• Intelligent Ophthalmology •

Interobserver variability in LOCS III cataract grading across experience levels: implications for AI training label quality

Zhen-Yu Wang, Xi-Jin Wu, Ran Zhang, Yu-Xuan Jia, En-Jie Li, Xue-Fei Ding, Pei-Lin Yue, Jing Tao, Jin-Da Wang, Xu-Dong Song

Beijing Tongren Eye Center, Beijing Tongren Hospital, Beijing Ophthalmology and Visual Science Key Lab, Capital Medical University, Beijing 100730, China

Co-first Authors: Zhen-Yu Wang and Xi-Jin Wu

Correspondence to: Xu-Dong Song and Jing Tao. Beijing Tongren Eye Center, Beijing Tongren Hospital, Beijing Ophthalmology and Visual Science Key Lab, Capital Medical University, No.1 Dongjiaominxiang Street, Dongcheng District, Beijing 100730, China. drxdsong@sina.com; taojing@mail.ccmu.edu.cn

Received: 2026-04-17 Accepted: 2026-07-27

Abstract

● **AIM:** To evaluate the variability in cataract grading based on the Lens Opacities Classification System III (LOCS III) among ophthalmologists with different experience levels, and identify which type of annotated data may serve as high-quality labels for artificial intelligence (AI) model development.

● **METHODS:** Ophthalmologists were grouped by clinical experience (resident, senior, expert) and independently graded nuclear, cortical, and posterior subcapsular cataract (PSC) according to LOCS III. Consistency was analyzed at multiple levels: label consistency (presence/absence of opacity), grading consistency (strict and tolerant), and cross-subtype consistency using presence scores (agreement on cataract detection) and pairwise exact-match proportions (agreement on identical LOCS III grade). Intragroup reliability was assessed using intraclass correlation coefficients (ICCs), while intergroup and cross-subtype comparisons were conducted using Friedman and Kruskal-Wallis tests with Bonferroni-corrected post-hoc analyses.

● **RESULTS:** A total of 138 participants with a mean age of 66.64±10.09y (77 females) diagnosed with cataracts were included in the study, yielding 644 anterior segment slit-lamp photographs under mydriatic conditions. Each image was annotated independently by six ophthalmologists, including two residents, two attending-level seniors, and two experts. Experts achieved the highest intragroup agreement (ICC=0.780), while senior ophthalmologists and residents showed comparable moderate agreement. Intergroup comparison showed significantly higher label and grading consistency between senior-expert pairs than between resident-senior or resident-expert pairs, especially for cortical cataracts and PSC (P<0.05). Among subtypes, nuclear cataracts achieved the highest overall consistency, while cortical cataracts and PSC showed lower agreement, likely due to diffuse or obscured opacities and image quality variability (P<0.001).

● **CONCLUSION:** Consistency in LOCS III-based cataract grading varies with clinical experience and cataract subtype. Expert and senior ophthalmologists provide more stable annotations, and nuclear cataract images demonstrate superior interobserver agreement. These findings emphasize the value of expert consensus and subtype-specific data curation for improving AI-based cataract grading models.

● **KEYWORDS:** cataract grading; LOCS III; interobserver consistency; ophthalmologist experience level; artificial intelligence

DOI:10.18240/ijo.2026.12.01

Citation: Wang ZY, Wu XJ, Zhang R, Jia YX, Li EJ, Ding XF, Yue PL, Tao J, Wang JD, Song XD. Interobserver variability in LOCS III cataract grading across experience levels: implications for AI training label quality. *Int J Ophthalmol* 2026;19(12):2313-2320

INTRODUCTION

Cataract is the leading cause of blindness and visual impairment globally, especially among the aging population[1-5]. In 2020, cataract accounted for an estimated 15 million cases of blindness and an additional 79 million cases of moderate to severe visual impairment among adults over 50 years old[6]. Accurately assessing lens opacity is crucial for clinical decision-making, including the selection of surgical timing and approaches. To address the need for standardized grading, the Lens Opacities Classification System III (LOCS

III) has become the most widely used system for assessment of nuclear opalescence, nuclear color, cortical, and posterior subcapsular cataract (PSC) using slit-lamp images[7]. However, despite the widespread use and standardization of LOCS III, the grading of cataract is still subjective and highly dependent on the evaluator's clinical experience and familiarity with the classification criteria. Especially, this subjective variability in interpretation across ophthalmologists with different experience, such as residents, attending physicians, and senior consultants, may lead to significant inconsistency in clinical assessments and research datasets. This issue is particularly critical in the training of artificial intelligence (AI) models in the context of the rapid development of AI-assisted diagnosis.

AI is a technology based on deep learning and machine learning, which has shown potential practicality in the field of image recognition in recent years[8-15]. Based on this technology, research on AI-assisted cataract classification support systems has gradually emerged[16-22]. The performance of deep learning models, particularly those based on supervised learning, highly depends on the quality and consistency of labeled data. If there is subjective bias or consistency difference in manual annotation due to different experience levels across annotators, the resulting training noise may significantly affect model accuracy, reliability, and generalizability in clinical applications. According to several studies, the diagnostic accuracy of AI models trained and published in the past 5y is roughly between 80% and 99%[16-19,23-24]. Therefore, it is important and urgent to critically evaluate the reliability of annotations with different experience, and to determine which type of annotated data is more suitable as the basic label for AI training.

