Identification of Risk Factors for Glaucoma Progression in Free-Text Clinical Notes using a Local Large Language Model

Anshul Bhatnagar, MD;[1] Rafael Scherer MD, PhD;[1] Gustavo A. Samico, MD;[1,2] Rohit Muralidhar, MD;[1] Naomi E. Gutkind, MD;[1] Vitoria Palazoni, MD;[1] Felipe A. Medeiros, MD, PhD[1]; Swarup S. Swaminathan, MD[1]

Affiliations:
[1] Bascom Palmer Eye Institute, University of Miami, Miami, FL, USA
[2] Department of Ophthalmology and Visual Sciences, Escola Paulista de Medicina Universidade Federal de São Paulo, São Paulo, Brazil

Meeting Presentation: Submitted for presentation at American Glaucoma Society Annual Meeting 2026, Association for Research in Vision and Ophthalmology Annual Meeting 2026

Financial Support: NIH EY036593 (FAM), NIH K23 EY033831 (SSS)

Conflicts of Interest: None

Financial Disclosures: AB: none. RS: Redcheck (F), Eyetec SlitSmart (P). RM: none. GAS: none. RM: none. NEG: none. VP: none. FAM: Abbvie (C), Annexon (C); Astellas (C); Carl Zeiss Meditec (C), Enavate Sciences (C), Galimedix (C); Heidelberg Engineering (F); InjectSense, Inc. (C), nGoggle Inc. (P), Novartis (F); ONL Therapeutics (C), Perfuse Therapeutics (C), Perceive Bio (C), Stealth Biotherapeutics (C); Stuart Therapeutics (C), Thea Pharmaceuticals (C), Reichert (C, F). SSS: Abbvie (C), Elios Vision (C), Lumata Health (C, E)

Running Header: LLM for Glaucoma Risk Factors from Notes

Correspondence: Swarup S. Swaminathan, MD, Bascom Palmer Eye Institute, 900 NW 17th St, Miami, FL. E-mail: sswaminathan@med.miami.edu

Keywords: large language model, artificial intelligence, electronic health records, glaucoma, risk factors

Word Count: 3,575

Abstract

Purpose: To evaluate the performance of a large language model (LLM) in identifying medication non-adherence, visit non-adherence, and family history of glaucoma (FHoG) in clinical notes from the electronic health record (EHR).

Methods: We extracted clinical notes of 1,250 glaucoma-related encounters between 2014 and 2024 and structured EHR family history field data from the Bascom Palmer Ophthalmic Repository, with 125 randomly selected notes (10%) used for prompt development and excluded from analysis. Two fellowship-trained glaucoma specialists labeled notes for evidence of non-adherence and FHoG. We utilized MedGemma-27B-text-it, a specialized medical LLM, to identify medication non-adherence, visit non-adherence, and FHoG. We calculated accuracy, sensitivity, and specificity of LLM performance for each task, Jaccard index for FHoG, and mean squared error (MSE) of number of family members with glaucoma.

Results: Prevalence of medication non-adherence, visit non-adherence, and FHoG were 7.3%, 4.7%, and 29.2%, respectively. LLM accuracy was 0.91 (sensitivity: 0.96; specificity: 0.91) for medication non-adherence and 0.96 (sensitivity: 0.97; specificity: 0.94) for visit non-adherence. For FHoG, LLM accuracy was 0.98 (sensitivity: 0.99; specificity: 0.99) with Jaccard index of 0.99, while EHR family history field accuracy and Jaccard index were 0.49 and 0.75, respectively. LLM and EHR MSE in quantifying the number of relatives with glaucoma were 0.05±0.56 and 0.85±1.80, respectively ($p<0.001$).

Conclusions: LLMs identified non-adherence to medication and visit schedules as well as degree of FHoG in clinical notes with high accuracy.

Translational Relevance: Local LLM pipelines can enable large-scale research into glaucoma risk factors that are unavailable in discrete EHR fields.

Introduction

Glaucoma is a leading cause of irreversible blindness in the United States and abroad.[1,2] Given glaucoma progresses asymptomatically, longitudinal monitoring and treatment are essential to prevent disease worsening. Key risk factors for progressive disease include older age, African ancestry, Latino ethnicity, elevated intraocular pressure (IOP), and thin central corneal thickness.[2] Medication non-adherence, visit non-adherence, and family history of glaucoma (FHoG) are also significantly associated with glaucoma disease progression. Studies have shown that glaucoma patients who are lost to follow-up or unable to follow prescribed medication regimens have significantly poorer visual outcomes compared with treatment-adherent patients.[3–6] Newman-Casey et al demonstrated that in a randomized clinical trial, glaucoma patients who were non-adherent to medications had a faster rate of visual field loss compared to adherent patients.[4] Separately, genetic and familial studies have shown that a FHoG can drastically increase the risk of glaucoma and worse visual outcomes, and that this risk rises exponentially when more relatives are diagnosed with glaucoma. Patients with one sibling with glaucoma had an increased glaucoma incidence risk of 2.31, while those with four or more affected siblings with glaucoma had an increased incidence risk of 26.66.[7,8]

Electronic health record (EHR) systems have created a unique opportunity to develop large real-world datasets for research endeavors. EHR phenotyping refers to the concept of developing patient cohorts with a specific condition; additional extracted patient data can further characterize phenotypes. Data regarding specific glaucoma risk

factors reside in discrete structured data elements in EHRs, such as age, self-reported ethnicity, tonometry, and pachymetry.[9–11] These discrete data can be extracted efficiently, but key glaucoma risk factors such as medication and visit non-adherence are buried in unstructured clinical notes, complicating large-scale analysis. Information about these risk factors can currently only be gathered through manual chart review, which is error prone and impractical in larger clinical research initiatives.[12] Although EHRs have formal medication lists or family history elements, research has repeatedly shown that such fields are often incorrect or outdated, and that unstructured clinical notes are often the most accurate source.[13–16] Similarly, visit attendance data in EHRs is unreliable, as patient attendance is not well described by no-show percentages.

Prior studies have attempted to utilize natural language processing to study clinical notes, including work on extracting examination findings and diagnosis codes from clinician notes.[17] Natural language models (NLM) are artificial intelligence algorithms that are trained not just to identify fields but also interpret text. Given that approximately 80% of all EHR data exists in an unstructured, narrative form,[18] NLM models are becoming increasingly recognized for their ability to distill key information from free-text notes.[17] Prior NLM research has focused on the use of large language models (LLMs), such as ChatGPT and BERT, which can effectively interpret ophthalmic information and free-text.[19] Nonetheless, most LLMs face constraints in clinical settings involving protected health information (PHI), primarily because they require cloud-based processing and data sharing via public application programming interfaces. In countries with strict health privacy laws such as the United States, this limitation may preclude

widespread adoption when analyzing clinical notes, as such text may contain PHI. Data security is paramount in such instances.

In contrast, a medium-sized LLM can run locally within a secure network, which is crucial when handling sensitive text.[20] In this study, we aimed to develop a local LLM pipeline that could accurately determine whether glaucoma patients had a history of medication non-adherence, visit non-adherence, or FHoG per free-text clinical notes. Such a tool could not only help researchers rapidly extract these important risk factors from EHRs for real-world data research, but also facilitate risk-stratification of patients in a clinical setting.

Methods

Data Collection, Note Sampling, and Verification

This study was approved by the Institutional Review Board at the University of Miami. The requirement for informed consent was waived because of the retrospective nature of the study. The procedures and protocols followed during the study adhere to the Declaration of Helsinki and comply with the Health Insurance Portability and Accountability Act (HIPAA) for maintaining patient confidentiality and integrity.

