AMERICAN ACADEMY OF OPHTHALMOLOGY®

Comparison of Diagnosis Codes to Clinical Notes in Classifying Patients with Diabetic Retinopathy

Sean Yonamine, MPH,[1,2] Chu Jian Ma, MD, PhD,[1] Rolake O. Alabi, MD, PhD,[1] Georgia Kaidonis, MBBS, PhD,[1] Lawrence Chan, MD,[1] Durga Borkar, MD,[3] Joshua D. Stein, MD, MS,[4] Benjamin F. Arnold, PhD,[1,5,6] Catherine Q. Sun, MD[1,5]

Purpose: Electronic health records (EHRs) contain a vast amount of clinical data. Improved automated classification approaches have the potential to accurately and efficiently identify patient cohorts for research. We evaluated if a rule-based natural language processing (NLP) algorithm using clinical notes performed better for classifying proliferative diabetic retinopathy (PDR) and nonproliferative diabetic retinopathy (NPDR) severity compared with International Classification of Diseases, ninth edition (ICD-9) or 10th edition (ICD-10) codes.

Design: Cross-sectional study.

Subjects: Deidentified EHR data from an academic medical center identified 2366 patients aged ≥18 years, with diabetes mellitus, diabetic retinopathy (DR), and available clinical notes.

Methods: From these 2366 patients, 306 random patients (100 training set, 206 test set) underwent chart review by ophthalmologists to establish the gold standard. International Classification of Diseases codes were extracted from the EHR. The notes algorithm identified positive mention of PDR and NPDR severity from clinical notes. Proliferative diabetic retinopathy and NPDR severity classification by ICD codes and the notes algorithm were compared with the gold standard. The entire DR cohort (N = 2366) was then classified as having presence (or absence) of PDR using ICD codes and the notes algorithm.

Main Outcome Measures: Sensitivity, specificity, positive predictive value (PPV), negative predictive value, and F1 score for the notes algorithm compared with ICD codes using a gold standard of chart review.

Results: For PDR classification of the test set patients, the notes algorithm performed better than ICD codes for all metrics. Specifically, the notes algorithm had significantly higher sensitivity (90.5% [95% confidence interval 85.7, 94.9] vs. 68.4% [60.4, 75.3]), but similar PPV (98.0% [95.4−100] vs. 94.7% [90.3, 98.3]) respectively. The F1 score was 0.941 [0.910, 0.966] for the notes algorithm compared with 0.794 [0.734, 0.842] for ICD codes. For PDR classification, ICD-10 codes performed better than ICD-9 codes (F1 score 0.836 [0.771, 0.878] vs. 0.596 [0.222, 0.692]). For NPDR severity classification, the notes algorithm performed similarly to ICD codes, but performance was limited by small sample size.

Conclusions: The notes algorithm outperformed ICD codes for PDR classification. The findings demonstrate the significant potential of applying a rule-based NLP algorithm to clinical notes to increase the efficiency and accuracy of cohort selection for research.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2024;4:100564*

Supplemental material available at www.ophthalmologyscience.org.

Electronic health records (EHRs) contain a vast amount of patient data. Most studies using EHR data rely on diagnosis or billing codes to identify diseases, mainly using International Classification of Diseases, ninth edition (ICD-9) or 10th edition (ICD-10) codes. However, these codes were developed primarily for billing and reimbursement purposes.[1] International Classification of Diseases codes can be broad and imprecise; reliance on them alone could lead to misclassification of patients for disease cohorts and lead to unreliable study results.[2−4] Furthermore, these codes may lack important detailed information about disease, such as eye laterality and disease severity, if not coded or coded incorrectly by the clinician.[2,5,6]

To our knowledge, only 1 study has assessed the accuracy of ICD codes for stage of diabetic retinopathy (DR) using EHR data.[7] The study used a single institution's EHR and found that ICD codes for nonproliferative DR (NPDR) and proliferative DR (PDR) had high accuracy when

https://doi.org/10.1016/j.xops.2024.100564
ISSN 2666-9145/24

compared with manual chart review by physicians. International Classification of Diseases, 10th edition codes were more reliable in correctly identifying DR compared with ICD-9 codes, especially when distinguishing between NPDR and PDR. Both ICD-9 and ICD-10 code accuracy were noticeably lower when identifying the stage of NPDR severity (i.e., mild, moderate, or severe) compared with chart review.

For cohort selection, the alternative to using diagnosis codes is to perform manual chart review to identify patients in the EHR. Manual data abstraction often serves as a gold standard for disease identification and data extraction.[8] The main limitations of manual data abstraction are the time-intensive nature and impracticality for large-scale databases that include millions of patients. Even though manual chart review is considered the gold standard, it is not guaranteed to have perfect accuracy because of the possibility for human error.[9] To fully utilize large-scale EHR data for research in an efficient manner, we need accurate and automated classification methods for cohort selection and data extraction.

