Clinical Ophthalmology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Comparison of Machine and Human Expert Evaluation of Capsulorrhexis Creation Performance Through Analysis of Surgical Video Recordings

Anvesh Annadanam[1], Ethan Kahana[1], Chris Andrews[1], Alexa R Thibodeau[1], Shahzad I Mian[1], Bradford L Tannen[1], Nambi Nallasamy[1,2]

[1]Department of Ophthalmology and Visual Sciences, University of Michigan, Ann Arbor, MI, USA; [2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

Correspondence: Nambi Nallasamy, Department of Ophthalmology and Visual Sciences, University of Michigan, 1000 Wall Street, Ann Arbor, MI, 48105, USA, Tel +1 734-763-8122, Email nnallasa@med.umich.edu

Purpose: Achieving competency in cataract surgery is an essential component of ophthalmology residency training. Video-based analysis of surgery can change training through its objective, reliable, and timely assessment of resident performance.

Methods: Using the Image Labeler application in MATLAB, the capsulorrhexis step of 208 surgical videos, recorded at the University of Michigan, was annotated for subjective and objective analysis. Two expert surgeons graded the creation of the capsulorrhexis based on the International Council of Ophthalmology's Ophthalmology Surgical Competency Assessment Rubric: Phacoemulsification (ICO-OSCAR:phaco) rating scale and a custom rubric (eccentricity, roundness, size, centration) that focuses on the objective aspects of this step. The annotated rhexis frames were run through an automated analysis to obtain objective scores for these components. The subjective scores were compared using both intra and inter-rater analyses to assess the consistency of a human-graded scale. The subjective and objective scores were compared using intraclass correlation methods to determine relative agreement.

Results: All rhexes were graded as 4/5 or 5/5 by both raters for both items 4 and 5 of the ICO-OSCAR:phaco rating scale. Only roundness scores were statistically different between the subjective graders (mean difference = −0.149, p-value = 0.0023). Subjective scores were highly correlated for all components (>0.6). Correlations between objective and subjective scores were low (0.09 to 0.39).

Conclusion: Video-based analysis of cataract surgery presents significant opportunities, including the ability to asynchronously evaluate performance and provide longitudinal assessment. Subjective scoring between two raters was moderately correlated for each component.

Keywords: capsulorrhexis, artificial intelligence, surgical training, cataract surgery

Introduction

Cataract surgery is the most commonly performed ophthalmic surgery and one of the most widely performed operations in the world.[1] Achieving competency in cataract surgery is an essential component of ophthalmology residency training. The assessment of surgical proficiency in a rigorous, objective manner, however, remains an elusive goal. Traditional approaches are fraught with limitations, including subjectivity, lack of multiple raters, and lack of longitudinal observation.[2]

Video-based analysis of surgical performance has the ability to fundamentally change surgical training and is beginning to be used across several surgical disciplines.[3] For cataract surgery in particular, it presents significant opportunities, including the ability to asynchronously evaluate performance, use multiple raters, and provide longitudinal assessment. However, manual editing and subsequent review of videos by experts can be inefficient, cumbersome, and often involves a small sample size.

Received: 15 October 2023
Accepted: 11 March 2024
Published: 27 March 2024

The application of machine learning and artificial intelligence to cataract surgical videos presents a unique opportunity to automate video-based competency analysis, generate continuous data from each surgical step, and help develop robust assessment tools that correlate with successful high-quality patient outcomes. Recent research has demonstrated the utility of machine learning and deep learning algorithms for automated identification of the steps of cataract surgery.[4] Additionally, computer vision-based motion analysis has been shown to provide robust measurement of instrument movement and discriminate among different surgeon skill levels.[5–7]

In this study, we focused on the analysis of capsulorrhexis creation, a critical and challenging step in cataract surgery.[8] Currently, resident performance of this step may be evaluated using validated structured rating scales, such as the International Council of Ophthalmology's Ophthalmology Surgical Competency Assessment Rubric:Phacoemulsification (ICO-OSCAR: phaco).[9] We also developed a custom rubric that focuses on objective components of capsulorrhexis creation, including centration, size, eccentricity, and roundness, which are all measurable by a computer.

We aimed to compare expert surgeon grading of videos of capsulorrhexis creation based on ICO-OSCAR:phaco to our custom rubric. We also evaluated how expert surgeon grading based on this custom rubric compares to computer-based automated measurements for each of the four components of the rubric.