This study used the Bascom Palmer Ophthalmic Repository (BPOR), which contains the data of over 70,000 patients evaluated for glaucoma at the Bascom Palmer Eye Institute (BPEI) in Miami, Florida.[21] The database contains demographics, testing

data, EHR-derived measurements, and clinical notes for included subjects. All data were extracted from Epic (Epic Systems, Verona, WI). We identified all signed clinical notes associated with an outpatient encounter that were written by a fellowship-trained glaucoma specialist between 2014 and 2024. Notes were only included if associated with an International Classification of Diseases (ICD)-10 glaucoma visit diagnosis (H40.X). Notes were excluded if the encounter type was hybrid, preoperative, postoperative, procedural, or consult given likely minimal detail regarding longitudinal glaucoma care. Notes that were fewer than 200 characters were also excluded. A total of 1,250 notes were randomly selected stratified by clinician to ensure equal representation of note styles. During note verification, eight notes were found to be attestations for trainee clinical visits and thus were excluded from further analysis. Only one note per subject was included. For each subject, the formal structured family history section was extracted directly from the EHR and filtered for mention of “glaucoma”. Notes were stored on a HIPAA-compliant secure Box server (Box, Redwood City, CA), which is authorized by the University of Miami to store PHI.

Note Labeling and Prompt Engineering

A total of 125 randomly selected notes (10% of the full dataset) stratified by clinician were used to guide prompt engineering and were subsequently excluded from further analysis. These notes were used to create “few shot” examples, provide guidance regarding approaches to handling conflicting data, and generate a list of abbreviations observed in notes. Text of the three prompts is provided in the **Supplemental Material**.

Two fellowship-trained glaucoma specialists (GAS and NEG) labeled the remaining 90% of notes for evidence of current medication non-adherence (yes/no or not mentioned), visit non-adherence (yes/no or not mentioned), and FHoG (yes/no/not mentioned). For the non-adherence tasks, “not mentioned” was grouped with “no” (i.e., binary classification), as physicians would likely only document cases of non-adherence. For the FHoG task, “not mentioned” was considered as a distinct category under the assumption that providers would actively document both positive and negative FHoG. If a positive FHoG was documented, graders were asked to identify the family members with glaucoma, if discussed in the note. In cases of disagreement between the graders, a third fellowship-trained glaucoma specialist (SSS) served as an adjudicator. This process produced a ‘ground truth’ standard for each extraction task to subsequently assess model performance.

Model Architecture, Setup, and Processing

We used MedGemma-27B-text-it, a specialized medical LLM developed by Google (Alphabet Inc., Mountain View, CA). This artificial intelligence model has been specifically trained to analyze medical text, such as clinician notes.[22] To ensure no exposure of PHI, this LLM was installed and run on a local computer in a secure location at BPEI and housed within a virtual environment. All notes were processed with their corresponding encounter dates to provide temporal context. The model was implemented using the Hugging Face Transformers library with PyTorch backend. To maximize computational speed and efficiency, the model was configured to run with bfloat16 precision instead of float32. To ensure deterministic outputs, meaning that the

model should return the same output when given the same input across devices or time, we disabled memory-efficient and flash self-attention mechanisms; we enabled mathematical self-attention instead.

We modified the key generation parameters (temperature=0.4, top_p=0.9, max_new_tokens=200) of the model to balance output consistency with appropriate variability. Generated outputs were parsed to extract structured JavaScript Object Notation (JSON) data with error handling for any malformed responses. The model was run without any formal training or fine-tuning to avoid significantly changing the native behavior of the model. Output tokens were reviewed, particularly in cases of false negatives and false positives, to provide insight into the model's rationale for classification.

Statistical Analysis

Model outputs were compared against the adjudicated standard for each note. To determine model performance for each data extraction task, accuracy, sensitivity, and specificity were calculated. For FHoG, overall accuracy was calculated across all three valid responses. The Jaccard index was calculated for FHoG to assess family member matching. The Jaccard index evaluates partial and exact matches between the predicted and true classifications, measuring similarity by dividing the size of their intersection by the size of their union. Regarding relatives with glaucoma, we converted the listed family members into a numeric estimate from the graded labels, LLM output, and EHR structured field. If the output was "Yes" with no family members listed, the value was listed as one. If the output included some indication of plurality (e.g.,

“multiple,” “cousins”) with no further details, the value was listed as two. We then calculated the mean squared error (MSE) of the LLM and EHR field when compared to the graded labels to determine their accuracy. Only subjects with a positive family history per the graded label were included in the MSE calculations to avoid overestimating accuracy of this task (i.e., an artificially low MSE estimate).

Gwet’s AC1 and the raw agreement rate were calculated to evaluate concordance rates between graders. Gwet’s AC1 is a useful assessment metric and alternative to Cohen’s kappa coefficient when prevalence of a condition is low. The Wilson score interval was used to calculate 95% confidence intervals (CI) for accuracy, sensitivity, and specificity. For Gwet’s AC1, 95% confidence intervals were estimated using bootstrap resampling with 1,000 iterations. MSE of LLM versus EHR structured field assessments were compared using the nonparametric Wilcoxon signed-rank test. A p-value <0.05 was considered statistically significant. All statistical analyses were conducted using Stata version 18 (StataCorp, College Station, TX, USA).

Results

A total of 1,117 glaucoma-related encounter notes were used to analyze model performance. The LLM was able to successfully evaluate all notes without computational error. Prevalence of medication and visit non-adherence was 7.3% (n=81) and 4.7% (n=53), respectively, based on graded labels. Clinicians documented FHoG in 29.2% (n=326) of all subjects. Inter-rater reliability was excellent across all

assessments. For medication non-adherence, Gwet's AC1 was 0.914 (95% CI, 0.887–0.936) with a raw agreement of 95.7%. For visit non-adherence, the AC1 was 0.961 (95% CI, 0.943–0.977) with a raw agreement of 98.0%. For FHoG, the AC1 was 0.913 (95% CI, 0.886–0.928) with a raw agreement of 93.5%.

For classifying medication non-adherence, the LLM had an accuracy of 0.91 (95% CI: 0.89-0.93), with sensitivity and specificity of 0.96 (95% CI: 0.90-0.99) and 0.91 (95% CI: 0.89-0.92), respectively (**Figure 1**). For classifying visit non-adherence, the LLM had an accuracy of 0.96 (95% CI: 0.95-0.97), with sensitivity and specificity of 0.97 (95% CI: 0.95-0.97) and 0.94 (95% CI: 0.84-0.98), respectively (**Figure 2**).

For FHoG, LLM accuracy was 0.98 (95% CI: 0.97-0.99; **Figure 3A**), with sensitivity of 0.99 (95% CI: 0.98-0.99) and specificity of 0.99 (95% CI: 0.98-0.99). In terms of family member matching, the Jaccard index was 0.99 (95% CI: 0.98-0.99). In contrast, the structured field for family history in the EHR had an accuracy of only 0.49 (95% CI: 0.46-0.52; **Figure 3B**), significantly lower than that of the LLM ($p<0.001$). Sensitivity and specificity of the EHR structured field were 0.76 (95% CI: 0.71-0.80) and 0.81 (95% CI: 0.78-0.83), respectively. The Jaccard index was 0.75 (95% CI: 0.73-0.78).

According to the clinicians' assessment, subjects with positive FHoG had a median of one affected relative (range: 1-4; **Figure 4**). The MSE of the LLM when estimating the number of relatives with glaucoma was 0.05±0.56. The MSE of the EHR structured field was much larger, 0.85±1.80 ($p<0.001$). In 2.1% of notes, the graded label noted a negative FHoG but the subject was noted to have a positive FHoG per the

structured EHR field. In 11.5% of notes, the graded label for FHoG in the clinical note was “not mentioned” but the subject had a positive FHoG per the structured EHR field. **Table 1** contains all performance metrics of all models for comparison.