Incorporating unstructured data (i.e., clinical notes, imaging reports, and examination findings) from EHRs can likely improve classification of disease. Classification for systemic conditions, such as rheumatoid arthritis, ulcerative colitis, and systemic lupus erythematous, have demonstrated that combining unstructured with structured data can improve algorithm sensitivity and positive predictive value (PPV) compared with using each data source alone.[10,11] In ophthalmology, a classification algorithm for exfoliation syndrome using structured and unstructured data accurately identified all cases and may have outperformed the clinician grader.[12]

In this study, we developed a rule-based natural language processing (NLP) classification algorithm using clinical notes (notes algorithm) and compared its performance in classifying DR type and severity to ICD-9 and ICD-10 codes (diagnosis codes) using physician manual chart review as the gold standard. We assessed the accuracy of classifying patients with NPDR, PDR, and different severity stages of NPDR using these methods.

Methods

Data Sources

The University of California, San Francisco (UCSF) Institutional Review Board approved this study and issued a waiver of informed consent for all subjects. This study followed the tenets of the Declaration of Helsinki. We obtained structured data from the UCSF De-Identified Clinical Data Warehouse (De-ID CDW), which has deidentified EHR data for all UCSF patients. Data in the De-ID CDW are based on the Epic Caboodle Data Warehouse and are updated monthly. Dates are shifted by up to 365 days in the De-ID CDW and protected health information is removed according to the Safe Harbor Method. We used a limited data set version of the De-ID CDW to obtain real dates.

The Machine Redacted Notes are deidentified free text from the UCSF EHR that are available for research and available through the De-ID CDW.[13] Unstructured data included clinical note text and metadata, including date, provider type, note type, encounter type, and department. We included notes from eye providers (i.e., ophthalmologist or optometrist) in the Department of Ophthalmology and Francis I. Proctor Foundation that were from an office visit, hospital encounter, or procedure visit. The De-ID CDW data were last accessed on March 12, 2024.

Subjects

The DR cohort included patients ≥18 years of age who had ≥1 completed in-person visit with an eye provider at UCSF between June 1, 2012 and June 1, 2022 (Fig 1). We included patients who had ≥1 ICD-9 or ICD-10 coded diagnosis of type 1 or 2 diabetes mellitus. Patients were excluded if their date of DR diagnosis was before June 1, 2012 (date UCSF EHR transitioned to Epic) given incomplete data in the De-ID CDW prior to this date. Patients were excluded if they did not have any deidentified clinical notes. International Classification of Diseases, ninth edition codes were used for all encounters before October 1, 2015 and ICD-10 codes for all encounters on or after October 1, 2015.[14,15]

Training and Test Sets

Using ICD codes for DR (Supplemental Table 1, available at www.opthalmologyscience.org), we prescreened patients to develop our DR cohort (N = 2366 patients). Then, we selected a random sample of 100 patients with PDR or NPDR by ICD code for our training set to develop and train the notes algorithm. This training set consisted of 31 patients with PDR and 69 patients with NPDR based on ICD code to approximate the proportion of patients with PDR/NPDR in our DR cohort. The visit encounters of these 100 patients also encompassed a broad range of note styles to allow for a diverse representation of linguistic patterns for effective NLP algorithm training. Two ophthalmologists (R.A. and G.K.) independently performed manual chart review on the entire training set of 100 patients (200 eyes) to determine the gold standard label. Each of the 200 eyes was categorized into no DR, NPDR, or PDR based on if they had any evidence of the disease in the EHR. The interrater reliability was assessed between the 2 ophthalmologists using the Cohen's kappa. A third ophthalmologist (C.Q.S.) independently adjudicated any differences between the 2 initial reviewers' gradings. All 3 reviewers were masked to the algorithm results and to each other's gradings.

The test set included a random sample of 206 patients (412 eyes) with PDR or NPDR from the DR cohort, excluding patients used in the training set. The decision to have a larger test set than training set, enriched for patients with PDR, was driven by the goal of evaluating the note algorithm's performance on a more diverse dataset with variations in severity. Three ophthalmologists (C.J.M., L.C., and C.Q.S.) manually chart reviewed a third of the patients in the test set each. Because of the high intergrader reliability between the initial 2 reviewers in the training set (unweighted Cohen's kappa was 0.86 and the weighted was 0.94), each patient in the test set was only reviewed by 1 ophthalmologist.

Note Algorithm Development

We developed a rule-based NLP classification algorithm using the UCSF deidentified notes. The clinical note texts underwent formatting, text cleaning, and assessment and plan segmentation. Natural language processing and negation were applied using scispaCy[16] and NegEx[4] respectively to the clinical note texts to determine disease classification for PDR or NPDR.

