Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prognostic performance of public large language models in primary rhegmatogenous retinal detachment

Carlos Campo-Beamud

carloscampobeamud@yahoo.es

Hospital General Universitario de Ciudad Real

Almudena Moreno-Martinez

Complejo Hospitalario Universitario de Albacete

Esther Campo-Beamud

Hospital General Universitario de Ciudad Real

Francisco Javier Gómez-Romero

Hospital General Universitario de Ciudad Real

Cristina Blanco-Marchite

Complejo Hospitalario Universitario de Albacete

Raquel Maroto

Complejo Hospitalario Universitario de Albacete

Alfredo Julian Fernández Ruíz

Hospital General Universitario de Ciudad Real

Sergio Copete

Complejo Hospitalario Universitario de Albacete

Research Article

Keywords: Retinal Detachment, Artificial Intelligence, Large Language Models, GPT-5, ChatGPT, Gemini, DeepSeek, Visual Prognosis

Posted Date: June 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9932285/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Purpose: To evaluate the prognostic performance of four publicly accessible large language models—ChatGPT-4o, ChatGPT-5, Gemini, and DeepSeek-R1—in forecasting postoperative best-corrected visual acuity (BCVA) and estimating recurrent rhegmatogenous retinal detachment (reRRD) risk after primary rhegmatogenous retinal detachment surgery.

Methods: In this multicenter retrospective cohort study, 206 patients who underwent surgical repair for primary rhegmatogenous retinal detachment at 2 tertiary hospitals were included. Structured preoperative clinical data were submitted to each model to forecast 1-year BCVA and estimate reRRD probability. Visual performance was assessed by the proportion of forecasts within ±0.2 logMAR and by intraclass correlation coefficients (ICC). Anatomical risk estimation was assessed using the area under the receiver operating characteristic curve (AUC), the Youden index, and decision curve analysis.

Results: ChatGPT-5 achieved the highest clinically acceptable BCVA forecast accuracy within ±0.2 logMAR (82.7%), followed by Gemini (81.1%), ChatGPT-4o (72.9%), and DeepSeek (70.4%), with ICC values of 0.82, 0.81, 0.76, and 0.70, respectively. For anatomical prognosis, ChatGPT-4o, ChatGPT-5, and Gemini showed outstanding population-level discrimination for reRRD (AUC 0.953, 0.947, and 0.942, respectively), whereas DeepSeek showed lower but still acceptable performance (AUC 0.785). A 30% threshold was selected for operational reRRD risk stratification based on concordant findings from the Youden index and decision curve analysis.

Conclusion: Publicly accessible large language models, particularly ChatGPT-5 and Gemini, showed clinically meaningful performance in postoperative visual forecasting, especially when assessed by clinically acceptable BCVA accuracy within ±0.2 logMAR. They may also provide adjunctive value for reRRD risk stratification, although anatomical findings should be interpreted more cautiously because of the limited number of recurrence events.

Key Message

What is known

- Postoperative visual recovery and recurrent rhegmatogenous retinal detachment after primary retinal detachment surgery are influenced by multiple preoperative and surgical factors, including macular status, baseline visual acuity, detachment duration, proliferative vitreoretinopathy, lens status, tamponade, and surgical complexity.
- Artificial intelligence models have shown potential for ophthalmic diagnostic and prognostic tasks, but the ability of publicly accessible large language models to estimate visual and anatomical prognosis after rhegmatogenous retinal detachment surgery remains insufficiently characterized.

What is new

- In this multicenter retrospective cohort of 206 primary rhegmatogenous retinal detachment cases, publicly accessible large language models generated clinically acceptable 1-year postoperative visual acuity forecasts in a substantial proportion of eyes, with the best performance observed for ChatGPT-5 and Gemini 3 Flash.
- ChatGPT-4o, ChatGPT-5, and Gemini 3 Flash showed strong population-level discrimination for recurrent rhegmatogenous retinal detachment risk, whereas DeepSeek-R1 showed lower performance.
- These models may have adjunctive value for postoperative counseling, risk communication, and follow-up prioritization, but their outputs remain insufficient for standalone case-level clinical decision-making.

INTRODUCTION

Large language models (LLMs) have emerged as versatile artificial intelligence tools capable of processing structured clinical information and generating context-sensitive outputs with potential relevance for clinical decision support and medical informatics applications [1]. In ophthalmology, these systems have shown promising performance in diagnostic reasoning, patient information, and surgical planning tasks, although their prognostic utility remains insufficiently characterized [2–8].

Rhegmatogenous retinal detachment (RRD) is a vision-threatening condition whose incidence is increasing, partly in parallel with the rising prevalence of myopia [9–12]. Despite successful anatomical repair in many cases, postoperative outcomes remain highly variable. Visual recovery depends on multiple preoperative and perioperative factors, while recurrent rhegmatogenous retinal detachment (reRRD) remains one of the main causes of surgical failure and postoperative uncertainty [13–24]. More accurate prognostic tools could therefore improve postoperative counseling, risk communication, and follow-up prioritization in routine vitreoretinal practice.

Artificial intelligence approaches, including dedicated machine learning and deep learning systems, have already shown potential for forecasting visual and anatomical outcomes after RRD surgery [9, 24]. However, these systems generally depend on task-specific development, curated datasets, imaging-based pipelines, or dedicated computational infrastructure, which may limit their broad implementation across heterogeneous clinical settings. By contrast, publicly accessible LLMs offer a different paradigm: they can operate using routinely available structured preoperative data, without requiring local model training or image acquisition workflows, potentially reducing implementation barriers and increasing applicability across centers, particularly those with more limited diagnostic or technical resources [1, 9, 24, 25].