Discussion

We demonstrate in this study that a tailored, secure, locally-installed LLM pipeline can perform key extraction tasks from EHR-derived clinical notes with high accuracy. The model was 91% accurate at identifying medication non-adherence, 96% accurate at identifying visit non-adherence, and 97% accurate at detecting FHoG. These risk factors are associated with worse glaucoma outcomes but have traditionally been difficult to study in glaucoma research due to their inaccessibility. By converting unstructured, narrative information to discrete, structured data that can be batch-extracted and analyzed, LLMs have the potential to aid not only research endeavors but also identify high-risk glaucoma patients in large practice settings. To our knowledge, this work represents the first study to develop a locally-installed LLM pipeline that identifies and interprets risk factor data from free-text clinical notes in ophthalmology.

The LLM was highly sensitive and specific (0.96 and 0.91, respectively) for the detection of medication non-adherence within clinical notes. Similarly, the model performed with high sensitivity and specificity (0.97 and 0.94, respectively) in identifying visit non-adherence in clinical notes. **Table 2** contains examples of false positive and false negative misclassifications for both medication and visit non-adherence. Medication adherence misclassification appeared to be due to etiologies such as difficulty in interpreting past versus current non-adherence, lack of clarity regarding

whether a patient self-discontinued a medication due to side effect or without reason, or an elevated IOP with a discussion of the importance of compliance in the clinician's note. Visit adherence misclassification was often due to misinterpretations regarding changes in clinical providers, missed follow-up with non-glaucoma clinicians, or lack of clarity regarding where the patient was supposed to follow-up (e.g., at BPEI or with an external provider). In contrast to FHoG classification, the non-adherence tasks required greater interpretation of note text, naturally leading to reduced accuracy (**Table 1**). While false positives were more common than false negatives, a false positive (i.e., declaring non-adherence when patient was adherent) is likely less consequential than a false negative (i.e., declaring adherence when patient was non-adherent). To this end, the sensitivities of these models remain robust, reflecting low false negative rates. One could also argue that the false positives in each task may actually identify patients whose actions suggest potential adherence challenges (e.g., forgetting use of drops on day of clinic visit), albeit they were not formally classified as non-adherent by graders.

Medication non-adherence is challenging to assess using EHR medication lists, as these fields are often inaccurate for ophthalmic medications. One study of glaucoma patients demonstrated that a third of all patients had at least one glaucoma medication mismatch between the formal EHR medication list and the clinical text note.[15] These findings suggest that ophthalmologists and optometrists do not update EHR medication lists frequently, despite 90% of ophthalmologists acknowledging the importance of assessing glaucoma medication adherence during clinic visits.[23] Similarly, claims databases may not accurately reflect medication adherence, as patients may fill

prescriptions as requested but still use medications incorrectly.[24] Consequently, free-text clinical notes are likely the most accurate source for such information.

Visit non-adherence is another significant risk factor for glaucoma progression that is often understudied and challenging to track through structural mechanisms in EHRs. Although raw no-show percentage (number of no-show visits / number of scheduled clinic visits), can be extracted from EHRs, this statistic does not always accurately reflect true visit adherence. Such percentages can be deceivingly low. For example, glaucoma patients that are lost to follow-up for years after missing just one visit will have a no-show percentage that does not reflect their extended lack of follow-up, which puts them at significant risk of disease progression.[25] Second, such statistics are often calculated across multiple specialties in a hospital system within the EHR and thus may not be specific to glaucoma visit adherence, which can be significantly decreased compared to that of other ophthalmic subspecialties.[26] Again, clinic notes may serve as the most accurate source of information regarding visit adherence specific to glaucoma care. Analysis of these additional patient dimensions via LLMs could facilitate further characterization of glaucoma phenotypes in EHR data.[27] In addition, larger organizations could create registries of high-risk glaucoma patients identified using such an LLM pipeline. These individuals may benefit from specialized and targeted interventions to improve visit adherence, such as personalized letters, telephone reminders, and social worker support.[28,29] Use of these registries could lead to improved visual outcomes for patients that might otherwise be lost to follow-up in larger practices.

FHoG is another key risk factor for glaucoma progression that is purportedly tracked within structured family history EHR fields. While such fields can be easily extracted, prior work in neurology and obstetrics/gynecology has shown that such lists are often inaccurate and infrequently updated.[14,30] Our study confirmed that with respect to glaucoma, EHR structured family history data were accurate less than half the time. In contrast, only 13.6% of patients had a positive FHoG per the structured EHR field that was missed in the clinical note (labeled as “no” or “not mentioned”). It is important to note that since the clinical note was treated as the ground truth, the LLM output had a potential advantage over the EHR field, which could make such comparisons challenging. However, given ophthalmologists and optometrists typically use their notes to track relevant family history details and are less likely to use the structured EHR field, our findings suggest that the EHR field may not be the optimal source of eye-related family history, particularly for research purposes.

In contrast, the LLM model exhibited high sensitivity and specificity (0.99 for both) in detecting FHoG from clinical notes. While the non-adherence tasks required more interpretation, FHoG assessment was likely an easier task involving the identification and reporting of relevant information, leading to strong model performance (**Table 1**). When assessing the number of family members with glaucoma, the LLM maintained excellent accuracy (MSE 0.05±0.56). This finding is particularly important given that the risk of glaucoma rises exponentially with the number of affected relatives with glaucoma.[8] LLMs may allow for accurate, quantified assessments of FHoG, which can provide valuable details for research work in glaucoma genetics or disease

progression assessments. EHR glaucoma phenotypes could potentially incorporate quantification of FHoG to yield more specific cohorts.

NLMs have become increasingly popular in medicine for their ability to interpret free narrative text, although their use thus far for ophthalmic purposes has been limited.[17] Prior ophthalmic research has used NLMs to predict glaucoma progression, identify cataract surgery complications, determine slit lamp-examination findings, and recognize ophthalmic diagnoses from clinic notes.[17,31–36] However, fewer studies have used NLMs to examine risk factors for glaucoma progression. Lin et al. used a natural language processing tool to evaluate if medication adherence was mentioned in clinical notes.[15] However, this smaller study differs from our analysis as it solely identified the presence of adherence information; no further interpretation of the adherence text was pursued.

In contrast to the focus on cloud-based LLMs in medical research, our use of a nimble, locally installed LLM provides various advantages. Most critically, the latter can be operated on a local network without the need for cloud architecture.[20] LLMs often require data to be shared across networks via third-party application programming interfaces. Many countries such as the United States have strict laws regarding PHI, which could limit widespread use of cloud-based LLMs in clinical or health research processes.[37] Our study demonstrated that a locally-installed LLM can achieve excellent performance in such tasks, even without extensive model training. Studies traditionally use a train/test approach to evaluate model performance, but accurate assessments can be challenging when data are sparse and prevalence of the target condition is

low.[38–40] In this work, we utilized a “few-shot” approach, with 10% of notes used for prompt engineering and the remainder utilized to evaluate model performance. A local LLM pipeline is uniquely suited to implementation within clinical workflows when working with sensitive data.

