

LLM-Based Phenotyping of Glaucoma from Unstructured Clinical Notes

Nicholas Solages BS¹, Gustavo A. Samico, MD^{1,2}, Rafael Scherer MD, PhD¹, Naomi E. Gutkind, MD^{1,3}, Janet Kang BS⁴, Felipe A. Medeiros, MD, PhD¹, Swarup S. Swaminathan, MD¹

1. Bascom Palmer Eye Institute, University of Miami Miller School of Medicine, Miami, FL
2. Department of Ophthalmology and Visual Sciences, Escola Paulista de Medicina Universidade Federal de São Paulo, São Paulo, SP, Brazil
3. Olive View – UCLA Medical Center, Sylmar, CA
4. Miller School of Medicine, University of Miami, Miami, FL

Background

Electronic health records (EHRs) are the primary data source for observational research within the OHDSI network, yet structured OMOP Common Data Model (CDM) fields frequently fail to capture clinically precise phenotypic information. Glaucoma research illustrates this limitation acutely: cohort definitions have historically relied on ICD-10-CM billing codes optimized for administrative workflows rather than diagnostic accuracy, producing imprecise characterization of glaucoma diagnosis, subtype, and disease severity. Inaccurate phenotyping can distort prevalence estimates, compromise cohort validity, and impair the reliability of stage-stratified or treatment-comparative analyses.

Prior NLP approaches to ophthalmology clinical notes have relied on task-specific pipelines, domain-specific pretraining (e.g., OphthaBERT), or conventional rule-based extraction, limiting their scalability and generalizability across institutions. Large language models (LLMs) offer a prompt-engineering paradigm capable of complex, contextual inference from free-text clinical narratives without specialty-specific model development, making them a scalable alternative for improving the reliability of OMOP-structured phenotypic data.

Methods

We conducted a retrospective analysis of 1,242 clinical notes from glaucoma-related encounters (2014–2024) extracted from the Bascom Palmer Ophthalmic Repository, containing data from over 70,000 patients. Notes were filtered to outpatient encounters with finalized documentation and glaucoma-related ICD codes (H40.X), excluding notes under 200 characters. Two fellowship-trained glaucoma specialists independently adjudicated each note across three extraction tasks: (1) glaucoma diagnosis (glaucoma / suspect / neither), (2) glaucoma type (17 categories including POAG, NTG, PXG, PACG etc.), and (3) disease severity (mild / moderate / severe / other). Discordant cases were resolved by a third specialist adjudicator. Inter-grader agreement was quantified using Gwet's AC1.

The dataset was partitioned into development (10%), validation (10%), and test (80%) sets, stratified by clinician to maintain proportionally even representation of each of the 20 contributing clinicians' writing styles and documentation conventions across partitions. Since no fine-tuning was performed, development and validation sets supported prompt engineering only, prioritizing a larger held-out test set for a stable performance estimate; prompts were frozen before testing (final prompts provided in Supplementary Materials). Stratification was applied at

the clinician level only, as balancing all three extraction tasks' subclasses simultaneously — particularly low-prevalence type/severity categories (<5%) — was not achievable within one partition scheme. Five state-of-the-art LLMs — Claude Opus 4.6, DeepSeek-V3.2, GPT-5.2, Grok 4.1, and Qwen 3.6-35B-A3B — were accessed via Microsoft Azure AI Foundry within HIPAA-compliant containers. Model performance was benchmarked using accuracy, sensitivity, specificity, and F1-score. ICD-10-CM codes (the structured-data equivalent available in OMOP-standardized datasets) and ClinicalBioBERT (a prior-generation clinical transformer, evaluated on identical raw unprocessed notes) were assessed against the same reference standard.

Results

Inter-grader agreement was strong across all three tasks: Gwet's AC1 = 0.930 (95% CI: 0.917–0.945) for diagnosis, 0.917 (95% CI: 0.904–0.930) for type classification, and 0.901 (95% CI: 0.884–0.916) for severity staging, establishing a high-quality reference standard for downstream benchmarking.

For glaucoma diagnosis, LLM accuracy ranged from 94.4% to 97.5% across models. Claude achieved the highest accuracy (97.5%; 95% CI: 96.7–98.1%), followed by GPT (96.2%), DeepSeek (96.0%), Qwen (95.5%), and Grok (94.4%). F1-scores for glaucoma detection ranged from 95.4% to 98.9%.