Methods

Video Collection

The BigCat database comprises 208 cataract surgical videos collected at a tertiary care academic center between 2020 and 2021.[7] Approval from the University of Michigan Institutional Review Board (IRB) was obtained (HUM00160950), and it was determined by the IRB that informed consent was not required due to the study's retrospective nature and the anonymized data utilized in the study. The study was performed in accordance with the tenets of the Declaration of Helsinki. Patients undergoing cataract surgery were pre-medicated with topical tropicamide 1% and phenylephrine 2.5% for preoperative mydriasis and underwent surgery with intravenous sedation as per the discretion of the anesthesia provider. Femtosecond laser surgeries were excluded from analysis due to the focus on manual performance of the capsulorrhexis. Cases with inadequate visualization of instruments were also excluded. Eyes that had undergone prior laser keratorefractive surgeries were not specifically excluded. Extraneous frames were trimmed from before and after the core surgery steps, which included from just before paracentesis creation to just after hydration of wounds. The source resolution was 1920×1080 pixels with a frame rate of 30 frames per second.

Capsulorrhexis Tracing

For each surgery in the database, a trained evaluator (EK) identified a frame visualizing the completed capsulorrhexis immediately prior to hydrodissection without any other instruments to obscure the surgical view. Adequate visualization of the limbus and pupil was confirmed. A frame meeting these criteria in which the first Purkinje image was closest to the center of the pupil was selected for each video. A total of 208 frames, one for each surgery, were selected. The Image Labeler application in MATLAB (The MathWorks, Natick, MA, USA) was then used to trace the anterior capsular opening in each frame manually using a Wacom One drawing pad (DTC133W0A, libwacom 1.3, Wacom Co., Ltd.; Saitama, Japan) (Figure 1). All pixels on and within the traced contour were classified as belonging to the capsulotomy, while all other pixels were classified as not belonging to the capsulotomy. In this manner, 208 ground truth binary masks for anterior capsulotomies were created, one for each surgery included.

Video Grading

We developed a custom four-category grading rubric for capsulorrhexis creation that was used subjectively by the expert raters as well as objectively by our computer analysis algorithm. Graded between 0 and 10, the four components were defined as follows: centration – the extent to which the capsulorrhexis is centered with regard to the intraoperative dilated pupil; size – the closeness to which the size of the capsulorrhexis is ideal for the size of the optic (approximately 5 mm in this dataset); roundness – the extent to which the circumference of the capsulorrhexis follows the path of a circle; and

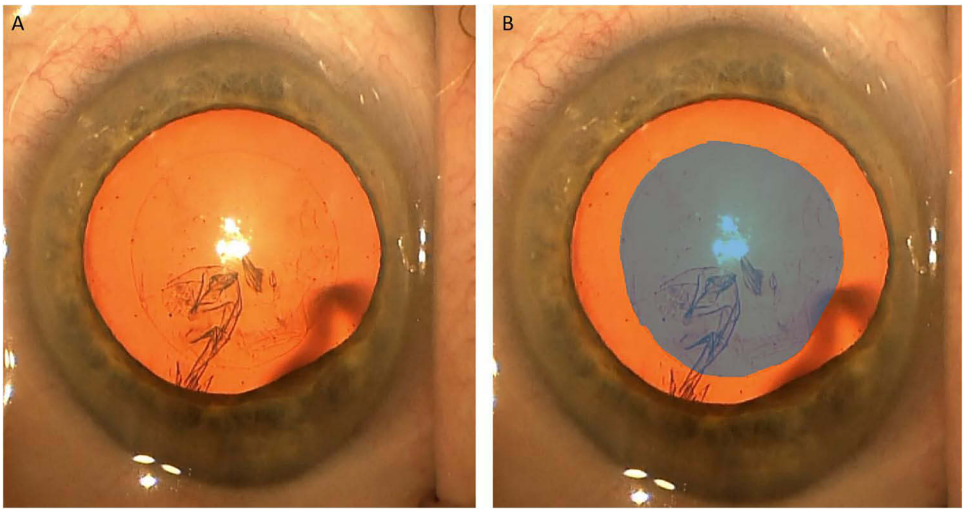

Figure 1 Capsulorrhexis frame before tracing (**A**) and after tracing (**B**).

eccentricity – the extent to which the major and minor axes of the completed capsulorrhexis matched. Eccentricity score is high when eccentricity is low.