Our study has natural limitations. First, the notes analyzed in this study were from one institution, which may limit the generalizability of our LLM pipeline. However, it is important to be mindful of institution-specific writing styles and abbreviations in clinical notes, which may limit the accuracy of using the same prompt at different institutions. Rather, our work suggests that individual institutions may need to follow a similar “few-shot” approach to capture institution-specific writing patterns. This work and the prompts provided in this study may serve as a foundation for initial testing at different institutions with further prompt engineering. Future work could also involve extracting a standardized output from a LLM pipeline run locally at individual institutions, which could be incorporated into larger multi-institutional studies regarding glaucoma progression. Second, the ideal LLM can only be as accurate as clinician notes, which may miss key information discussed during patient encounters.[41,42] If clinicians are prone to under-reporting adherence challenges in their notes, the LLM will follow suit.

This study demonstrated that a domain-adapted, locally-installed LLM pipeline can identify medication non-adherence, visit non-adherence, and FHoG including quantification of family members with high accuracy. Data regarding these risk factors for glaucoma progression do not reside in structured EHR fields that can be easily extracted, making them challenging to include in glaucoma data science research.

Implementation of a local, secure LLM, which offers many healthcare-specific advantages, may allow for rapid extraction and interpretation of unstructured data, potentially even supporting the creation of novel variables in large public datasets. Local LLM pipelines have the potential to improve how clinical researchers utilize sensitive free text to extract valuable data and to potentially support targeted patient interventions in clinical settings.

Acknowledgements

None

References

1. Tham YC, Li X, Wong TY, Quigley HA, Aung T, Cheng CY. Global Prevalence of Glaucoma and Projections of Glaucoma Burden through 2040: A Systematic Review and Meta-Analysis. *Ophthalmology*. 2014;121(11):2081-2090. doi:10.1016/j.ophtha.2014.05.013
2. Allison K, Patel D, Alabi O. Epidemiology of Glaucoma: The Past, Present, and Predictions for the Future. *Cureus*. 2020;12(11):e11686. doi:10.7759/cureus.11686
3. Oltramari L, Mansberger SL, Souza JMP, de Souza LB, de Azevedo SFM, Abe RY. The association between glaucoma treatment adherence with disease progression and loss to follow-up. *Sci Rep*. 2024;14(1):2195. doi:10.1038/s41598-024-52800-2
4. Newman-Casey PA, Niziol LM, Gillespie BW, Janz NK, Lichter PR, Musch DC. The Association between Medication Adherence and Visual Field Progression in the Collaborative Initial Glaucoma Treatment Study (CIGTS). *Ophthalmology*. 2020;127(4):477-483. doi:10.1016/j.ophtha.2019.10.022
5. Williams AM, Schempf T, Liu PJ, Rosdahl JA. Loss to Follow up among Glaucoma Patients at a Tertiary Eye Center over 10 Years: Incidence, Risk Factors, and Clinical Outcomes. *Ophthal Epidemiol*. 2023;30(4):383-391. doi:10.1080/09286586.2022.2127787
6. Williams AM, Liang HW, Lin HHS. Loss to Follow-Up and Risk of Incident Blindness among Patients with Glaucoma in the IRIS® Registry. *Ophthalmol Glaucoma*. 2025;0(0). doi:10.1016/j.ogla.2025.05.001
7. Green CM, Kearns LS, Wu J, et al. How significant is a family history of glaucoma? Experience from the Glaucoma Inheritance Study in Tasmania. *Clin Exp Ophthalmol*. 2007;35(9):793-799. doi:10.1111/j.1442-9071.2007.01612.x
8. Li X, Sundquist J, Zöller B, Sundquist K. Familial Risks of Glaucoma in the Population of Sweden. *J Glaucoma*. 2018;27(9):802-806. doi:10.1097/IJG.0000000000001013
9. Leske MC, Wu SY, Hennis A, Honkanen R, Nemesure B. Risk Factors for Incident Open-angle Glaucoma: The Barbados Eye Studies. *Ophthalmology*. 2008;115(1):85-93. doi:10.1016/j.ophtha.2007.03.017
10. Coleman AL, Miglior S. Risk Factors for Glaucoma Onset and Progression. *Surv Ophthalmol*. 2008;53(6, Supplement):S3-S10. doi:10.1016/j.survophthal.2008.08.006
11. Herndon LW, Weizer JS, Stinnett SS. Central corneal thickness as a risk factor for advanced glaucoma damage. *Arch Ophthalmol*. 2004;122(1):17-21. doi:10.1001/archopht.122.1.17
12. Feng JE, Anoushiravani AA, Tesoriero PJ, et al. Transcription Error Rates in Retrospective Chart Reviews. *Orthopedics*. 2020;43(5):e404-e408. doi:10.3928/01477447-20200619-10

13. Yang X, Zhang H, He X, Bian J, Wu Y. Extracting Family History of Patients From Clinical Narratives: Exploring an End-to-End Solution With Deep Learning Models. *JMIR Med Inform*. 2020;8(12):e22982. doi:10.2196/22982

14. Polubriaginof F, Tatonetti NP, Vawdrey DK. An Assessment of Family History Information Captured in an Electronic Health Record. *AMIA Annu Symp Proc*. 2015;2015:2035-2042.

15. Lin WC, Chen JS, Kaluzny J, Chen A, Chiang MF, Hribar MR. Extraction of Active Medications and Adherence Using Natural Language Processing for Glaucoma Patients. *AMIA Annu Symp Proc*. 2022;2021:773-782.

16. Ashfaq HA, Lester CA, Ballouz D, Errickson J, Woodward MA. Medication Accuracy in Electronic Health Records for Microbial Keratitis. *JAMA Ophthalmol*. 2019;137(8):929-931. doi:10.1001/jamaophthalmol.2019.1444

17. Chen JS, Baxter SL. Applications of natural language processing in ophthalmology: present and future. *Front Med*. 2022;9. doi:10.3389/fmed.2022.906554

18. Murdoch TB, Detsky AS. The Inevitable Application of Big Data to Health Care. *JAMA*. 2013;309(13):1351-1352. doi:10.1001/jama.2013.393

19. Tan TF, Thirunavukarasu AJ, Campbell JP, et al. Generative Artificial Intelligence Through ChatGPT and Other Large Language Models in Ophthalmology: Clinical Applications and Challenges. *Ophthaly Sci*. 2023;3(4):100394. doi:10.1016/j.xops.2023.100394

20. Magnini M, Aguzzi G, Montagna S. Open-source small language models for personal medical assistant chatbots. *Intell-Base Med*. 2025;11:100197. doi:10.1016/j.ibmed.2024.100197

21. Gallo Afflitto G, Swaminathan SS. Racial-ethnic disparities in concurrent rates of peripapillary & macular OCT parameters among a large glaucomatous clinical population. *Eye (Lond)*. 2024;38(14):2711-2717. doi:10.1038/s41433-024-03103-3

22. Sellergren A, Kazemzadeh S, Jaroensri T, et al. MedGemma Technical Report. *arXiv*. Preprint posted online July 12, 2025. doi:10.48550/arXiv.2507.05201

23. Stewart WC, Konstas AGP, Pfeiffer N. Patient and Ophthalmologist Attitudes Concerning Compliance and Dosing in Glaucoma Treatment. *J Ocul Pharmacol Ther*. 2004;20(6):461-469. doi:10.1089/jop.2004.20.461

24. Glassberg MB, Trygstad T, Wei D, Robinson T, Farley JF. Accuracy of Prescription Claims Data in Identifying Truly Nonadherent Patients. *JMCP*. 2019;25(12):1349-1356. doi:10.18553/jmcp.2019.25.12.1349

25. Kamdar A, Yohannes GB, Swaminathan SS. Association Between Sociodemographic Risk Factors and No-Show Propensity in a Glaucoma Population Before and During COVID-19 Pandemic. *J Glaucoma*. 2025;34(8):e41. doi:10.1097/IJG.0000000000002550