In this study, we conducted a prospective analysis of interobserver consistency differences in LOCS III-based cataract grading among ophthalmologists of varying experience levels, including resident, senior, and expert groups. Through systemic analysis, we aim to quantify the degree of consistency and stability in LOCS III score dimensions among different ophthalmologists, and to evaluate the consistency of different subtypes of cataract ratings by different doctors to explore suitable and reliable annotation source for developing AI models.

PARTICIPANTS AND METHODS

Ethical Approval All the procedures provided in this study adhered to the principles of the Declaration of Helsinki and were approved by Ethics Committee of Beijing Tongren Hospital, China (TREC2024-KY058). Written informed consent was obtained from all participants prior to inclusion in the study.

Participants The inclusion criteria were: 1) aged 18y or older; 2) diagnosis of age-related cataract and suitable for phacoemulsification with intraocular lens implantation; 3) willing to participate in the study and able to provide written informed consent. The exclusion criteria were: 1) inability to cooperate with medical history collection or complete ophthalmic examinations; 2) presence of severe systemic diseases or serious ocular pathologies.

Data Collection All the slit-lamp photographs of anterior segment of cataract patients were taken between April 2024 and November 2024 at Beijing Tongren Hospital. All the patients were in mydriatic conditions. For slit-beam and diffuse-illuminated images, a 30° angle was maintained between the illumination arm and the viewing arm. Slit-beam mode was applied to evaluate nuclear opalescence (NO) and nuclear color (NC) while diffuse illumination mode was used to assess cortical opacities. For retro-illuminated photographs which emphasized posterior subcapsular opacities, a frontal viewing angle was adopted with coaxial illumination aligned with the optical axis.

Then, all slit-lamp photographs were systematically organized, numbered, and stored in an electronic database. A double-data-entry process was employed to ensure the completeness and accuracy of the recorded information. Data of patients with missing information and images with low quality were excluded.

Cataract Labelling All slit-lamp photographs were systematically organized and uploaded to a standardized annotation platform. According to the LOCS III, the annotators were instructed to independently classify and grade each image in terms of nuclear cataract, cortical cataract, and PSC. Each image was annotated independently by six ophthalmologists with different levels of clinical experience, including two residents, two attending-level seniors, and two experts. For a given patient and cataract type, a label of "present" was assigned if at least one of the two raters within a group assigned a non-zero grade, while "absent" indicated that neither did. If both raters provided a grade, the group median was used; if only one did, that value was retained; otherwise, the entry was treated as missing. Annotators worked independently and were blinded to each other's results. Images were presented in a randomized order. All annotations were subsequently used for consistency analysis and label quality assessment.

Intragroup and Intergroup Consistency Evaluation To evaluate the intragroup and intergroup consistency among the three experience levels within each cataract subtype, we employed a two-level evaluation criterion including label consistency and grading consistency. Label consistency refers to whether two annotators have consistent judgments on the presence of a certain cataract type (such as nuclear, cortical, or PSC), calculated as the number of patients with the same

label divided by the total number of annotated cases. For cases with matched labels, grading consistency was further assessed to determine the extent of agreement on cataract severity according to LOCS III scores. We used two calculation methods: 1) strict grading consistency was defined as the proportion of identical LOCS III grades (N_{match}) among cases where both groups provided grading (intersection-based, N_{both}), which could be exhibited as: strict grading consistency=N_{match}/N_{both}; 2) tolerant grading consistency was defined as the proportion of identical grades among all cases where at least one group provided a grade (union-based, N_{any}), which could be exhibited as: tolerant grading consistency=N_{match}/N_{any}.

For intragroup consistency, after calculation of label and grading consistency, intraclass correlation coefficient (ICC) was further used to quantify the reliability of cataract grading between two annotators within each experience group (expert, senior, and resident). Based on a two-way random-effects model with absolute agreement [ICC(2,1)], ICC values were calculated for overall cataract grading as well as for each cataract subtype. The ICC values were interpreted according to conventional thresholds: <0.5 (poor), 0.5–<0.75 (moderate), 0.75–<0.9 (good), and ≥0.9 (excellent) reliability. The 95% confidence intervals (CIs) were calculated for each ICC.

