Few interviewers interviewed all applicants in each year, while others interviewed only part of the applicants. The ICC can only be calculated for a cohort of raters if they all interviewed the same applicants. To account for this, two analyses were performed. The first subset analysis included only the interviewers who interviewed the majority of the applicants in a given year while the second subset analysis included all interviewers who assessed more than 10 applicants in a given year.

Levene's test was used to assesses the equality of variances for each interviewer across the years [10]. To ensure the validity of Levene's test, interviewers who had the highest numbers of interviews with applicants were included. The standard deviations (SDs) for each interviewer were calculated across the years 2016–2021. In addition, Levene's test was performed to compare the variances in the year 2021 to the variances of the previous years (2016–2020) to assess any effect of the virtual interview format compared to the classical in-person interviews. A p value ≤ 0.05 was considered to be statistically significant and indicated that the increase (I) or the decrease (D) in the variability from one year to the next was unlikely to be due to chance.

3. Results

A total of 214 applicants were interviewed during the period of 2016–2021 and included in the analysis, Table 1. The size of the interviewed cohort each year varied between 30 and 44, while the number of raters varied between 18 and 24. The average applicant rating ranged from 1.77 to 1.92. The ICC consistently indicated poor inter-rater reliability across the years in both subset analyses (ICC < 0.5, P < 0.05), Table 2. In each year, the subset including interviewers who assessed the majority of applicants tended to have relatively higher ICC when compared to the subset including all interviewers who assessed more than 10 applicants, Table 2.

Variability of each interviewer in a given year was estimated using standard deviation of scores submitted by that interviewer in that year. Six out of ten interviewers demonstrated significant changes (increase or decrease) in variability across the years (2016–2021), ($p < 0.05$). Changes in score variability for the remaining four interviewers did not achieve statistical significance, (p greater than 0.05).When comparing the scores submitted in 2021 (virtual interviews) with the scores submitted in the previous years (2016–2020), only two interviewers had less variability in 2021, Table 3. Among 60% of the raters, variability of scores significantly changed from year to year, ($p < 0.05$).

Table 1
Applicant and rater characteristics per year.

Year	No. of Applicants	No. of Raters	Mean Applicant Rating
2021	44	24	1.83
2020	33	24	1.80
2019	41	24	1.77
2018	30	23	1.92
2017	33	19	1.89
2016	33	18	1.81

4. Discussion

Neurosurgery is considered among the most competitive residencies [11]. Appropriate selection of trainees is of particular interest given the high stress of practice, the critically ill patient population, extended length of training, the relatively small size of most residency programs and the high-volume work demands placed on residents. Attrition rate among neurosurgical residents can be as high as 14% [12]. Accordingly, prior studies have investigated the current tools used in the assessment of candidates [13,14]. Much emphasis was placed on the interview process, and prior work has questioned the validity and reliability of unstructured interviews [15,16].

Our analysis revealed poor agreement among different raters on the same applicant(s) within each year. The variability of scores for a single applicant by different interviewers is inevitable since interviewers cannot be perfectly objective [16-18]. In fact, one of the reasons that protocols necessitate the inclusion of multiple interviewers is to reduce the effect of unrecognized underlying interviewer bias. However, there should be a certain degree of agreement among the interviewers on each applicant in order to consider the interview process as a reliable and valid tool to stratify applicants [16-18]. This has also been validated by previous studies in different specialties [8,15,17,19].

In addition, Levene's test demonstrated that the variance/SDs of scores submitted by each rater significantly changed (increased or decreased) from year to year. While we do not expect the absence of this variability even in ideal interviews, the significance and consistency of this variability among most raters can be an indicator of poor reliability

Table 2
Inter-rater reliability for interviewers within different years.