26. Thompson AC, Thompson MO, Young DL, et al. Barriers to Follow-Up and Strategies to Improve Adherence to Appointments for Care of Chronic Eye

Diseases. *Invest Ophthalmol Vis Sci*. 2015;56(8):4324-4331. doi:10.1167/iovs.15-16444

27. Yang S, Varghese P, Stephenson E, Tu K, Gronsbell J. Machine learning approaches for electronic health records phenotyping: a methodical review. *J Am Med Inform Assoc*. 2023;30(2):367-381. doi:10.1093/jamia/ocac216

28. Leiby BE. A Randomized Trial to Improve Adherence to Follow-up Eye Examinations Among People With Glaucoma. *Prev Chronic Dis*. 2021;18. doi:10.5888/pcd18.200567

29. Pizzi LT, Tran J, Shafa A, et al. Effectiveness and Cost of a Personalized Reminder Intervention to Improve Adherence to Glaucoma Care. *Appl Health Econ Health Policy*. 2016;14(2):229-240. doi:10.1007/s40258-016-0231-8

30. Rybinski M, Dai X, Singh S, Karimi S, Nguyen A. Extracting Family History Information From Electronic Health Records: Natural Language Processing Analysis. *JMIR Med Inform*. 2021;9(4):e24020. doi:10.2196/24020

31. Wang SY, Tseng B, Hernandez-Boussard T. Deep Learning Approaches for Predicting Glaucoma Progression Using Electronic Health Records and Natural Language Processing. *Ophthalmol Sci*. 2022;2(2):100127. doi:10.1016/j.xops.2022.100127

32. Jalamangala Shivananjaiah SK, Kumari S, Majid I, Wang SY. Predicting near-term glaucoma progression: An artificial intelligence approach using clinical free-text notes and data from electronic health records. *Front Med*. 2023;10:1157016. doi:10.3389/fmed.2023.1157016

33. Zheng C, Luo Y, Mercado C, et al. Using natural language processing for identification of herpes zoster ophthalmicus cases to support population-based study. *Clin Exp Ophthalmol*. 2019;47(1):7-14. doi:10.1111/ceo.13340

34. Stein JD, Rahman M, Andrews C, et al. Evaluation of an Algorithm for Identifying Ocular Conditions in Electronic Health Record Data. *JAMA Ophthalmol*. 2019;137(5):491-497. doi:10.1001/jamaophthalmol.2018.7051

35. Liu L, Shorstein NH, Amsden LB, Herrinton LJ. Natural language processing to ascertain two key variables from operative reports in ophthalmology. *Pharmacoepidemiol Drug Saf*. 2017;26(4):378-385. doi:10.1002/pds.4149

36. Shaheen A, Afflitto GG, Swaminathan SS. ChatGPT-Assisted Classification of Postoperative Bleeding Following Microinvasive Glaucoma Surgery Using Electronic Health Record Data. *Ophthalmol Sci*. 2025;5(1):100602. doi:10.1016/j.xops.2024.100602

37. Ong JCL, Chang SYH, William W, et al. Ethical and regulatory challenges of large language models in medicine. *Lancet Digit Health*. 2024;6(6):e428-e432. doi:10.1016/S2589-7500(24)00061-X

38. Lee YM, Bacchi S, Macri C, Tan Y, Casson R, Chan WO. Ophthalmology Operation Note Encoding with Open-Source Machine Learning and Natural Language Processing. *Ophthalmic Res*. 2023;66(1):928-939. doi:10.1159/000530954

39. Tan Y, Bacchi S, Casson RJ, Selva D, Chan W. Triaging ophthalmology outpatient referrals with machine learning: A pilot study. *Clin Exp Ophthalmol*. 2020;48(2):169-173. doi:10.1111/ceo.13666

40. Bernstein IA, Koornwinder A, Hwang HH, Wang SY. Automated Recognition of Visual Acuity Measurements in Ophthalmology Clinical Notes Using Deep Learning. *Ophthalmol Sci*. 2024;4(2):100371. doi:10.1016/j.xops.2023.100371

41. Weiner SJ, Wang S, Kelly B, Sharma G, Schwartz A. How accurate is the medical record? A comparison of the physician's note with a concealed audio recording in unannounced standardized patient encounters. *J Am Med Inform Assoc*. 2020;27(5):770-775. doi:10.1093/jamia/ocaa027

42. Valikodath NG, Newman-Casey PA, Lee PP, Musch DC, Niziol LM, Woodward MA. Agreement of Ocular Symptom Reporting Between Patient-Reported Outcomes and Medical Records. *JAMA Ophthalmol*. 2017;135(3):225-231. doi:10.1001/jamaophthalmol.2016.5551

Table 1. Summary of large language model performance metrics for classification tasks with 95% confidence intervals.

Metric	Medication Non-Adherence	Visit Non-Adherence	FHoG	FHoG (EHR Structured Field)
Accuracy	0.91 (95% CI: 0.89–0.93)	0.96 (95% CI: 0.95–0.97)	0.98 (95% CI: 0.97–0.99)	0.49 (95% CI: 0.46–0.52)
Sensitivity	0.96 (95% CI: 0.90–0.99)	0.97 (95% CI: 0.95–0.97)	0.99 (95% CI: 0.98–0.99)	0.76 (95% CI: 0.71–0.80)
Specificity	0.91 (95% CI: 0.89–0.92)	0.94 (95% CI: 0.84–0.98)	0.99 (95% CI: 0.98–0.99)	0.81 (95% CI: 0.78–0.83)
MSE	-	-	0.05 ± 0.56	0.85 ± 1.80

EHR: Electronic Health Record, FHoG: Family History of Glaucoma, MSE: Mean Squared Error

Table 2. Examples of medication and visit non-adherence misclassification by the large learning model (LLM). The output tokens provide insight into the rationale for decision-making by the LLM. Any names, dates, or other potentially identifying information have been replaced by asterisks.

Classification Task	Note	Graded Label	LLM Label	LLM Output Tokens
Medication Non-Adherence	Impression: 1. Primary angle closure glaucoma LEFT- peripheral iridectomy absent, moderate, superior visual field defect, sparing fixation. Target pressure unknown (insufficient information/follow up), cannot comment on target today. IOP with 3 meds in the teens, probably not dangerous. Had multiple long discussion about need for LPI, patient is still "not ready." HVF **/**/**slightly worse than **/**/** with average IOP of 19. Doubt target not low enough rather than noncompliance or intermittent angle closure. 2. Phthisis, right eye (injury), cosmetically excellent with scleral shell. Seen *** (ocularist). Cataract, left eye, visually insignificant. 4. Blunt trauma, left eye. No obvious sequelae. Repeat DFE deferred given angle closure and absence of patent iridotomy. Plan: 1. Continue Xalatan 0/1 and Cosopt 0/2. Emphasized compliance. Discussed acute angle closure symptoms. 2. Return in 3 months, VA, IOP. Plan to repeat HVF 24-2 left eye in 6 months, sooner if IOP high in the interim. 3. Patient will call me if desires laser sooner than next visit (LPI left eye).	Yes (Decreased Adherence)	No (Adherent)	…The note explicitly states "Emphasized compliance." This indicates that adherence was discussed, but it does *not* state that the patient *is* non-adherent. It's a general emphasis, not a report of non-adherence…The note speculates about non-compliance as a *possible* reason for the IOP, but doesn't confirm it… The note mentions emphasizing compliance but does not provide any specific evidence or statement indicating that the patient *is* currently non-adherent with their IOP-lowering medications (Xalatan and Cosopt)…
Medication Non-	First visit with ***: **/**/**** Assessment: #Primary open angle glaucoma, mild OS Component of steroid response? POHx:	No (Adherent)	Yes (Decreased	… Is it due to side effects? Yes, "due