Cross-subtype Consistency Metrics To determine which cataract subtype exhibited the highest reliability across different raters and was therefore more suitable for AI-based annotation, we calculated two image-level metrics across all slit-lamp photographs: presence score and pairwise exact-match proportion. The presence score was used to quantify the overall detectability for each cataract subtype. For each image, the grading result of each of the six ophthalmologists was encoded as 1 if the corresponding cataract subtype (nuclear, cortical, or PSC) was considered present; if not present, it was encoded as 0. The scores of each ophthalmologist were added up to obtain the presence score of each subtype in each image, ranging from 0 to 6. Subsequently, statistical comparisons were conducted based on the distribution of cataract subtype presence scores. Furthermore, the pairwise exact-match proportion was calculated to evaluate agreement in detailed LOCS III grading for each image. For all 15 possible annotator pairs, a pair was considered an exact match when both raters provided a grade and the scores were identical. The final score for each image was the proportion of matched pairs among all 15 possible combinations. This metric and further statistical analysis reflect the grading consistency of each cataract subtype.

Statistical Analysis Pairwise label consistency and grading consistency (strict and tolerant) among the three experience levels were analyzed using the Friedman test. When Friedman test was significant, post-hoc pairwise comparisons were performed using the Wilcoxon signed-rank test with Bonferroni correction. Cross-subtype comparisons of mean consistency were conducted using the Kruskal-Wallis *H* test, followed by Dunn's post hoc test with Bonferroni correction when appropriate. Label and grading consistency, presence score, and pairwise exact-match proportion were reported as median with 95%CI. All statistical analyses were performed using the SciPy and scikit-posthocs packages in Python, and a two-tailed *P* value<0.05 was considered statistically significant. Adjusted *P* values<0.017 (α=0.05/3) were considered statistically significant for Bonferroni correction.

Table 1 Baseline clinical characteristics of the patients

Variable	Values
Total patients, *n*	138
Images, *n*	644
Age (y, mean±SD)	66.64±10.09
Female, *n* (%)	77 (55.80)
Male, *n* (%)	61 (44.20)
Photographic conditions, *n*	
Diffuse illumination images	149
Slit-beam illumination images	354
Retro-illumination images	141

RESULTS

Baseline Clinical Characteristics of the Patients A total of 138 patients diagnosed with cataracts were included in the study, yielding 644 anterior segment slit-lamp photographs under mydriatic conditions. The mean age of patients was 66.64±10.09y, with 55.80% female participants (*n*=77). Each eye was photographed in three illumination modes: diffuse illumination, slit-beam, and retro-illumination. The total number of images obtained for each mode was 149, 354, and 141, respectively. Baseline clinical characteristics of the patients were shown in Table 1.

Intragroup Annotation Consistency The intragroup comparison (Figure 1) demonstrated that annotation consistency varied considerably across experience levels. Experts consistently achieved the highest agreement across all cataract subtypes, while senior and resident raters showed lower and more variable agreement.

For nuclear cataract, the expert group showed perfect label consistency (1.000) and high grading consistency (0.746 for both metrics). Seniors followed with strong label agreement (0.949) but moderate grading consistency (~0.57), whereas residents showed the lowest reliability across all metrics.

Cortical cataract analysis revealed a similar trend: experts maintained superior grading consistency (strict: 0.574) compared to seniors and residents. Interestingly, seniors achieved the highest label consistency (0.918), suggesting strong detection skills despite lower grading precision than experts.

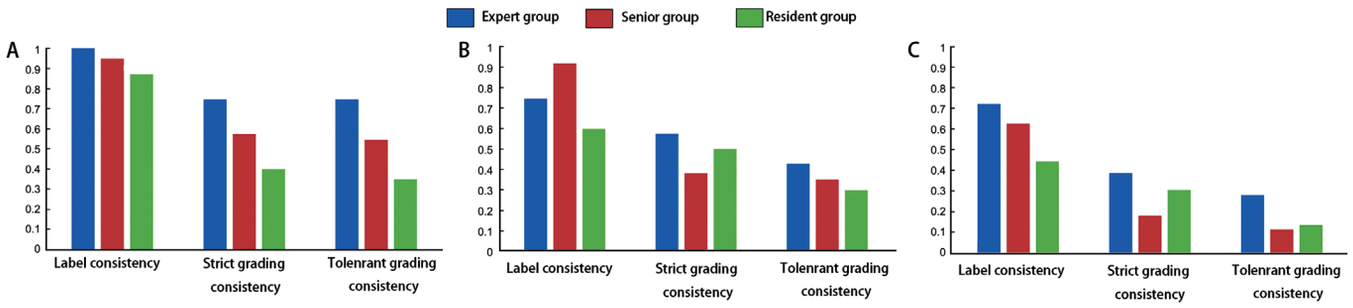


Figure 1 Intragroup label and grading consistency in LOCS III-based cataract grading among ophthalmologists of different experience levels
A: Intragroup consistency in nuclear cataract subtype; B: Intragroup consistency in cortical cataract subtype; C: Intragroup consistency in posterior subcapsular cataract subtype. LOCS III: Lens Opacities Classification System III.

