**Introduction:**

*Background*

Electronic Health Records (EHRs) have significantly changed how healthcare is delivered by gathering diverse and comprehensive patient data. Healthcare data includes both structured and unstructured information. Structured data includes various medical codes for diagnoses and procedures. Unstructured data contains clinical notes as text.

Proliferative Diabetic Retinopathy (PDR) is the advanced stage of diabetic retinopathy and a leading cause of severe vision loss. It is characterized by the abnormal blood vessel growth in the retina, which can lead to vision loss and blindness. Timely identification of incident PDR is crucial for preventing complications, optimizing treatment, and improving patient outcomes. The progression from non-proliferative diabetic retinopathy (NPDR) to PDR represents a critical clinical endpoint often used in research studies to evaluate disease progression and treatment effectiveness.

However, pinpointing new cases of PDR using EHR data is not without its challenges. The structured data alone can cause issues in diagnostic coding and incomplete medical histories, especially for patients receiving care from different healthcare providers. While unstructured clinical notes hold important insights, extracting useful information can be tough. Fortunately, recent advancements in machine learning and AI present opportunities for better identifying PDR cases by blending structured and unstructured data.

*Research Goal*

This study evaluates classification methods using structured and unstructured EHR data to determine which most accurately identifies incident PDR cases for use in clinical research and practice.

**Methods:**

*Data Source*

The data source we used was data from the UCSF De-Identified Clinical Data Warehouse (De-ID CDW), which has de-identified EHR data for all UCSF patients. The limited data set version of the De-ID CDW was used to obtain real dates. The machine-redacted notes are de-identified free text from the UCSF EHR and are available for research through the De-ID CDW. Unstructured data included clinical note text and metadata, including date, provider type, note type, encounter type, and department. We included notes from eye providers in the department of Ophthalmology and Francis I. Proctor Foundation.

*Participants*

The inclusion criteria were patients aged ≥ 18 years with at least one encounter with a diagnosis of diabetic retinopathy (DR) by an eye care provider and available de-identified clinical notes. Patients were excluded if they did not have any de-identified clinical notes. International Classification of Diseases, ninth edition (ICD-9), was used for all encounters before October 1, 2015, and ICD-10 codes for all encounters on or after October 1, 2015. There were three hundred twenty-one random patients with DR who underwent manual chart review by physicians, and they were used as the gold standard. Among these, 193 patients had a confirmed diagnosis of PDR. Of those 193 patients, 51 were identified with incident PDR based on the physician's chart review.

*Gold Standard*

To establish a gold standard for incident proliferative diabetic retinopathy, a manual chart review was conducted by one to two ophthalmologists. Incident PDR was defined as the first documented evidence of PDR diagnosis in the EHR. Dates with incomplete date information, like missing month or day, were excluded to ensure accuracy. A 90-day window was used to evaluate date agreement with other classification methods, accounting for variability in documentation and potential delays in diagnosis or coding within the EHR.

*Classification Methods*

Six distinct methods were evaluated to identify incident PDR. Each method was implemented separately and then assessed against the gold standard:

(1) ICD-Only (No Lookback): First ICD-9/10 diagnosis code for PDR in the EHR with no lookback period.

This first method identifies the earliest occurrence of a PDR diagnosis using ICD-9 (362.02) or ICD-10 (H35.35x) codes. The first encounter with a PDR code was labeled as the incident date, with no additional filtering.

(2) ICD + 1 Year Lookback Period (All Departments): First ICD-9/10 code for PDR with ≥ 1 prior encounter in any department at least one year before.

This second method adds a lookback window of ≥ 365 days, requiring that the patient had at least one prior UCSF encounter in any department without a PDR diagnosis before the first PDR-coded encounter.

(3) ICD + 1 Year Lookback Period (Ophthalmology Only): First ICD-9/10 code for PDR with ≥ 1 prior encounter in an ophthalmology or Proctor department at least one year before.

This third method is the same as the second one, but it limits the lookback requirement to encounters in eye care settings. This makes it stricter and more reliable to filter encounters by a specific department.

(4) Rule-based NLP: detection of “PDR” based on keyword and pattern matching in the assessment/plan section of the de-identified clinical notes.