In detail, clinical note texts were cleaned and formatted for standardization by correcting Unicode characters; converting all characters to lowercase; expanding abbreviations (e.g., PDR to proliferative diabetic retinopathy and NPDR to nonproliferative

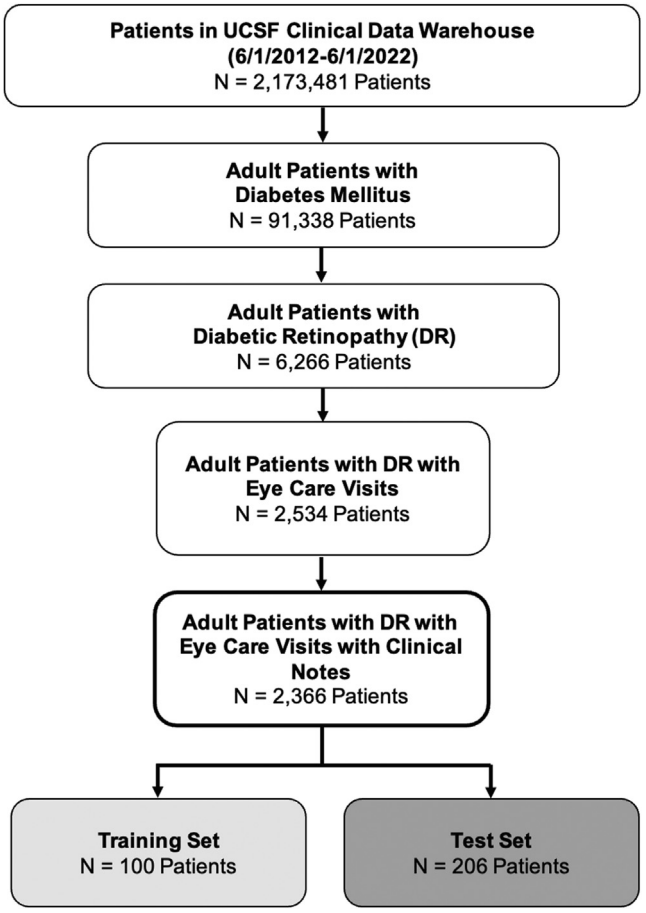

Figure 1. Study workflow and cohort definition. DR = diabetic retinopathy; UCSF = University of California, San Francisco.

diabetic retinopathy); and removing headers, footers, section breaks, blank lines, stray punctuations, and extra spaces using Python3 RegEx.[17] Assessment and plan segmentation was then performed to focus on this area of the note text. Part of speech tagging, dependency parsing, named entity recognition, tokenization, and sentence segmentation was applied to cleaned notes using scispaCy, a Python package using the spaCy model to process biomedical and clinical text.[16] We then screened for mention of "proliferative retinopathy" or "proliferative diabetic retinopathy" for PDR and "non-proliferative retinopathy," "non-proliferative diabetic retinopathy," "non proliferative retinopathy," "non proliferative diabetic retinopathy," "nonproliferative retinopathy," "nonproliferative diabetic retinopathy," or "diabetic retinopathy" for NPDR. NegEx was used to identify negations of disease.[4] Our notes algorithm tagged each note with positive or negative findings of NPDR and PDR. For simplicity, no mention of retinopathy and negative mention of retinopathy were both classified as negative finding of retinopathy. If a note had both positive and negative mentions of retinopathy, the algorithm would default to tagging the note with positive finding of retinopathy. Similarly, a note that had positive mentions of both NPDR and PDR was classified as PDR. For each patient, the individual note results were aggregated to determine if there was any presence of NPDR or PDR.

Notes that were labeled as NPDR were further separated into 1 of 4 severity types: mild, moderate, severe, or unspecified. Mild, moderate, and severe classifications were performed by a simple free-text search of the words "mild," "moderate," or "severe" within 10 words before or after a positive mention of NPDR within the clinical note. We selected 10 words based on a review of notes in the training set. An NPDR-labeled note with no mention of mild, moderate, or severe was classified as "unspecified." If an NPDR-labeled note was classified with multiple NPDR severities, the most advanced NPDR severity was used. These methods were finalized based on results from the training set. The notes algorithm was developed using Python 3.9.7.

Algorithm Evaluation and Testing

The diagnosis codes and gold standard (manual chart review) diagnosed NPDR or PDR at the eye level. Since the notes algorithm was unable to identify laterality, the diagnosis codes and gold standard were categorized at the patient level instead for comparison. All training and test set patients were categorized into no DR, NPDR, or PDR based on if they had any evidence of the disease in either eye in the entire EHR. For NPDR severity, the most severe NPDR diagnosis in either eye at the most recent visit was used to classify each patient in all 3 methods.

The holdout test set consisted of 206 random patients from the DR cohort that were distinct from the training set. Using the test set, the ICD code and notes algorithm were applied and compared with the gold standard to determine the sensitivity, specificity, PPV, NPV, and F1 score ($2\times[\text{Sensitivity}\times\text{PPV}]/[\text{Sensitivity}+\text{PPV}]$) for classification of PDR and NPDR severity