This study evaluates the prognostic performance of four publicly accessible LLMs—ChatGPT-4o, Gemini 3 Flash, DeepSeek-R1, and ChatGPT-5—for forecasting postoperative best-corrected visual acuity and estimating the risk of recurrent detachment after primary RRD surgery. Using structured, non-image-based preoperative clinical data, we aimed to determine whether these openly accessible tools could

provide clinically relevant postoperative risk stratification. Their accessibility and image-free design could make them especially relevant in resource-constrained clinical environments [9, 24].

METHODS

Study Design

This multicenter retrospective cohort study included 206 consecutive patients with primary RRD who underwent surgery at two tertiary hospitals in central Spain (University General Hospital of Ciudad Real and Albacete University Hospital Complex) between January 2024 and January 2025. All patients had a minimum follow-up of 12 months.

Outcome

The main objective was to evaluate the prognostic performance of four publicly accessible LLMs—ChatGPT-4o, Gemini 3 Flash, DeepSeek-R1, and ChatGPT-5—in forecasting postoperative BCVA and estimating the risk of reRRD.

Inclusion and Exclusion Criteria

Eligible participants were adults (≥ 18 years) diagnosed with primary RRD who underwent surgical repair via pars plana vitrectomy (PPV), with or without adjunctive scleral buckling or phacoemulsification, and received intraocular tamponade with gas or silicone oil. Exclusion criteria included incomplete clinical records, follow-up < 12 months, tractional or exudative retinal detachment, previous retinal surgeries or ocular trauma, and significant pre-existing macular pathology.

Variables and Data Preparation

To approximate expert-level reasoning, each LLM was first exposed to a standardized literature-based briefing summarizing key prognostic factors associated with postoperative outcomes in RRD [11, 13–19]. This procedure was designed to provide uniform contextual information without modifying model parameters. The full content of this standardized briefing is provided in Supplementary Material.

For each patient, a standardized case summary was created incorporating preoperative variables known to influence RRD outcomes. These included age, sex, systemic and ocular comorbidities, preoperative BCVA, symptom duration, macular status, number and morphology of retinal breaks, presence and grade of proliferative vitreoretinopathy (PVR), lens status, surgical technique, tamponade agent, and surgeon experience (categorized as beginner [< 5 years], intermediate [5–10 years], or expert [> 10 years]).

The BCVA was defined as the value recorded closest to 12 ± 1 months postoperatively and converted from Snellen to logMAR for analysis. A clinically acceptable forecast error was predefined as ± 0.2

logMAR (approximately 2 ETDRS lines). Anatomical outcome was assessed as the occurrence of reRRD within 12 months. For anatomical prognostication, model outputs were analyzed primarily as continuous risk estimates. In addition, threshold-based risk stratification analyses were performed to identify clinically useful probability cutoffs. The final operational threshold was set at 30%, based on concordant findings from the Youden index and Decision Curve Analysis (DCA).

Chatbot Prompting Procedure

Each chatbot was queried independently using identical structured inputs. The prompts simulated clinical reasoning and included the following instruction: *"Given the following clinical data, forecast the 1-year postoperative best-corrected visual acuity (logMAR) and the probability (in %) of recurrent rhegmatogenous retinal detachment. Provide numeric responses only."*

No prompt iteration, fine-tuning, or feedback adjustment was permitted. Model outputs were recorded manually and archived systematically for audit purposes. Programmatic API access was not used.

LLM Configuration

Four publicly accessible LLMs were evaluated: ChatGPT-4o (OpenAI, San Francisco, USA), ChatGPT-5 (OpenAI, San Francisco, USA), Gemini 3 Flash (Google DeepMind, London, UK), and DeepSeek-R1 (DeepSeek AI, Beijing, China). Models were selected on the basis of public accessibility and their potential relevance for real-world clinical implementation. All models were accessed through their official web-based interfaces between 20 and 31 December 2025 using default platform parameters. Each case was submitted in a new session without reused conversation history, and no regeneration or requerying was permitted. Only fully anonymized structured clinical case summaries were entered into the platforms, and no directly identifiable patient information was submitted.

Error Handling and Exploratory Analysis

The BCVA forecasts that substantially exceeded expected postoperative values without clinical justification were flagged as visual overestimations ("hallucinations"). Forecast plausibility was reviewed independently by two retina specialists, with discrepancies resolved by consensus. Cases with forecast deviations ≥ 0.4 logMAR (approximately 4 ETDRS lines) were analyzed descriptively as large forecast errors. Response time per case was also recorded.

To explore variable influence on model forecasts, synthetic cases were generated by modifying single variables in isolation while holding others constant. These results were used to estimate heuristic feature weightings, presented in Supplementary Material.

Statistical Analysis

Continuous variables are reported as mean ± standard deviation (SD), and categorical variables as frequencies with 95% confidence intervals (CI). Agreement between forecast and actual BCVA was assessed using intraclass correlation coefficients (ICC), interpreted as poor (< 0.5), moderate (0.5–0.75), good (0.75–0.9), and excellent (> 0.9) [26]. Visual forecast performance was primarily evaluated by the proportion of forecasts within ± 0.2 logMAR and by ICC.