This fourth method uses a previously developed natural language processing pipeline to identify incident PDR using de-identified clinical notes from UCSF’s CDW. The algorithm focused on parsing and analyzing the assessment and plan (A/P) section of each note, as this section typically contains the provider’s diagnostic summary and treatment plan, which are key locations for documenting new diagnoses [3].

Notes were first preprocessed to standardize text formatting. This included converting text to lowercase, expanding abbreviations (e.g., “PDR” to “proliferative diabetic retinopathy”), and removing headers, footers, and extraneous whitespace or punctuation. The A/P section was extracted using regular expressions, and only notes with valid A/P content were retained.

We used the scispaCy library for biomedical NLP, including sentence segmentation and named entity recognition. The NegEx module from negspacy was applied to flag negated mentions of PDR. The algorithm then screened each A/P section for positive, non-negated mentions of PDR, using a pattern-matching approach that included phrases such as “proliferative diabetic retinopathy” and “proliferative retinopathy.” If a note contained such a mention, the note date was recorded as the incident PDR date for that patient.

Patients were classified as having incident PDR if they had at least one note with a positive, non-negated mention of PDR. If multiple positive notes existed, the earliest note date was used. Notes without a valid A/P section or without qualifying mentions were excluded from the NLP output.

(5) ICD + NLP: Combines the best-performing ICD method (from 1-3) with the rule-based NLP output to refine incident case detection.

This fifth method combines the best-performing ICD-based approach with the rule-based NLP method. The best ICD method was selected based on how closely its predicted incident PDR dates aligned with the gold standard dates from manual chart review across the patient cohort. To construct the hybrid cohort, we looked at flagging patients as having PDR if either the selected ICD method or the rule-based NLP method identified them. If either method produced a date that matched the gold standard within the defined window, the patient was correctly identified.

(6) Generative AI model: A large language model (LLM) that was prompt-engineered to identify incident PDR based on full-note context.

This sixth method used UCSF’s Versa chat interface to classify clinical notes using the GPT‑4o model in a secure, HIPAA-compliant environment. Each patient’s assessment/plan note was submitted using a structured zero-shot prompt that asked the model to (1) determine whether proliferative diabetic retinopathy (PDR) was diagnosed, (2) extract the date of first diagnosis, and (3) provide relevant contextual keywords. Notes were processed individually via manual interaction with the chat interface; no API or scripting was used. This approach allowed flexible, contextual interpretation of clinical language while maintaining data security within UCSF’s infrastructure. A sample prompt used in the Versa chat interface is provided in the Supplementary Materials.

*Statistical Analysis*

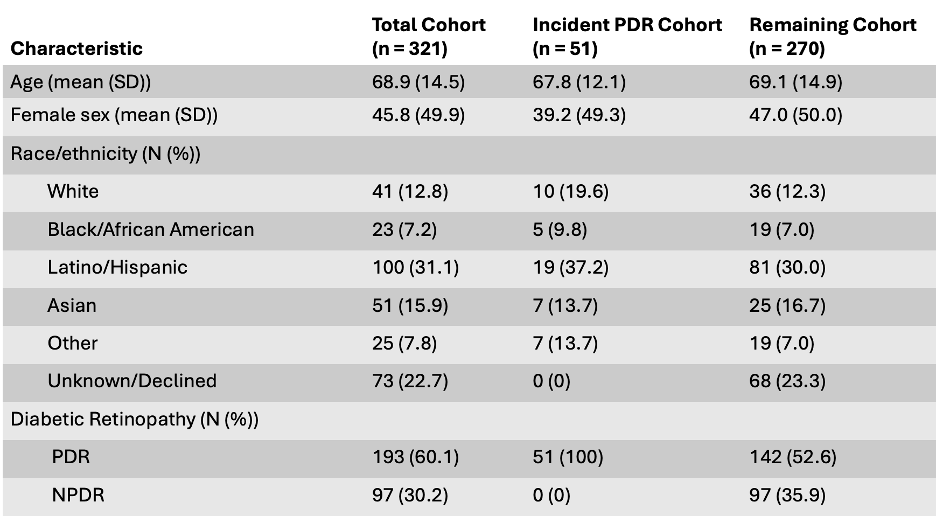
All data extraction was taken from the UCSF De-Identified Clinical Data Warehouse using Azure SQL Server. All statistical analyses were carried out using R Version 4.4.0. The NLP alorgithm was written in Python Verison 3.12.7.