Table 1. Baseline Characteristics of Training and Test Set Patients

Variables	Training Set N = 100	Test Set N = 206
Age (mean, SD)	61.6 (13.4)	58.1 (13.4)
Female sex (mean, SD)	47 (47.0)	91 (44.2)
Race/ethnicity (N, %)		
White	38 (38.0)	38 (18.4)
Black/African American	16 (16.0)	15 (7.3)
Latino/Hispanic	18 (18.0)	80 (38.8)
Asian	18 (18.0)	32 (15.5)
Other	8 (8.0)	26 (12.6)
Unspecified	2 (2.0)	15 (7.3)
Diabetes mellitus* (N, %)		
Type 1	5 (5.0)	15 (7.3)
Type 2	95 (95.0)	191 (92.7)
Diabetic retinopathy* (N, %)		
PDR	31 (31.0)	129 (62.6)
NPDR	69 (69.0)	77 (37.4)
NPDR severity* (N, %)		
Mild	22 (31.9)	20 (26.0)
Moderate	9 (13.0)	7 (9.1)
Severe	0 (0.0)	11 (14.3)
Unspecified	38 (55.1)	39 (50.6)
Health insurance (N, %)		
Medicare	41 (41.0)	75 (36.4)
Medicaid	15 (15.0)	77 (37.4)
Private	39 (39.0)	43 (20.9)
Self-pay or no insurance	5 (5.0)	11 (5.3)

N = number; NPDR = nonproliferative diabetic retinopathy; PDR = proliferative diabetic retinopathy; SD = standard deviation.
*By International Classification of Diseases ninth or 10th edition code.

types. Then, the notes algorithm and ICD codes were applied to the entire 2366 patients from the DR cohort to identify the prevalence of PDR among patients with DR in our EHR, enabling a direct comparison of the total number of patients identified by each approach.

Statistical Analysis

Bootstrapping with replacement was done 1000 times to create 95% confidence intervals for PPV, NPV, sensitivity, specificity, and F1-scores. To test interrater reliability for the training set, the Cohen's kappa was interpreted as follows: <0 indicating no agreement, 0 to 0.20 as none to slight, 0.21 to 0.40 as fair, 0.41 to 0.60 as moderate, 0.61 to 0.80 as substantial, and 0.81 to 1.00 as almost perfect agreement.[18] All statistical analyses were conducted in R version 4.3.1 (R Foundation for Statistical Computing).

Results

The entire DR cohort consisted of 2366 patients with 1660 NPDR patients (70.2%) and 706 PDR patients (29.8%) by ICD codes. The training and test sets consisted of distinct and random selections of 100 patients and 206 patients extracted from the DR cohort (Table 1).

Table 2 shows the results of how each classification method performed against the gold standard of manual chart review for determining presence or absence of PDR. Since all 206 patients in the test set had a diagnosis of either NPDR or PDR by ICD codes, we chose to report the results for classifying PDR only. The notes algorithm performed the best for all metrics with sensitivity of 90.5%, specificity of 93.8%, PPV of 98.0%, NPV of 75.0%, and F1 score of 0.941. The notes algorithm compared with the combined ICD-9/10 codes had significantly higher sensitivity, NPV, and F1 score, but similar PPV and specificity. The ICD-9 codes performed significantly worse than ICD-10 codes for sensitivity, NPV, and F1 score.

For NPDR severity classification, the notes algorithm performed similarly compared with the combined ICD-9/10 codes for all metrics and all severity except for specificity in the "mild" group (Table 3). The only classification method that resulted in a "good" F1 score of >0.7 was the notes algorithm for mild NPDR (F1 score 0.743) and severe NPDR (F1 score 0.769). The notes algorithm for moderate and unspecified NPDR had F1-scores of <0.5. In a sensitivity analysis, moderate and severe NPDR were combined and evaluated, since this is a common threshold for treating DR in screening protocols.[19] We did not find any significant difference between the notes algorithm and the combined ICD9/10 codes for moderate/severe NPDR across all metrics.

The entire 2366 patients from the DR cohort were classified as having presence or absence of PDR by ICD-9/10 codes and the notes algorithm. The ICD-9/10 codes alone identified 706 patients (29.8% of DR cohort) with evidence of PDR and had 90.2% overlap with the notes algorithm.

Table 2. Comparison of Classification Method Against Gold Standard for Presence (or Absence) of Proliferative Diabetic Retinopathy among Test Set Patients

Classification Method (vs. Gold Standard)	F1 Score [95% CI]*	PPV [95% CI]*	NPV [95% CI]*	Sensitivity [95% CI]*	Specificity [95% CI]*
Notes algorithm (N = 206)	0.941 [0.910, 0.966]	98.0% [95.4%, 100%]	75.0% [63.4%, 85.9%]	90.5% [85.7%, 94.9%]	93.8% [86.0%, 100%]
Diagnosis codes					
ICD-9/ICD-10 (N = 206)	0.794 [0.734, 0.842]	94.7% [90.3%, 98.3%]	45.7% [36.0%, 55.9%]	68.4% [60.4%, 75.3%]	87.5% [77.4%, 95.9%]
ICD-9 (N = 40)	0.596 [0.222, 0.692]	82.4% [53.1%, 100%]	30.4% [12.8%, 39.9%]	46.7% [15.0%, 61.5%]	70.0% [33.3%, 100%]
ICD-10 (N = 166)	0.836 [0.771, 0.878]	96.9% [92.9%, 100%]	50.7% [39.9%, 61.2%]	73.4% [64.3%, 79.4%]	92.1% [82.9%, 100%]

CI = confidence interval; ICD-9 = International Classification of Disease, ninth edition; ICD-10 = International Classification of Disease, 10th edition; N = number; NPV = negative predictive value; PPV = positive predictive value.
*Confidence intervals estimated using a nonparametric bootstrap.