For glaucoma type classification (15 categories), LLM accuracy ranged from 94.0% to 97.1%, with Claude leading (97.1%). ICD-10-CM codes demonstrated a substantially lower overall accuracy of 88.9% and aggregate F1-score of 63.8%. Per-category ICD diagnostic performance is presented in Table 1; common subtypes (POAG, NTG) were relatively well-captured, while rarer subtypes (uveitic, angle-recession, PACG) showed markedly lower positive predictive value. Encounters ICD-coded as generic 'Suspect' showed bidirectional information loss: 9.8% (38/389) of such encounters had a specific glaucoma type documented in the note that ICD coding collapsed into a non-specific designation, while conversely, 74.5% (458/615) of encounters with no specific type documented in the note were nonetheless assigned a specific ICD type not clearly supported by the note.

For severity staging, LLM accuracy ranged from 94.0% to 95.2% with F1 scores reaching 96.3%. ICD-10-CM codes yielded only 69.2% overall accuracy. For moderate disease specifically, ICD coding achieved an F1-score of just 47.2%, reflecting a positive predictive value of 35.9%. Per-category ICD diagnostic performance is presented in Table 2; misclassification of moderate cases was disproportionately concentrated in the "Other" (19.8%) and "Mild" (13.2%) categories, illustrating this granularity loss.

ClinicalBioBERT, evaluated on identical raw notes, performed well below all LLMs across every task: 68.0% accuracy (macro F1 0.44; Cohen's κ : 0.31) for diagnosis, failing entirely on "Neither" (F1: 0.00); 52.0% accuracy (macro F1 0.09) for type, with 12/17 categories at F1 0.00; and 53.0% accuracy (macro F1 0.22) for severity, with mild/moderate/severe each at F1 0.22–0.24.

Conclusion

Secure cloud-based LLMs substantially outperform ICD-10-CM codes — and are expected to outperform prior-generation clinical NLP transformers — in extracting glaucoma diagnosis, type, and severity from free-text EHR notes. The magnitude of the ICD-severity gap (69.2% vs. ~95% for LLMs) has direct implications for OHDSI: cohorts built from structured billing fields alone risk systematic phenotyping errors that propagate through downstream comparative effectiveness and risk prediction analyses. Because the study cohort was defined by pre-existing H40.X ICD codes (including suspect-level codes, which are required to justify glaucoma-related testing or treatment), reported sensitivity and PPV reflect performance conditional on ICD-based case ascertainment. Given the largely asymptomatic nature of glaucoma, uncoded true or false cases are expected to be rare, though formal evaluation in a non-ICD-filtered cohort remains a direction for future work.

These results reflect an operational constraint of BERT-class architectures: without curated preprocessing, section segmentation, and class re-balancing strategies, their representational capacity does not transfer to the complexity and distributional imbalance of raw, multi-category clinical text. Critically, this is not a methodological limitation, but rather the central point of comparison — the preprocessing overhead BERT-class models require represents a substantial deployment barrier for scalable phenotyping pipelines, while the LLM pipeline achieved 29.5, 45.1, and 42.2 percentage-point gains over ClinicalBioBERT for diagnosis, type, and severity, respectively, through prompt engineering alone.

By converting unstructured notes into reproducible phenotypic variables, LLM-based workflows can improve cohort construction across multi-site studies without specialty-specific model development, operating within HIPAA-compliant infrastructure and deployable through prompt engineering alone — lowering barriers to reliable real-world evidence generation from routine clinical documentation.

Table 1. Diagnostic performance of ICD-10-CM codes for glaucoma type classification.