	Interviewers who assessed the majority of applicants in given year			All interviewers who assessed more than 10 applicants in a given year		
	No. of applicants	ICC [95%CI]	p-value	No. of applicants	ICC [95%CI]	p-value
2021	41/44	0.328 [0.171, 0.505]	p < 0.001	29/44	0.311 [0.166, 0.502]	p < 0.001
2020	33/33	0.330 [0.115, 0.552];	p = 0.001	26/33	0.166 [0.025, 0.374]	p = 0.009
2019	39/41	0.342 [0.182, 0.524]	p < 0.001	25/41	0.244 [0.100, 0.449]	p < 0.001
2018	30/30	0.529 [0.316, 0.716]	p < 0.001	20/30	0.491 [0.290, 0.707]	p < 0.001
2017	33/33	0.364 [0.149, 0.581]	p < 0.001	21/33	0.238 [0.067, 0.479]	p = 0.002
2016	33/33	0.453 [0.241, 0.651]	p < 0.001	16/33	0.157 [-0.013, 0.437]	p = 0.037

Table 3
P-values for Levene's test comparing year-to-year, overall, and zoom differences in variability. "I" indicates a significant increase while "D" indicates a significant decrease in variability from one year to the next.

Rater	2016–2017	2017–2018	2018–2019	2019–2020	2020–2021	Overall	Pre-zoom-2021
1	0.573	0.606	**0.025 (D)**	0.125	0.314	0.146	0.152
2	0.133	0.948	**0.039 (D)**	0.325	0.882	0.249	0.661
3	**0.002 (I)**	0.597	0.802	0.491	0.375	**0.033**	0.143
4	0.586	0.091	**0.027 (I)**	**0.011 (D)**	0.438	**0.038**	0.139
5	0.464	0.116	**0.046 (D)**	0.075	**0.023 (D)**	**0.032**	0.194
6	0.34	**0.01 (I)**	**0.001 (D)**	0.67	0.317	**0.002**	0.601
7	0.816	**<0.001 (D)**	0.118	0.117	0.962	**<0.001**	0.054
8	0.214	0.34	**0.003 (D)**	**0.001 (I)**	**0.006 (D)**	**0.001**	**0.012**
9	0.071	0.938	0.494	0.195	0.369	0.202	0.737
10	0.554	0.455	0.336	0.242	0.346	0.421	0.833

of the interview process from a statistical standpoint [10,20]. Of note, if we assume that the scores generated in the interview process for each year are normally distributed, it should not be expected to have significant changes in the standard deviations and variances from year to year [21,22].

Of note, prior studies have investigated other limitations of the residency interviews. Stephenson-Famy and colleagues [8] conducted a review of thirty-four studies and found that residency interviews did not predict subsequent clinical performance in internship or residency, particularly with a traditional unstructured interview format. In another survey of neurosurgery applicants, Zuckerman et al [14] found that the students perceived the most frequently discussed topics during interviews were not useful. Limoges et al [23] found that 94% of neurosurgical residency applicants during 2018 and 2019 reported being asked at least one inappropriate or potentially illegal question during interviews and these inappropriate questions negatively affected program rankings.

Several strategies have been suggested to improve the residency applicants' selection process. Several authors have demonstrated that use of structured interviews may increase the reliability and validity of the process [13,14,24-27]. Blouin et al [16] found that a structured interview tool provided good interrater reliability for admission to an emergency medicine residency program. Lubelski et al [13] investigated the use of standardized personality assessment tools in the selection of neurosurgery residents. The authors found that the use of these assessment tools can have the potential to provide insight into an applicant's future performance. In addition, our analysis showed that the subset including interviewers who assessed the majority of applicants tended to have relatively higher ICC in each year. This can indicate that interviewing more applicants might increase the inter-rater reliability among interviewers and possibly improves the overall selection process. However, it is important to mention that this comparison did not achieve statistical significance since the confidence intervals of ICC values were wide and overlapping. This is probably due to the relatively small sample size (number of applicants interviewed by same raters).

4.1. Virtual interviews

When comparing the scores submitted in 2021 (virtual interviews) with the scores submitted in the prior years (in-person interviews), we found that two interviewers had less variability in 2021, suggesting a possible advantage of virtual interviews when compared to in-person interviews. The COVID-19 pandemic had made virtual interviews the sole option for the last match. Several studies showed that virtual interview represents a potential opportunity to minimize expenditure and maximize convenience, flexibility, and time utilization [28-31]. Others, however, argue that the traditional in-person interview allows more interaction with faculty and residents on interview day. The National Resident Matching Program Director survey in 2018 demonstrates that such interactions were cited to be important by 96% of residency program directors, with a mean importance rating of 4.8 on a 5-point scale [32].