Subjective Assessment

Two expert cataract surgeons (BT and NN) graded the capsulorrhexis creation step of the surgical videos using items 4 (capsulorrhexis: commencement of flap and follow-through) and 5 (capsulorrhexis: formation and circular completion) of the ICO-OSCAR:phaco scale[9] as well as the custom rubric. They each re-graded 30 randomly selected frames to assess intrarater agreement using the custom rubric only. They were masked to the identity of the original surgeon performing the cataract surgery as well as to each other's scoring.

Objective Assessment

The computer-based objective grading system used the annotated capsulorrhexis tracings that passed human inspection to generate automated measurements on a continuous scale of 0 to 1 for each of the components of the custom rubric. The centration score was computed based on the difference between the centroid of the limbus and the centroid of the capsulotomy. The size score was computed based on the deviation of the equivalent diameter of the capsulotomy region from the optimal diameter (considered to be 5.0 mm for this study involving intraocular lenses of 6.0 mm diameter). Size was computed relative to the annotated horizontal corneal diameter (WTW), which was assumed to be equivalent to the mean across all patients at our center.[10] Roundness was computed as (4*pi*area)/(perimeter^2). Eccentricity was computed as the ratio of the distance between the foci of the ellipse fitting the capsulotomy region and its major axis length. Since the objective measurements were on a different scale than the subjective ratings, only associations between the objective and subjective ratings were studied.

Statistical Analysis

Although the subjective scoring scale was 0 to 10, the majority of scores were 6 to 10. For Kappa agreement analysis, any grade lower than 6 was recorded as 6. Weighted Cohen's Kappa and intraclass correlation coefficients were calculated to measure agreement between the subjective graders. The polychoric correlation coefficient was calculated as a measure of association between the subjective graders. The calculations were repeated for intrarater analysis. Only correlational analysis (via polyserial correlation coefficient) was used to compare objective and subjective scores, as they are measured on different scales. The restriction of analyses to associations rather than a form of exact agreement ensures validity of the correlation estimates in the context of our study. Each of these statistics was calculated for each custom variable.

Table 1 Interrater Analysis (n=208)

	Centration	Size	Roundness	Eccentricity
Mean difference	0.072	0.024	−0.15	−0.087
95% CI	−0.028 to 0.17	−0.082 to 0.13	−0.24 to −0.054	−0.19 to 0.013
p-value	0.16	0.66	0.0023 *	0.089
Correlation	0.73	0.63	0.76	0.76
95% CI	0.66 to 0.80	0.53 to 0.72	0.69 to 0.83	0.69 to 0.83

Note: *p<0.05.
Abbreviation: CI, confidence interval.

Results

The only significant difference in means (Table 1) was found in the roundness scores (NN scored higher than BT). Grader scores were highly correlated (>0.6, Table 1 and Figure 2).

Graders' mean of regraded scores was higher than original mean scores for several variables (Table 2). For BT, the mean difference in scores was significantly different for centration and roundness. Correlation was less than 0.5 for all categories except eccentricity. For NN, the mean differences in scores were significantly different for centration, roundness, and eccentricity. Correlation was less than 0.5 for all categories except roundness.

Correlation coefficients between the average of the two surgeon scores and machine scores for all four categories are shown in Table 3. Correlations between objective and subjective scores were low (0.09 to 0.39, Table 3 and Figure 3).

For item 4 (capsulorrhexis: commencement of flap and follow-through) on the ICO-OSCAR:phaco rubric, BT gave a score of 5/5 to 206 videos (99%) and 4/5 to 2 videos (1%). NN gave a score of 5/5 to 203 videos (97.6%) and 4/5 to 5 videos (2.4%). For item 5 (capsulorrhexis: formation and circular completion), both graders gave the same scores for all videos (207 videos (99.5%) scored 5/5, 1 video (0.5%) scored 4/5).