We evaluated the performance of each classification method using the following metrics: sensitivity, specificity, positive predictive value, negative predictive value, and F1-score. Each metric was calculated by comparing method-identified cases of incident PDR against the gold standard derived from manual chart review. To estimate 95% confidence intervals for all performance metrics, we applied bootstrapping with replacement, repeated 1,000 times.

**Results:**

*Demographics Table*

Based on the manual chart review, the entire cohort consisted of 321 patients with 193 (60.1%) confirmed PDR. Of these, 51 (15.9%) were confirmed to have incident proliferative diabetic retinopathy. The remaining 270 (84.1%) patients either had PDR or only non-proliferative diabetic retinopathy (NPDR). The mean age of the total cohort was 68.9 years (SD 14.5), with the incident PDR group being slightly younger at 67.8 years (SD 12.1). Female patients comprised 45.8% of the overall cohort and 39.2% of the incident PDR group. Latino/Hispanic patients comprised the largest racial/ethnic subgroup across all the divides.

******

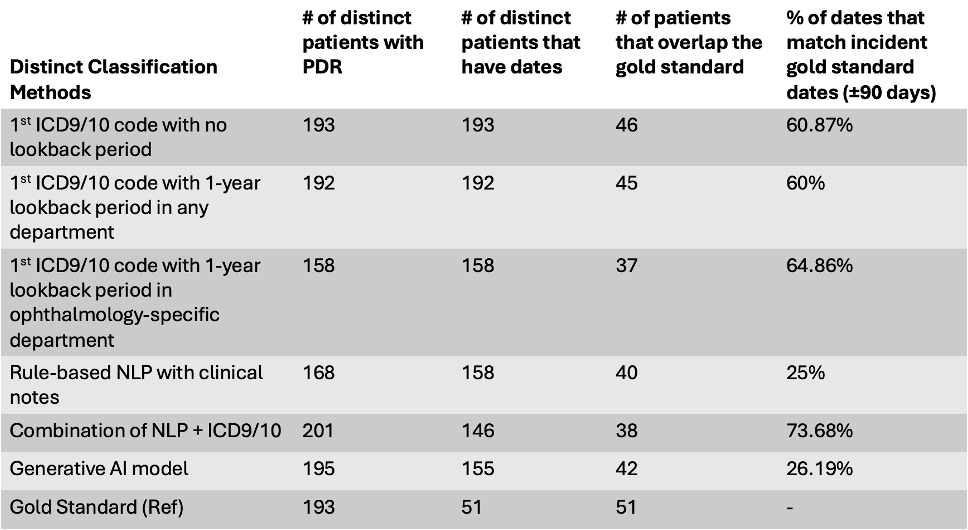
*Comparing First PDR Dates Across Methods Table*

This table summarizes the performance of each classification method in identifying incident PDR and the alignment of predicted incident dates with the gold standard. It includes: (1) the number of patients identified with PDR regardless of timing, (2) the subset for whom an incident date was assigned, (3) how many of those overlapped with the gold standard, and (4) the percentage of overlapping cases whose predicted dates fell within ±90 days of the gold standard date.

Among the 51 confirmed incident PDR cases in the gold standard, we evaluated how many were correctly identified and whether the predicted date of diagnosis fell within a ±90-day window of the gold standard incident date. The ICD-9/10 with no lookback period identified 193 patients with PDR. The method using the first ICD-9/10 code with a 1-year lookback restricted to ophthalmology departments produced the highest temporal accuracy, with 64.86% of cases aligning with the gold standard dates. This filtering reduced potential false positives from PDR cases that may have had prior undocumented diagnoses.

The rule-based NLP method identified 168 patients with clinical note mentions of PDR, and 158 of these had valid dates. While it overlapped with 40 gold standard cases, only 25% of those dates matched the gold standard timing. Despite identifying 195 patients and overlapping with 42 gold standard cases, the generative AI model also had a low temporal agreement at 26.2%, suggesting limitations in how well the model inferred the proper timing of incident diagnoses from free-text clinical context.