Adherence	Started on Latanoprost OS in Sept **** for glaucoma. OHTN and CME while on Latanoprost in Nov ****. Surg hx: remote hx of PPV/SO/SB OD, s/p enucleation OD **** Laser hx: none Tmax: unk Glaucoma FHx: none Trauma?: soccer to head causing RD in OD as a teenager Systemic B-blocker?: no Low BP/Migraines?: no Steroid use?: no Med Allergies: n/a CCT: prosthetic 506 **/**/**Not protective Gonio: OS open to SS/CB 2+ pigment **/**/**Target IOP: ~21 HVF 24-2c **/**/** OS: reliable, non-specific sup/inf depression; baseline OCT RNFL **/**/**Adequate SS; inferior thinning, confounded by VMT/CME; baseline **/**/**: IOP at target OU Current meds: Latanoprost 0/1 **/**/**: IOP elevated to 30 Current meds: PF 0/2 ON photos: inferior notch **/**/**: IOP at target Current meds: finished PF, self dc'ed Cosopt 0/2 due to burning Resolved CME after holding Latanoprost Plan: - Start Timolol 0/2 - Continue to hold PGA due to CME - Repeat HVF OS next visit # superior RRD OS s/p PPV/EL/retinotomy/C3F8 **/**/**Retina attached - Follows with Dr. *** #CME OS - resolved Attributed to PGA Resolved off Latanoprost #Anopthalmos for blind painful eye OD h/o RD s/p PPV w/ SO/SB S/p enucleation 19 mm Su-Por implant **/**/** (Dr.***) Monocular precautions #DES OS Status **/**/**: Established diagnosis- Stable - continue ATs QID, ointment QHS #PCO OS Status **/**/**: New diagnosis Not visually significant - Observe RTC: 4 months + VA/IOP + HVF OS		Adherence)	to burning". However, the *act* of self-discontinuation *is* considered decreased adherence, even if prompted by a side effect. The criteria state "Self-discontinuation against medical advice". While the *reason* is a side effect, the *action* is self-discontinuation…
Medication Non-Adherence	Impression 1. Ocular hypertension OD>OS IOP 27-29/20-22 today off medications IOP in the mid-to-high teens in Chicago from ****-**** (patient brought records) except for one IOP of 24 mmHg (Tonopen) on **/**/** CCT 571/568 Goniscopy: open 360 HVF (**/**/**): OD reliable and stable with scattered inferior defects; OS reliable and stable with scattered inferonasal defects OCT rNFL (**/**/**): wnl OU Optic nerve/fundus photos (**/**/**): ONH appearance stable today Family history: None per patient 2. Choroidal nevus of	No (Adherent)	Yes (Decreased Adherence)	…The note explicitly states the patient's IOP is elevated (27-29 OD, 20-22 OS) *today* and that they are "off medications". This implies they were previously prescribed IOP-lowering medications but are not currently

	the left eye Followed by Dr. *** 3. Nuclear sclerotic cataracts, both eyes Not visually significant 4. Refractive error, both eyes Stable 5. Blepharitis, both eyes Stable 6. Diplopia Followed by Dr. *** and evaluated in Chicago, IL Plan - Start latanoprost qhs OU - Return to clinic 4-6 weeks			taking them. … The patient was off their glaucoma medications at the time of this visit, leading to elevated IOP. This constitutes a period of non-use against medical advice (even if the reason isn't stated, being "off medications" when IOP is elevated implies non-adherence or discontinuation). **Conclusion:** The note clearly indicates the patient was not taking their prescribed IOP-lowering medications at the time of the visit, leading to elevated IOP. This fits the criteria for decreased adherence…
Visit Non-Adherence	Impression: POAG, mild OU Found by optometry to have ocular HTN in ****. Evaluated by Dr. *** at *** in **** and found to have IOP within normal range and full HVFs with plan for f/u in 6 months.. Was then evaluated by Dr. *** **/****with IOP OU of 30 and re-referred to ***. She has a mother with glaucoma and follows with Dr. ***. Field in **** showed bilateral nasal defects consistent with glaucoma and nerves consistent with glaucoma. Decision to treat and laser chosen as initial modality. SLT nasal OD done **/**/**and OS done **/**/**Today, **/**/**: · IOP OD 19, 20 / IOP OS 18, 19 Applanation, Applanation	Yes (Decreased Adherence)	No (Adherent)	…"Evaluated by Dr. *** at BPEI in **** and found to have IOP within normal range and full HVFs with plan for f/u in 6 months." * "Was evaluated by Dr. *** **/****with IOP OU of 30 and re-referred to BPEI". Patient has been adherent to visit

	· Using no meds · Fields are stable · Nerves look like photos Her IOP is still lower than pre laser and her field and exam are stable; will continue off meds			schedule…
Visit Non-Adherence	ASSESSMENT/PLAN: H/o of MVA On ASA Daily Primary open angle glaucoma OU Followed by Dr. *** in the past (last ****) and subsequently by Dr. *** -ONH appearance stable compared with **** -Baseline RNFL performed (**/**/**) with superior and inferior thinning OD > OS, stable (**/**/**). GCC sup thinning OD, full OS (**/**/**) -Mother with glaucoma Cataracts OU Impression Mr. *** is a *** y.o. male Today, **/**/**: VA OD sc: cc:20/30 VA OS sc: cc:20/30 IOP OD 11 / IOP OS 9 at 8:50 AM Applanation Ocular medications: Latanoprost 1/1 LD Cosopt 2/2 LD 1. Visually significant cataracts (OU OD>OS) ADL's affected by vision. Plan for combined surgery GATT 360 with cataract extraction with Phaco and intraocular lens placement in the OD eye Consider CE in OS in the future … Disposition Plan for CE in OD w/ conventional lens Plano target on TBD. Consider future CE OS. RTC: For POD1 Advised pt to discontinue ASA 1 week prior to surgery OD	No (Adherent)	Yes (Decreased Adherence)	… The note mentions "Followed by Dr. *** in the past (last ****) and subsequently by Dr. ***". This indicates a significant gap in care since ****. 3. **Interpret the finding:** The patient was last seen by another provider in ****. The current encounter is in ****…
Visit Non-Adherence	Uncontrolled POAG Associated with atopic dermatitis, although the patient has not used corticosteroids around the eyes for several years Preoperative IOP was 50 mm Hg in the right eye with admittedly poor compliance s/p BGI 350 mm2 / patch graft OD (**/**/**) s/p LSL OD (**/**/**) Using PF q2h and atropine bid IOP of 10/14 on bilateral latanoprost, alphagan, cosopt, pilo BID, and Diamox 250 TID PO. Outside GCC with possible progression OS. Follicular Conjunctivitis OU -2/2 to Alphagan -discontinue Plan: -Continue your glaucoma drops (Cosopt,pilo, Xalatan) in both eyes. -Stop the Alphagan OU -Decrease to diamox 125 QID -If IOP is still uncontrolled consider incisional surgery in the left eye. "improved" vs **/**/**, possibly better at hvf test today No obvious change in	No (Adherent)	Yes (Decreased Adherence)	…The phrase "admittedly poor compliance" is a strong indicator of adherence issues, although it's specifically about medication. The question asks about *visit* adherence. The note doesn't explicitly state the patient missed visits, but it is more likely given poor compliance…

	appearance of visual fields or advanced RNFL loss Plan: Continue Xal 1/1 RTC 6 months. Follow up: Return in about 6 months			