Discussion

In this study, we sought to determine whether using an objective computer-based approach to the grading of the capsulorrhexis step of cataract surgery could improve the consistency and accuracy of the assessment of ophthalmic resident surgeons. We created a custom rubric to compare scores from two expert cataract surgeons to the objective scores determined by a computer program. The capsulorrhexis step was chosen to analyze for its importance in cataract surgery and generally well accepted technique and shape.[11]

The surgeon scores correlated well for all custom rubric items, but the lowest correlation was for size, as explained below. Conversely, intrarater correlation for both graders was poor, which has been shown in other fields as well.[12,13] Correlation between surgeon and machine assessment was also poor. One possible reason for this discrepancy is the difference in interpretation of rubric metrics by the subjective graders compared to what was programmed for the objective grader. Overall, these results suggest that interpretation of surgeon skill can be highly variable depending on the observer and circumstance, even if performance is the same. An objective measurement would not be affected by the same limitations.

There was especially low correlation between objective and subjective size scores. A likely explanation is the multidimensionality by which humans judge size relative to an object's surroundings compared to a computer program. Surgeons may use varying levels of zoom and focus which can distort perception of the recording. Objective assessment measures the rhexis in comparison to the limbus, annotated in each frame, which likely improved consistency of the scoring. In future iterations, use of the individual biometric WTW for each eye undergoing surgery rather than a population mean would help further improve the accuracy of the objective size calculation. A recent study has explored new ways to record surgical videos that may improve future analysis.[14]

Another study aimed to use machine learning to provide objective analysis of an entire cataract surgery through instrument detection without prior manual annotation.[15] While successful in detecting tools, their model struggled with accurately and consistently grading surgeries. It performed best when the surgeries were