The combination method merged patients identified by either the best-performing ICD method (ophthalmology-specific department) or the rule-based NLP method. This approach identified 201 patients, with 146 having dates and 38 overlapping with the gold standard. This method achieved the highest temporal match rate of all approaches, with 73.68% of overlapping cases aligning within the ±90-day window.



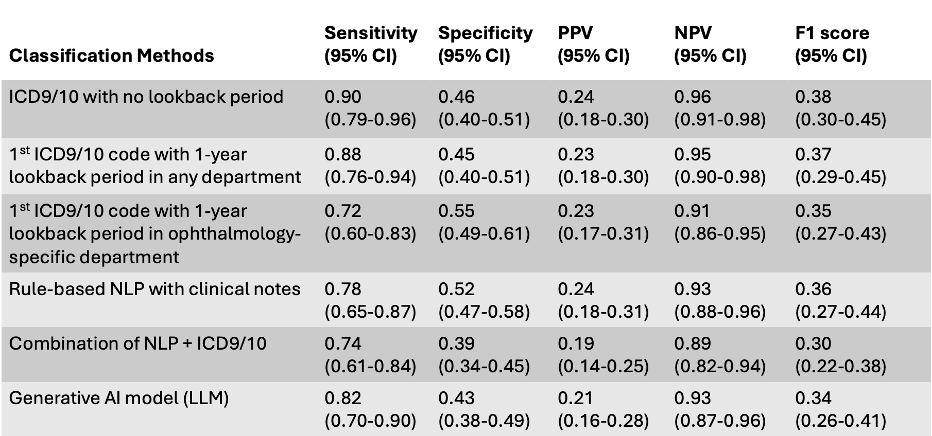
*Performance Metrics for Identifying Gold Standard Incident Dates Table*

Table 3 shows performance metrics for each classification method in identifying incident PDR cases, as defined by the gold standard manual chart review. Metrics include sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score, all with corresponding 95% confidence intervals. These were calculated using binary classification logic, with the gold standard as the reference label and each method’s identified cases as predictions.

The highest sensitivity was observed in the ICD9/10 method without a lookback period (0.90 [0.79–0.96]), reflecting its ability to capture most gold standard cases. However, this came at the cost of low specificity (0.46), due to the inclusion of many PDR cases being misclassified as incident. The 1-year lookback any department method slightly decreased sensitivity to 0.88, with a comparable specificity of 0.45. The ophthalmology-specific ICD lookback method offered a more balanced trade-off, with sensitivity of 0.72 and specificity of 0.55, reflecting improved exclusion of cases but at the cost of missing some true positives. PPV remained modest at 0.23 due to the relatively large number of false positives.

The rule-based NLP method yielded sensitivity of 0.78 and specificity of 0.52, comparable to ICD-only methods, but still suffered from limited PPV (0.24) due to overidentification in ambiguous contexts. The combination method (NLP + ICD) achieved moderate sensitivity (0.74) but had the lowest specificity (0.39) and PPV (0.19), reflecting the broadened inclusion criteria. Despite this, given its higher sensitivity, its overall F1 score (0.30) remained in range with the other methods.

The generative AI model had a sensitivity of 0.82, showing strong case capture, but with low specificity (0.43) and PPV (0.21). Its F1 score of 0.34 [0.26–0.41] indicated moderate balance but had concerns about its ability to distinguish incident cases based solely on contextual interpretation. Across all methods, negative predictive value (NPV) remained high (≥0.89), driven by the relatively low prevalence of incident PDR in the overall cohort.



**Discussion:**

This study evaluated six classification methods for identifying incident proliferative diabetic retinopathy using structured and unstructured EHR data. Structured ICD-based methods showed high sensitivity but often misclassified cases as incident due to limited historical context. In contrast, NLP and large language model (LLM) approaches relied on information from the clinical notes to identify diagnoses, but had trouble determining the exact timing of initial diagnosis. Among structured methods, the ICD approach with a 1-year lookback limited to ophthalmology departments offered the best balance of sensitivity, specificity, and temporal accuracy. However, it lacked the granularity to capture the clinical transition from non-proliferative to proliferative disease.