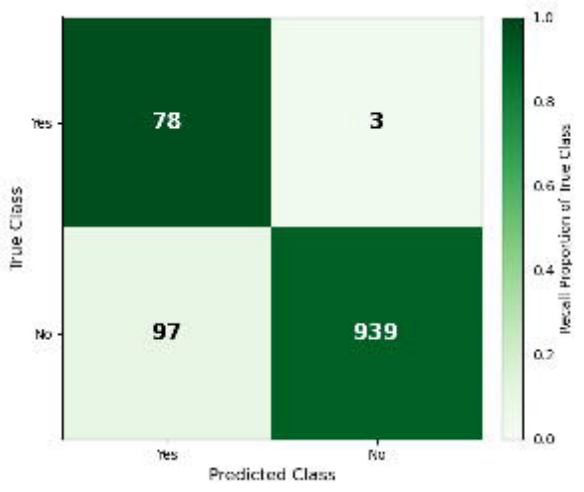
Figure Legends

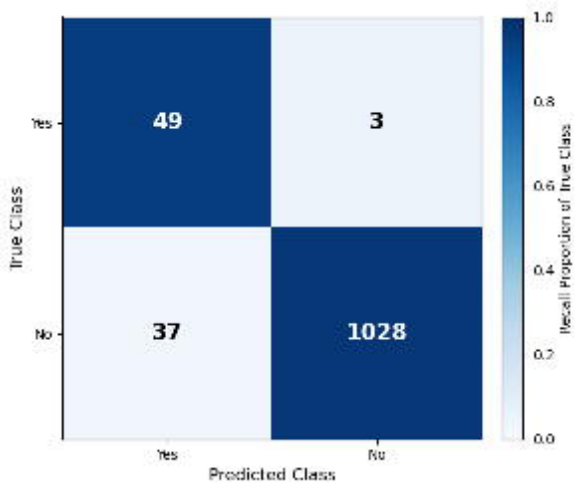
Figure 1. Confusion matrix for medication non-adherence. Classification was binary ("yes" or "no / not mentioned").

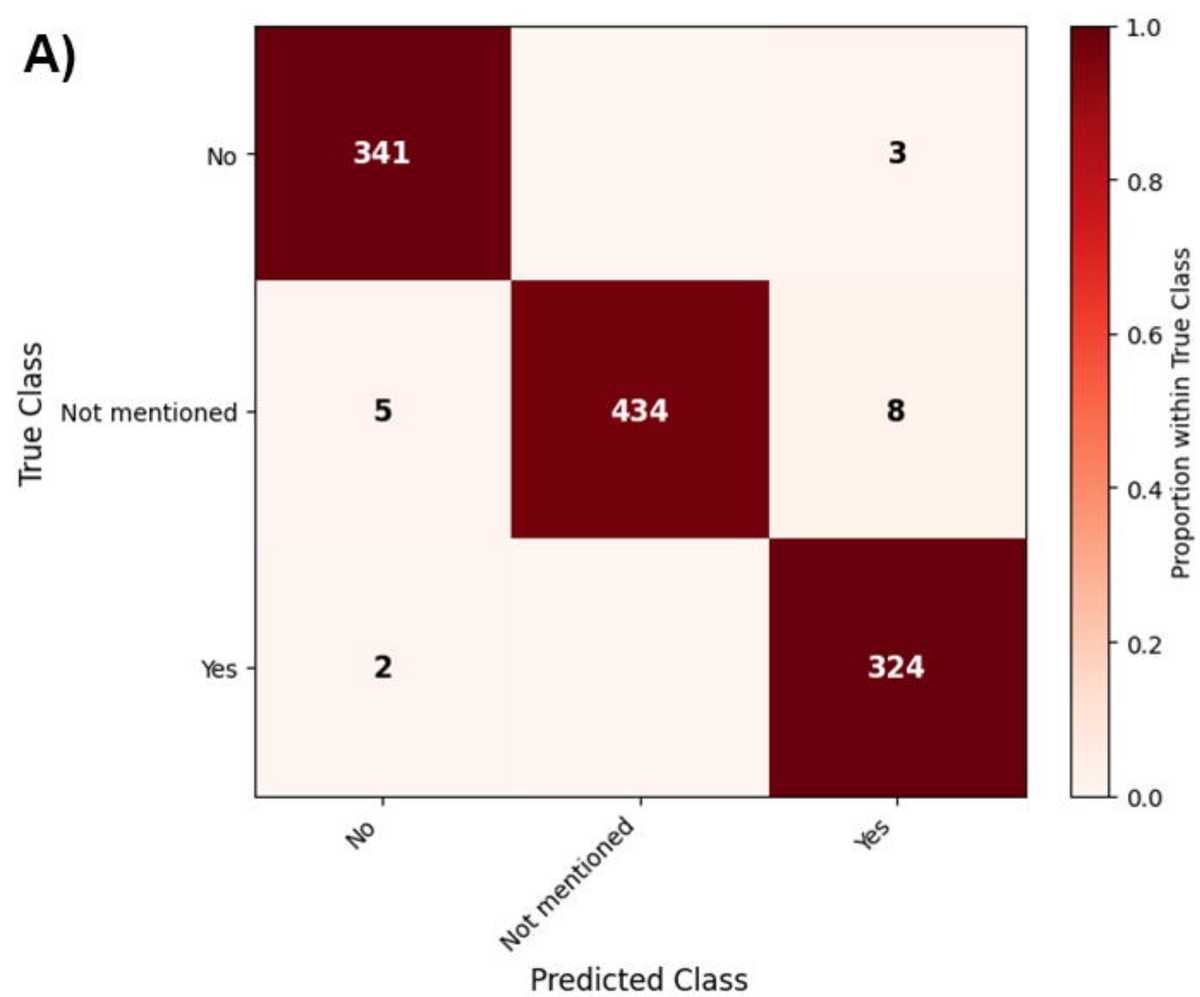
Figure 2. Confusion matrix for visit non-adherence. Classification was binary ("yes" or "no / not mentioned").