For anatomical prognostication, discrimination of continuous reRRD risk estimates was evaluated using receiver operating characteristic (ROC) curve analysis and the area under the curve (AUC), classified as no discrimination (0.5), poor/limited (0.6–0.7), acceptable (0.7–0.8), excellent (0.8–0.9), or outstanding (≥ 0.9) [27].

To identify the statistically optimal probability threshold for reRRD classification, the Youden index was calculated. To assess the clinical utility of threshold-based risk stratification, DCA was performed across a range of threshold probabilities, comparing model-based strategies against “treat-all” and “treat-none” approaches [28]. The final operational threshold of 30% was selected because it was supported by both the Youden index and DCA. All statistical analyses were performed using RStudio (version 4.1.0; R Foundation for Statistical Computing, Vienna, Austria).

Ethical Considerations

This study was approved by the Institutional Review Boards of both participating institutions. All procedures adhered to the principles of the Declaration of Helsinki. Informed consent was waived due to the retrospective and anonymized nature of the data.

RESULTS

A total of 206 eyes with primary RRD were included in the analysis. Baseline anatomical characteristics are summarized in Table 1. Among these, 84 eyes (40.8%) presented with macular detachment at diagnosis, and 23 eyes (11.2%) developed reRRD within the first postoperative year.

Table 1

Baseline anatomical characteristics of the study population. Baseline distribution of retinal detachment subtypes and selected anatomical features in the 206 included eyes. RRD = rhegmatogenous retinal detachment; reRRD = recurrent rhegmatogenous retinal detachment; PVR = proliferative vitreoretinopathy.

Retinal detachment subtype	N (%)
Single break	117 (56.8%)
Multiple breaks	67 (32.5%)
Giant retinal tears or retinal dialysis	22 (10.7%)
Any grade of vitreoretinal proliferation	40 (19.4%)
Macula-off retinal detachment	84 (40.8%)
reRRD	23 (11.2%)

Visual Prognosis

At 12 months, exact BCVA forecasts were achieved in 99 eyes (48.1%) by ChatGPT-5, 92 eyes (44.7%) by Gemini, 46 eyes (22.3%) by DeepSeek, and 40 eyes (19.4%) by ChatGPT-4o. A clinically acceptable forecast error, defined as a forecast within ± 0.2 logMAR, was achieved in 82.7% of eyes for ChatGPT-5, 81.1% for Gemini, 72.9% for ChatGPT-4o, and 70.4% for DeepSeek. Thus, ChatGPT-5 and Gemini showed the highest clinically acceptable visual forecast accuracy within ± 0.2 logMAR.

Agreement between forecast and actual BCVA, assessed using ICC, was good for ChatGPT-5 (ICC = 0.82; 95% CI: 0.76–0.86), Gemini (ICC = 0.81; 95% CI: 0.76–0.85), and ChatGPT-4o (ICC = 0.76; 95% CI: 0.69–0.81), and moderate for DeepSeek (ICC = 0.70; 95% CI: 0.62–0.76). These findings indicate that ChatGPT-5 and Gemini provided the most reliable visual forecasts overall.

Forecast Errors and “Hallucinations”

Large visual forecast errors (≥ 0.4 logMAR) were observed across all models and were frequently associated with postoperative complications not available in the preoperative input data. These included cystoid macular edema, persistent subretinal fluid, visually significant cataract remaining unoperated within the first postoperative year, posterior capsule opacification without YAG capsulotomy, subretinal fibrosis, and one case of neurotrophic corneal ulcer. In a subset of cases, no clear clinical explanation for the magnitude of the error was identified.

A subset of forecasts was categorized as clinically implausible visual overestimations (“hallucinations”). These occurred in 6 cases with ChatGPT-4o, 3 cases with Gemini, 2 cases with ChatGPT-5, and 1 case with DeepSeek. These visual overestimations typically occurred in younger, non-myopic, phakic patients without major systemic comorbidities, with macula-on detachments and surgery performed by

experienced surgeons, suggesting a tendency to overweight favorable preoperative features in otherwise low-risk profiles.

Anatomical Prognosis

The ability of the LLMs to estimate the risk of reRRD was based on preoperative case-specific characteristics. Although none of the models supported sufficiently reliable case-level anatomical prediction for standalone clinical decision-making, their continuous risk estimates showed substantial differences in population-level discriminatory performance.

ChatGPT-4o achieved an AUC of 0.953 (95% CI: 0.926–0.981), ChatGPT-5 an AUC of 0.947 (95% CI: 0.878–1.000), and Gemini an AUC of 0.942 (95% CI: 0.873–1.000), indicating outstanding discrimination at the population level. DeepSeek showed lower but still acceptable discriminatory performance, with an AUC of 0.785 (95% CI: 0.721–0.849) (Fig. 1).

To identify a clinically relevant threshold for anatomical risk stratification, both the Youden index and DCA were applied. For both ChatGPT-5 and Gemini, the Youden index identified an optimal threshold of 30%, and DCA also showed peak net benefit at 0.29. Based on this concordance, 30% was selected as the operational threshold for reRRD risk stratification (Fig. 2).