Table 3. Comparison of Classification Methods Against Gold Standard for NPDR Severity

NPDR Severity	Classification Method (vs. Gold Standard)	F1 Score [95% CI]*	PPV [95% CI]*	NPV [95% CI]*	Sensitivity [95% CI]*	Specificity [95% CI]*
Mild (N = 22)†	Notes algorithm	0.743 [0.541, 0.889]	92.9% [75.0%, 100%]	74.2% [58.5%, 89.3%]	61.9% [38.9%, 83.3%]	95.8% [85.0%, 100%]
	ICD-9/ICD-10	0.564 [0.323, 0.727]	55.0% [31.3%, 76.9%]	63.6% [41.7%, 83.3%]	57.9% [31.6%, 80.0%]	60.9% [38.9%, 80.0%]
	ICD-9	NA	0%	33.3% [0%, 69.2%]	0%	66.7% [0%, 100%]
	ICD-10	0.647 [0.400, 0.813]	57.9% [33.3%, 79.0%]	75.0% [53.3%, 95.0%]	73.3% [47.1%, 95.0%]	60.0% [37.0%, 81.3%]
Moderate (N = 13)†	Notes algorithm	0.471 [0.182, 0.727]	66.7% [20.0%, 100%]	82.1% [68.6%, 92.9%]	36.4% [10.0%, 66.7%]	94.1% [84.6%, 100%]
	ICD-9/ICD-10	0.316 [0.111, 0.571]	50.0% [0%, 100%]	72.2% [56.7%, 85.3%]	23.1% [0%, 50.0%]	89.7% [77.3%, 100%]
	ICD-9	NA	NA	71.4% [33.3%, 88.9%]	0%	100% [100%, 100%]
	ICD-10	0.353 [0.125, 0.615]	50.0% [0%, 100%]	72.4% [54.5%, 88.2%]	27.3% [0%, 58.4%]	87.5% [72.7%, 100%]
Severe (N = 7)†	Notes algorithm	0.769 [0.400, 1.000]	83.3% [44.4%, 100%]	94.9% [86.9%, 100%]	71.4% [33.3%, 100%]	97.4% [90.9%, 100%]
	ICD-9/ICD-10	0.462 [0.167, 0.778]	50.0% [0%, 100%]	88.9% [77.5%, 97.6%]	42.9% [0%, 85.8%]	91.4% [80.6%, 100%]
	ICD-9	0.500 [0.286, 1.000]	33.3% [16.7%, 100%]	100% [100%, 100%]	100% [100%, 100%]	66.7% [22.2%, 100%]
	ICD-10	0.444 [0.182, 0.833]	66.7% [0%, 100%]	87.5% [77.5%, 97.3%]	33.3% [0%, 83.3%]	96.6% [87.9%, 100%]
Unspecified (N = 6)†	Notes algorithm	0.320 [0.100, 0.545]	21.1% [5.3%, 41.1%]	92.3% [79.5%, 100%]	66.7% [20.0%, 100%]	61.5% [47.2%, 76.9%]
	ICD-9/ICD-10	0.154 [0.105, 0.465]	10.0% [0%, 33.3%]	93.8% [84.0%, 100%]	33.3% [0%, 100%]	76.9% [62.8%, 89.5%]
	ICD-9	NA	0%	100% [100%, 100%]	NA	57.1% [14.3%, 100%]
	ICD-10	0.200 [0.133, 0.571]	14.3% [0%, 50.0%]	92.8% [81.8%, 100%]	33.3% [0%, 100%]	81.3% [67.7%, 93.6%]

CI = confidence interval; ICD-9 = International Classification of Disease, ninth edition; ICD-10 = International Classification of Disease, 10th edition; N = number; NA = not applicable; NPDR = nonproliferative diabetic retinopathy; NPV = negative predictive value; PPV = positive predictive value.
*Confidence intervals estimated using a nonparametric bootstrap.
†Reported based on gold standard classification.

The notes algorithm identified 768 patients (32.4%) with evidence of PDR with 82.9% overlap with the ICD-9/10 codes. With the notes algorithm serving as the gold standard because of its better performance on the test set, the ICD-9/10 result for classifying PDR had sensitivity of 82.9%, specificity of 95.7%, PPV of 90.2%, and NPV of 92.1%.