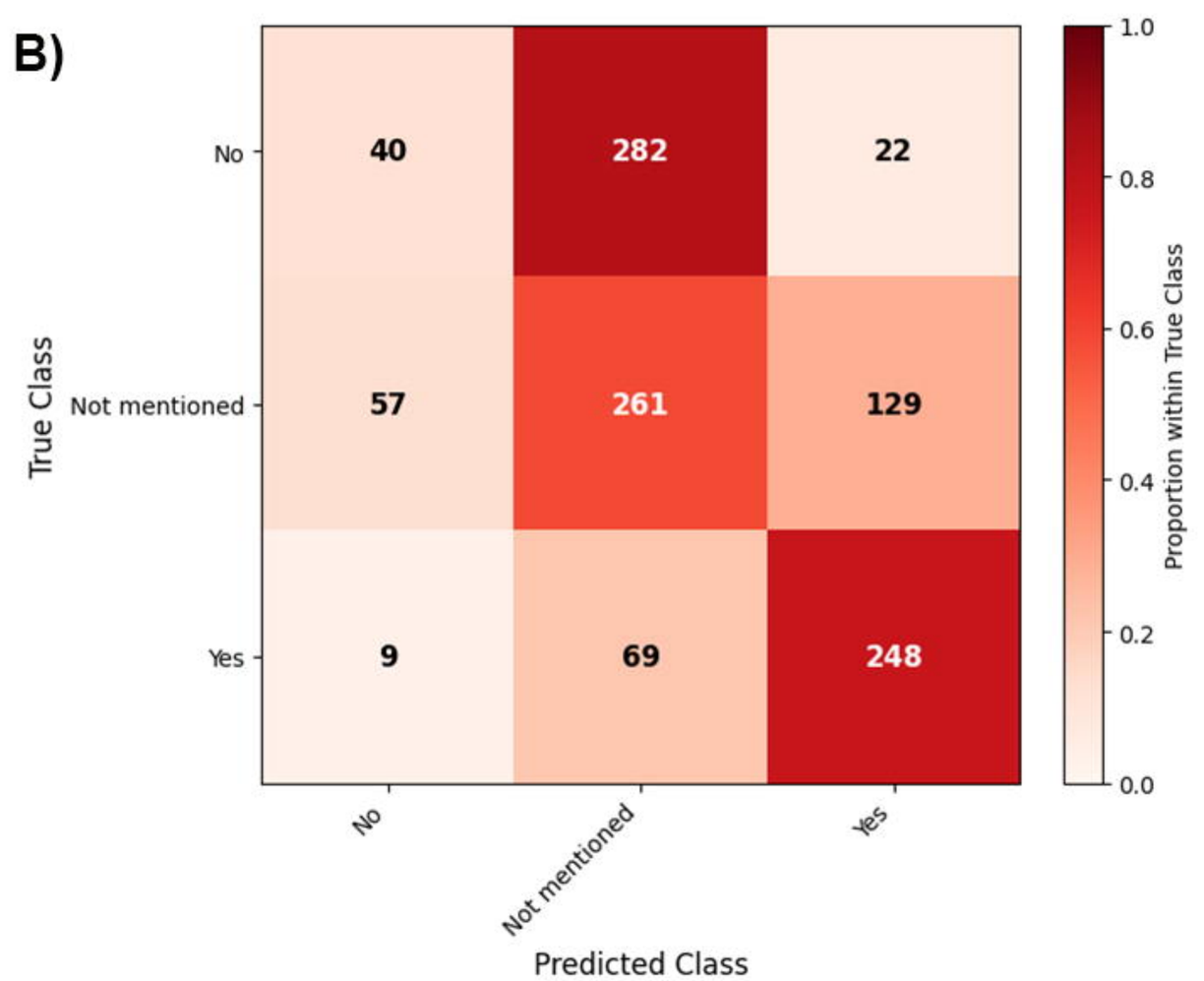
Figure 3. Confusion matrix for family history of glaucoma using A) large language model classification and B) electronic health record structured field classification. Classification was ternary ("yes", "no", or "not mentioned").

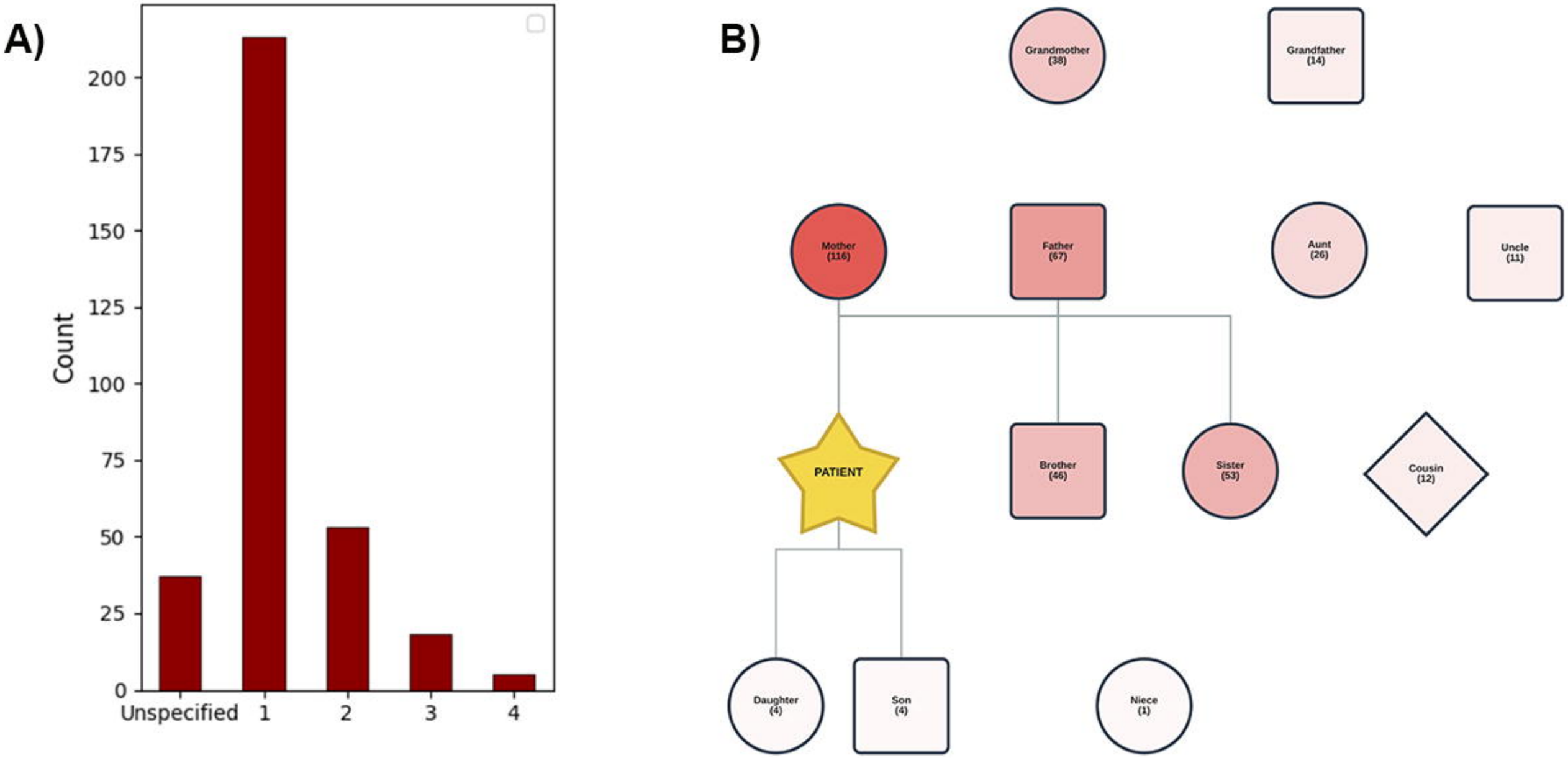
Figure 4. A) Histogram describing the distribution of the number of family members with glaucoma per graded labels. B) Diagram reflecting the frequency of relatives noted to have glaucoma per the graded labels. Darker colors reflect higher frequency (count indicated in parentheses).


78
3
97
939
Yes
No
True Class
Yes
No
Predicted Class
1.0
0.8
0.6
0.4
0.2
0.0
Recall Proportion of True Class


49
3
37
1028
Yes
No
True Class
Yes
No
Predicted Class
1.0
0.8
0.6
0.4
0.2
0.0
Recall Proportion of True Class


A)
No
Not mentioned
Yes
True Class
341
3
5
434
8
2
324
No
Not mentioned
Yes
Predicted Class
1.0
0.8
0.6
0.4
0.2
0.0
Proportion within True Class


B)
No
Not mentioned
Yes
True Class
40
282
22
57
261
129
9
69
248
No
Not mentioned
Yes
Predicted Class
1.0
0.8
0.6
0.4
0.2
0.0
Proportion within True Class


A)
Count
0
25
50
75
100
125
150
175
200
Unspecified
1
2
3
4
B)
Grandmother (38)
Grandfather (14)
Mother (116)
Father (67)
Aunt (26)
Uncle (11)
PATIENT
Brother (46)
Sister (53)
Cousin (12)
Daughter (4)
Son (4)
Niece (1)