Overall, there remains an unmet need to develop a reliable interviewing process that is capable of appropriately assessing applicants and accurately predicting their future performance. This is especially important given that other selection factors, such as the USMLE score, are not predictive of the applicants' performance as a resident.[33].

4.2. Limitations

The retrospective design and relatively small sample size are the major limitations of the present study. Our study only included a sample from a single institution thereby limiting generalizability. Additionally, only few faculty members interviewed *all* applicants each year, thereby potentially introducing an element of heterogeneity. To try to mitigate the effect, we conducted two subset analyses looking at the inter-rater reliability in both the total population as well as in just the group that interviewed all applicants. Both groups showed concordant results. Furthermore, although having two interviewers with less variability scores in 2021 may suggest a possible advantage of virtual interviews over in-person interviews, this is undoubtedly a weak and very preliminary evidence and further evaluation of the advantages and disadvantages of this process are warranted. Ultimately, it is important to recognize that there may be intangible values to the neurosurgery interview which may be harder to quantify – including developing professional relationships and obtaining a better understanding of the nature of a given program. Similarly, applicants open up differently to different interviewers, and different aspects of the applicant that become apparent may make them more or less appropriate for the individual training program. There may be value in the open conversational structure of the current interview process, but given the heterogeneity and poor reliability in the current process, programs may want to consider incorporating more structured and validated portions as well.

5. Conclusion

Our analysis found that the interview process for neurosurgery residency applicants has poor inter- and intra-rater reliability. The average scores vary from year to year. There is substantial variability in how faculty members rate a given applicant within each year. The virtual interviews may provide a cost-effective strategy to improve the reliability of the interview and application process. We suspect that other similar institutions may be encountering similar heterogeneity. Further research is encouraged to validate the results of this study as well as to find strategies for maximizing the reliability of the selection process.

6. Disclosure of funding

None.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment:

We would like to acknowledge Colleen Hickson – the medical training programs administrator at department of neurosurgery at Johns Hopkins University for her assistance in obtaining and tabulating the data.

We would like to acknowledge support for the statistical analysis from the National Center for Research Resources and the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health through Grant Number 1UL1TR001079.