The rule-based NLP method struggled with temporal precision because it extracted the date of the first non-negated mention of “PDR” in the assessment and plan section, which did not always reflect a new diagnosis. Most clinical notes lacked explicit phrases like “incident” or “new diagnosis,” leading to misattribution of initial timing. The LLM showed promise in interpreting contextual clues and sometimes produced dates closer to the gold standard. However, it was limited by incomplete documentation, particularly for patients diagnosed outside the UCSF system, where only the diagnosis year was included. In many cases, the LLM had to infer timing from vague or indirect language, which introduced variability in the predicted dates.

The gold standard was intentionally strict, making it difficult for any method to consistently match the actual incident date. Some discrepancies were likely due to data fragmentation or missing documentation. Of the 321 patients, only 296 had notes with usable assessment/plan sections, possibly excluding relevant information elsewhere in the EHR.

Despite these challenges, combining structured and unstructured methods proved valuable. The hybrid ICD + NLP method achieved the highest temporal alignment with the gold standard. Notably, while each method identified similar incident PDR cases, only 29 of the 51 gold standard cases were captured across all methods combined. This suggests that different subsets of patients are being identified, likely due to variations in documentation and data completeness. It highlights the importance of evaluating how many cases are identified and who is being identified, especially when methods inform clinical or research decision-making. Overall, while no single method was perfect, the combination of structured ICD data with rule-based NLP achieved the most accurate and consistent identification of incident PDR.

Limitations include a modest sample size that may be underpowered to detect slight differences, particularly between advanced methods. Clinical notes contain rich information but are difficult to parse without sophisticated tools[1]. Rule-based methods may miss nuance, while LLMs remain limited by data quality and context variability. Still, advancements in NLP and transformer-based models offer promising opportunities for improving phenotyping from unstructured data.

To build on these findings, future work should apply these methods to larger, multi-institutional datasets and test generalizability across different EHR systems. While structured ICD-based methods provided strong sensitivity, combining them with NLP approaches improved temporal precision and completeness. Hybrid strategies offer the most robust framework for real-world identification of incident clinical events.

**Supplementary Material:**

*VersaChat Prompt*

You are a physician reviewing a set of clinical assessment and plan notes for a patient. Your task is to determine whether the patient has been diagnosed with proliferative diabetic retinopathy (PDR) based on the information provided in the notes.

Please extract the following:

1. Does this patient have proliferative diabetic retinopathy (PDR)?

Answer Yes or No/Uncertain

1. If Yes, identify the first documented date of PDR diagnosis in the format YYYY-MM-DD.
2. Provide supporting context from the note, such as keywords, findings, or treatments that indicate the diagnosis.

If multiple notes are provided, focus on the first documented occurrence of PDR.

Here are the notes:

**References:**

[1] Fu JT, Sholle E, Krichevsky S, Scandura J, Campion TR. Extracting and classifying diagnosis dates from clinical notes: A case study. J Biomed Inform. 2020 Oct;110:103569. [[DOI](https://pubmed.ncbi.nlm.nih.gov/32949781/)]

[2] Schmier JK, Covert DW, Lau EC, Matthews GP. Medicare expenditures associated with diabetes and diabetic retinopathy. Retina 2009; 29:199–206 [[DOI](https://doi.org/10.1097/IAE.0b013e3181884f2d)]

[3] Sean Yonamine, Chu Jian Ma, Rolake O. Alabi, Georgia Kaidonis, Lawrence Chan, Durga Borkar, Joshua D. Stein, Benjamin F. Arnold, Catherine Q. Sun, Comparison of Diagnosis Codes to Clinical Notes in Classifying Patients with Diabetic Retinopathy, Ophthalmology Science, Volume 4, Issue 6, 2024, 100564, ISSN 2666-9145, [[DOI](https://doi.org/10.1016/j.xops.2024.100564)]

[4] Pan J, Lee S, Cheligeer C, Martin EA, Riazi K, Quan H, Li N. Integrating large language models with human expertise for disease detection in electronic health records. Comput Biol Med. 2025 Jun;191:110161. [[DOI](https://pubmed.ncbi.nlm.nih.gov/40198990/)]

[5] 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](https://pubmed.ncbi.nlm.nih.gov/30789656/)]

[6] University of California, San Francisco. UCSF Versa, Assistants, and API. UCSF Center for Intelligent Imaging. Published 2023. Accessed June 8, 2025.[[DOI](https://ai.ucsf.edu/platforms-tools-and-resources/ucsf-versa)]