Using a > 30% risk threshold, ChatGPT-5 classified 20 eyes as high risk, of which 19 developed reRRD. This corresponded to a sensitivity of 82.6%, a specificity of 99.5%, and a positive predictive value of 95.0%. Gemini showed a very similar classification profile, identifying 18 of 23 reRRD cases with only one false positive, whereas ChatGPT-4o identified 13 of 23 reRRD cases with two false positives. DeepSeek detected fewer recurrence cases overall and showed a less favorable threshold-based profile. Eyes classified above this threshold had markedly higher observed risk of reRRD than those classified at or below it, with an estimated relative risk of approximately 44 for ChatGPT-5, 24.6 for Gemini, and 12.0 for ChatGPT-4o.

At this threshold, ChatGPT-5 and Gemini showed the most favorable balance between discrimination and clinical utility, whereas ChatGPT-4o remained informative but less balanced, and DeepSeek demonstrated a lower net benefit profile. DCA further confirmed that the 30% threshold fell within a range of positive clinical net benefit for the best-performing models, supporting its use as a pragmatic threshold for postoperative triage and follow-up prioritization (Fig. 2).

To further explore model behavior, synthetic scenarios were generated by independently modifying single clinical variables while holding others constant. The resulting heuristic weights, summarized in Table 2, illustrate how each model differentially weighted prognostic factors such as macular status, PVR, chronicity, and surgeon experience when estimating reRRD risk. These exploratory findings highlight meaningful inter-model variability in risk integration.

Table 2

Estimated heuristic weights for recurrent rhegmatogenous retinal detachment risk. Approximate percentage increases in forecast reRRD risk after isolated synthetic modification of selected preoperative variables for each chatbot. ReRRD = recurrent rhegmatogenous retinal detachment; PVR = proliferative vitreoretinopathy; DM = diabetes mellitus; BCVA = best-corrected visual acuity; N/A = not applicable.

Risk Factor	ChatGPT-4o	ChatGPT-5	Gemini 3.0 Flash	DeepSeek-R1
Macula OFF	+ 10%	+ 5%	+ 5%	+ 5–10%
PVR grade A	+ 15%	+ 10%	+ 10%	+ 10–15%
PVR grade B-C	+ 20%	+ 15–20%	+ 15%	+ 15–20%
High myopia	+ 10%	+ 10%	+ 10%	+ 10%
Diabetes Mellitus (DM)	+ 5%	+ 5%	+ 5%	+ 5%
Age > 65 years	+ 5%	+ 5%	0%	+ 5%
Pseudophakia	+ 5%	+ 5%	+ 5%	+ 5%
Multiple tears (2)	+ 5%	+ 5%	+ 5%	+ 5%
Multiple tears (≥ 3)	+ 10%	+ 10%	+ 10%	+ 10%
RRD duration > 14 days	+ 5%	+ 5%	+ 5%	+ 5%
Inexperienced surgeon	+ 15%	+ 10%	+ 10%	+ 10–15%
Sex, eye involved, baseline BCVA	N/A	N/A	N/A	N/A

DISCUSSION

This study shows that publicly accessible LLMs can provide clinically relevant postoperative prognostic support after primary RRD surgery. Their most robust finding was the high rate of clinically acceptable BCVA forecasts within ± 0.2 logMAR, particularly with ChatGPT-5 and Gemini, while their anatomical outputs appeared more useful for population-level risk stratification than for case-level prediction.

For visual prognosis, the present findings should be interpreted in the context of established predictors of postoperative recovery after RRD repair, including macular status, preoperative BCVA, detachment duration, PVR, and surgical factors [11, 13–20, 23]. In our study, ChatGPT-5 and Gemini captured a meaningful proportion of this structured prognostic signal, although their value appeared greater for clinically acceptable approximation than for precise continuous numerical prediction, consistent with prior AI studies suggesting stronger performance in categorization than in fine-grained regression tasks [24, 29, 30].

The pattern of visual errors was also informative. Large forecast deviations were frequently associated with postoperative events not represented in the preoperative prompts, including cystoid macular edema, persistent subretinal fluid, delayed cataract surgery, posterior capsule opacification, and subretinal

fibrosis. Clinically implausible visual overestimations (“hallucinations”) mainly occurred in eyes with otherwise favorable baseline profiles, suggesting a tendency to overweight positive prognostic features when competing negative signals were insufficiently encoded, in line with broader concerns regarding LLM calibration and interpretability [4, 5, 9].

For anatomical prognosis, the key result was the marked difference between case-level anatomical prediction and population-level risk discrimination. Although none of the models should be interpreted as reliable standalone tools for deterministic individual anatomical prediction, ChatGPT-5, ChatGPT-4o, and Gemini all achieved outstanding AUC values for reRRD, indicating that their outputs were highly effective at ranking patients according to recurrence risk. Their value, therefore, lies less in predicting whether a specific eye will recur than in identifying eyes with meaningfully higher postoperative risk.

This is particularly relevant in light of prior AI work in RRD prognosis. Guo et al. showed that machine learning techniques can predict visual outcomes after RRD surgery, while Fung et al. demonstrated promising anatomical forecasting with deep learning models [9, 31]. Our findings do not suggest that public LLMs outperform dedicated systems, but rather that they may offer a different advantage: broader accessibility, substantially lower development and implementation barriers, and useful postoperative risk stratification from structured text-based preoperative summaries alone. This may be particularly relevant in centers with fewer technical, imaging, or computational resources, where developing, validating, and deploying dedicated deep learning pipelines may be less feasible [32, 33]. Furthermore, although OCT may improve prognostic precision after RRD surgery, it was not incorporated because it is not consistently available in all ophthalmic emergency settings. Restricting the model to routinely available structured clinical data was therefore intended to maximize portability across centers with different levels of diagnostic infrastructure, although this likely reduced precision in borderline cases [11, 19].