Glaucoma Type	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1-Score (%)	Accuracy (%)
POAG	90.5 (88.3 - 92.3)	89.9 (88.0 - 91.5)	85.8 (83.3 - 88.0)	93.3 (91.7 - 94.6)	88.1 (86.4 - 89.6)	90.1 (88.7 - 91.4)
NTG	89.0 (82.4 - 93.3)	99.0 (98.5 - 99.4)	86.3 (79.3 - 91.1)	99.2 (98.7 - 99.6)	87.6 (83.0 - 91.1)	98.4 (97.7 - 98.9)
PACG	84.8 (76.7 - 90.4)	98.1 (97.4 - 98.6)	71.2 (62.7 - 78.4)	99.1 (98.6 - 99.5)	77.4 (71.6 - 82.3)	97.4 (96.6 - 98.0)
Pseudoexfoliative Glaucoma	69.1 (59.3 - 77.4)	99.7 (99.3 - 99.9)	91.8 (90.5 - 92.7)	98.4 (97.8 - 98.9)	78.8 (72.1 - 84.3)	98.2 (97.5 - 98.7)
Pigmentary Glaucoma	70.0 (48.1 - 85.5)	99.9 (99.6 - 99.9)	87.5 (64.0 - 96.5)	99.7 (99.3 - 99.9)	77.8 (61.9 - 88.3)	99.6 (99.2 - 99.8)
Angle Recession Glaucoma	75.0 (40.9 - 92.9)	99.9 (99.6 - 100.0)	66.7 (35.4 - 87.9)	99.9 (99.6 - 100.0)	70.6 (46.9 - 86.7)	99.8 (99.4 - 99.9)
Uveitic Glaucoma	71.4 (50.0 - 86.2)	99.4 (99.0 - 99.7)	57.7 (39.0 - 74.5)	99.7 (99.3 - 99.9)	63.8 (49.5 - 76.0)	99.1 (98.6 - 99.5)

95% confidence intervals are reported in parentheses. Absent documentation in either gold standard or ICD code was considered negative for the specific condition. ICD-10-CM = International Classification of Diseases, 10th Revision, Clinical Modification; NPV = negative predictive value; NTG = Normal Tension Glaucoma; PACG = Primary Angle-Closure Glaucoma; PPV = positive predictive value; POAG = Primary Open-Angle Glaucoma

Table 2. Diagnostic performance of ICD-10-CM codes for glaucoma severity classification.

Severity	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1-Score (%)	Accuracy (%)
Mild	70.8 (65.6 - 75.6)	91.6 (90.2 - 92.9)	61.2 (56.1 - 66.1)	94.4 (93.2 - 95.4)	65.7 (62.0 - 69.2)	88.4 (86.9 - 89.7)
Moderate	68.7 (61.6 - 75.0)	87.6 (86.0 - 89.1)	35.9 (31.1 - 41.1)	96.5 (95.5 - 97.3)	47.2 (43.0 - 51.4)	85.9 (84.3 - 87.4)
Severe	71.4 (66.5 - 75.9)	95.1 (93.9 - 96.0)	76.1 (71.3 - 80.4)	93.8 (92.6 - 94.9)	73.7 (70.3 - 76.9)	90.8 (89.5 - 92.0)
Other	68.1 (65.4 - 70.8)	80.3 (77.5 - 82.8)	82.1 (79.6 - 84.4)	65.4 (62.5 - 68.3)	74.5 (72.6 - 76.3)	73.3 (71.4 - 75.2)

95% confidence intervals are reported in parentheses. Absent documentation in either gold standard or ICD code was considered negative for the specific condition. ICD-10-CM = International Classification of Diseases, 10th Revision, Clinical Modification; NPV = negative predictive value; PPV = positive predictive value.

References

1. Higgins BE, Leonard-Hawkhead B, Azuara-Blanco A. Quality of reporting electronic health record data in glaucoma: a systematic literature review. *Ophthalmol Glaucoma*. 2024;7(5):422-30.
2. Liebman DL, Johnson GE, Marte A, et al. Accuracy of ICD-10 glaucoma codes in a large academic practice. *Ophthalmol Glaucoma*. 2025;8(6):643-5.
3. Shah R, Moradi M, Eslami S, et al. OphthaBERT: automated glaucoma diagnosis from clinical notes. *medRxiv*. 2025.
4. Stein JD, An HS, Andrews CA, et al. Enhanced phenotype identification of common ocular diseases in real-world datasets. *Ophthalmology Science*. 2025;5(4).
5. Acuff KN, Radha Saseendrakumar B, Weinreb RN, Baxter SL. Assessment of missing data on glaucoma severity among participants in the NIH All of Us Research Program. *J Glaucoma*. 2025;34(1):39-46.