Table 2 Comparison of intergroup consistency across experience levels within each cataract subtype

Cataract subtype	Consistency measure	Resident–senior	Resident–expert	Senior–expert	Friedman $\chi^2(2)$	*P*
Nuclear	Label consistency	0.971 (95%CI: 0.941–0.993)	0.985 (95%CI: 0.962–1.000)	0.985 (95%CI: 0.963–1.000)	4.000	0.135
	Strict grading consistency	0.432 (95%CI: 0.349–0.519)	0.333 (95%CI: 0.256–0.411)	0.440 (95%CI: 0.356–0.523)	4.682	0.096
	Tolerant grading consistency	0.419 (95%CI: 0.341–0.500)	0.328 (95%CI: 0.252–0.405)	0.434 (95%CI: 0.348–0.515)	4.682	0.096
Cortical	Label consistency	0.649 (95%CI: 0.564–0.733)	0.638 (95%CI: 0.554–0.722)	0.940 (95%CI: 0.896–0.978)	72.043	<0.001[a]
	Strict grading consistency	0.253 (95%CI: 0.163–0.345)	0.337 (95%CI: 0.235–0.440)	0.341 (95%CI: 0.262–0.422)	13.765	0.001[a]
	Tolerant grading consistency	0.164 (95%CI: 0.102–0.226)	0.215 (95%CI: 0.144–0.288)	0.321 (95%CI: 0.246–0.397)	13.765	0.001[a]
PSC	Label consistency	0.472 (95%CI: 0.373–0.569)	0.586 (95%CI: 0.479–0.691)	0.755 (95%CI: 0.675–0.832)	58.514	<0.001[a]
	Strict grading consistency	0.412 (95%CI: 0.283–0.550)	0.235 (95%CI: 0.122–0.360)	0.181 (95%CI: 0.102–0.267)	4.200	0.122
	Tolerant grading consistency	0.194 (95%CI: 0.125–0.269)	0.138 (95%CI: 0.070–0.215)	0.136 (95%CI: 0.076–0.204)	4.200	0.122

[a]*P*<0.05 indicates statistical significance. CI: Confidence interval; PSC: Posterior subcapsular cataract.

PSCs were proved as the most challenging subtype. While experts performed better than other groups (label consistency: 0.721), overall agreement was low. Seniors and residents exhibited marked declines in consistency, with resident tolerant grading dropping to 0.135, indicating significant difficulty in standardizing PSC assessment among less experienced clinicians.

The intragroup ICC values further confirmed these findings (Figure 2). The expert group exhibited the highest overall consistency (ICC=0.780, 95%CI: 0.736–0.817); the senior group (ICC=0.554, 95%CI: 0.443–0.642) and the resident group (ICC=0.596, 95%CI: 0.530–0.655) showed comparable moderate reliability. Subtype stratification showed that nuclear cataracts yielded the strongest agreement (expert ICC=0.906), while PSCs showed the weakest reliability, particularly among seniors (ICC=0.391).

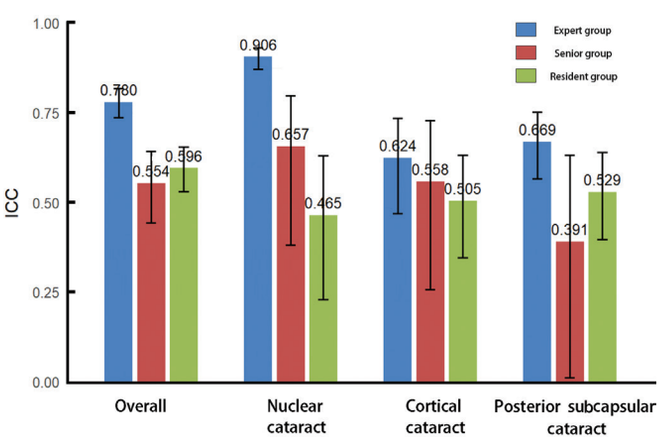


Figure 2 Intragroup ICC value of different groups ICC: Intraclass correlation coefficient.

Intergroup Annotation Consistency Comparisons of intergroup consistency were performed across different subtypes and the results were shown in Table 2.

For nuclear cataracts, all three experience levels demonstrated similarly high label consistency, with values exceeding 0.97 (resident–senior: 0.971; resident–expert: 0.985; senior–expert: 0.985). No significant difference was observed among the three groups [$\chi^2(2)$=4.000, *P*=0.135]. Regarding grading consistency, both strict and tolerant grading consistency showed moderate agreement (strict: 0.333–0.440; tolerant: 0.328–0.434) without significant group differences [$\chi^2(2)$=4.682, *P*=0.096].