The 30% threshold was supported by both the Youden index and DCA. The Youden index identified an optimal threshold of 30% for the best-performing models, while DCA showed peak or near-peak net benefit in the same range (Fig. 2), providing complementary statistical and clinical support for the selected cutoff. In practical terms, ChatGPT-5 and Gemini were able to separate groups with markedly different observed risks of reRRD, supporting their potential role in postoperative triage and follow-up prioritization rather than replacing clinician judgment [34].

This potential clinical usefulness extends beyond risk stratification alone. In routine practice, one of the most common postoperative questions asked by patients with RRD is how much vision they are likely to recover. Although these models are not sufficiently robust for definitive individualized counseling, approximate visual forecasting and anatomical risk stratification may still help clinicians provide more structured prognostic information, improve patient communication, and tailor the intensity of postoperative follow-up according to estimated risk.

Several limitations should be acknowledged, including the retrospective design, the relatively small number of reRRD events, the absence of external validation, and the use of general-purpose public LLMs through default web interfaces. The selected threshold, although supported by Youden and DCA, remains

internally derived and requires prospective confirmation. Nevertheless, the fact that ChatGPT-5 and Gemini produced meaningful prognostic stratification from anonymized, non-image-based clinical summaries suggests that publicly accessible LLMs may already have value as adjunctive tools for counseling, risk communication, and follow-up prioritization in routine vitreoretinal care.

In conclusion, publicly accessible LLMs—particularly ChatGPT-5 and Gemini 3 Flash—may have a clinically relevant role in postoperative prognostication after primary RRD surgery. Their most consistent contribution in this study was the ability to provide clinically acceptable postoperative visual forecasts, while their anatomical outputs may offer additional value for recurrence risk stratification at a population level. Although they are not sufficiently robust for standalone case-level decision-making, their accessibility, image-free design, and lower implementation burden make them potentially useful as adjunctive tools for postoperative counseling and follow-up planning.

Declarations

Conflict of Interest

The authors declare that they have no competing interests.

Ethics Approval

This study was approved by the Institutional Review Board of Hospital General Universitario de Ciudad Real and adhered to the tenets of the Declaration of Helsinki.

Consent to Participate

Not applicable, as this study used fully anonymised retrospective clinical data.

Consent for Publication

Not applicable.

Clinical trial number

Not applicable.

Funding

Not applicable.

Author Contribution

CCB conceived the study, designed the methodology, curated the dataset, and coordinated the overall research project. AMM contributed to data interpretation, critical revision of the clinical content, and refinement of the manuscript. ECB and FJGR performed data analysis and drafted the initial version of the manuscript. CBM and RM contributed to data collection and patient inclusion. AFJR provided senior clinical oversight, contributed to the interpretation of the results, and critically revised the manuscript for important intellectual content. SC supervised the research activity, contributed to the study design and methodological framework, and provided strategic guidance throughout the project. All authors reviewed and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Acknowledgements

None.

Declaration of generative AI and AI-assisted technologies in the writing process.

During the preparation of this work, the authors used ChatGPT-5 to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data Availability

The datasets generated and analysed during the current study are not publicly available because they contain clinical data derived from patient records, even though anonymised, and their unrestricted sharing was not included in the ethics approval. De-identified data may be made available from the corresponding author upon reasonable request and subject to institutional and ethical approval.

References

1. Romano MF, Shih LC, Paschalidis IC, Au R, Kolachalama VB (2023) Large language models in neurology research and future practice. Neurology 101:1058–1067. https://doi.org/10.1212/WNL.0000000000207967
2. Bahir D, Zur O, Attal L, Nujeidat Z, Knaanie A, Pikkel J, Mimouni M, Plopsky G (2024) Gemini AI vs. ChatGPT: a comprehensive examination alongside ophthalmology residents in medical knowledge. Graefes Arch Clin Exp Ophthalmol. https://doi.org/10.1007/s00417-024-06625-4
3. Conroy G, Mallapaty S (2025) How China created AI model DeepSeek and shocked the world. Nature 638:300–301. https://doi.org/10.1038/d41586-025-00259-0
4. Milad D, Antaki F, Milad J, Farah A, Khairy T, Mikhail D, Giguère CÉ, Touma S, Bernstein A, Szigiato AA, Nayman T, Mullie GA, Duval R (2024) Assessing the medical reasoning skills of GPT-4 in complex ophthalmology cases. Br J Ophthalmol 108:1398–1405. https://doi.org/10.1136/bjo-2023-325053

5. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, Landman A, Dreyer K, Succi MD (2023) Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res 25:e48659. https://doi.org/10.2196/48659
6. Carlà MM, Gambini G, Baldascino A, Giannuzzi F, Boselli F, Crincoli E, D'Onofrio NC, Rizzo S (2024) Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. Br J Ophthalmol 108:1457–1469. https://doi.org/10.1136/bjo-2023-325143
7. Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, Mansour HA, Abishek RM, Xu D, Sridhar J, Yonekawa Y, Kuriyan AE (2023) Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. Ophthalmol Retina 7:862–868. https://doi.org/10.1016/j.oret.2023.05.022
8. Strzalkowski P, Strzalkowska A, Chhablani J, Pfau K, Errera MH, Roth M, Schaub F, Bechrakis NE, Hoerauf H, Reiter C, Schuster AK, Geerling G, Guthoff R (2024) Evaluation of the accuracy and readability of ChatGPT-4 and Google Gemini in providing information on retinal detachment: a multicenter expert comparative study. Int J Retina Vitreous 10:61. https://doi.org/10.1186/s40942-024-00579-9
9. Guo H, Ou C, Wang G, Lu B, Li X, Yang T, Zhang J (2024) Prediction of visual outcome after rhegmatogenous retinal detachment surgery using artificial intelligence techniques. Transl Vis Sci Technol 13:17. https://doi.org/10.1167/tvst.13.5.17
10. Nielsen BR, Alberti M, Bjerrum SS, la Cour M (2020) The incidence of rhegmatogenous retinal detachment is increasing. Acta Ophthalmol 98:603–606. https://doi.org/10.1111/aos.14380
11. Cheng SF, Yang CH, Lee CH, Yang CM, Huang JS, Ho TC, Lin CP, Chen MS (2008) Anatomical and functional outcome of surgery of primary rhegmatogenous retinal detachment in high myopic eyes. Eye (Lond) 22:70–76. https://doi.org/10.1038/sj.eye.6702527
12. Ullrich M, Zwickl H, Findl O (2021) Incidence of rhegmatogenous retinal detachment in myopic phakic eyes. J Cataract Refract Surg 47:533. https://doi.org/10.1097/j.jcrs.0000000000000420
13. Ross WH (2002) Visual recovery after macula-off retinal detachment. Eye (Lond) 16:440–446. https://doi.org/10.1038/sj.eye.6700192
14. Doyle E, Herbert EN, Bunce C, Williamson TH, Laidlaw DAH (2007) How effective is macula-off retinal detachment surgery. Might good outcome be predicted? Eye (Lond) 21:534–540. https://doi.org/10.1038/sj.eye.6702260
15. Pastor JC, de la Rúa ER, Martín F (2002) Proliferative vitreoretinopathy: risk factors and pathobiology. Prog Retin Eye Res 21:127–144. https://doi.org/10.1016/s1350-9462(01)00023-4
16. Cowley M, Conway BP, Campochiaro PA, Kaiser D, Gaskin H (1989) Clinical risk factors for proliferative vitreoretinopathy. Arch Ophthalmol 107:1147–1151. https://doi.org/10.1001/archopht.1989.01070020213027
17. Nishimura S, Okuda T, Higashide T, Sugiyama K (2025) Preoperative retinal detachment variables causing macular detachment. Clin Ophthalmol 19:11–18. https://doi.org/10.2147/OPTH.S499960

18. Mason JO, Mason LB, Patel SA, McGwin G, Finley TA, Friedman DA, Pomerleau DL, Albert MA, Feist RM, Thomley ML, Mason JO (2016) Vitreoretinal surgical outcomes performed by supervised retinal fellows versus attending faculty surgeons. Retina 36:981–985. https://doi.org/10.1097/IAE.0000000000000828
19. Yorston D, Donachie PHJ, Laidlaw DA, Steel DH, Sparrow JM, Aylward GW, Williamson TH, Members of the BEAVRS Retinal Detachment Outcomes Group (2021) Factors affecting visual recovery after successful repair of macula-off retinal detachments: findings from a large prospective UK cohort study. Eye (Lond) 35:1431–1439. https://doi.org/10.1038/s41433-020-1021-y
20. Burton TC (1977) Preoperative factors influencing anatomic success rates following retinal detachment surgery. Trans Sect Ophthalmol Am Acad Ophthalmol Otolaryngol 83:OP499–OP505
21. Burton TC (1982) Recovery of visual acuity after retinal detachment involving the macula. Trans Am Ophthalmol Soc 80:475–497
22. Frings A, Markau N, Katz T, Stemplewitz B, Skevas C, Druchkiv V, Wagenfeld L (2016) Visual recovery after retinal detachment with macula-off: is surgery within the first 72 h better than after? Br J Ophthalmol 100:1466–1469. https://doi.org/10.1136/bjophthalmol-2015-308153
23. Tani P, Robertson DM, Langworthy A (1981) Prognosis for central vision and anatomic reattachment in rhegmatogenous retinal detachment with macula detached. Am J Ophthalmol 92:611–620. https://doi.org/10.1016/s0002-9394(14)74651-3
24. Geiger M, Smith JM, Lynch A, Patnaik JL, Oliver SCN, Dixon JA, Mandava N, Palestine AG, University of Colorado Retina Research Group (2020) Predictors for recovery of macular function after surgery for primary macula-off rhegmatogenous retinal detachment. Int Ophthalmol 40:609–616. https://doi.org/10.1007/s10792-019-01219-0
25. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L et al (2024) GPT-4 technical report. https://doi.org/10.48550/arXiv.2303.08774
26. Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 15:155–163. https://doi.org/10.1016/j.jcm.2016.02.012
27. Mandrekar JN (2010) Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol 5:1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d
28. Vickers AJ, Elkin EB (2006) Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak 26:565–574. https://doi.org/10.1177/0272989X06295361
29. Chen JZ, Li CC, Li SH, Su YT, Zhang T, Wang YS, Dou GR, Chen T, Wang XC, Zhang ZM (2023) A feasibility study for objective evaluation of visual acuity based on pattern-reversal visual evoked potentials and other related visual parameters with machine learning algorithm. BMC Ophthalmol 23:293. https://doi.org/10.1186/s12886-023-03044-7
30. Paul W, Burlina P, Mocharla R, Joshi N, Li Z, Gu S, Nanegrungsunk O, Lin K, Bressler SB, Cai CX, Kong J, Liu TYA, Moini H, Du W, Amer F, Chu K, Vitti R, Sepehrband F, Bressler NM (2023) Accuracy of artificial intelligence in estimating best-corrected visual acuity from fundus photographs in eyes

with diabetic macular edema. JAMA Ophthalmol 141:677–685. https://doi.org/10.1001/jamaophthalmol.2023.2271