Discussion

This study aimed to harness the vast amount of unstructured data available in the EHR and provide an alternative method to using ICD codes for identifying patients with DR. We developed a rule-based NLP classification algorithm using clinical notes. In addition, we directly compared the performance of ICD-9 and ICD-10 codes to the notes algorithm using manual chart review by ophthalmologists as the gold standard. Our findings showed that for classifying patients with PDR, the notes algorithm performed the best across all metrics compared with ICD codes. Specifically, the notes algorithm had a significantly higher sensitivity and NPV compared with ICD-9/10 codes for PDR classification and had a F1 score of 0.941, indicating near perfect classification (0 is worst, 1 is perfect). For NPDR severity classification, the performance of the notes algorithm was at best good for mild and severe NPDR severity. This is likely attributed to our small sample sizes when stratifying by NPDR severity types. Across the entire DR cohort, the notes algorithm identified 62 more patients with PDR than ICD-9/10 codes.

International Classification of Diseases, ninth edition or 10th edition codes can result in misclassification for certain diseases due to a number of factors such as human error, inadequate clinician training on the nuances of billing codes, poor EHR system design for diagnosis coding, and time-saving motivations for choosing broader codes, such as "unspecified."[20–22] With the transition from ICD-9 to ICD-10 codes in late 2015, increased granularity was introduced which theoretically allowed for greater accuracy with disease classification.[14,23]

Cai et al conducted a single-site retrospective cohort study comparing the accuracy of ICD-9 and ICD-10 codes with stages of DR.[7] The study concluded that ICD-10 codes were more accurate than ICD-9 codes, particularly in distinguishing between NPDR and PDR. This distinction was more pronounced during the later time-period after the transition to ICD-10 codes. Our study found similar results with ICD-10 codes having significantly better performance by F1 score and NPV when compared with ICD-9 codes for PDR and NPDR severity classifications. However, our ICD-9 and ICD-10 codes for classifying PDR performed worse than reported in Cai et al.[7] We had noticeably lower results for NPV (45.7% vs. 97.86%) and sensitivity (68.4% vs. 97.76%), respectively, which is likely attributable to higher numbers of false negatives in our study. This suggests instances where there was no ICD code listed for PDR, but the manual chart review (gold standard) identified the presence of PDR in the patient's EHR. One possible reason to explain this difference is that Cai et al excluded patients with concurrent retinal diagnoses that could confound the determination of DR or would also require anti-VEGF treatments.[7] These diagnoses included branch or central retinal vein occlusion, hypertensive retinopathy, ocular ischemic syndrome, and neovascular age-related macular degeneration.[7] Since we did not exclude those with confounding concurrent retinal diagnoses, our cohort may be more heterogenous in terms of ocular comorbidities, which could potentially explain why the ICD codes performed worse. Other possibilities include that these patients were mainly seen at UCSF for non-DR related issues, such as glaucoma or refraction, and the eye provider did not code PDR since it was not addressed during the visit.

Clinical notes contain more information than diagnosis or billing codes, but it is also more challenging to extract meaningful data from them. With the recent progression in NLP and text-mining tools, it has become easier to harness unstructured data in clinical research. In this study, our rule-based notes algorithm had a significantly higher sensitivity, NPV, and F1 score compared with ICD-9/ICD-10 codes. Of the 3 false-positive PDR classifications for the notes algorithm, 2 patients had notes discussing the general practice guidelines for PDR though no eyes actually had PDR. Of the 15 false-negative PDR classifications for the notes algorithm, all misclassification cases were due to the notes not mentioning any evidence of PDR, but the manual chart review (gold standard) of the entire EHR indicated that the patient had a positive mention of PDR. This discrepancy may be because of the use of other data sources, including historical records prior to 6/1/2012 (e.g., scanned EHR documents), imaging reports, or clinical examination findings.

When examining NPDR severity, the performance across all NPDR severity was mediocre to at best good for all classification methods. Our sample sizes were limited, ranging from 6 to 22 patients when grouped by NPDR severity type. When there are insufficient data, the model may struggle to form an accurate representation of the data and may have poor performance.[24] As a sensitivity analysis, we further assessed if the notes algorithm or ICD-9/10 codes over-called (i.e., higher severity compared with gold standard), under-called (i.e. lower severity compared with gold standard), or was unable to assess severity (i.e., unspecified NPDR severity) for those with a gold standard classification of mild, moderate, or severe NPDR. For our notes algorithm compared with the gold standard (N = 20 had discrepancies), there was inability to assess severity because of unspecified NPDR in the notes 75% of the time and over-call 25% of the time. For the ICD-9/10 codes compared with the gold standard (N = 25 had discrepancies), there was over-call 28% of the time, under-call 36% of the time, and inability to assess severity 32% of the time. Similar to our PDR findings, the notes algorithm was largely unable to assess NPDR severity accurately because of no mention of NPDR severity in the notes. For the gold standard, the reviewing physician may have used other data sources to help them determine NPDR severity, such as exam findings.