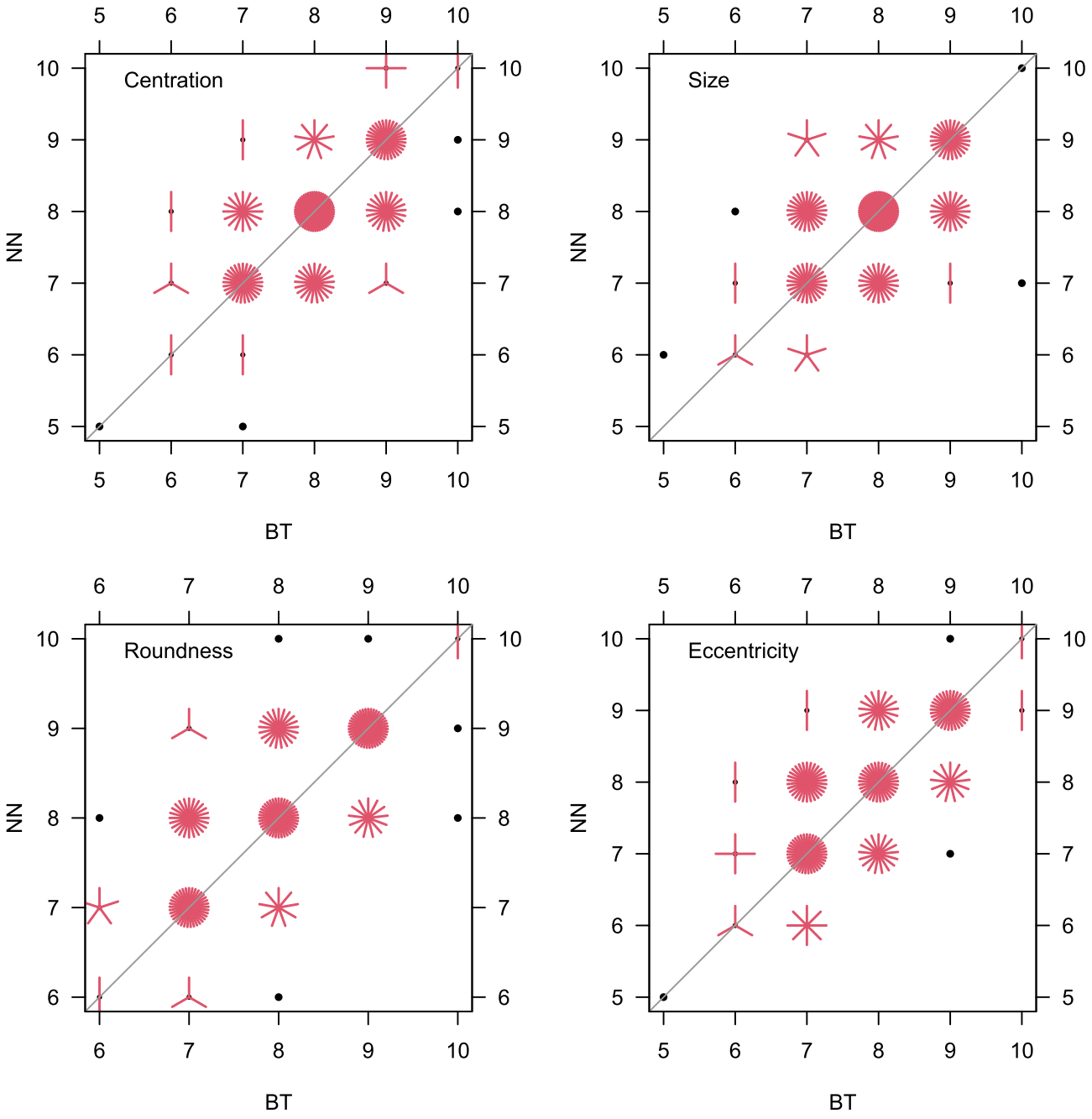

Figure 2 Sunflower plots showing the bivariate distributions of scores given by subjective graders for each measure. Each petal represents a single observation.

completed by a single surgeon and when given a single small clip of videos (framewise). This suggests the importance of focusing on a single step of surgery, such as the capsulorrhexis in our study, as well as the utility of pre-annotation, whether manual or automated, in order to properly grade a surgical video objectively.

In this study, we have shown that a more detailed custom rubric can increase the inconsistencies between human graders but allows for more precise points of feedback for an objective grader. It has been previously shown that the reliability of the scoring scale depends on the wording of the question.[16]

There are a few limitations to our study. Although statistical significance was seen for various comparisons performed in our analysis, a preliminary sample size estimate would be of value in future studies. Manual tracing of the completed capsulorrhexis using a drawing pad could introduce human-based variability. Future studies will

Table 2 Intrarater Analysis (n=208)

	Centration	Size	Roundness	Eccentricity
Mean difference				
BT 95% CI p-value	0.40 0.066 to 0.73 0.021*	0.33 −0.025 to 0.69 0.067	0.53 0.14 to 0.92 0.0089*	0.30 −0.042 to 0.64 0.083
NN 95% CI p-value	0.60 0.19 to 1.010 0.0057*	0.23 −0.087 to 0.55 0.067	0.53 0.24 to 0.82 0.00080*	0.50 0.087 to 0.91 0.019*
Correlation				
BT 95% CI	0.49 0.12 to 0.86	0.47 0.088 to 0.85	0.34 −0.11 to 0.78	0.61 0.29 to 0.94
NN 95% CI	−0.030 −0.49 to 0.43	0.44 0.046 to 0.84	0.75 0.52 to 0.98	0.28 −0.14 to 0.70

Note: *p<0.05.
Abbreviation: CI, confidence interval.

Table 3 Objective vs Subjective Analysis (n=208)

	Centration	Size	Roundness	Eccentricity
Correlation 95% CI	0.11 −0.27 to 0.25	0.090 −0.047 to 0.23	0.39 0.28 to 0.50	0.24 0.12 to 0.37

Abbreviation: CI, confidence interval.

include automatic tracing using a computer-based semantic segmentation system. Our custom rubric is not otherwise validated, but was used only for the purpose of comparing subjective and objective scores in this study. Additionally, both the objective and subjective scales did not take patient head or eye tilt into account during the grading process, which may marginally affect rubric items including eccentricity and centration. Given the available data, the present study did not evaluate the effects of viscoelastic fill, anterior chamber depth, or corneal curvature on the appearance of the capsulotomy. These factors may be accounted for in the future with additional biometric data and video processing methods. Additionally, the determination of the optimal centration of the capsulotomy is a complex challenge, and the selection of the optimal center for the capsulotomy could be adjusted in future work. The time taken for each step of surgery, as well as total surgical time, can be important to determine the proficiency of a training surgeon but were not assessed in this study, as only still images were analyzed.

Future studies are needed to correlate surgeon performance on objective scales with patient outcomes. In addition, the remaining eighteen components of the ICO-OSCAR:phaco rubric, which focus on steps other than capsulorrhexis creation, could be assessed with full surgical videos.