31. Fung THM, John NCRA, Guillemaut JY, Yorston D, Frohlich D, Steel DHW, Williamson TH, BEAVRS Retinal Detachment Outcomes Group (2023) Artificial intelligence using deep learning to predict the anatomical outcome of rhegmatogenous retinal detachment surgery: a pilot study. Graefes Arch Clin Exp Ophthalmol 261:715–721. https://doi.org/10.1007/s00417-022-05884-3
32. Ahmed MI, Spooner B, Isherwood J, Lane M, Orrock E, Dennison A (2023) A systematic review of the barriers to the implementation of artificial intelligence in healthcare. Cureus 15:e46454. https://doi.org/10.7759/cureus.46454
33. Tan TF, Thirunavukarasu AJ, Jin L, Lim J, Poh S, Teo ZL, Ang M, Chan RVP, Ong J, Turner A, Karlström J, Wong TY, Stern J, Ting DSW (2023) Artificial intelligence and digital health in global eye health: opportunities and challenges. Lancet Glob Health 11:e1432–e1443. https://doi.org/10.1016/S2214-109X(23)00323-6
34. Campo-Beamud C, Adan Ruiz A, Bastante Quijano J, Campo Beamud E, Gómez-Romero FJ, Fernández Ruíz AJ, Copete S (2026) Publicly available multimodal large language models for ocular surface infections: benchmarking against corneal specialists in triage, diagnosis and treatment. Br J. https://doi.org/10.1136/bjo-2025-328867. Ophthalmol:bjo-2025-328867