In contrast, significant intergroup differences were observed in cortical cataracts. Label consistency was substantially higher for the senior–expert (0.940; 95%CI, 0.896–0.978) than for resident–senior (0.649; 95%CI, 0.564–0.733) or resident–expert (0.638; 95%CI, 0.554–0.722), and the Friedman test confirmed a highly significant difference among pairs [$\chi^2(2)$=72.043, *P*<0.001]. In grading metrics, senior–expert again attained the highest strict (0.341) and tolerant (0.321) consistencies, whereas resident–senior showed the lowest

Table 3 Comparison of average consistency across cataract subtypes

Parameters	Presence score[a]	Pairwise exact-match[b]
Nuclear	5.652 (95%CI: 5.479–5.825)	0.570 (95%CI: 0.532–0.608)
Cortical	4.464 (95%CI: 4.206–4.721)	0.439 (95%CI: 0.391–0.487)
PSC	2.877 (95%CI: 2.520–3.233)	0.270 (95%CI: 0.226–0.314)
Kruskal-Wallis test	$\chi^2(2)=149.51$	$\chi^2(2)=89.58$
	$P<0.001$[c]	$P<0.001$[c]
Post-hoc significant comparisons		
Nuclear *vs* cortical	$P<0.001$[c]	$P<0.001$[c]
Nuclear *vs* PSC	$P<0.001$[c]	$P<0.001$[c]
Cortical *vs* PSC	$P<0.001$[c]	$P<0.001$[c]

[a]Presence score ranges from 0 to 6, representing the number of raters identifying the subtype; [b]Pairwise exact-match proportion ranges from 0 to 1, calculated across all 15 rater pairs; [c]$P<0.05$ indicates statistical significance for Kruskal-Wallis test. $P<0.017$ indicates statistical significance after Bonferroni correction. PSC: Posterior subcapsular cataract. All post-hoc Dunn's tests showed significant differences between all subtype pairs ($P<0.017$).

strict (0.253) and resident–senior the lowest tolerant (0.164); between-pair differences were significant for both strict and tolerant grading [Friedman $\chi^2(2)=13.765$, $P=0.001$]. These findings suggest that senior and expert annotators align substantially better with each other on cortical lesions than either does with residents.

For PSCs, while label consistency in senior–expert group (0.755; 95%CI, 0.675–0.832) was significantly higher compared to others [Friedman $\chi^2(2)=58.514$, $P<0.001$], grading consistency remained poor across all pairings [strict grading <0.42, Friedman $\chi^2(2)=4.200$, $P=0.122$]. This implies that although experts and seniors are more consistent in detecting PSC presence, their agreement on opacity severity remains limited.

Post-hoc Wilcoxon signed-rank tests further confirmed these findings, identifying significant differences (Bonferroni-adjusted $P<0.017$) primarily in cortical and PSC subtypes.

Comparison of Consistency Metrics Across Cataract Types Comparisons of consistency across cataract subtypes are summarized in Table 3. For the presence score, significant differences were observed among the three cataract types [Kruskal-Wallis $\chi^2(2)=149.51$, $P<0.001$]. Pairwise exact-match proportion, which reflects grading-level concordance across all annotators, a similar trend was noted [Kruskal-Wallis $\chi^2(2)=89.58$, $P<0.001$]. Furthermore, pairwise post hoc analysis using Dunn's test with Bonferroni correction confirmed that all the three subtypes differed significantly from each other on presence score and pairwise exact-match proportion (adjusted $P<0.017$), demonstrating that both cataract detectability and grading reproducibility vary across subtypes.

Specifically, the nuclear cataracts exhibited the highest presence score and exact-match proportion, followed by cortical cataracts, with PSC showing the lowest values. These findings indicated that nuclear cataracts demonstrated the greatest annotation consistency both in identifying their presence and in assigning the same LOCS III grade, whereas PSC remained the most challenging for consistent recognition.

DISCUSSION

LOCS III grading system is currently the gold standard for clinical assessment[7]. The classic certification protocol was established by Chylack *et al*[7] for ensuring interobserver consistency in clinical trials. The original program required graders to achieve high concordance with the Chylack Cataract Research Group (CCRG) using specific certification slide sets. However, in the context of modern large-scale AI development and clinical applications, it is challenging and uncommon in routine clinical settings to strictly adhere to the original centralized certification. Due to different experiences across ophthalmologists, interobserver variability may occur. Our study systematically investigated the interobserver variability in cataract grading based on the LOCS III across ophthalmologists with different clinical experience levels, and further compared the consistency of different cataract subtypes. According to our research, the overall label and grading consistency based on LOCS III grading was influenced by physician experience, with experts and senior ophthalmologists showing substantially higher agreement than residents. Besides, consistency was different across subtypes, with nuclear cataracts showing the highest presence and grading agreement, cortical cataracts intermediate agreement, and PSC the lowest. These results indicate that the reproducibility of LOCS III grading is determined by both experience of annotators and characteristics of the lesion itself, suggesting that the differences in experience and subtypes should be considered when constructing AI training data for development.