This study aims to provide new findings to the area of objective analysis of surgical performance. We have shown that the subjective ICO-OSCAR:phaco scale is limited in real-world assessment of surgical performance and that a machine learning-based objective scale that utilizes frame annotation can be used to assess cataract surgeries more accurately than other methods. The poor correlation between the objective and subjective grading suggests that interpretation of surgeon skill can be highly variable depending on the observer and circumstance even if performance is the same. These findings indicate that machine-computed objective ratings can improve the analysis of surgical performance.

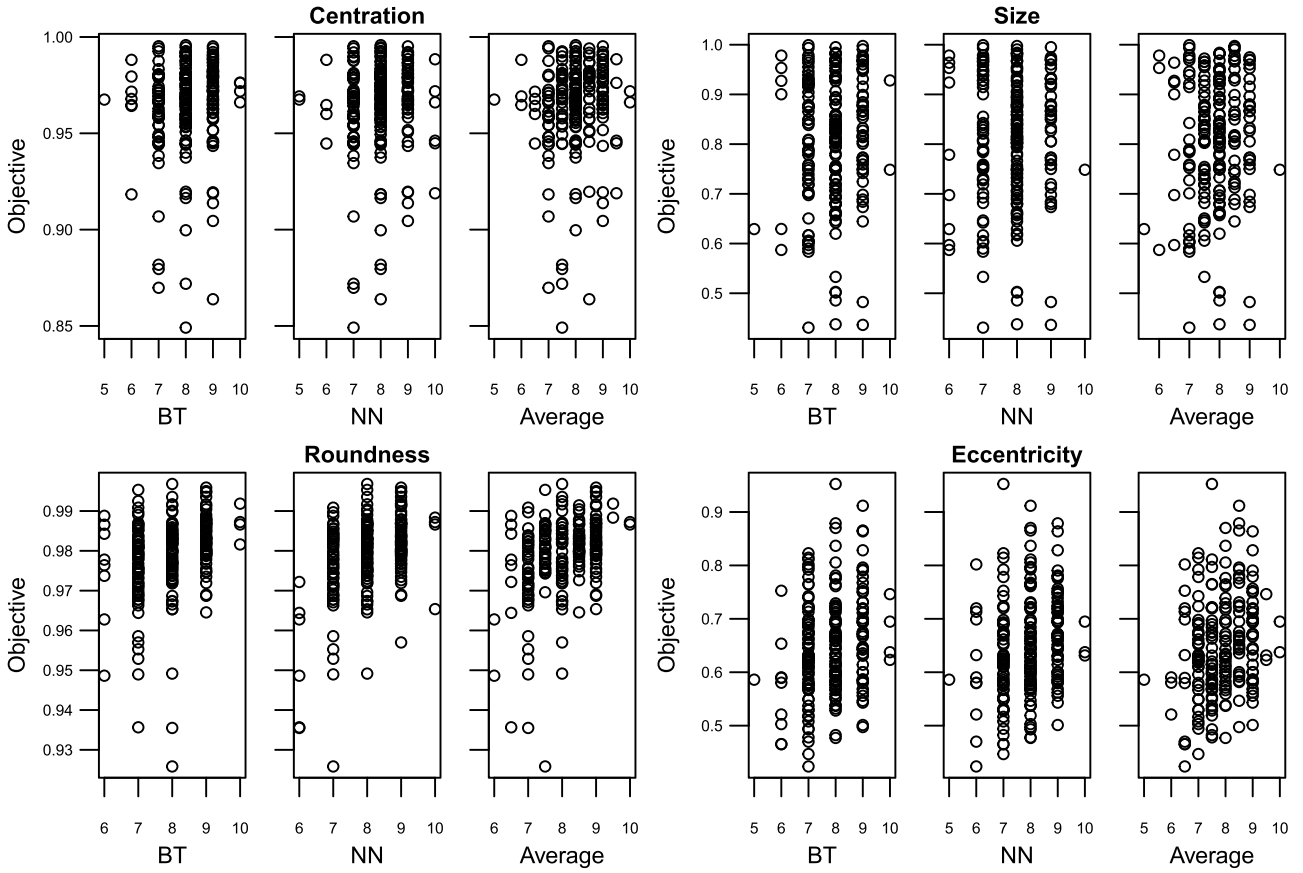

Figure 3 Scatterplots showing bivariate distributions of objective score with each grader's subjective score and the average of the subjective scores. Each circle represents a single observation.