Figures

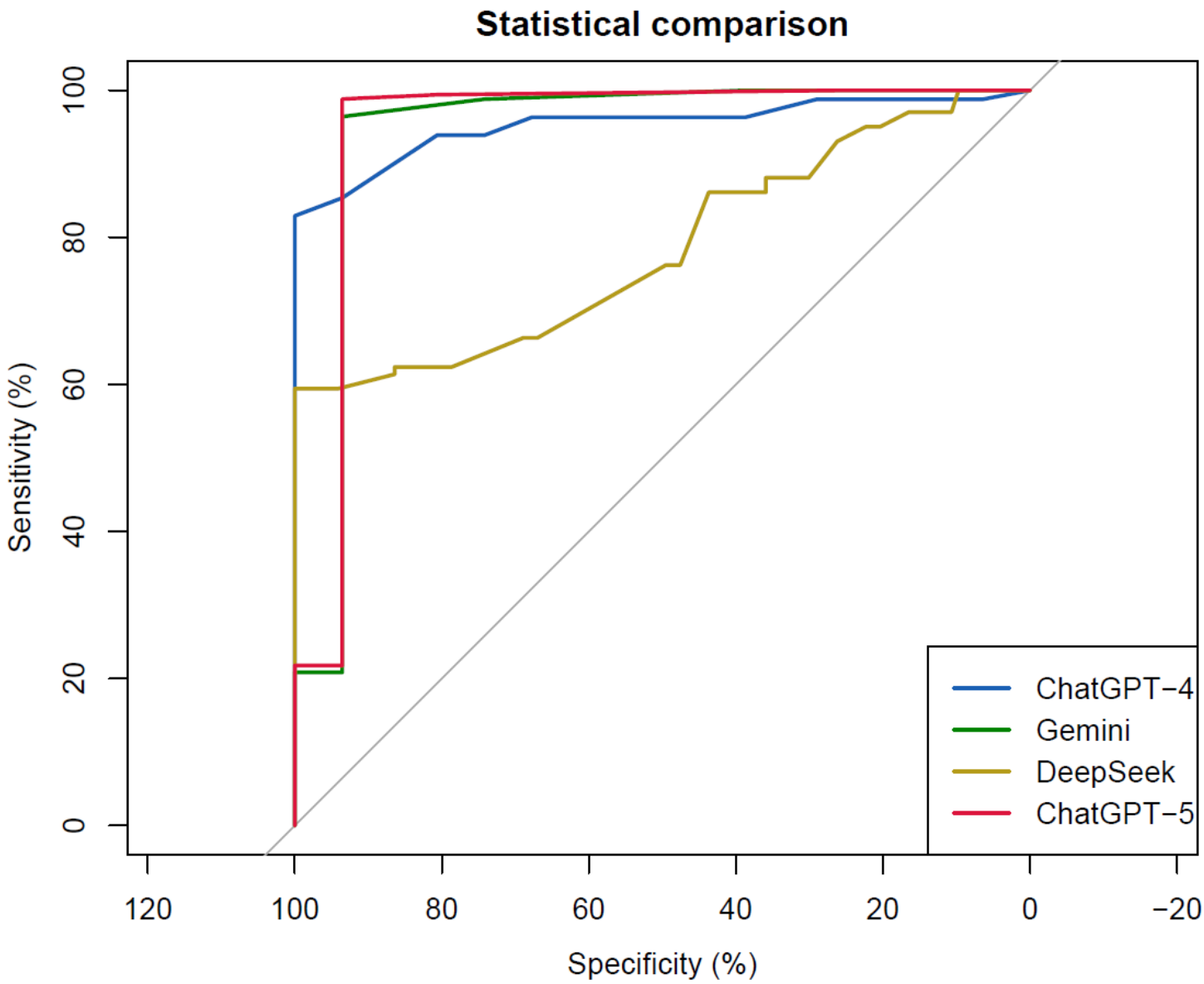


Figure 1

Receiver operating characteristic (ROC) curves for chatbot-based risk estimation of recurrent rhegmatogenous retinal detachment (reRRD). ROC curves showing model-level discrimination for reRRD within 12 months. ChatGPT-4o, GPT-5, and Gemini showed outstanding performance (AUC = 0.953, 0.947, and 0.942, respectively), whereas DeepSeek showed lower but still acceptable discrimination (AUC = 0.785). AUC = area under the curve; reRRD = recurrent rhegmatogenous retinal detachment.

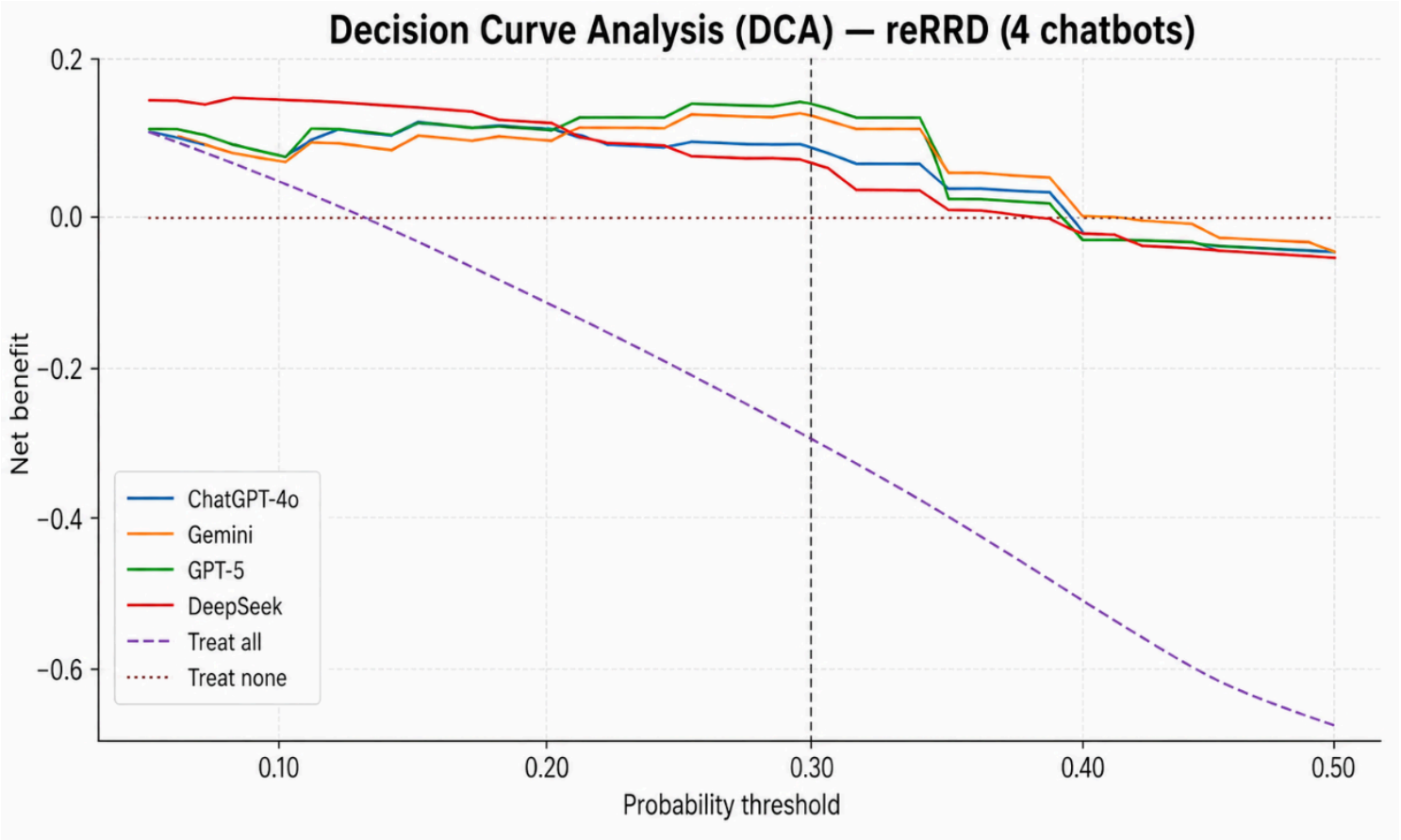


Figure 2

Decision curve analysis (DCA) for recurrent rhegmatogenous retinal detachment (reRRD) risk forecasts by the four evaluated chatbots. Net benefit is shown across a range of threshold probabilities for ChatGPT-4o, Gemini, GPT-5, and DeepSeek, alongside the default “treat all” and “treat none” strategies. The vertical dashed line marks the 30% probability threshold selected for operational risk stratification, supported by concordant findings from the Youden index and DCA. At this threshold, GPT-5 and Gemini showed the most favorable balance between clinical utility and discrimination.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Materialcomplemt.docx