The study has several limitations. First, the data only represent 1 academic institution. Because of the potential for protected health information leakage, it is currently challenging to obtain deidentified clinical notes from other institutions. However, the Subjective, Objective, Assessment and Plan note

is a widely used method for documentation at follow-up visits by physicians.[25] Based on this note structure, physicians typically document the diagnoses under the assessment and plan section, which is where our notes algorithm screened for mention of keywords for diseases in the note. Second, while the gold standard chart review by physicians reviewed the entire EHR, including clinical notes and exam findings, our notes algorithm only used clinical notes from patient encounters because of the lack of eye examination data in our deidentified database because of data redaction for protected health information. We are currently working to reduce the data redaction in the examination free text to use eye examination data in the future.

Third, there could be bias based on how we prescreened our DR cohort using ICD-9/10 codes initially. Since ICD-9/10 codes are currently the only tool readily available for classification, we elected to use it to estimate the ratio of cases to controls in the UCSF EHR for our training and test sets. If we had not increased the minority class (PDR) in our training and test sets, then we could have class imbalance and potentially worse model performance.[26] If we had simply taken a random sample of patients in the EHR using the prevalence for PDR that is present in the general United States population with diabetes mellitus (~5%), then we could have had very few cases (PDR) and many controls (no PDR), which would have led to worse class imbalance and potentially poor model performance. Since we prescreened with ICD-9/10 codes, the performance of the ICD-9/10 codes and the notes algorithm may have been higher than if we had taken a random sample from the EHR. Furthermore, the PPV and NPV may be lower in cohorts with reduced prevalence of the disease.

Fourth, the notes algorithm did not have the ability to determine laterality. The clinical text mentioned laterality in a variety of ways, which made it difficult to develop a rule-based approach to accurately diagnose disease at the eye level. This is a notable limitation of the notes algorithm as laterality is an important aspect of ophthalmic research. Lastly, the notes algorithm was a rule-based algorithm instead of a more complex machine learning approach or large language model. A machine learning incorporated algorithm may improve DR classification performance by identifying more complex patterns in clinical notes with positive disease findings such as laterality and NPDR severity.[27,28] Since rule-based algorithms are easier to deploy, we decided to start with this approach, but future directions include applying a large language model and comparing with manual chart review.

In conclusion, we demonstrated the significant potential of a rule-based NLP classification algorithm using clinical notes to identify patients with PDR, outperforming the traditional approach of using ICD codes. These findings highlight the limitations of ICD codes and the growing importance of harnessing unstructured clinical data. Using an NLP algorithm approach on clinical notes increases the efficiency and accuracy of cohort identification for research compared with using ICD codes alone. Future directions include more advanced machine learning models, incorporating additional data sources, expanding applicability across institutions, leveraging structured and unstructured data in a complementary manner, and addressing limitations in sample size and data collection methods to further improve classification of DR and other ocular disease cohorts for research.

Footnotes and Disclosures

Originally received: January 8, 2024.
Final revision: May 31, 2024.
Accepted: June 10, 2024.
Available online: June 14, 2024. Manuscript no. XOPS-D-24-00009.

[1] Department of Ophthalmology, University of California, San Francisco, California.

[2] Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland.

[3] Department of Ophthalmology, Duke University, Durham, North Carolina.

[4] Department of Ophthalmology and Visual Sciences, University of Michigan, Ann Arbor, Michigan.

[5] F.I. Proctor Foundation, University of California, San Francisco, California.

[6] Institute for Global Health Sciences, University of California, San Francisco, California.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s):

D.B.: Consultant – AbbVie/Allergan, Glaukos, Genentech, Iveric Bio, and Verana Health; Honoraria – Iveric Bio.

J.D.S: Grants – NEI R01 EY032475, NEI R01 EY034444, Research to Prevent Blindness, Abbvie Pharmaceuticals, Janssen Pharmaceuticals, and Ocular Therapuetix.

B.F.A.: COAST Study DSMB, honoraria – US National Eye Institute.

C.Q.S.: Grants – National Eye Institute, All May See Foundation, and UCSF Senate Grant.

This work was supported in part by the following grants: National Institutes of Health [NEI K23 EY032637], National Institutes of Health [NIH-NEI P30 EY002162 – UCSF Core Grant for Vision Research], Research to Prevent Blindness unrestricted grant, New York, NY.

HUMAN SUBJECTS: Human subjects data were included in this study. The UCSF Institutional Review Board approved this study and issued a waiver of informed consent for all subjects. This study followed the tenets of the Declaration of Helsinki.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Yonanime, Ma, Borkar, Stein, Arnold, Sun

Data collection: Yonanime, Ma, Alabi, Kaidonis, Chan, Sun

Analysis and interpretation: Yonanime, Borkar, Arnold, Sun

Obtained funding: Sun

Overall responsibility: Yonanime, Ma, Alabi, Kaidonis, Chan, Borkar, Stein, Arnold, Sun

Abbreviations and Acronyms:
De-ID CDW = Deidentified Clinical Data Warehouse; **DR** = diabetic retinopathy; **EHR** = electronic health record; **ICD** = International Classification of Diseases; **NLP** = natural language processing; **NPDR** = nonproliferative diabetic retinopathy; **NPV** = negative predictive value; **PDR** = proliferative diabetic retinopathy; **PPV** = positive predictive value; **UCSF** = University of California, San Francisco.