Funding

This work was supported in part by the Graduate Medical Education Innovations Fund (NN, BT), The Doctors Company Foundation (NN, BT), NIH K12EY022299 (NN), and Fogarty/NIH D43TW012027 (NN).

Disclosure

Dr Nambi Nallasamy reports a patent 63/445,053 pending. The authors have no other conflicts of interest in this work.

References

1. Thoughts on Cataract Surgery: 2015. Available from: https://www.reviewofophthalmology.com/article/thoughts-on–cataract-surgery-2015. Accessed November 22, 2021.
2. Lee AG, Oetting T, Beaver HA, Carter K. The ACGME Outcome Project in ophthalmology: practical recommendations for overcoming the barriers to local implementation of the national mandate. *Surv Ophthalmol*. 2009;54(4):507–517. doi:10.1016/J.SURVOPHTHAL.2009.04.004
3. Vedula SS, Ishii M, Hager GD. Objective assessment of surgical technical skill and competency in the operating room. *Annu Rev Biomed Eng*. 2017;19. doi:10.1146/annurev-bioeng-071516-044435
4. Yu F, Silva Croso G, Kim TS, et al. Assessment of automated identification of phases in videos of cataract surgery using machine learning and deep learning techniques. *JAMA Network Open*. 2019;2(4):e191860. doi:10.1001/JAMANETWORKOPEN.2019.1860
5. Bouget D, Allan M, Stoyanov D, Jannin P. Vision-based and marker-less surgical tool detection and tracking: a review of the literature. *Med Image Anal*. 2017;35:633–654. doi:10.1016/J.MEDIA.2016.09.003
6. Smith P, Tang L, Balntas V, et al. "PhacoTracking": an evolving paradigm in ophthalmic surgical training. *JAMA Ophthalmol*. 2013;131(5):659–661. doi:10.1001/JAMAOPHTHALMOL.2013.28
7. Matton N, Qalieh A, Zhang Y, et al. Analysis of cataract surgery instrument identification performance of convolutional and recurrent neural network ensembles leveraging BigCat. *Transl Vis Sci Technol*. 2022;11(4):1. doi:10.1167/tvst.11.4.1
8. Al-Jindan M, Almarshood A, Yassin SA, Alarfaj K, Al Mahmood A, Sulaimani NM. Assessment of learning curve in phacoemulsification surgery among the eastern province ophthalmology program residents. *Clin Ophthalmol*. 2020;14:113–118. doi:10.2147/OPTH.S241250

9. Golnik KC, Beaver H, Gauba V, et al. Cataract surgical skill assessment. *Ophthalmology*. 2011;118(2). doi:10.1016/J.OPHTHA.2010.09.023
10. Li T, Stein J, Nallasamy N. Evaluation of the Nallasamy formula: a stacking ensemble machine learning method for refraction prediction in cataract surgery. *Br J Ophthalmol*. 2022;107(8):1066–1071. doi:10.1136/bjophthalmol-2021-320599
11. Sharma B, Abell R, Arora T, Antony T, Vajpayee R. Techniques of anterior capsulotomy in cataract surgery. *Indian J Ophthalmol*. 2019;67(4):450. doi:10.4103/ijo.IJO_1728_18
12. Farzin B, Fahed R, Guilbert F, et al. Early CT changes in patients admitted for thrombectomy. *Neurology*. 2016;87(3):249–256. doi:10.1212/WNL.0000000000002860
13. Massey SL, Shou H, Clancy R, et al. Interrater and intrarater agreement in neonatal electroencephalogram background scoring. *J Clin Neurophysiol*. 2019;36(1):1–8. doi:10.1097/WNP.0000000000000534
14. Thia BC, Wong NJ, Sheth SJ. Video recording in ophthalmic surgery. *Surv Ophthalmol*. 2019;64(4):570–578. doi:10.1016/j.survophthal.2019.01.005
15. Ruzicki J, Holden M, Cheon S, Ungi T, Egan R, Law C. Use of machine learning to assess cataract surgery skill level with tool detection. *Ophthalmol Sci*. 2023;3(1):100235. doi:10.1016/j.xops.2022.100235
16. Smith RJ, McCannel CA, Gordon LK, et al. Evaluating teaching methods of cataract surgery: validation of an evaluation tool for assessing surgical technique of capsulorhexis. *J Cataract Refract Surg*. 2012;38(5):799–806. doi:10.1016/j.jcrs.2011.11.046