For experience-related differences in annotation consistency, we systematically analyzed intragroup and intergroup consistency of different ophthalmologists. According to our results, the expert group ophthalmologists achieved

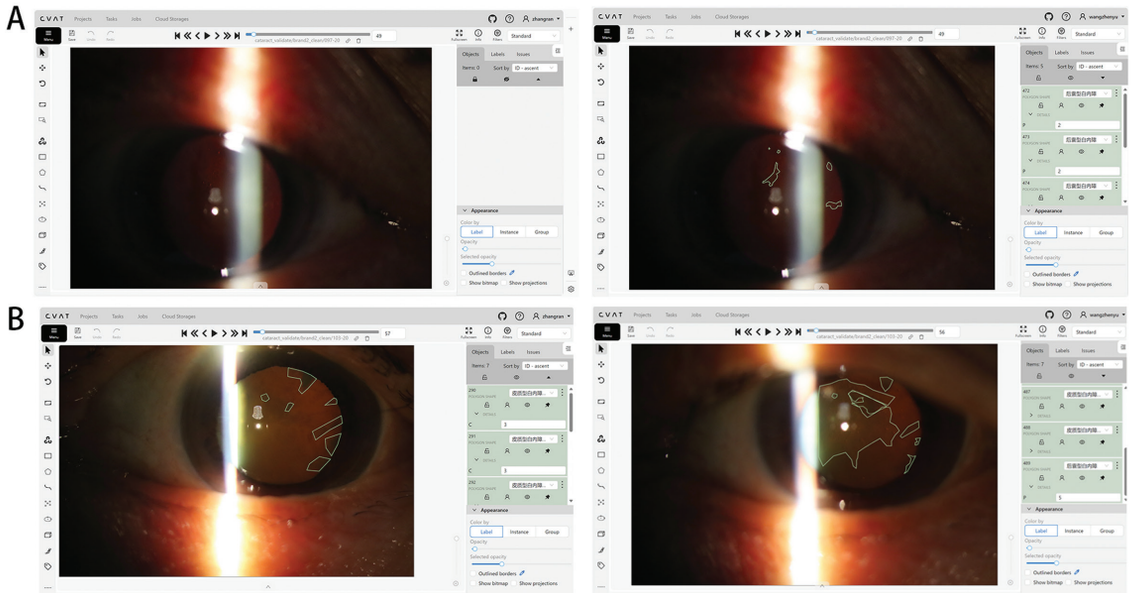


Figure 3 Representative examples illustrating differences in cataract recognition between residents and experts A: In low-definition slit-lamp images, residents failed to detect lens opacity, whereas experts successfully identified the cataract; B: In cases with cortical or posterior subcapsular opacity partially obscured by reflection or overlap, residents often missed the lesion, while experts recognized it accurately.

markedly higher label consistency than residents, particularly for cortical and PSC cataracts. Besides, expert–senior pairs consistently achieved higher label and grading consistency than resident–senior or resident–expert pairs, indicating that more experienced annotators are not only more consistent in identifying the presence of lens opacity but also in recognizing subtle morphological variations. In contrast, residents exhibited greater variability and lower agreement, reflecting the challenges faced by less experienced clinicians when interpreting lens opacities with diffuse, overlapped and irregular patterns, or when facing low-definition images (Figure 3). This pattern reflects not only increased familiarity with the grading system but also fundamental differences in how visual information is perceived and interpreted. Expert ophthalmologists are better able to integrate nuclear color, cortical texture, and posterior subcapsular density, forming a more holistic judgment aligned with the LOCS III grading standard. In contrast, junior doctors may rely more on individual visual features, leading to greater intra-observer noise and inter-observer variability[25]. Furthermore, repeated exposure to standardized grading scales during clinical practice may strengthen experts' internal references, enabling more consistent translation of visual impressions into quantitative grades. Similar results have been reported in other imaging-based fields, such as diabetic retinopathy and glaucoma screening, where diagnostic reproducibility improved with training and experience[26-29]. These findings highlight that interobserver consistency in cataract grading is not only a matter of technical familiarity, but also reflects a perceptual-cognitive learning process. This discovery emphasizes the need to strengthen structured training and calibration among readers when building AI training datasets to ensure the stability and reproducibility of manual annotation.

In addition to experience level, we also detected differences in consistency across cataract subtypes. Our results revealed that the consistency of nuclear cataracts achieved the highest interobserver agreement, followed by cortical and PSC subtypes, in both the presence score and pairwise exact-match proportion. The high reproducibility of nuclear recognition and grading (presence score: 5.652, pairwise exact-match: 0.570) can be attributed to its well-defined and centrally located opacity, which remains consistently visible under varying illumination angles, slit widths, or pupil diameters[30]. Such central localization and strong contrast make both lesion detection and LOCS III grading more reliable across experience levels. In contrast, cortical opacities often occur peripherally as wedge-shaped or radial spokes and are highly dependent on slit-beam angle and pupil size. Inadequate dilation frequently causes peripheral spokes to be partially or completely obscured by the iris, resulting in inconsistent judgments of both presence and extent. PSC opacities, which are located at the posterior capsule, are often small but visually significant[31]. Their visibility depends heavily on image quality involved retroillumination, exposure, and optical alignment. Inappropriate lighting or partial occlusion caused by opacity of the anterior cortex can mask PSC, while backscatter artifacts may mimic or obscure real lesions, making detection particularly experience-dependent.