Keywords:
Clinical notes, Diabetic retinopathy, ICD-9, ICD-10, Natural language processing.

Correspondence:
Catherine Q. Sun, MD, Department of Ophthalmology, 490 Illinois Street, San Francisco, CA 94158. E-mail: catherine.sun@ucsf.edu.

References

1. Meyer H. Coding complexity: US Health Care gets ready for the coming of ICD-10. *Health Aff*. 2011;30:968–974.
2. Palestine AG, Merrill PT, Saleem SM, et al. Assessing the precision of ICD-10 codes for uveitis in 2 electronic health record systems. *JAMA Ophthalmol*. 2018;136:1186–1190.
3. Mainor AJ, Morden NE, Smith J, et al. ICD-10 coding will challenge researchers- caution and collaboration may reduce measurement error and improve comparability over time. *Med Care*. 2019;57:e42–e46.
4. Chapman WW, Bridewell W, Hanbury P, et al. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*. 2001;34:301–310.
5. Stein JD, Lum F, Lee PP, et al. Use of health care claims data to study patients with ophthalmologic conditions. *Ophthalmology*. 2014;121:1134–1141.
6. Leshno A, Tsamis E, Harizman N, et al. The ICD-10 glaucoma severity score underestimates the extent of glaucomatous optic nerve damage. *Am J Ophthalmol*. 2022;244:133–142.
7. Cai CX, Michalak SM, Stinnett SS, et al. Effect of ICD-9 to ICD-10 transition on accuracy of codes for stage of diabetic retinopathy and related complications: results from the CODER study. *Ophthalmol Retina*. 2021;5:374–380.
8. Yin AL, Guo WL, Sholle ET, et al. Comparing automated vs. manual data collection for COVID-specific medications from electronic health records. *Int J Med Inf*. 2022;157:104622.
9. McKenzie J, Rajapakshe R, Shen H, et al. A semiautomated chart review for assessing the development of radiation pneumonitis using natural language processing: diagnostic accuracy and feasibility study. *JMIR Med Inform*. 2021;9:e29241.
10. Barnado A, Casey C, Carroll RJ, et al. Developing electronic health record algorithms that accurately identify patients with systemic lupus erythematosus. *Arthritis Care Res*. 2017;69:687–693.
11. Liao KP, Cai T, Savova GK, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ*. 2015;350:h1885.
12. Stein JD, Rahman M, Andrews C, et al. Evaluation of an algorithm for identifying ocular conditions in electronic health record data. *JAMA Ophthalmol*. 2019;137:491–497.
13. Norgeot B, Muenzen K, Peterson TA, et al. Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes. *NPJ Digit Med*. 2020;3:1–8.
14. Hirsch JA, Nicola G, McGinty G, et al. ICD-10: history and context. *AJNR Am J Neuroradiol*. 2016;37:596–599.
15. Transition to ICD-10. DOL. Available at: http://www.dol.gov/agencies/owcp/FECA/ICD10transition. Accessed September 21, 2023.
16. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics; 2019:319–327.
17. Van Rossum G, Drake Jr FL. The Python standard library, text processing services, re — regular expression. Available at: https://docs.python.org/3/library/re.html; 2023. Accessed July 1, 2024.
18. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med*. 2012;22:276–282.
19. Lee SY. Diabetic retinopathy screening. Available at: https://eyewiki.aao.org/Diabetic_Retinopathy_Screening; 2023. Accessed July 1, 2024.
20. Horsky J, Drucker EA, Ramelson HZ. Accuracy and completeness of clinical coding using ICD-10 for ambulatory visits. *AMIA Annu Symp Proc*. 2018;2017:912–920.
21. Chuen VL, Chan ACH, Ma J, et al. Assessing the accuracy of international classification of diseases (ICD) coding for delirium. *J Appl Gerontol*. 2022;41:1485–1490.
22. O'Malley KJ, Cook KF, Price MD, et al. Measuring diagnoses: ICD code accuracy. *Health Serv Res*. 2005;40:1620–1639.
23. Sivashankaran S, Borsi JP, Yoho A. Have ICD-10 coding practices changed since 2015? *AMIA Annu Symp Proc*. 2020;2019:804–811.
24. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. *JMIR Med Inform*. 2020;8:e17984.
25. Podder V, Lew V, Ghassemzadeh S. SOAP notes. Available at:. In: *StatPearls*. StatPearls Publishing; 2024. http://www.ncbi.nlm.nih.gov/books/NBK482263/. Accessed May 31, 2024.
26. Lin WJ, Chen JJ. Class-imbalanced classifiers for high-dimensional data. *Brief Bioinformatics*. 2013;14:13–26.
27. Jamian L, Wheless L, Crofford LJ, Barnado A. Rule-based and machine learning algorithms identify patients with systemic sclerosis accurately in the electronic health record. *Arthritis Res Ther*. 2019;21:305.
28. Mykowiecka A, Marciniak M, Kupść A. Rule-based information extraction from patients' clinical data. *J Biomed Inform*. 2009;42:923–936.