References

[1] Deloney LA, Perrot LJ, Lensing SY, Jambhekar K. Radiology resident recruitment: A study of the impact of web-based information and interview day activities. Acad Radiol 2014;21(7):931–7.
[2] Long T, Dodd S, Licatino L, Rose S. Factors important to anesthesiology residency applicants during recruitment. J Educ Perioper Med 2017;19(2):E604.
[3] Schneeweiss R, Bergman J, Clayton J. Characteristics of the residency interview process preferred by medical student applicants. J Fam Pract 1982;15(4):669–72.
[4] 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.
[5] Agarwal N, Choi PA, Okonkwo DO, Barrow DL, Friedlander RM. Financial burden associated with the residency match in neurological surgery. J Neurosurg 2017;126(1):184–90.
[6] Gardner AK, D'Onofrio BC, Dunkin BJ. Can We Get Faculty Interviewers on the Same Page? An Examination of a Structured Interview Course for Surgeons. J Surg Educ 2018;75(1):72–7.
[7] Hern HG, Trivedi T, Alter HJ, Wills CP. How prevalent are potentially illegal questions during residency interviews? A follow-up study of applicants to all specialties in the national resident matching program. Acad Med 2016;91(11):1546–53.
[8] Stephenson-Famy A, Houmard BS, Oberoi S, Manyak A, Chiang S, Kim S. Use of the Interview in Resident Candidate Selection: A Review of the Literature. J Grad Med Educ 2015;7(4):539–48.
[9] Koo TK, Li MYA. Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropractic Med 2016;15(2):155–63.
[10] Zimmerman DW. A note on preliminary tests of equality of variances. Br J Math Stat Psychol 2004;57(1):173–81.
[11] Lubelski D, Xiao R, Mukherjee D, Ashley WW, Witham T, Brem H, et al. Improving medical student recruitment to neurosurgery. J Neurosurg JNS 2020;133(3):848–54.
[12] Lynch G, Nieto K, Puthenveettil S, Reyes M, Jureller M, Huang JH, et al. Attrition rates in neurosurgery residency: analysis of 1361 consecutive residents matched from 1990 to 1999. J Neurosurg JNS 2015;122(2):240–9.
[13] Lubelski D, Healy AT, Friedman A, Ferraris D, Benzel EC, Schlenk R. Correlation of personality assessments with standard selection criteria for neurosurgical residency applicants. J Neurosurg JNS 2016;125(4):986–94.
[14] Zuckerman SL, Limoges N, Yengo-Kahn AM, Graffeo CS, Chambless LB, Chitale R, et al. The neurosurgery residency interview: assessing applicant perspectives on question content, utility, and stress. J Neurosurg JNS 2021;134(6):1974–82.
[15] Peeters MJ, Serres ML, Gundrum TE. Improving reliability of a residency interview process. Am J Pharm Educ 2013;77(8):168.
[16] Blouin D. Reliability of a structured interview for admission to an emergency medicine residency program. Teach Learn Med 2010;22(4):246–50.
[17] Bandiera G, Regehr G. Reliability of a structured interview scoring instrument for a Canadian postgraduate emergency medicine training program. Acad Emerg Med 2004;11(1):27–32.
[18] Blouin D, Day AG, Pavlov A. Comparative reliability of structured versus unstructured interviews in the admission process of a residency program. J Grad Med Edu 2011;3(4):517–23.
[19] Prager JD, Myer CM, Hayes KM, Myer CM, Pensak ML. 3rd, and Pensak ML. Improving methods of resident selection. Laryngoscope 2010;120(12):2391–8.
[20] Cohen Y. Estimating the Intra-Rater Reliability of Essay Raters. Front Educ 2017;2.
[21] Altman DG, Bland JM. Statistics notes: the normal distribution. BMJ 1995;310(6975):298.
[22] Krithikadatta J. Normal distribution. J Conservative Dentistry : JCD 2014;17(1):96–7.
[23] Limoges N, Zuckerman SL, Chambless LB, Benzil DL, Cruz A, Borden JH, et al. Neurosurgery Resident Interviews: The Prevalence and Impact of Inappropriate and Potentially Illegal Questions. Neurosurgery 2021;89(1):53–9.
[24] Patrick LE, Altmaier EM, Kuperman S, Ugolini KA. structured interview for medical school admission, Phase 1: initial procedures and results. Acad Med 2001;76(1):66–71.
[25] Altmaier EM, Smith WL, O??halloran CM, Franken EA. Jr.The predictive utility of behavior-based interviewing compared with traditional interviewing in the selection of radiology residents. Invest Radiol 1992;27(5):385–9.
[26] Campion MA, Pursell ED, Brown BK. Structured interviewing: Raising the psychometric properties of the employment interview. Pers Psychol 1988;41(1):25–42.
[27] Campion MA, Palmer DK, Campion JEA. review of structure in the selection interview. Pers Psychol 1997;50(3):655–702.
[28] Vadi MG, Malkin MR, Lenart J, Stier GR, Gatling JW, Applegate RL. Comparison of web-based and face-to-face interviews for application to an anesthesiology training program: a pilot study. Int J Med Educ 2016;7:102–8.
[29] Shah SK, Arora S, Skipper B, Kalishman S, Timm TC, Smith AY. Randomized evaluation of a web based interview process for urology resident selection. J Urol 2012;187(4):1380–4.
[30] Daram SR, Wu R, Tang SJ. Interview from anywhere: feasibility and utility of web-based videoconference interviews in the gastroenterology fellowship selection process. Am J Gastroenterol 2014;109(2):155–9.
[31] Asaad M, Rajesh A, Kambhampati PV, Rohrich RJ, Maricevich R. Virtual Interviews During COVID-19: The New Norm for Residency Applicants. Ann Plast Surg 2021;86(4):367–70.
[32] Results of the 2018 NRMP program director survey. NRMP. 2018. Available at: https://www.nrmp.org/wp-content/uploads/2018/07/NRMP-2018-Program-Director-Survey-for-WWW.pdf. Accessed May 24, 2020.
[33] McGaghie WC, Cohen ER, Wayne DB. Are United States Medical Licensing Exam Step 1 and 2 Scores Valid Measures for Postgraduate Medical Residency Selection Decisions? Acad Med 2011;86(1):48–52.