Even under standardized slit-lamp imaging, residual variations in illumination direction, pupil dilation, and exposure can

disproportionately affect cortical and PSC visibility, whereas nuclear opacities remain robust. These findings emphasize that, beside observer experience, strict image acquisition standardization is essential to reduce non-pathological labeling noise in AI datasets. Nuclear cataracts—with their consistent appearance and well-defined grading criteria—may serve as the most stable and reliable target for AI model training, while cortical and PSC datasets require additional imaging quality control and expert verification to ensure comparable data quality.

This study systematically quantified the variability in LOCS III-based cataract grading across ophthalmologists with different experience levels and evaluated the stability of grading among different cataract subtypes. The results revealed important influences of observer experience and characteristics of cataract subtypes on labeling consistency, offering practical insights for building high-quality training datasets in AI model development. The innovation of this study is mainly reflected in the following two aspects: first, for the first time, we have constructed a cross-labeled database containing three levels of clinical experience (resident group, senior group, and expert group) under a unified digital slit lamp imaging system, which can accurately depict the systematic impact of clinical experience on grading consistency; second, by extending consistency analysis across different cataract subtypes, our study provides evidence that nuclear cataracts represent the most reliable annotation source for AI training, providing new ideas for optimizing subtype-specific data strategies.

Still, there are some limitations that should be acknowledged in our study. First, although imaging illumination and acquisition protocols were standardized, minor variations in pupil dilation, lens reflection, or exposure may still introduce image-level biases. Second, while the sample size in our study has significantly increased compared to previous studies, the cases of some subtypes (such as PSC) are still relatively limited, which could affect the stability of statistical effects. Most importantly, the LOCS III system heavily relies on subjective clinical estimation rather than precise physical measurement. Our analysis was based solely on manual LOCS III annotations and did not incorporate validation using objective control measures. Therefore, our designation of “expert” grading represents the highest level of clinical consensus and reproducibility, rather than an absolute objective standard. Future study should aim to integrate objective imaging modalities to establish a more reliable reference standard for AI model training.

In summary, our study quantitatively evaluated the consistency of LOCS III-based cataract grading among ophthalmologists with different experience levels. Our findings revealed that while grading variability exists across experience levels, senior and expert annotators demonstrated higher reliability and stability than residents. Among cataract subtypes, nuclear opacities achieved the highest interobserver consistency, indicating that such images may serve as more dependable sources for AI training data. In contrast, cortical and posterior subcapsular opacities exhibited greater variability, likely due to image quality, illumination angle, and lesion visibility. By analyzing both experience-dependent and subtype-dependent variability, our findings provide practical guidance for optimizing expert annotation strategies and selecting high-quality reference datasets for AI-based cataract classification systems.

ACKNOWLEDGEMENTS

Authors’ Contributions: Wang ZY and Wu XJ contributed equally: conceptualization, methodology, data collection, and drafting of the manuscript. Zhang R: data collection and annotation. Jia YX, Li EJ, Ding XF, and Yue PL: image grading and data curation. Tao J and Wang JD: supervision and critical revision. Song XD: conceptualization, supervision, funding acquisition, and final approval of the manuscript. All authors read and approved the final version.

Data Availability Statement: The datasets are available from the corresponding author on reasonable request.

AI-Generated Content Disclosure: No AI or AI-assisted technologies were used in the preparation of this manuscript.

Foundations: Supported by Beijing Nova Program (No.20240484601); Beijing Natural Science Foundation (No.7242038; No.7264252).

Conflicts of Interest: Wang ZY, None; **Wu XJ,** None; **Zhang R,** None; **Jia YX,** None; **Li EJ,** None; **Ding XF,** None; **Yue PL,** None; **Tao J,** None; **Wang JD,** None; **Song XD,** None.

REFERENCES

1 Flaxman SR, Bourne RRA, Resnikoff S, *et al*. Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis. *Lancet Glob Heal* 2017;5(12):e1221-e1234.

2 Ang MJ, Afshari NA. Cataract and systemic disease: a review. *Clin Exp Ophthalmol* 2021;49(2):118-127.

3 Chang WC, Lee CH, Chiou SH, *et al*. Proteomic analysis of aqueous humor proteins in association with cataract risks: diabetes and smoking. *J Clin Med* 2021;10(24):5731.

4 Chen SP, Woreta F, Chang DF. Cataracts: a review. *JAMA* 2025;333(23):2093-2103.

5 Mishra D, Kashyap A, Srivastav T, *et al*. Enzymatic and biochemical properties of lens in age-related cataract versus diabetic cataract: a narrative review. *Indian J Ophthalmol* 2023;71(6):2379-2384.

6 GBD 2019 Blindness and Vision Impairment Collaborators, Vision Loss Expert Group of the Global Burden of Disease Study. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. *Lancet Glob Health* 2021;9(2):e144-e160.

7 Chylack LT Jr, Wolfe JK, Singer DM, *et al*. The Lens Opacities Classification System III. The Longitudinal Study of Cataract Study Group. *Arch Ophthalmol* 1993;111(6):831-836.

8 Wu XH, Liu LX, Zhao LQ, *et al*. Application of artificial intelligence in anterior segment ophthalmic diseases: diversity and standardization. *Ann Transl Med* 2020;8(11):714.

9 He R, Jie PY, Wang ZF, *et al*. Artificial intelligence-assisted real-time nasopharyngeal cancer diagnostic model enhances rhinologist performance: a prospective multi-reader study. *Ann Med* 2026;58(1):2620335.

10 Liu Q, Zhang CH, Li PQ, *et al*. Artificial intelligence for precision management of epithelial ovarian cancer: a comprehensive review. *Front Med (Lausanne)* 2025;12:1713629.

11 Booth BB, Khare SK, Blanes-Vidal V, *et al*. Higher diagnostic accuracy of an AI model for colposcopy compared with conventional and digital colposcopic evaluation. *J Low Genit Tract Dis* 2026;30(2):118-123.

12 Lu MZ, Yang KL, Deng XX, *et al*. Artificial intelligence in clinical diagnosis and treatment of dry eye: workflows, effectiveness, and evaluation. *J Curr Ophthalmol* 2024;36(4):315-324.

13 Wan C, Lv XX, Hu XY, *et al*. A novel system for measuring eyeball rotation angle based on color fundus photographs in natural head position. *Transl Vis Sci Technol* 2025;14(8):25.

14 Chong PL, Vaigeshwari V, Mohammed Reyasudin BK, *et al*. Integrating artificial intelligence in healthcare: applications, challenges, and future directions. *Future Sci OA* 2025;11(1):2527505.

15 Oshika T. Artificial intelligence applications in ophthalmology. *JMA J* 2025;8(1):66-75.

16 Xiao ZJ, Zhang XQ, Zheng BF, *et al*. Multi-style spatial attention module for cortical cataract classification in AS-OCT image with supervised contrastive learning. *Comput Methods Programs Biomed* 2024;244:107958.

17 Mai ELC, Chen BH, Su TY. Innovative utilization of ultra-wide field fundus images and deep learning algorithms for screening high-risk posterior polar cataract. *J Cataract Refract Surg* 2024;50(6):618-623.

18 Xie H, Li ZW, Wu CC, *et al*. Deep learning for detecting visually impaired cataracts using fundus images. *Front Cell Dev Biol* 2023;11:1197239.

19 Lu Q, Wei L, He WW, *et al*. Lens Opacities Classification System III-based artificial intelligence program for automatic cataract grading. *J Cataract Refract Surg* 2022;48(5):528-534.

20 Wu X, Hu LX, Xiao ZJ, *et al*. Explainable machine learning framework for cataracts recognition using visual features. *Vis Comput Ind Biomed Art* 2025;8(1):3.

21 Tognetto D, Giglio R, Vinciguerra AL, *et al*. Artificial intelligence applications and cataract management: a systematic review. *Surv Ophthalmol* 2022;67(3):817-829.

22 Gutierrez L, Lim JS, Foo LL, *et al*. Application of artificial intelligence in cataract management: current and future directions. *Eye Vis* (*Lond*) 2022;9(1):3.

23 Hu SM, Luan XZ, Wu H, *et al*. ACCV: automatic classification algorithm of cataract video based on deep learning. *Biomed Eng Online* 2021;20(1):78.

24 Wu X, Xu D, Ma T, *et al*. Artificial intelligence model for antiinterference cataract automatic diagnosis: a diagnostic accuracy study. *Front Cell Dev Biol* 2022;10:906042.

25 Tan AC, Loon SC, Choi H, *et al*. Lens Opacities Classification System III: cataract grading variability between junior and senior staff at a Singapore hospital. *J Cataract Refract Surg* 2008;34(11):1948-1952.

26 Antonio-Aguirre B, Gadiraju A, Ownagh V, *et al*. Novel artificial intelligence applications for pediatric retina. *Curr Opin Ophthalmol* 2025:36(6):456-461.

27 Marra KV, Chen JS, Nudleman E, *et al*. Review of retinopathy of prematurity management in the anti-VEGF era: evolving global paradigms, persistent challenges and our AI-assisted future. *Clin Exp Ophthalmol* 2025;53(9):1202-1217.

28 Pourjavan S, Bourguignon GH, Marinescu C, *et al*. Evaluating the influence of clinical data on inter-observer variability in optic disc analysis for AI-assisted glaucoma screening. *Clin Ophthalmol* 2024;18:3999-4009.

29 Cassottana P, Iester C, Bonzano L, *et al*. Quantitative evaluation of OCT angiography images in healthy and glaucomatous subjects through a novel approach: exploring inter-image variability. *Eye* (*Lond*) 2024;38(10):1903-1907.

30 Zhang Y, Zhang YY, Zhang JT, *et al*. Intraoperative quantitative crystalline lens nuclear opacities analysis based on crystalline lenSx platform. *BMC Ophthalmol* 2024;24(1):206.

31 Richardson RB, Ainsbury EA, Prescott CR, *et al*. Etiology of posterior subcapsular cataracts based on a review of risk factors including aging, diabetes, and ionizing radiation. *Int J Radiat Biol* 2020;96(11): 1339-1361.